Pattern Recognition 92 (2019) 231–245

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Feature-guided Gaussian mixture model for image matching

Jiayi Ma[a,b,c], Xingyu Jiang[a], Junjun Jiang[d], Yuan Gao[e,*]

[a] Electronic Information School, Wuhan University, Wuhan, 430072, China
[b] Beijing Advanced Innovation Center for Intelligent Robots and Systems, Beijing Institute of Technology, Beijing, 10081, China
[c] Hubei Key Laboratory of Advanced Control and Intelligent Automation for Complex Systems, Wuhan, 430074, China
[d] School of Computer Science, China University of Geosciences, Wuhan, 430074, China
[e] Tencent AI Laboratory, Shenzhen, 518057, China

ARTICLE INFO

Article history:
Received 17 June 2017
Revised 6 February 2019
Accepted 1 April 2019
Available online 1 April 2019

Keywords:
Image matching
Feature-guided
Gaussian mixture model
Local geometric constraint
Semi-supervised EM

ABSTRACT

This study proposes a novel feature-guided Gaussian mixture model (FG-GMM) for image matching, which generally requires matching two sets of feature points extracted from the provided images. The problem is formulated as the estimation of a feature-guided mixture of densities: a GMM is fitted to one point set, in which both the centers and local features of the Gaussian densities are constrained to coincide with another point set. The said problem is solved under a unified maximum-likelihood framework, in which an iterative semi-supervised expectation-maximization algorithm initialized by the confident feature correspondence is also implemented. This algorithm is flexible and has a general scope, which can handle both rigid and non-rigid image transformations. The transformation in the non-rigid case is specified in a reproducing kernel Hilbert space, and a sparse approximation is adopted to accomplish rapid implementation. Extensive experiments on different real images show that the proposed approach consistently outperforms other state-of-the-art methods, which validates its robustness.

© 2019 Elsevier Ltd. All rights reserved.

1. Introduction

Establishing reliable correspondence between two images is a key problem in computer vision and pattern recognition; this issue is a critical prerequisite in different applications, such as 3D reconstruction, content-based image retrieval, tracking, image fusion, super-resolution, and object detection and recognition [1–7]. This study formulates the aforementioned scenario as a matching problem between two sets of discrete points; each point is an image feature extracted by a feature detector and has a local image descriptor, such as the scale invariant feature transform (SIFT) [8].

Different methods have been proposed to address this matching problem in the past few decades. A popular strategy involves the construction of a set of putative point correspondence based on a *similarity constraint*, which requires that points can only match with other points that have similar descriptors. False correspondences are then removed, and the transformation parameters (either rigid or non-rigid) are estimated robustly based on a *geometric constraint*, in which the matches must satisfy an underlying geometrical requirement [1,9]. Several examples of this strategy include the hypothesize-and-verify random sample consensus (RANSAC) and analogous algorithms [10–12] (based on parametric models), as well as the smooth motion field interpolation methods [1,13,14] (based on non-parametric models). However, the putative set in the first step commonly contains only a small part of all the existing true correspondence instances [15,16]. This quantity reduces further for low-quality or small overlapping images, which can result in an inadequate correspondence that computes the transformation parameters in the second step. Therefore, developing a technique that can preserve most of the existing true matches is largely advantageous.

Estimating the point correspondence and spatial transformation simultaneously instead of computing these two variables individually is another popular strategy [17,18]. These methods commonly involve an iteration process that alternates between the correspondence and transformation estimation. The iterated closest point (ICP) algorithm [19] is one of the more popular point matching approaches based on the aforementioned strategy. It adopts nearest-neighbor relationships to assign a binary correspondence and then employs the estimated correspondence in refining the transformation. The nearest point strategy of ICP is alternatively replaced with soft assignments [17]. Several probabilistic methods have also been introduced recently [18,20,21], where the matching is formulated as the estimation of a mixture of densities utilizing Gaussian mixture models (GMMs). The said matching is then solved through an

* Corresponding author.
E-mail addresses: jyma2010@gmail.com (J. Ma), jiangx.y@whu.edu.cn (X. Jiang), junjun0595@163.com (J. Jiang), ethan.y.gao@gmail.com (Y. Gao).

https://doi.org/10.1016/j.patcog.2019.04.001
0031-3203/© 2019 Elsevier Ltd. All rights reserved.

iterative expectation-maximization (EM) method. The aforementioned methods generate a correspondence matrix between the two original feature sets, and as a result, they do not lose any true matches. However, the feature points in these methods are generally treated as pure spatial coordinates. In particular, the feature descriptors are entirely discarded, which can easily lead to a suboptimal solution when severely degraded data, such as a large outlier percentage, are obtained. Therefore, incorporating the local appearance information of feature points into the formulation is necessary, which helps establish better point correspondence.

This study proposes a novel feature-guided GMM (FG-GMM) to address the problem of robust image matching. This new formulation can incorporate local feature information and preserve most of the existing true matches in an image pair. Specifically, we formulate point matching as the estimation of an FG mixture of densities. A GMM is fitted to one point set, in which both the centers and local features of the Gaussian densities are constrained to coincide with another point set. The said problem is solved under a unified maximum-likelihood framework with a semi-supervised EM algorithm, which is initialized by the confident feature correspondence. The proposed algorithm is flexible and has a general scope, which can handle both rigid and non-rigid image transformations. Thus, it can solve different real-world matching tasks. The transformation in the non-rigid case is modeled in a functional space called reproducing kernel Hilbert space (RKHS) [22], in which the transformation function has an explicit kernel representation. We also provide a fast implementation based on sparse approximation to improve the computational efficiency. The qualitative and quantitative experiments on different real images demonstrate that our method can generate more correct correspondence instances and accomplish better matching accuracy compared with other state-of-the-art methods.

This article is an extension of our earlier published work [23]. Our primary new contributions are detailed as follows: First, we extended our FG-GMM from non-rigid matching to rigid and affine matching, and thus, our method can more flexibly and generally solve different real-world matching problems. Second, we extensively reviewed the up-to-date related work and pointed out the strengths of the proposed method. Third, we presented more theoretical derivations and implementation details of our method, which can better explain why and how our method works. Lastly, we added several different datasets from the remote sensing, medical imaging, computer vision, and multimedia communities for a comprehensive experimental evaluation. We also applied our FG-GMM to the content-based image retrieval problem, which further demonstrates the effectiveness of our proposed method.

2. Related work

This section briefly reviews the background material applied as reference for the current study. This material includes two method types: the first type establishes a set of putative correspondence and then removes false matches, whereas the second type solves a correspondence matrix between point sets.

2.1. Two-step strategy-based methods

The matching problem has a combinatorial nature, thereby creating a large matching space of all the possible matches. A simple problem of matching N points to other N points can lead to a total of $N!$ permutations even without considering the outliers [15]. A popular strategy to establish a reliable point correspondence and address the said issue involves two steps [1]: (i) computing a set of putative correspondence, and (ii) then removing the outliers that utilize geometrical constraints. Putative correspondence instances are obtained in the first step by pruning the set of all possible point correspondence. This scenario is achieved by computing feature descriptors [8,24,25] at the points and removing the matches between points whose descriptors are excessively dissimilar. Lowe [8] proposed a distance ratio method that compares the ratio between the nearest and next-nearest neighbors against a predefined threshold to filter out unstable matches. Pele and Michael [26] further applied the earth mover's distance to replace the Euclidean distance in [8] to measure the similarity between descriptors and improve the matching accuracy. Guo and Cao [27] proposed a triangle constraint, which can performs better in exploring putative correspondence in terms of quantity and accuracy compared with the distance ratio. Hu et al. [27] proposed the local selection of a suitable descriptor for each feature point instead of employing a global descriptor during putative correspondence construction. A cascade scheme has been suggested to prevent the loss of true matches, which can significantly enhance the correspondence number [15,28,29].

Different methods have been proposed in the past decades, including statistical regression methods, resampling methods, non-parametric interpolation methods, and graph matching methods, to remove false matches from the putative set in the second step. Statistics literature shows that the methods that minimize the L_1 norm are more robust and can resist a larger proportion of outliers compared with quadratic L_2 norms [30,31]. Chen et al. [32] proposed the utilization of alternate Hough and inverted Hough transforms for robust feature matching, which can attain mutual verification of relevant correspondence. Liu et al. [33,34] proposed a regression method based on adaptive boosting learning for 3D rigid matching. Maier et al. [35] recently introduced a guided matching scheme based on statistical optical flow; the said researchers obtained promising results in terms of both accuracy and efficiency. The most popular resampling method is RANSAC, which has several variants such as MLESAC [11] and PROSAC [12]. These methods adopt a hypothesize-and-verify approach and attempt to obtain the smallest possible outlier-free subset to estimate a provided parametric model by resampling. The resampling methods rely on a predefined parametric model, which become less efficient when the underlying image transformation is non-rigid; these methods also tend to severely degrade if the outlier proportion becomes large [13]. Several non-parametric interpolation methods [1,13,36,37] have recently been introduced to address these issues. These methods commonly interpolate a non-parametric function by applying the prior condition, in which the motion field associated with the feature correspondence is slow-and-smooth. In addition, Ma et al. [38,39] introduced a locality preserving strategy which can produce accurate matching results within only several milliseconds. Graph matching is another technique to solve the matching problem; several representative studies include spectral matching [40], dual decomposition [41], mode-seeking [15,42], deformable graph matching [43], adjacency tensor matching [44], and graph shift (GS) [45]. Graph matching provides considerable flexibility to the object model and delivers robust matching and recognition. However, it suffers from similar drawbacks of its non-polynomial-hard nature. Recently, some learning-based methods have been developed for feature matching such as learning to find good correspondences (LFGC) [46], which aims to train a multilayer perceptron from a set of putative matches and the camera intrinsics under a parametric geometrical constraint, to label the testing correspondences as inliers or outliers.

2.2. Correspondence matrix-based methods

Formulating this problem in terms of a correspondence matrix between points along with a parametric or non-parametric geometric constraint is another strategy for point correspondence. One of the best-known point matching approaches that follow

this strategy is ICP [19]. ICP alternatively assigns a binary correspondence utilizing nearest-neighbor relationships; it then performs least square (LS) transformation estimation utilizing the estimated correspondence until a local minimum is reached. Guo et al. [47] applied an ICP variant to range image registration. Liu et al. [48] introduced a feature guided model for retinal image registration based on an affine transformation. Chui and Rangarajan [17] established a general framework for non-rigid matching called TPS-RPM, which replaces the nearest point strategy of ICP with soft assignments within a continuous optimization framework that involves deterministic annealing. Boughorbel et al. [49] introduced the Gaussian fields into rigid registration, which was later generalized to the non-rigid setting in [50] and [51]. The image feature matching based on Gaussian fields has also been investigated recently in [52] and [53]. The registration problem has also been solved by employing a robust estimator such as L_2E [14,20,54], which attempts to obtain a robust transformation estimate. Point matching has commonly been solved by probabilistic methods in recent years [18,21,55,56]. Ge et al. [57,58] specifically proposed a global-local topology preservation method based on the coherent point drift (CPD) to cope with highly articulated deformation [18]. These methods formulate matching as the estimation of a mixture of densities utilizing GMMs, which is solved within the maximum-likelihood framework and EM algorithm. In order to simultaneously incorporate the global and local priors, Yang et al. [59] further proposed a robust method called global and local mixture distance with thin plate spline (GLMDTPS) which has achieved promising results. Alternatively, Zhang et al. [60] proposed to use a dual-feature for registration of point sets, where the global-local structure is preserved by using two regularization terms.

The aforementioned methods have been successfully implemented in many scenarios. However, the two-step strategy loses true correspondence, whereas the correspondence matrix-based methods do not employ local appearance information. The present study proposes a novel FG-GMM formulation with a local geometric constraint and provides an optimization strategy based on the semi-supervised EM technique to address the said issues.

3. Method

This section describes the proposed matching algorithm. We start by introducing the FG-GMM formulation to register feature sets with associated descriptors and then provide the optimization method based on semi-supervised EM. We then present a local geometric constraint to ensure the well-posedness of the problem, followed by the estimation of spatial transformation that includes rigid, affine, and non-rigid models. Finally, we analyze the computational complexity and provide the implementation details of our method.

3.1. Feature-guided gaussian mixture model

Suppose that we obtain two feature sets extracted from two provided images: a model feature set $\{\mathcal{X}, \mathcal{S}_\mathbf{x}\}$ and a target feature set $\{\mathcal{Y}, \mathcal{S}_\mathbf{y}\}$, where $\mathcal{X} = \{\mathbf{x}_n\}_{n=1}^N$ and $\mathcal{Y} = \{\mathbf{y}_m\}_{m=1}^M$ are 2D column vector sets, which indicate the spatial positions of feature points. $\mathcal{S}_\mathbf{x} = \{S(\mathbf{x}_n)\}_{n=1}^N$ and $\mathcal{S}_\mathbf{y} = \{S(\mathbf{y}_m)\}_{m=1}^M$ are the associated feature descriptor vector sets. We attempt to establish accurate correspondence instances between the two feature sets and simultaneously solve the spatial transformation $\mathcal{T}$ to align the two original images accordingly.

Point matching can be formulated as the estimation of a mixture of densities without considering the associated feature descriptors. A GMM is fitted to the target points $\mathcal{Y}$ such that the centroids of the Gaussian densities are constrained to coincide with the transformed model points $\mathcal{T}(\mathcal{X})$ [18,20,21]. Let $\mathcal{Z} = \{z_m \in \mathbb{N}_{N+1} : m \in \mathbb{N}_M\}$ be a set of latent variables, where each variable z_m assigns a target point $\mathbf{y}_m$ to a GMM centroid $\mathcal{T}(\mathbf{x}_n)$ (if $z_m = n$, $1 \le n \le N$) or to an additional outlier class (if $z_m = N+1$). The GMM probability density function can then be defined as follows:

$$p(\mathbf{y}_m) = \sum_{n=1}^{N+1} P(z_m = n)\, p(\mathbf{y}_m | z_m = n). \tag{1}$$

We generalize the formulation in this study to register feature sets with associated descriptors. In particular, let π_{mn} be the membership probability of the GMM, which is generally assumed to be equal for all GMM components in the original formulation (i.e., $\pi_{mn} = \frac{1}{N}$, $\forall m \in \mathbb{N}_M, n \in \mathbb{N}_N$) [18,21]. We instead assign its value based on the associated feature descriptor vectors $\mathcal{S}_\mathbf{x}$ and $\mathcal{S}_\mathbf{y}$. Hence, we first match $\mathcal{S}_\mathbf{x}$ and $\mathcal{S}_\mathbf{y}$ according to a descriptor similarity constraint, such as comparing the distance of the closest neighbor to that of the second-closest neighbor (i.e., distance ratio) and matching them if the distance ratio is below a predefined threshold t [8]. We then assign $\pi_{mn} = \tau$ if $S(\mathbf{x}_n)$ is matched to $S(\mathbf{y}_m)$, where parameter τ, $0 \le \tau \le 1$, can be considered as the confidence of a feature correspondence. We set the remaining elements of $\{\pi_{mn}\}_{m=1,n=1}^{M,N}$ to either $(1-\tau)/(N-1)$ or $1/N$, so that they satisfy $0 \le \pi_{mn} \le 1$ together with $\forall m$, $\sum_{n=1}^N \pi_{mn} = 1$. Note that the matched correspondence instances can be contaminated by some false correspondences and generally contain only a small part of the true correspondence instances [16].

A popular assumption for point matching is the equal isotropic covariance $\sigma^2\mathbf{I}$ on all GMM components and uniform distribution $1/a$ for the outliers [18,55]. We denote the set of unknown parameters as $\boldsymbol{\theta} = \{\mathcal{T}, \sigma^2, \gamma\}$, where $\gamma \in [0, 1]$ is the outlier percentage. The mixture model in Eq. (1) then takes the following form:

$$\begin{aligned} p(\mathbf{y}_m|\boldsymbol{\theta}) &= \gamma\frac{1}{a} + (1-\gamma)\sum_{n=1}^{N}\pi_{mn}\mathcal{N}(\mathbf{y}_m|\mathcal{T}(\mathbf{x}_n), \sigma^2\mathbf{I}) \\ &= \gamma\frac{1}{a} + (1-\gamma)\sum_{n=1}^{N}\frac{\pi_{mn}}{2\pi\sigma^2}e^{-\frac{\|\mathbf{y}_m - \mathcal{T}(\mathbf{x}_n)\|^2}{2\sigma^2}}. \end{aligned} \tag{2}$$

The parameter set $\boldsymbol{\theta}$ can be estimated by maximizing the likelihood or minimizing the negative log-likelihood as follows:

$$\mathcal{L}(\boldsymbol{\theta}|\mathcal{Y}) = -\sum_{m=1}^{M}\ln p(\mathbf{y}_m|\boldsymbol{\theta}), \tag{3}$$

where we implemented the i.i.d. data assumption. The correspondence probability between the two features $\{\mathbf{x}_n, S(\mathbf{x}_n)\}$ and $\{\mathbf{y}_m, S(\mathbf{y}_m)\}$ can be defined as the posterior probability of the GMM centroid given the target point: $P(z_m = n|\mathbf{y}_m) = \pi_{mn}p(\mathbf{y}_m|z_m = n)/p(\mathbf{y}_m)$. The transformation $\mathcal{T}$ can thus be obtained from the optimal solution $\boldsymbol{\theta}^*$.

3.2. The semi-supervised EM algorithm

The parameters of the mixture model can be estimated in several ways, such as the EM algorithm, gradient descent, and variational inference. The EM algorithm [61] is a technique to learn and infer in the context of latent variables. This algorithm alternates between the expectation step (E-step) and the maximization step (M-step). We follow a standard notation [62] and remove several terms that are independent of $\boldsymbol{\theta}$. Considering the negative log-likelihood function (Eq. (3)), the complete-data log-likelihood is then expressed as follows:

$$\begin{aligned} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}}) &= M_\mathbf{P}\ln\sigma^2 - M_\mathbf{P}\ln(1-\gamma) - (M - M_\mathbf{P})\ln\gamma \\ &\quad + \frac{1}{2\sigma^2}\sum_{m=1}^{M}\sum_{n=1}^{N}P(z_m = n|\mathbf{y}_m, \boldsymbol{\theta}^{\text{old}})\|\mathbf{y}_m - \mathcal{T}(\mathbf{x}_n)\|^2, \end{aligned} \tag{4}$$

where $M_{\mathbf{P}} = \sum_{m=1}^{M}\sum_{n=1}^{N} P(z_m = n|\mathbf{y}_m, \boldsymbol{\theta}^{\text{old}}) \le M$.

E-Step: This step attempts to estimate the posterior distributions of the latent variables (i.e., $p_{mn} = P(z_m = n|\mathbf{y}_m, \boldsymbol{\theta}^{\text{old}})$) by applying the current estimated parameters, $\boldsymbol{\theta}^{\text{old}}$. Given that we have some confident feature correspondence obtained according to the associated descriptors, we select the semi-supervised EM [63] over the original EM. We particularly compute p_{mn} according to the following rules:

(i) The target features $\{\mathbf{y}_m\}$ with known correspondence are expected to serve as anchors that lead the EM iteration to avoid or alleviate getting trapped into the local minima. Thus, we set the following:

$$p_{mn} = \pi_{mn}, \quad 1 \le n \le N. \tag{5}$$

(ii) The posterior distribution for the target features $\{\mathbf{y}_m\}$ with unknown correspondence can be computed by applying Bayes rule as follows:

$$\begin{aligned} p_{mn} &= \frac{P(\mathbf{y}_m|z_m = n, \boldsymbol{\theta}^{\text{old}})P(z_m = n|\boldsymbol{\theta}^{\text{old}})}{P(\mathbf{y}_m|\boldsymbol{\theta}^{\text{old}})} \\ &= \frac{\pi_{mn} e^{-\frac{\|\mathbf{y}_m - \mathcal{T}(\mathbf{x}_n)\|^2}{2\sigma^2}}}{\sum_{k=1}^{N} \pi_{mk} e^{-\frac{\|\mathbf{y}_m - \mathcal{T}(\mathbf{x}_k)\|^2}{2\sigma^2}} + \frac{2\gamma\pi\sigma^2}{(1-\gamma)a}}, \quad 1 \le n \le N. \end{aligned} \tag{6}$$

The posterior distribution p_{mn} is a soft assignment, which indicates the degree to which the target feature $\{\mathbf{y}_m, S(\mathbf{y}_m)\}$ coincides with the model feature $\{\mathbf{x}_n, S(\mathbf{x}_n)\}$ under the current estimated parameters, $\boldsymbol{\theta}^{\text{old}}$.

M-Step: We compute the revised parameters as $\boldsymbol{\theta}^{\text{new}} = \arg\max_{\boldsymbol{\theta}} \mathcal{Q}(\boldsymbol{\theta}, \boldsymbol{\theta}^{\text{old}})$. Considering the derivatives of $\mathcal{Q}(\boldsymbol{\theta})$ with respect to γ and σ^2, as well as setting them to zero, we thus obtain the following:

$$\gamma = 1 - M_{\mathbf{P}}/M, \tag{7}$$

$$\sigma^2 = \frac{\sum_{m=1}^{M}\sum_{n=1}^{N} p_{mn}\|\mathbf{y}_m - \mathcal{T}(\mathbf{x}_n)\|^2}{2M_{\mathbf{P}}}. \tag{8}$$

Estimating the variance σ^2 with a large initial value is conceptually similar to deterministic annealing [17], which applies the solution of an easy (e.g., smoothed) problem to recursively provide the initial conditions to increasingly more difficult problems. However, these approaches differ in several aspects, such as not requiring any annealing schedule. Maximization $\mathcal{Q}(\boldsymbol{\theta})$ with respect to $\mathcal{T}$ is a complicated procedure, which will be discussed in the proceeding sections.

We obtain the estimated spatial transformation $\mathcal{T}$ once the semi-supervised EM converges. We can then superpose the transformed model image on the target image to produce a mosaic image. The feature correspondence can also be computed based on the posterior distribution $\{p_{mn}\}_{m=1,n=1}^{M,N}$. However, the posterior distribution p_{mn} can suffer from outliers because the known correspondence can contain false correspondence. We update p_{mn} associated with the known correspondences one more time by utilizing Eq. (6) rather than Eq. (5) to address the said issue. We then obtain the correspondence set $\mathcal{I}$ with a predefined threshold η:

$$\mathcal{I} = \{(m, n) : p_{mn} > \eta, m \in \mathbb{N}_M, n \in \mathbb{N}_N\}. \tag{9}$$

We observed that the posterior probabilities of the samples in practice are primarily (over 99%) either smaller than 0.01 or larger than 0.99 after the EM iteration converges. Therefore, the selection of η is not a priority in our method.

Convergence analysis: The objective function (3) is not convex, so any algorithm that determines its global minimum is unlikely. However, a stable local minimum is often sufficient for many practical applications. Thus, our strategy is to apply the semi-supervised EM technique, which adopts known correspondence to assign the latent variables. It makes the known correspondence serve as anchors, so it can avoid getting trapped into the local minima during the EM iteration. The adaptive estimate of the variance σ^2 also initializes the variance σ^2 with a large initial value. It then utilizes the semi-supervised EM algorithm. The objective function becomes convex in a large region that can filter out many unstable shallow local minima if the σ^2 value is sufficiently large. Hence, we can likely determine a suitable minimum for large variance. As σ^2 decreases, the objective function tends to change smoothly, which makes employing the old minimum as the initial value more helpful in converging to a new suitable minimum. Therefore, we can more probably reach a stable local minimum as the iterations continue. This scenario is conceptually similar to deterministic annealing [14,17], which adopts the solution of an easy problem to recursively provide initial conditions to increasingly more difficult problems.

3.3. Local geometric constraint

The spatial transformation $\mathcal{T}$ is estimated by minimizing a weighted empirical error $\mathcal{Q}(\mathcal{T}) = \frac{1}{2\sigma^2}\sum_{m=1}^{M}\sum_{n=1}^{N} p_{mn}\|\mathbf{y}_m - \mathcal{T}(\mathbf{x}_n)\|^2$ according to the objective function (4). This condition is not tractable because the feature sets generally suffer from noise and outliers. The problem will also become large in the non-rigid case because the solution of $\mathcal{T}$ is not unique. The rough structures of two feature sets extracted from an image pair must generally be similar. For example, most neighboring feature points cannot move independently under deformation because of physical constraints. This scenario is particularly beneficial when the images involve non-rigid or discontinuous motions [64]. Therefore, developing a local geometrical constraint that regularizes the feature correspondence can establish accurate matches, which is beneficial for computing the spatial transformation.

We introduce an efficient scheme similar to the locally linear embedding algorithm [57,65,66] to impose a local geometric constraint such as a regularizer. The said scheme is proposed as a nonlinear dimensionality reduction method to preserve the local neighborhood structure in a low-dimensional manifold. First, the K nearest neighbors for each point in $\mathcal{X}$ are searched. An $N \times N$ weight matrix is denoted by $\mathbf{W}$, and $\mathbf{W}_{ij} = 0$ is enforced if $\mathbf{x}_j$ does not belong to the neighbor set of $\mathbf{x}_i$. Second, the reconstruction errors measured by the cost function (10) are minimized under a constraint, in which the rows of the weight matrix sum are equal to one: $\forall i, \sum_{j=1}^{N}\mathbf{W}_{ij} = 1$ with $\mathbf{W}_{ij}$ being nonnegative:

$$\mathcal{E}(\mathbf{W}) = \sum_{i=1}^{N}\|\mathbf{x}_i - \sum_{j=1}^{N}\mathbf{W}_{ij}\mathbf{x}_j\|^2. \tag{10}$$

The optimal weight $\mathbf{W}_{ij}$ can be obtained by solving an LS problem. Third, the local geometry of each model point after the transformation $\mathcal{T}$ can be preserved by minimizing a transforming cost term $\sum_{i=1}^{N}\|\mathcal{T}(\mathbf{x}_i) - \sum_{j=1}^{N}\mathbf{W}_{ij}\mathcal{T}(\mathbf{x}_j)\|^2$. Combining this term with $\mathcal{Q}(\mathcal{T})$ yields the following minimizing problem:

$$\begin{aligned} \Psi(\mathcal{T}) = &\frac{1}{2\sigma^2}\sum_{m=1}^{M}\sum_{n=1}^{N} p_{mn}\|\mathbf{y}_m - \mathcal{T}(\mathbf{x}_n)\|^2 \\ &+ \lambda\sum_{i=1}^{N}\|\mathcal{T}(\mathbf{x}_i) - \sum_{j=1}^{N}\mathbf{W}_{ij}\mathcal{T}(\mathbf{x}_j)\|^2. \end{aligned} \tag{11}$$

The said problem is composed of an empirical error term and a regularized transforming cost term with a parameter $\lambda > 0$ that controls the trade-off between them.

3.4. Estimation of spatial transformation

We next consider the modelling of the spatial transformation $\mathcal{T}$. The relationships between image pairs in image matching tasks, such as image stitching/mosaicing, are commonly modeled by rigid or affine transformations [66]. This scenario is appropriate because of the following factors: (i) these images are often captured at a long range (e.g., remote sensing images), and then they can be approximately considered as planar scenes; (ii) complex non-rigid models can easily lead to large error accumulation during constructing large panoramas. Moreover, the simple rigid or affine transformations are also more preferable in matching low-quality image pairs to avoid over-fitting because of a lack of reliable correspondence. However, the scenes in matching tasks such as object/shape recognition, medical image registration, and image retrieval often involve transformations that cannot be approximated by a simple linear model (e.g., different poses, non-rigid deformations, and irregular movements). A relatively complex non-rigid model is more preferable in this case [67]. The proposed formulation in the present study is independent of the transformation model, and it can handle most common geometric distortions in the image matching problem. We specify the transformation $\mathcal{T}$ for rigid, affine, and non-rigid cases separately to solve it utilizing Eq. (11).

Rigid matching: For rigid matching, we define the transformation as $\mathcal{T}(\mathbf{x}_n) = s\mathbf{R}\mathbf{x}_n + \mathbf{t}$, where $\mathbf{R}$ is a 2×2 rotation matrix, $\mathbf{t}$ is a 2×1 translation vector, and s is a scaling parameter. By considering that $\mathbf{R}$ is orthogonal and the constraint $\forall i,\ \sum_{j=1}^{N}\mathbf{W}_{ij} = 1$, the objective function in Eq. (11) becomes

$$\begin{aligned}\Psi(\mathbf{R},\mathbf{t},s) = &\frac{1}{2\sigma^2}\sum_{m=1}^{M}\sum_{n=1}^{N} p_{mn}\|\mathbf{y}_m - s\mathbf{R}\mathbf{x}_n - \mathbf{t}\|^2 \\ &+\lambda\sum_{i=1}^{N}\|s(\mathbf{x}_i - \sum_{j=1}^{N}\mathbf{W}_{ij}\mathbf{x}_j)\|^2, \qquad (12)\end{aligned}$$

$$\text{s.t. } \mathbf{R}^{\mathrm{T}}\mathbf{R} = \mathbf{I},\quad \det(\mathbf{R}) = 1.$$

Note that the first term is similar to the absolute orientation problem [18,68], which is defined as $\min\sum_{n=1}^{N}\|\mathbf{y}_n - s\mathbf{R}\mathbf{x}_n - \mathbf{t}\|^2$. The solutions of $\mathbf{t}$ and s are straightforward, while the solution of $\mathbf{R}$ is complicated due to the additional constraints. To obtain the closed form solution, we consider the following lemma [69].

Lemma 1. *Let $\mathbf{R}$ be an unknown $D\times D$ rotation matrix and $\mathbf{B}$ be a known $D\times D$ real square matrix. Let $\mathbf{USV}^{\mathrm{T}}$ be a Singular Value Decomposition (SVD) of $\mathbf{B}$, where $\mathbf{UU}^{\mathrm{T}} = \mathbf{VV}^{\mathrm{T}} = \mathbf{I}$ and $\mathbf{S} = d(s_i)$ is a diagonal matrix with $s_1 \geq, \ldots, \geq s_D \geq 0$. Then the optimal rotation matrix $\mathbf{R}$ that maximizes $tr(\mathbf{B}^{\mathrm{T}}\mathbf{R})$ is $\mathbf{R} = \mathbf{UDV}^{\mathrm{T}}$, where $\mathbf{D} = d(1,\ldots,1,\det(\mathbf{UV}^{\mathrm{T}}))$.*

To solve the rotation matrix $\mathbf{R}$, we rewrite the objective function (12) so that it has the form $\mathrm{tr}(\mathbf{B}^{\mathrm{T}}\mathbf{R})$. To this end, we first eliminate the translation parameter $\mathbf{t}$. Taking derivative of Ψ with respect to $\mathbf{t}$ and setting it to zero, we obtain:

$$\mathbf{t} = \frac{1}{M_{\mathbf{P}}}\mathbf{Y}^{\mathrm{T}}\mathbf{P}\mathbf{1} - \frac{1}{M_{\mathbf{P}}}s\mathbf{R}\mathbf{X}^{\mathrm{T}}\mathbf{P}^{\mathrm{T}}\mathbf{1} = \mu_{\mathbf{y}} - s\mathbf{R}\mu_{\mathbf{x}}, \qquad (13)$$

where $\mathbf{X} = (\mathbf{x}_1,\cdots,\mathbf{x}_N)^{\mathrm{T}}$, $\mathbf{Y} = (\mathbf{y}_1,\cdots,\mathbf{y}_M)^{\mathrm{T}}$, $\mathbf{P}$ is an $M\times N$ matrix with the (m,n)th element being p_{mn}, $\mu_{\mathbf{x}}$ and $\mu_{\mathbf{y}}$ are the mean vectors defined as:

$$\mu_{\mathbf{x}} = \frac{1}{M_{\mathbf{P}}}\mathbf{X}^{\mathrm{T}}\mathbf{P}^{\mathrm{T}}\mathbf{1},\quad \mu_{\mathbf{y}} = \frac{1}{M_{\mathbf{P}}}\mathbf{Y}^{\mathrm{T}}\mathbf{P}\mathbf{1}. \qquad (14)$$

Substituting $\mathbf{t}$ back into the objective function and omitting the terms that are independent of $\mathbf{R}$ and s, we obtain:

$$\begin{aligned}\Psi(\mathbf{R},s) = &\frac{1}{2\sigma^2}\mathrm{tr}(s^2\hat{\mathbf{X}}^{\mathrm{T}}\mathrm{d}(\mathbf{P}^{\mathrm{T}}\mathbf{1})\hat{\mathbf{X}} - 2s\hat{\mathbf{Y}}^{\mathrm{T}}\mathbf{P}\hat{\mathbf{X}}\mathbf{R}^{\mathrm{T}}) \\ &+\lambda\cdot\mathrm{tr}(s^2\mathbf{X}^{\mathrm{T}}\mathbf{Q}\mathbf{X}), \qquad (15)\end{aligned}$$

where $\hat{\mathbf{X}} = \mathbf{X} - \mathbf{1}\mu_{\mathbf{x}}^{\mathrm{T}}$ and $\hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{1}\mu_{\mathbf{y}}^{\mathrm{T}}$ are centered point matrices, $\mathbf{Q} = (\mathbf{I}-\mathbf{W})^{\mathrm{T}}(\mathbf{I}-\mathbf{W})$. Specifically, we consider the term related to $\mathbf{R}$, which has the form:

$$\Psi(\mathbf{R}) = -\frac{s}{\sigma^2}\mathrm{tr}((\hat{\mathbf{Y}}^{\mathrm{T}}\mathbf{P}\hat{\mathbf{X}})^{\mathrm{T}}\mathbf{R}). \qquad (16)$$

Therefore, by applying Lemma 1, the optimal $\mathbf{R}$ of the problem in Eq. (12) is given by

$$\mathbf{R} = \mathbf{UDV}^{\mathrm{T}}, \qquad (17)$$

where $\mathbf{U}$ and $\mathbf{V}$ can be obtained from $\mathbf{USV}^{\mathrm{T}} = \mathrm{svd}(\hat{\mathbf{Y}}^{\mathrm{T}}\mathbf{P}\hat{\mathbf{X}})$, and $\mathbf{D} = \mathrm{d}(1,\det(\mathbf{UV}^{\mathrm{T}}))$.

To solve the scaling parameter s, we equate the corresponding derivative of Eq. (15) to zero and obtain

$$s = \frac{\mathrm{tr}((\hat{\mathbf{Y}}^{\mathrm{T}}\mathbf{P}\hat{\mathbf{X}})^{\mathrm{T}}\mathbf{R})}{\mathrm{tr}(\hat{\mathbf{X}}^{\mathrm{T}}\mathrm{d}(\mathbf{P}^{\mathrm{T}}\mathbf{1})\hat{\mathbf{X}}) + 2\lambda\sigma^2\mathrm{tr}(\mathbf{X}^{\mathrm{T}}\mathbf{Q}\mathbf{X})}. \qquad (18)$$

Affine matching: Compared to the rigid case, affine matching is simpler since the optimization is unconstrained. We define the transformation as $\mathcal{T}(\mathbf{x}_n) = \mathbf{A}\mathbf{x}_n + \mathbf{t}$, where $\mathbf{A}$ is a 2×2 affine matrix, and $\mathbf{t}$ is a 2×1 translation vector. The objective function Eq. (11) then becomes

$$\begin{aligned}\Psi(\mathbf{A},\mathbf{t}) = &\frac{1}{2\sigma^2}\sum_{m=1}^{M}\sum_{n=1}^{N} p_{mn}\|\mathbf{y}_m - \mathbf{A}\mathbf{x}_n - \mathbf{t}\|^2 \\ &+\lambda\sum_{i=1}^{N}\|\mathbf{A}(\mathbf{x}_i - \sum_{j=1}^{N}\mathbf{W}_{ij}\mathbf{x}_j)\|^2. \qquad (19)\end{aligned}$$

The solution of $\mathbf{t}$ is similar to the rigid case. The solution of $\mathbf{A}$ can be obtained by directly taking the partial derivative of Ψ, setting it to zero, and solving the resulting linear system of equations. The optimal $\mathbf{t}$ and $\mathbf{A}$ are given by:

$$\mathbf{t} = \mu_{\mathbf{y}} - \mathbf{A}\mu_{\mathbf{x}}, \qquad (20)$$

$$\mathbf{A} = (\hat{\mathbf{Y}}^{\mathrm{T}}\mathbf{P}\hat{\mathbf{X}})(\hat{\mathbf{X}}^{\mathrm{T}}\mathrm{d}(\mathbf{P}^{\mathrm{T}}\mathbf{1})\hat{\mathbf{X}} + 2\lambda\sigma^2\mathbf{X}^{\mathrm{T}}\mathbf{Q}\mathbf{X})^{-1}. \qquad (21)$$

Non-rigid matching: We define the transformation $\mathcal{T}$ as the initial position plus a displacement function $\mathbf{f}$: $\mathcal{T}(\mathbf{x}) = \mathbf{x} + \mathbf{f}(\mathbf{x})$, where $\mathbf{f}$ is modeled by requiring it to lie within a specific functional space $\mathcal{H}$, namely a vector-valued RKHS [70] (associated with a particular kernel), as described in detail in the appendix. We define $\mathcal{H}$ by a matrix-valued kernel $\Gamma:\mathbb{R}^2\times\mathbb{R}^2\to\mathbb{R}^{2\times 2}$, and a diagonal Gaussian kernel $\Gamma(\mathbf{x}_i,\mathbf{x}_j) = \kappa(\mathbf{x}_i,\mathbf{x}_j)\cdot\mathbf{I} = e^{-\beta\|\mathbf{x}_i-\mathbf{x}_j\|^2}\cdot\mathbf{I}$ is chosen in this paper. Thus we have the following theorem.

Theorem 1. *The optimal solution of the objective function (11) in the non-rigid case is given by*

$$\mathcal{T}(\mathbf{x}) = \mathbf{x} + \mathbf{f}(\mathbf{x}) = \mathbf{x} + \sum_{n=1}^{N}\Gamma(\mathbf{x},\mathbf{x}_n)\mathbf{c}_n, \qquad (22)$$

with the coefficient set $\{\mathbf{c}_n : n\in\mathbb{N}_N\}$ determined by a linear system

$$(\mathrm{d}(\mathbf{P}^{\mathrm{T}}\mathbf{1}) + 2\lambda\sigma^2\mathbf{Q})\mathbf{\Gamma}\mathbf{C} = \mathbf{P}^{\mathrm{T}}\mathbf{Y} - (\mathrm{d}(\mathbf{P}^{\mathrm{T}}\mathbf{1}) + 2\lambda\sigma^2\mathbf{Q})\mathbf{X}, \qquad (23)$$

where $\mathbf{C} = (\mathbf{c}_1,\cdots,\mathbf{c}_N)^{\mathrm{T}}$, $\mathbf{\Gamma}\in\mathbb{R}^{N\times N}$ is the so-called Gram matrix with $\mathbf{\Gamma}_{ij} = \kappa(\mathbf{x}_i,\mathbf{x}_j) = e^{-\beta\|\mathbf{x}_i-\mathbf{x}_j\|^2}$.

Proof. For any given reproducing kernel Γ, we can define a unique RKHS $\mathcal{H}_N$ as in Eq. (31) in the appendix. Let $\mathcal{H}_N^{\perp}$ be a subspace of $\mathcal{H}$,

$$\mathcal{H}_N^{\perp} = \{\mathbf{f}\in\mathcal{H} : \mathbf{f}(\mathbf{x}_n) = 0, n\in\mathbb{N}_N\}. \qquad (24)$$

From the reproducing property, i.e. Remark 1, $\forall\mathbf{f}\in\mathcal{H}_N^{\perp}$

$$\left\langle \mathbf{f}, \sum_{n=1}^{N}\mathbf{\Gamma}(\cdot,\mathbf{x}_n)\mathbf{c}_n\right\rangle_{\mathcal{H}} = \sum_{n=1}^{N}\langle\mathbf{f}(\mathbf{x}_n),\mathbf{c}_n\rangle = 0. \qquad (25)$$

Thus $\mathcal{H}_N^{\perp}$ is the orthogonal complement of $\mathcal{H}_N$; then every $\mathbf{f} \in \mathcal{H}$ can be uniquely decomposed in components along and perpendicular to $\mathcal{H}_N$: $\mathbf{f} = \mathbf{f}_N + \mathbf{f}_N^{\perp}$, where $\mathbf{f}_N \in \mathcal{H}_N$ and $\mathbf{f}_N^{\perp} \in \mathcal{H}_N^{\perp}$. That is to say, $\forall \mathbf{f} \in \mathcal{H}$, we have $\mathbf{f}(\mathbf{x}_n) = \mathbf{f}_N(\mathbf{x}_n)$. Therefore, the optimal displacement function $\mathbf{f}$ comes from the space $\mathcal{H}_N$, and hence the optimal solution of the objective function (11) has the form (22).

To solve the coefficient set $\mathbf{C}$, we consider the terms of Ψ that are related to $\mathbf{C}$ and rewrite them in matrix form:

$$\begin{aligned}\Psi(\mathbf{C}) = &\frac{1}{2\sigma^2}\mathrm{tr}(\mathbf{C}^{\mathrm{T}}\boldsymbol{\Gamma}\mathrm{d}(\mathbf{P}^{\mathrm{T}}\mathbf{1})\boldsymbol{\Gamma}\mathbf{C} + 2\mathbf{C}^{\mathrm{T}}\boldsymbol{\Gamma}\mathrm{d}(\mathbf{P}^{\mathrm{T}}\mathbf{1})\mathbf{X} \\ &- 2\mathbf{C}^{\mathrm{T}}\boldsymbol{\Gamma}\mathbf{P}^{\mathrm{T}}\mathbf{Y}) + \lambda\mathrm{tr}(\mathbf{C}^{\mathrm{T}}\boldsymbol{\Gamma}\mathbf{Q}\boldsymbol{\Gamma}\mathbf{C} + 2\mathbf{C}^{\mathrm{T}}\boldsymbol{\Gamma}\mathbf{Q}\mathbf{X}). \end{aligned} \tag{26}$$

Taking derivative of Eq. (26) with respect to $\mathbf{C}$ and setting it to zero, we obtain the linear system in Eq. (23). Thus the coefficient set $\{\mathbf{c}_n : n \in \mathbb{N}_N\}$ of the optimal solution is determined by the linear system (23). □

Fast Implementation. The non-rigid model requires at least $O(N^3)$ computational complexity because it requires solving the linear system (23), which can cause a significant computational problem in the case of large-scale feature sets. Consequently, we adopt a sparse approximation and randomly select only a subset of size L input points $\{\tilde{\mathbf{x}}_l\}_{l=1}^{L}$ to have non-zero coefficients in the solution expansion (Eq. (22)). This approach follows [71,72] who determined that this approximation works properly, and that simply selecting a random subset of the input points in this manner performs equally with more sophisticated and time-consuming methods. Thus, we seek a solution as follows:

$$\mathbf{f}(\mathbf{x}) = \sum_{l=1}^{L} \Gamma(\mathbf{x}, \tilde{\mathbf{x}}_l)\mathbf{c}_l. \tag{27}$$

The selected point set $\{\tilde{\mathbf{x}}_l\}_{l=1}^{L}$ is somewhat analogous to control points. The linear system (23) becomes the following by utilizing the sparse approximation:

$$\begin{aligned}&\mathbf{E}^{\mathrm{T}}(\mathrm{d}(\mathbf{P}^{\mathrm{T}}\mathbf{1}) + 2\lambda\sigma^2\mathbf{Q})\mathbf{E}\mathbf{C}^s \\ &= \mathbf{E}^{\mathrm{T}}\mathbf{P}^{\mathrm{T}}\mathbf{Y} - \mathbf{E}^{\mathrm{T}}(\mathrm{d}(\mathbf{P}^{\mathrm{T}}\mathbf{1}) + 2\lambda\sigma^2\mathbf{Q})\mathbf{X}, \end{aligned} \tag{28}$$

where the coefficient matrix $\mathbf{C}^s = (\mathbf{c}_1, \cdots, \mathbf{c}_L)^{\mathrm{T}} \in \mathbb{R}^{L\times 2}$, and $\mathbf{E} \in \mathbb{R}^{N\times L}$ with $\mathbf{E}_{ij} = \kappa(\mathbf{x}_i, \tilde{\mathbf{x}}_j) = e^{-\beta\|\mathbf{x}_i - \tilde{\mathbf{x}}_j\|^2}$.

We call our proposed matching algorithm as FG-GMM, which is summarized in Algorithm 1.

3.5. Computational complexity

The time complexity is near $O((K+N)\log N)$ by utilizing the k-d tree [73] to search the K nearest neighbors for each point in $\mathbf{X}$. The time complexity of obtaining the weight matrix $\mathbf{W}$ is $O(K^3N)$ according to Eq. (10) because each row of $\mathbf{W}$ can be solved separately with $O(K^3)$ time complexity. The time complexities of solving the transformations for the rigid and affine cases are both $O(KN+MN)$, so the total time complexities for rigid and affine matching are both $O(K^3N + MN + N\log N)$. The space complexities for rigid and affine matching are both $O(KN+MN)$ because of the memory requirements for storing the weight matrix $\mathbf{W}$ and posterior distribution matrix $\mathbf{P}$.

The time complexity of solving the linear system (23) is $O(KN + MN + N^3)$ for the non-rigid case, so the total complexity can be written as $O(K^3N + MN + N^3)$. The space complexity scales are $O(KN + MN + N^2)$ because of the memory requirements for storing the Gram matrix $\boldsymbol{\Gamma}$, as well as $\mathbf{W}$ and $\mathbf{P}$. The time complexity to solve the linear system (28) decreases to $O(L^3 + L^2N + LMN + KN))$ by applying the sparse approximation. Therefore, the total time complexity is $O(L^3 + L^2N + LMN + K^3N + N\log N)$. The space complexity decreases to $O(MN + KN + LN)$ because of the memory requirements for storing $\mathbf{P}$, $\mathbf{E}$ and $\mathbf{W}$.

Algorithm 1: The proposed FG-GMM algorithm.

Input: An image pair, parameters t, τ, K, λ, η, β, L
Output: Correspondence set $\mathcal{I}$, spatial transformation $\mathcal{T}$

```
Extract two feature sets using SIFT: {X, S_x}, {Y, S_y};
Match X and Y using S_x and S_y with a distance ratio threshold t;
Assign the membership probability π_mn;
switch transformation model do
    case rigid
        Initialize R = I, t = 0, s = 1;
    case affine
        Initialize A = I, t = 0;
    case non-rigid
        Initialize C = 0;
        Construct matrix Γ or E using definition of Γ;
endsw
Set a to the volume of the output space;
Initialize γ, p_mn = π_mn, σ^2 (using (8);
Search the K nearest neighbors for each point in X;
Compute W by minimizing the cost function (10);
repeat
    E-step:
    Update P by Eqs.~(5) and (6);
    M-step:
    switch transformation model do
        case rigid
            Compute X̂ and Ŷ according to Eq.~(14);
            Compute USV^T = svd(Ŷ^T P X̂);
            Update R, s, t by Eqs.~(17), (18) and (13);
        case affine
            Compute X̂ and Ŷ according to Eq.~(14);
            Update A, t by Eqs.~(21) and (20);
        case non-rigid
            Update C based on linear system (23) or (28);
    endsw
    Update σ^2 and γ by Eqs. (8) and (7);
until Q converges;
Correspondence set I is determined by Eq. (9);
Transformation T is obtained by estimated parameters.
```

Table 1
Computational complexities of our FG-GMM algorithm.

	Rigid	Affine	Non-rigid	Non-rigid (fast)
Time	$K^3N + N^2$	$K^3N + N^2$	$K^3N + N^3$	$K^3N + LN^2$
Space	N^2	N^2	N^2	N^2

We generally have $M \approx N$ and $M, N \gg L, K$. Thus, the complexities can be simplified as listed in Table 1. The time and space complexities are all quadratic with respect to the scale of the provided correspondence set, where our fast implementation can lower the time complexity from cubic to quadratic in the non-rigid case. This condition is significant for large-scale problems.

3.6. Implementation details

The performance of feature matching algorithms largely depends on the coordinate system where feature points are expressed. We utilize data normalization to control this condition. We specifically perform linear rescaling to allow the spatial positions of the two feature point sets to have zero mean and unit variance. The constant a of the uniform distribution in Eq. (2) is the area of the second image (i.e., the range of $\mathbf{y}_m$), which must

be set according to the data normalization. The experiments are performed on a laptop with 2.5 GHz Intel Core CPU, 8 GB memory, and MATLAB code, and all the codes were realized without special optimization such as parallel computing or streaming single instruction multiple data extensions.

Parameter setting. Eight parameters are primarily adopted in our method: t, τ, K, λ, η, γ, β and L. Parameter t is the distance ratio threshold utilized to establish the initial correspondence based on feature descriptors. Parameter τ is applied to assign the membership probability π_{mn}, which is the confidence of a known correspondence. Parameter K controls the number of nearest neighbors for linear reconstruction. Parameter λ controls the influence of the local geometrical constraint on the transformation $\mathcal{T}$. Parameter η is a threshold, which is adopted to decide on the correctness of a correspondence. Parameter γ reflects our initial assumption on the inlier amount in the correspondence sets. Parameters β and L are employed in our non-rigid matching algorithm, where the former determines how wide the range of interaction between feature points, whereas the latter is the required number of control points for sparse approximation. We tune the parameters on several image pairs in our experiments to attain their best performance and remain unchanged in all other experiments. We determined that many of the parameters can accomplish suitable performance at different values, such as t, τ, λ, η, and γ. Specifically, we set $t = 0.8$, $\tau = 0.9$, $K = 15$, $\lambda = 1000$, $\eta = 0.5$, $\gamma = 0.9$, $\beta = 0.1$ and $L = 15$, throughout this paper.

4. Experimental results

We test the performance of our proposed algorithm on real images. The open source VLFEAT toolbox [74] is employed to determine the putative correspondence of SIFT [8]. SIFT is a well-known method to detect and describe local features in images that can be applied to perform reliable matching between different views of an object or scene. This method is invariant to uniform scaling and orientation, as well as partially invariant to affine distortion and illumination changes. We compute the ground truth transformation by utilizing manually selected feature correspondence instances of LS fitting. We then apply the same overlap error criterion in [1] to determine the match correctness. Experimental results are evaluated by precision and the number of identified correct matches. Precision is defined as the ratio of the identified correct match number and the preserved match number. Thus,

$$Precision = \frac{\#\text{identified correct matches}}{\#\text{preserved matches}}. \tag{29}$$

We compare our FG-GMM algorithm with five other state-of-the-art matching algorithms, such as RANSAC [10], ICF [13], VFC [1], LFGC [46] and CPD [18]. These five algorithms are chosen due to that they are representatives of five different types of matching methods. In particular, RANSAC is a classic resampling method, ICF is a regression method, VFC is an interpolation method based on a slow-and-smooth prior, LFGC is a deep learning-based method, and CPD is correspondence matrix-based method. We implement ICF and tune all parameters accordingly to determine the optimal settings. The other four methods are implemented by adopting publicly available codes. The parameters of the six methods are fixed throughout all the experiments. The following sections present the experiments on rigid, affine, and non-rigid image pairs, respectively.

4.1. Results on rigid image pairs

We first test the capability of our FG-GMM in handling rigid deformation. We apply a dataset that consists of 65 image pairs categorized as follows: color-infrared aerial photograph image pairs with small overlap areas and SPOT image pairs that represent the same area captured at different times. The images sizes are from 1391×1374 to 3086×2865, which were captured over Eastern Illinois, USA (from the Erdas example data[1]) and Shanghai, China.

We first provide intuitive results of our FG-GMM on two typical image pairs as presented in Fig. 1. The first image pair contains a small overlap, whereas the second involves extremely local illumination changes. Therefore, establishing reliable feature correspondence is relatively challenging. Our results are presented at the bottom row, which demonstrates that our FG-GMM can produce many correct feature matches. These results are beneficial for many remote sensing applications, such as image mosaic and change detection. We also present the results of RANSAC [10], a classic and widely adopted method for image matching, at the top row for a performance comparison. We can observe that the RANSAC results are acceptable because the rigid transformation is relatively easy to estimate. However, RANSAC operates on a set of putative correspondence, which suffers from missing true correspondence. Hence, the numbers of identified correct matches are much smaller compared with our FG-GMM.[2]

The statistics of the precision and identified correct match numbers of RANSAC, ICF, VFC, LFGC, CPD and our FG-GMM on the entire dataset are shown in Fig. 2. All methods can generate accurate matches on most image pairs because the rigid transformation is relatively simple. This condition can be observed from the precision curves, where all methods have 100% precision on most of the image pairs. However, when the outlier percentage in the putative correspondence is extremely large for RANSAC, ICF, VFC and LFGC, the transformation estimate can fail. By contrast, our FG-GMM is more robust. Although its precisions are slightly lower on several image pairs, its average precision is the largest and it can also identify more correct matches. CPD fails on most image pairs because it ignores the feature descriptor information. Hence, it severely degrades in the case of a large percentage of false matches, which frequently occurs in low overlap or low-quality images. This scenario also justifies the reasonability of incorporating local image features in our formulation.

The average runtimes of the six methods on the test data are listed in Table 2, where we have excluded the cost of SIFT feature extraction for all methods. LFGC is the most efficient method which requires only dozens of milliseconds. RANSAC is also very efficient because the few parameters involved in a rigid model significantly decreases the iteration number. Nevertheless, the performance of our FG-GMM is still acceptable, which is similar to ICF and VFC. However, CPD is slightly inefficient because it requires repeated iterations to converge in the case of badly degraded data.

4.2. Results on affine image pairs

We then test the capability of our FG-GMM in handling affine deformation. Thus, we conduct experiments on retinal pairs with different imaging modalities. Given that such image pairs are commonly obtained with similar viewpoints, an affine model is more preferable[3] Registering multimodal retinal images is a relatively challenging task because of the large homogeneous texture-

[1] The dataset is available at: http://download.intergraph.com/downloads/erdas-imagine-2013-2014-example-data.

[2] A possible solution to retain more true matches is to enlarge the size of the putative set; however, this can rapidly decrease the correct match percentage in the putative set and severely degrade the matching performance [15].

[3] Note that the affine model cannot accurately approximate the transformation here because the feature points are generally located on a hemisphere, such as the eyeball. However, the extracted feature points generally contain a few suitable matches and many outliers because of the low-quality multimodal retinal images. Therefore, complex models such as non-rigid functions can easily become trapped in overfitting.

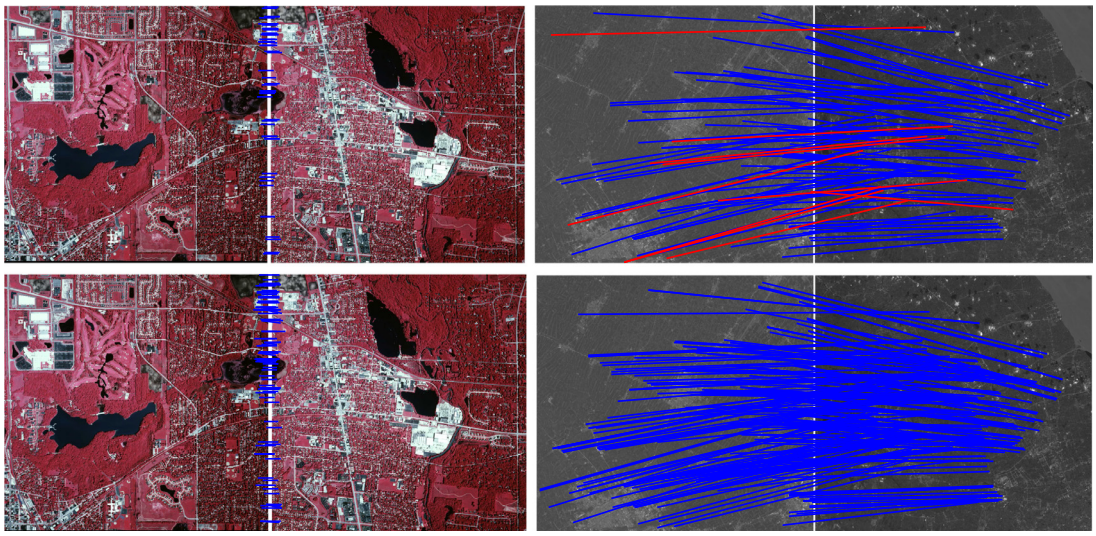

Fig. 1. Matching results of RANSAC [10] (top) and our FG-GMM (bottom) on two typical remote sensing image pairs.

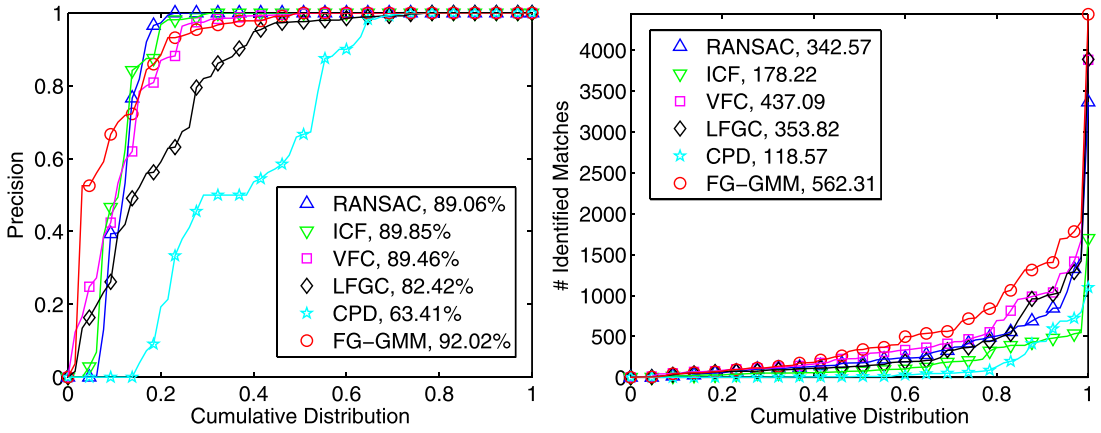

Fig. 2. Precision (left) and identified correct match number (right) of RANSAC [10], ICF [13], VFC [1], CPD [18] and our FG-GMM with respect to the cumulative distribution on the 65 remote sensing image pairs. The numbers in the boxes represent the average precision and the average number of identified correct matches. In addition, a point on the curve with coordinate (x, y) denotes that there are $100^{*}x$ percents of image pairs which have precision or identified match numbers no more than y.

Table 2
Average Run Times of different methods on the 65 remote sensing image pairs. Bold indicates the best result.

	RANSAC [10]	ICF [13]	VFC [1]	LFGC [46]	CPD [18]	FG-GMM
Time (s)	0.26	1.52	2.04	**0.04**	6.81	2.09

less/nonvascular regions, non-uniform intensity/contrast distributions, and different pathologies that cause degradation. We select two groups of retinal images that involve red-free and fundus autofluorescence as applied in [75]. The images have a resolution range of 640×480 to 1280×960 pixels. We construct 200 image pairs from these images for quantitative evaluation.

We again initially provide intuitive results of our FG-GMM on two typical image pairs as presented in Fig. 3. Results show that our method can generate many suitable matches, even with low-quality images that have extreme noise and pathology that cause degradation in the right pair. We also present the RANSAC results [10] at the top row for a performance comparison. We can observe that the identified feature matches are much fewer; the feature matching procedure of RANSAC completely fails in the second pair because of the extreme noise, which results in few true matches and high percentage of outliers.

The precision statistics and identified correct match numbers of our FG-GMM and the five other methods on the entire dataset are shown in Fig. 4. Our method can evidently produce the best results for both evaluation criteria. Our curves are almost consistently above those for all the other methods, and the average correct match number of our method is about twice to thrice those of RANSAC and VFC. The sufficient correct matches can guarantee the accuracy of the transformation estimation. VFC and LFGC have a stable performance, and they perform slightly better than RANSAC for the averages of the evaluation criteria. CPD again completely fails on most image pairs, which generated the worst precision and worst identified correct match number.

The average runtimes (excluding the cost of SIFT feature extraction) of the six methods on the test data are shown in Table 3. The performance has a similar trend to that on the rigid dataset. Note that the runtimes on this dataset are lowered for most methods

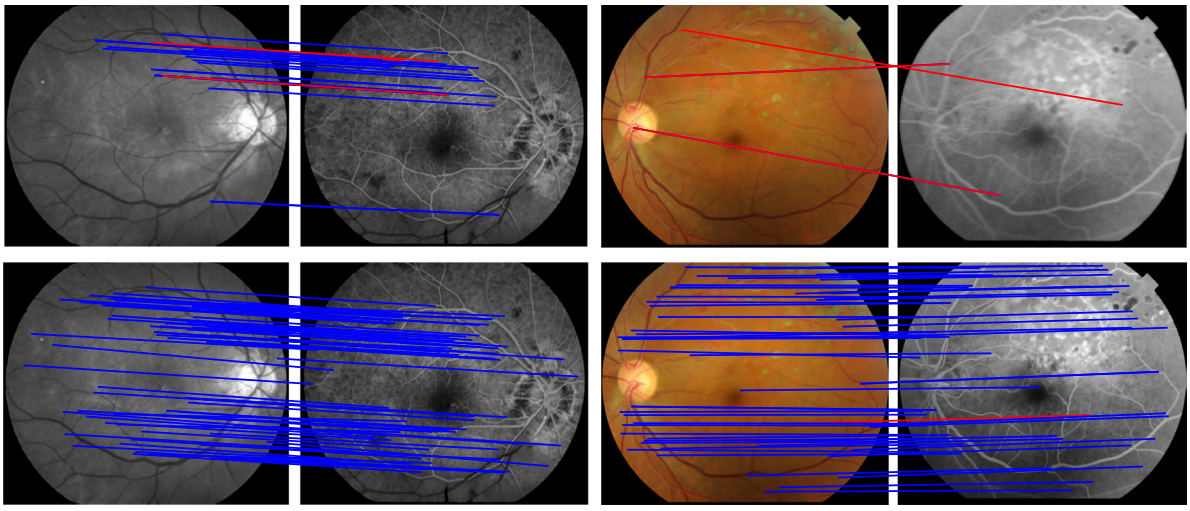

Fig. 3. Matching results of RANSAC [10] (top) and our FG-GMM (bottom) on two typical multimodal retinal image pairs.

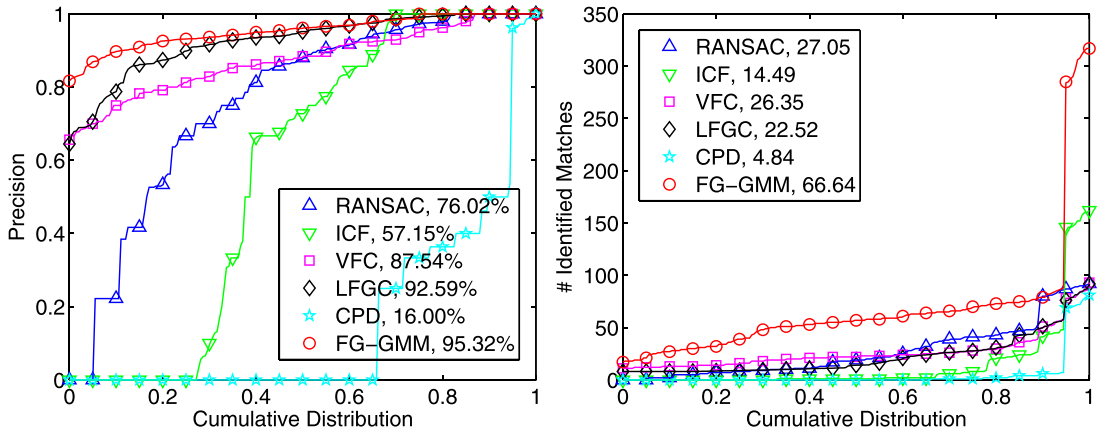

Fig. 4. Precision (left) and identified correct match number (right) of RANSAC [10], ICF [13], VFC [1], LFGC [46], CPD [18] and our FG-GMM with respect to the cumulative distribution on the 200 multimodal retinal image pairs.

Table 3
Average Run Times of different methods on the 200 multimodal retinal image pairs. Bold indicates the best result.

	RANSAC [10]	ICF [13]	VFC [1]	LFGC [46]	CPD [18]	FG-GMM
Time (s)	0.49	0.27	0.57	**0.02**	5.12	1.06

because the feature numbers extracted from the low-quality retinal images are significantly decreased.

4.3. Results on non-rigid image pairs

The capability of our FG-GMM in handling non-rigid deformation is tested in this section. We conduct experiments on the dataset of Mikolajczyk et al. [76], which contains 40 image pairs either of planar scenes or captured by camera in a fixed position during acquisition. Therefore, these images always obey homography. The ground truth homographies are supplied by the dataset. The dataset contains eight folders, in which the images involve viewpoint change, scale and rotation, image blur, light change, and JPEG compression. Several examples are shown in Fig. 5.

The statistics of the precision and the identified correct match number for RANSAC, ICF, VFC, LFGC and our FG-GMM with fast implementation are shown in Fig. 6. We do not report the CPD results in this section because it fails on most of the image pairs. The results show that our FG-GMM has the best average precision (95.21%) and largest average identified correct match number (922.34), followed by VFC and RANSAC. Note that RANSAC suitably works on this dataset because the image transformation satisfies a parametric model such as homography.

Table 4
Average Run Times of different methods on the dataset of Mikolajczyk et al. [76]. Bold indicates the best result.

	RANSAC [10]	ICF [13]	VFC [1]	LFGC [46]	FG-GMM
Time (s)	3.54	2.12	5.47	**0.19**	16.35

Table 4 provides the average runtimes (excluding the cost of SIFT feature extraction) of the five methods on the test data. The performance in this case has a similar trend, and the runtimes of all methods are relatively longer compared with the results of the other test data. This scenario is due to the average number of extracted SIFT features for an image in this dataset being approximately 2,630, which must be a large-scale problem for image matching. We also report the original version of our FG-GMM

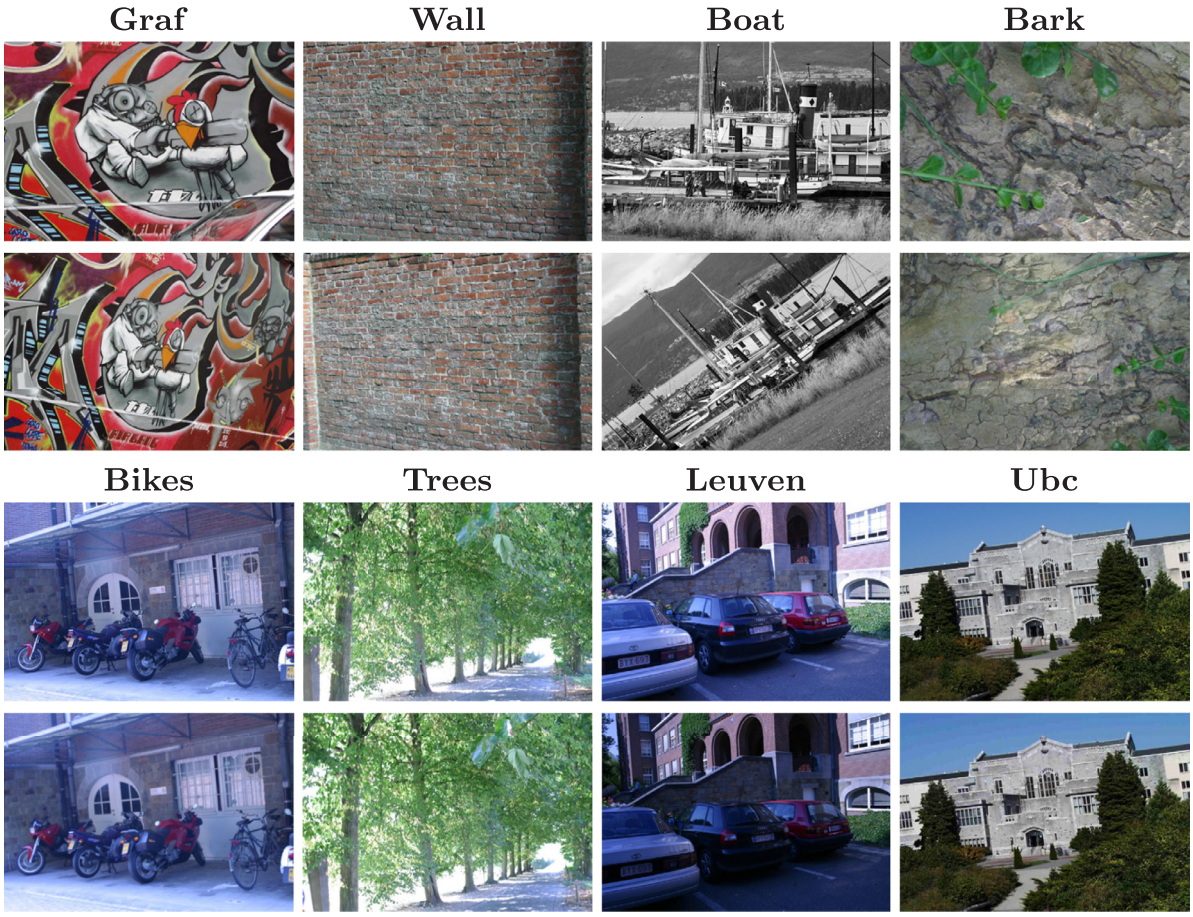

Fig. 5. Examples of images in the dataset [76].

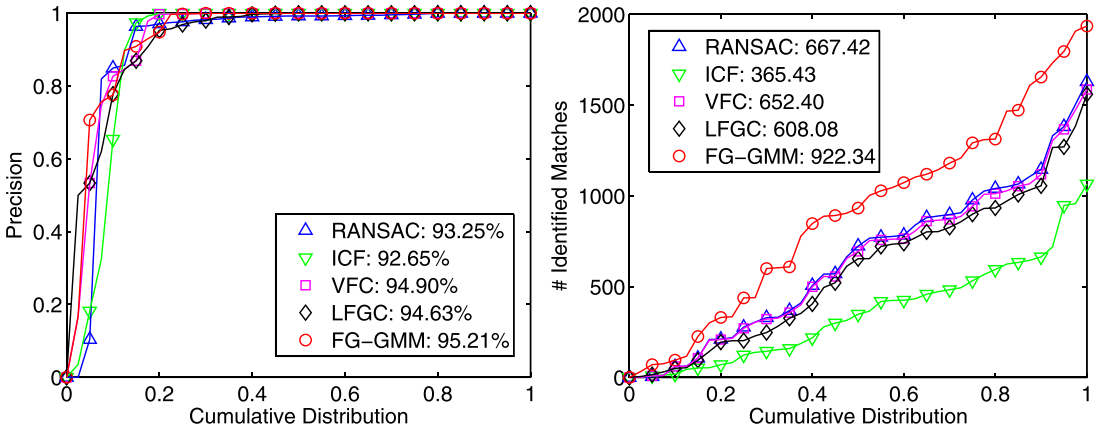

Fig. 6. Precision (left) and identified correct match number (right) of RANSAC [10], ICF [13], VFC [1], LFGC [46] and our FG-GMM with respect to the cumulative distribution on the dataset of Mikolajczyk et al. [76].

on this dataset. The average precision and identified correct match number are approximately 95.84% and 920.15, respectively, which are similar to the results of the fast FG-GMM. However, the average runtime increases to approximately 34.23 s per image pair. Therefore, the fast implementation can significantly lower the computational complexity without sacrificing accuracy.

Note that the homography is essentially a linear function, which is one of the simplest forms of non-rigid transformation. However, obtaining the ground truth transformations of image pairs that involve deformable objects with non-rigid/nonlinear motions is unlikely because the transformation model is unknown and generally complex. Therefore, establishing the ground truth feature matches is difficult. No non-rigid image dataset that contains ground truth feature matches is publicly available to the best of our knowledge. We also conduct experiments on several typical image pairs that involving deformable objects to test our FG-GMM in such a challenging case, where the match correctness is determined by manual checking. The intuitive performance of our FG-GMM is shown in Fig. 7, where the precision values are 97.85%, 99.07% and 99.29%. The motion fields related to the three image pairs are provided

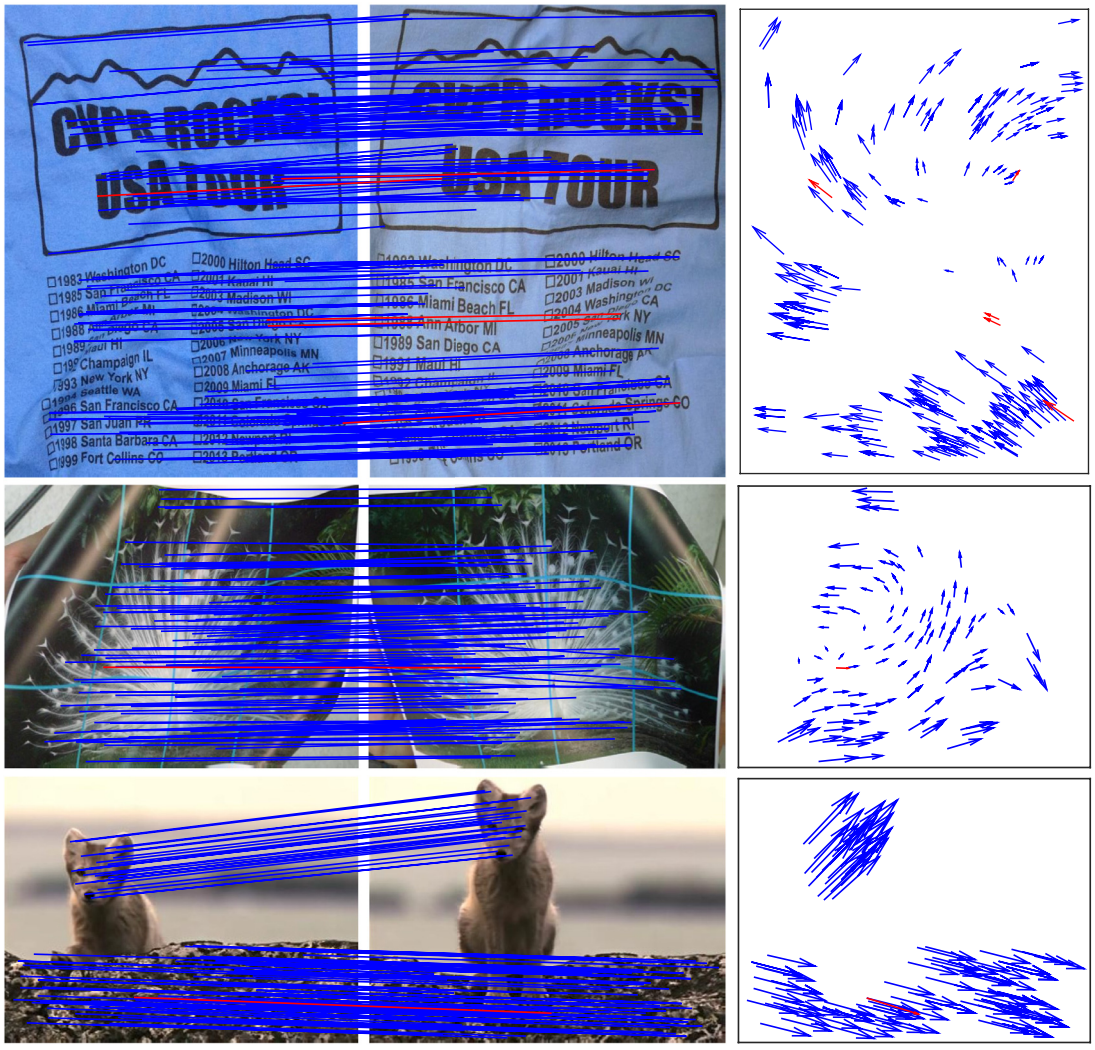

Fig. 7. Results of FG-GMM on three typical image pairs (e.g., *T-shirt*, *Peacock* and *Fox*) involving deformable objects. The precisions and identified correct match numbers are (97.85%, 288), (99.07%, 107), and (99.29%, 139). Blue and red lines/arrows indicate correct and false matches, respectively. The right column is the corresponding motion fields, where the head and tail of each arrow correspond to the positions of feature points in two images. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Table 5
Comparison of precisions and preserved correct match numbers on image pairs involving deformable objects.

	T-shirt	*Peacock*	*Fox*
RANSAC [10]	(89.21%, 124)	(96.61%, 57)	(97.33%, 73)
ICF [13]	(95.00%, 76)	(97.67%, 42)	(98.70%, 76)
VFC [1]	(96.18%, 126)	(98.44%, 63)	(98.94%, 93)
LFGC [46]	(95.04%, 115)	(96.72%, 59)	(96.74%, 89)
CPD [18]	(90.00%, 45)	(96.92%, 63)	(95.31%, 61)
FG-GMM	(**97.85%**, **288**)	(**99.07%**, **107**)	(**99.29%**, **139**)

at the right column of Fig. 7. We can observe the relatively large degree of the non-rigid deformation, where different parts of the scenes have different motion manners. However, the variation in the motion field is slow-and-smooth, which guarantees that our method suitably works in this case.

The results of the five other state-of-the-art methods are shown in Table 5. Our FG-GMM evidently has consistently better precisions and can identify more true matches. RANSAC has satisfying precision values because it can identify a majority of the putative correspondence instances that satisfy a geometric constraint. However, the geometric constraint is based on a parametric model (e.g., homography in our experiments), which may not approximate the real non-rigid deformation properly with a complex deformation. This scenario can be observed from the *T-shirt* pair with a larger degree of deformation, in which RANSAC has a much lower precision. By contrast, the deep learning-based method LFGC and the two non-parametric-based methods ICF and VFC have better precision values. Our evaluation again shows that the CPD method completely fails on all the three pairs (i.e., we omit the detailed results in this section for clarity). We alternatively test CPD on two feature sets obtained from the putative sets, such as those in RANSAC, ICF, and VFC. The results are listed in Table 5, which are still relatively unsuitable compared with VFC. In particular, VFC only needs to remove false matches from a putative set, and it has initial correspondence information unlike in CPD, which can solve the matching problem.

4.4. Ablation experiments

Note that in our model, we use the semi-supervised EM algorithm rather than the standard EM algorithm to optimize the

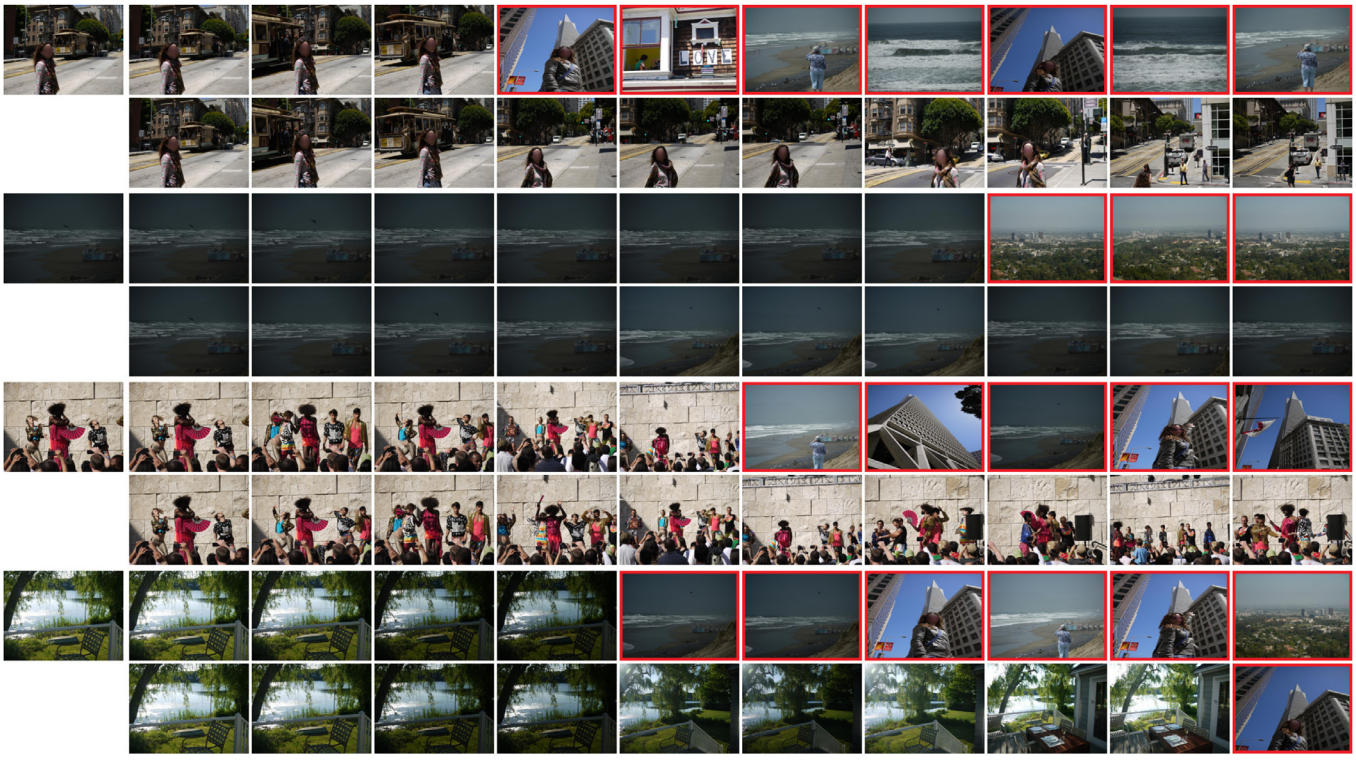

Fig. 8. Schematically illustration of the image retrieval results in every two rows. For each group of results, the left image is the query image, and the rest 10 images are the 10 most similar images retrieved by ICF [13] (the top row) and our FG-GMM (the bottom row) which is listed in descending order. Red boxes indicate false results.

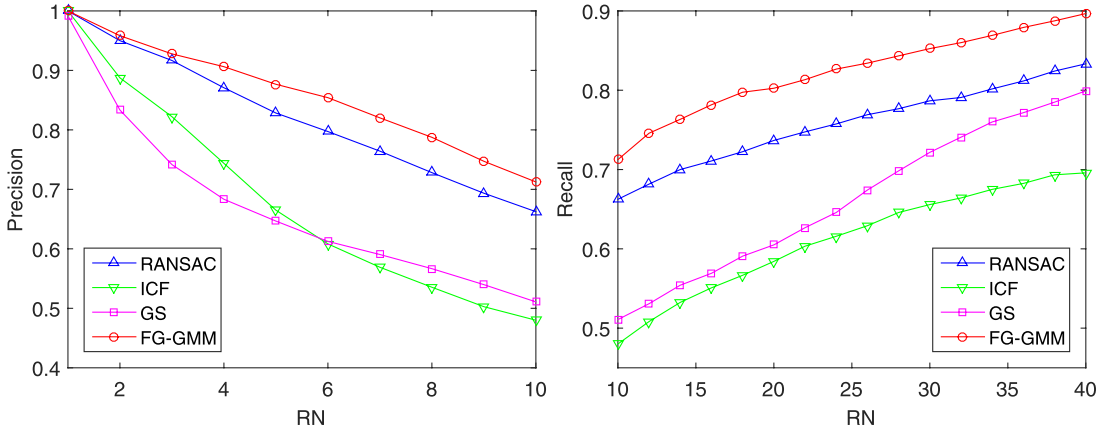

Fig. 9. Precisions (left) and recalls (right) of RANSAC [10], ICF [13], GS [45] and our FG-GMM with respect to *RN*, i.e., the required number of images to be retrieved for a given image.

likelihood in Eq. (4), which can avoid or alleviate getting trapped into the local minima during the EM iteration, and hence promotes the matching performance. In addition, we also use the local geometric constraint to regularize the solution of spatial transformation $\mathcal{T}$, so that we can obtain a stable and meaningful solution when there exists noise, outliers, or non-rigid deformations. In this section, we conduct ablation experiments to verify the effectiveness of the semi-supervised EM and local geometric constraint. To this end, we consider the following two scenarios.

On the one hand, we use the original EM to optimize the likelihood in Eq. (4), and fix all the other settings as the same as our FG-GMM. In particular, we update the posterior distribution p_{mn} using only Eq. (6), without using Eq. (5) to serve as anchors. On the other hand, we set $\lambda = 0$ in Eq. (11) and fix all the other settings as the same as our FG-GMM, which means to abandon the local geometric constraint. We test the two scenarios on all the three datasets, including the rigid, affine and non-rigid datasets. The average precisions and identified correct match numbers of two scenarios on the three datasets are reported in Table 6. In addition, we also provide the statistics of our FG-GMM for comparison.

Table 6
Ablation study of our FG-GMM. The pairs in the table are the average precisions and average preserved correct match numbers of three scenarios on the three datasets in Figs. 2, 4 and 6. Scenario 1 means using the original EM instead of semi-supervised EM, scenario 2 means without using the local geometric constraint, and scenario 3 denotes our FG-GMM.

	Rigid	*Affine*	*Non-rigid*
Scenario 1	(90.36%, 546.57)	(93.94%, 62.79)	(92.33%, 898.43)
Scenario 2	(91.13%, 551.28)	(88.70%, 57.31)	(81.56%, 832.68)
Scenario 3	(**92.02%**, **562.31**)	(**95.32%**, **66.64**)	(**95.21%**, **922.34**)

From the results, we see that both the semi-supervised EM and local geometric constraint play an important role in improving the matching performance. In particular, for the local geometric constraint, its importance becomes evident as the transformation becomes complex.

4.5. Application to near-duplicate image retrieval

Finally, we test our FG-GMM for near-duplicate image retrieval and compare it with RANSAC [10], ICF [13], and GS [45] on the California-ND dataset [77]. We select all of the 12 classes that have 10 or more images. We also randomly select 10 images for evaluation in each class. Therefore, the test data contains 120 images that generate a total of 7,260 image pairs. We run the matching algorithms on all 7,260 image pairs and utilize the number of preserved matches as the similarity between image pairs. We then return a ranked list for a provided image according to its similarities with every other image in the dataset. The performance is also characterized by precision and recall, where precision is defined as the ratio of the retrieved correct image number and total retrieved image number. The recall is defined as the ratio of the retrieved correct image number and total correct image number. We denote the required image number to be retrieved for a provided image as RN. The precision is valid for $RN \leq 10$, and the recall is valid for $RN \geq 10$ because each class contains 10 images. We utilize our FG-GMM with the non-rigid model for testing because the images in this dataset involve deformable objects with non-rigid motions.

Several typical retrieval results are shown in Fig. 8 and compared with ICF to provide some intuitive performance analysis of our FG-GMM. Results show that ICF can only retrieve several correct images. The matching score of ICF becomes severely degraded when an image pair of a similar scene involves large viewpoint or pose changes. Hence, it fails to retrieve the image. By contrast, our FG-GMM is more robust to these distortions, and it can retrieve most of the correct images at the top of the ranking list.

The statistic retrieval results of the four methods in the dataset are presented in Fig. 9. Our FG-GMM evidently outperforms all other methods and obtains the best precision and recall, followed by RANSAC. Specifically, the average retrieved correct image numbers of RANSAC, ICF, GS, and our FG-GMM for $RN = 10$ are approximately 6.63, 4.80, 5.11 and 7.13, respectively.

We also measure the retrieval performance of the so-called bulls-eye score [78], which is defined as the ratio of the total number of correct images among the 20 most similar images to the highest possible number (i.e., 10). The best possible rate is 100%. The bulls-eye scores of RANSAC, ICF, GS, and our FG-GMM are approximately 73.67%, 58.42%, 60.50% and 80.25%, respectively. Our method again evidently showcases the best performance.

5. Conclusion

This paper reports on a proposed FG-GMM for robust image matching that undergoes either rigid or non-rigid transformation. A key characteristic of our approach is that it can preserve more true feature matches and incorporate local feature information during matching. The semi-supervised EM algorithm is introduced to solve the problem, which is formulated as maximum-likelihood estimation. We also provide an efficient implementation of our method to decrease the computational complexity without significantly lowering the matching quality.

Experiments on publicly available datasets demonstrate that our approach generates superior results compared with those of other state-of-the-art methods. On the one hand, our FG-GMM seeks feature correspondences from the original feature sets rather than removing outliers from a putative match set, which is able to avoid loss of 'true feature matches' and hence generates more feature matches. On the other hand, the feature guided strategy ensures good initialization of our method and hence it can converge to a satisfying result even when the data is badly degraded, for example, there are very few true matches or a great many false matches in the data. Therefore, our method has particular advantages in matching low-quality (e.g., strong noise or low resolution) images, small overlap images, images with complex non-rigid deformations, etc. We also have shown that FG-GMM is beneficial for a real-world visual task such as content-based image retrieval.

Our method in this paper aims to establish accurate correspondences between two sets of feature points, where each point is associated with a local image descriptor. This image descriptor is used to assign the membership probability of the GMM in our formulation, and hence plays a pivotal role in our feature-guided matching. In fact, our method can also be used to address the point set registration problem where each point only consists of a spatial position. This is because that we can construct a descriptor for each point based on its neighborhood structure with respect to other points in the point set. For example, the shape context feature [24] is a preferred descriptor in such case to achieve the goal of feature-guided matching. In addition, our method is not influenced by the dimension of the input data, and hence it also applies to 3D matching problem such as 3D point cloud registration. We leave these interesting extensions for future work.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant No. 61773295, in part by the Beijing Advanced Innovation Center for Intelligent Robots and Systems under Grant 2016IRS15, and in part by the 111 project under Grant B17040.

Appendix. Vector-valued reproducing kernel Hilbert space

We review the basic theory of vector-valued reproducing kernel Hilbert space, and for further details and references we refer to [70,79].

Let $\mathcal{X}$ be a set, for example, $\mathcal{X} \subseteq \mathbb{R}^P$, $\mathcal{Y}$ a real Hilbert space with inner product (norm) $\langle \cdot, \cdot \rangle$, $(\|\cdot\|)$, for example, $\mathcal{Y} \subseteq \mathbb{R}^D$, and $\mathcal{H}$ a Hilbert space with inner product (norm) $\langle \cdot, \cdot \rangle_{\mathcal{H}}$, $(\|\cdot\|_{\mathcal{H}})$, where $P = D = 2$ in our problem. Note that a norm can be induced by an inner product, for example, $\forall \mathbf{f} \in \mathcal{H}$, $\|\mathbf{f}\|_{\mathcal{H}} = \sqrt{\langle \mathbf{f}, \mathbf{f} \rangle_{\mathcal{H}}}$. And a Hilbert space is a real or complex inner product space that is also a complete metric space with respect to the distance function induced by the inner product. Thus a vector-valued RKHS can be defined as follows.

Definition 1. A Hilbert space $\mathcal{H}$ is an RKHS if the evaluation maps $ev_{\mathbf{x}} : \mathcal{H} \to \mathcal{Y}$ (i.e., $ev_{\mathbf{x}}(\mathbf{f}) = \mathbf{f}(\mathbf{x})$) are bounded, i.e., if $\forall \mathbf{x} \in \mathcal{X}$ there exists a positive constant $C_{\mathbf{x}}$ such that

$$\|ev_{\mathbf{x}}(\mathbf{f})\| = \|\mathbf{f}(\mathbf{x})\| \leq C_{\mathbf{x}} \|\mathbf{f}\|_{\mathcal{H}}, \quad \forall \mathbf{f} \in \mathcal{H}. \tag{30}$$

A reproducing kernel $\Gamma : \mathcal{X} \times \mathcal{X} \to \mathcal{B}(\mathcal{Y})$ is then defined as: $\Gamma(\mathbf{x}, \mathbf{x}') := ev_{\mathbf{x}} ev_{\mathbf{x}'}^*$, where $\mathcal{B}(\mathcal{Y})$ is the Banach space of bounded linear operators (i.e., $\Gamma(\mathbf{x}, \mathbf{x}')$, $\forall \mathbf{x}, \mathbf{x}' \in \mathcal{X}$) on $\mathcal{Y}$, for example, $\mathcal{B}(\mathcal{Y}) \subseteq \mathbb{R}^{D \times D}$, and $ev_{\mathbf{x}}^*$ is the adjoint of $ev_{\mathbf{x}}$. We have the following two properties about the RKHS and kernel.

Remark 1. The kernel Γ reproduces the value of a function $\mathbf{f} \in \mathcal{H}$ at a point $\mathbf{x} \in \mathcal{X}$. Indeed, $\forall \mathbf{x} \in \mathcal{X}$ and $\mathbf{y} \in \mathcal{Y}$, we have $ev_{\mathbf{x}}^* \mathbf{y} = \Gamma(\cdot, \mathbf{x})\mathbf{y}$, so that $\langle \mathbf{f}(\mathbf{x}), \mathbf{y} \rangle = \langle \mathbf{f}, \Gamma(\cdot, \mathbf{x})\mathbf{y} \rangle_{\mathcal{H}}$.

Remark 2. An RKHS defines a corresponding reproducing kernel. Conversely, a reproducing kernel defines a unique RKHS.

More specifically, for any $N \in \mathbb{N}$, $\{\mathbf{x}_n\}_{n=1}^N \subseteq \mathcal{X}$, and a reproducing kernel Γ, a unique RKHS can be defined by considering the

completion of the space

$$\mathcal{H}_N = \left\{ \sum_{n=1}^{N} \Gamma(\cdot, \mathbf{x}_n)\mathbf{c}_n : \mathbf{c}_n \in \mathcal{Y} \right\}, \tag{31}$$

with respect to the norm induced by the inner product

$$\langle \mathbf{f}, \mathbf{g} \rangle_{\mathcal{H}} = \sum_{i,j=1}^{N} \langle \Gamma(\mathbf{x}_j, \mathbf{x}_i)\mathbf{c}_i, \mathbf{d}_j \rangle \quad \forall \mathbf{f}, \mathbf{g} \in \mathcal{H}_N, \tag{32}$$

where $\mathbf{f} = \sum_{i=1}^{N} \Gamma(\cdot, \mathbf{x}_i)\mathbf{c}_i$ and $\mathbf{g} = \sum_{j=1}^{N} \Gamma(\cdot, \mathbf{x}_j)\mathbf{d}_j$.

References

[1] J. Ma, J. Zhao, J. Tian, A.L. Yuille, Z. Tu, Robust point matching via vector field consensus, IEEE Trans. Image Process. 23 (4) (2014) 1706–1721.
[2] X. Guo, X. Cao, Good match exploration using triangle constraint, Pattern Recognit. Lett. 33 (7) (2012) 872–881.
[3] J. Jiang, C. Chen, J. Ma, Z. Wang, Z. Wang, R. Hu, SRLSP: a face image super-resolution algorithm using smooth regression with local structure prior, IEEE Trans. Multimedia 19 (1) (2016) 27–40.
[4] Z. Wang, P. Yi, K. Jiang, J. Jiang, Z. Han, T. Lu, J. Ma, Multi-Memory Convolutional Neural Network for Video Super-Resolution, IEEE Trans. Image Process. 28 (5) (2019) 2530–2544.
[5] J. Ma, Y. Ma, C. Li, Infrared and visible image fusion methods and applications: a survey, Inf. Fusion 45 (2019) 153–178.
[6] S. Ryu, S. Kim, K. Sohn, LAT: local area transform for cross modal correspondence matching, Pattern Recognit. 63 (2017) 218–228.
[7] J. Ma, W. Yu, P. Liang, C. Li, J. Jiang, FusionGAN: a generative adversarial network for infrared and visible image fusion, Inf. Fusion 48 (2019) 11–26.
[8] D. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[9] X. Tan, C. Sun, X. Sirault, R. Furbank, T.D. Pham, Feature matching in stereo images encouraging uniform spatial distribution, Pattern Recognit. 48 (8) (2015) 2530–2542.
[10] M.A. Fischler, R.C. Bolles, Random sample consensus: a paradigm for model fitting with application to image analysis and automated cartography, Commun. ACM 24 (6) (1981) 381–395.
[11] P.H.S. Torr, A. Zisserman, MLESAC: a new robust estimator with application to estimating image geometry, Comput. Vis. Image Understand. 78 (1) (2000) 138–156.
[12] O. Chum, J. Matas, Matching with PROSAC - progressive sample consensus, in: Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2005, pp. 220–226.
[13] X. Li, Z. Hu, Rejecting mismatches by correspondence function, Int. J. Comput. Vis. 89 (1) (2010) 1–17.
[14] J. Ma, W. Qiu, J. Zhao, Y. Ma, A.L. Yuille, Z. Tu, Robust L_2E estimation of transformation for non-rigid registration, IEEE Trans. Signal Process. 63 (5) (2015) 1115–1129.
[15] C. Wang, L. Wang, L. Liu, Progressive mode-seeking on graphs for sparse feature matching, in: Proc. Eur. Conf. Comput. Vis., 2014, pp. 788–802.
[16] J. Ma, Y. Ma, J. Zhao, J. Tian, Image feature matching via progressive vector field consensus, IEEE Signal Process. Lett. 22 (6) (2015) 767–771.
[17] H. Chui, A. Rangarajan, A new point matching algorithm for non-rigid registration, Comput. Vis. Image Understand. 89 (2003) 114–141.
[18] A. Myronenko, X. Song, Point set registration: coherent point drift, IEEE Trans. Pattern Anal. Mach. Intell. 32 (12) (2010) 2262–2275.
[19] P.J. Besl, N.D. McKay, A method for registration of 3-d shapes, IEEE Trans. Pattern Anal. Mach. Intell. 14 (2) (1992) 239–256.
[20] B. Jian, B.C. Vemuri, Robust point set registration using gaussian mixture models, IEEE Trans. Pattern Anal. Mach. Intell. 33 (8) (2011) 1633–1645.
[21] R. Horaud, F. Forbes, M. Yguel, G. Dewaele, J. Zhang, Rigid and articulated point registration with expectation conditional maximization, IEEE Trans. Pattern Anal. Mach. Intell. 33 (3) (2011) 587–602.
[22] N. Aronszajn, Theory of reproducing kernels, Trans. Amer. Math. Soc. 68 (3) (1950) 337–404.
[23] J. Ma, J. Jiang, Y. Gao, J. Chen, C. Liu, Robust image matching via feature guided gaussian mixture model, in: Proc. IEEE Int. Conf. Multimedia Expo, 2016, pp. 1–6.
[24] S. Belongie, J. Malik, J. Puzicha, Shape matching and object recognition using shape contexts, IEEE Trans. Pattern Anal. Mach. Intell. 24 (24) (2002) 509–522.
[25] R.B. Rusu, N. Blodow, M. Beetz, Fast point feature histograms (FPFH) for 3d registration, in: Proc. IEEE Int. Conf. Robot. Autom., 2009, pp. 3212–3217.
[26] O. Pele, M. Werman, A linear time histogram metric for improved SIFT matching, in: Proc. Eur. Conf. Comput. Vis., 2008, pp. 495–508.
[27] Y.-T. Hu, Y.-Y. Lin, H.-Y. Chen, K.-J. Hsu, B.-Y. Chen, Matching images with multiple descriptors: an unsupervised approach for locally adaptive descriptor selection, IEEE Trans. Image Process. 24 (12) (2015) 5995–6010.
[28] M. Cho, K.M. Lee, Progressive graph matching: making a move of graphs via probabilistic voting, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2012, pp. 398–405.
[29] Y.-T. Hu, Y.-Y. Lin, Progressive feature matching with alternate descriptor selection and correspondence enrichment, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 346–354.
[30] P.J. Huber, Robust Statistics, John Wiley & Sons, New York, 1981.
[31] P.J. Rousseeuw, A. Leroy, Robust Regression and Outlier Detection, John Wiley & Sons, New York, 1987.
[32] H.-Y. Chen, Y.-Y. Lin, B.-Y. Chen, Robust feature matching with alternate hough and inverted hough transforms, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2013, pp. 2762–2769.
[33] Y. Liu, L. Dominicis, B. Wei, L. Chen, R. Martin, Regularization based iterative point match weighting for accurate rigid transformation estimation, IEEE Trans. Vis. Comput. Graph. 21 (9) (2015) 1058–1071.
[34] Y. Liu, H. Liu, R.R. Martin, L. De Dominicis, R. Song, Y. Zhao, Accurately estimating rigid transformations in registration using a boosting-inspired mechanism, Pattern Recognit. 60 (2016) 849–862.
[35] J. Maier, M. Humenberger, M. Murschitz, O. Zendel, M. Vincze, Guided matching based on statistical optical flow for fast and robust correspondence analysis, in: Proc. Eur. Conf. Comput. Vis., 2016, pp. 101–117.
[36] G. Wang, Z. Wang, Y. Chen, X. Liu, Y. Ren, L. Peng, Learning coherent vector fields for robust point matching under manifold regularization, Neurocomputing 216 (2016) 393–401.
[37] J. Ma, J. Wu, J. Zhao, J. Jiang, H. Zhou, Q.Z. Sheng, Nonrigid point set registration with robust transformation learning under manifold regularization, IEEE Trans. Neural Netw. Learn. Syst. (to be published, doi:10.1109/TNNLS.2018.2872528a).
[38] J. Ma, J. Zhao, J. Ziang, H. Zhou, X. Guo, Locality preserving matching, Int. J. Comput. Vis. 127 (5) (2019) 512–531.
[39] J. Ma, J. Jiang, H. Zhou, J. Zhao, X. Guo, Guided locality preserving feature matching for remote sensing image registration, IEEE Trans. Geosci. Remote Sens. 56 (8) (2018) 4435–4447.
[40] M. Leordeanu, M. Hebert, A spectral technique for correspondence problems using pairwise constraints, in: Proc. IEEE Int. Conf. Comput. Vis., 2005, pp. 1482–1489.
[41] L. Torresani, V. Kolmogorov, C. Rother, Feature correspondence via graph matching: Models and global optimization, in: Proc. Eur. Conf. Comput. Vis., 2008, pp. 596–609.
[42] M. Cho, K.M. Lee, Mode-seeking on graphs via random walks, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2012, pp. 606–613.
[43] F. Zhou, F. De la Torre, Deformable graph matching, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2013, pp. 2922–2929.
[44] X. Yang, H. Qiao, Z.-Y. Liu, Point correspondence by a new third order graph matching algorithm, Pattern Recognit. 65 (2017) 108–118.
[45] H. Liu, S. Yan, Common visual pattern discovery via spatially coherent correspondence, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2010, pp. 1609–1616.
[46] K.M. Yi, E. Trulls, Y. Ono, V. Lepetit, M. Salzmann, P. Fua, Learning to find good correspondences, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 2666–2674.
[47] Y. Guo, F. Sohel, M. Bennamoun, J. Wan, M. Lu, An accurate and robust range image registration algorithm for 3d object modeling, IEEE Trans. Multimed. 16 (5) (2014) 1377–1390.
[48] C. Liu, J. Ma, Y. Ma, J. Huang, Retinal image registration via feature-guided gaussian mixture model, JOSA A 33 (7) (2016) 1267–1276.
[49] F. Boughorbel, A. Koschan, B. Abidi, M. Abidi, Gaussian fields: a new criterion for 3d rigid registration, Pattern Recognit. 37 (7) (2004) 1567–1571.
[50] J. Ma, J. Zhao, Y. Ma, J. Tian, Non-rigid visible and infrared face registration via regularized gaussian fields criterion, Pattern Recognit. 48 (3) (2015) 772–784.
[51] G. Wang, Z. Wang, Y. Chen, Q. Zhou, W. Zhao, Context-aware gaussian fields for non-rigid point set registration, in: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 5811–5819.
[52] G. Wang, Y. Chen, X. Zheng, Gaussian field consensus: a robust nonparametric matching method for outlier rejection, Pattern Recognit. 74 (2018) 305–316.
[53] Q. Ma, X. Du, J. Wang, Y. Ma, J. Ma, Robust feature matching via gaussian field criterion for remote sensing image registration, J. Real-Time Image Process. (2018) 1–14.
[54] G. Wang, Z. Wang, Y. Chen, W. Zhao, A robust non-rigid point set registration method based on asymmetric gaussian representation, Comput. Vis. Image Understand. 141 (2015) 67–80.
[55] J. Ma, J. Zhao, A.L. Yuille, Non-rigid point set registration by preserving global and local structures, IEEE Trans. Image Process. 25 (1) (2016) 53–64.
[56] K. Yang, A. Pan, Y. Yang, S. Zhang, S.H. Ong, H. Tang, Remote sensing image registration using multiple image features, Remote Sens. 9 (6) (2017) 581.
[57] S. Ge, G. Fan, M. Ding, Non-rigid point set registration with global-local topology preservation, in: Proc. IEEE Conf. Comput. Vis. Pattern Recog. Workshops, 2014, pp. 245–251.
[58] S. Ge, G. Fan, Articulated non-rigid point set registration for human pose estimation from 3d sensors, Sensors 15 (7) (2015) 15218–15245.
[59] Y. Yang, S.H. Ong, K.W.C. Foong, A robust global and local mixture distance based non-rigid point set registration, Pattern Recognit. 48 (1) (2015) 156–173.
[60] S. Zhang, K. Yang, Y. Yang, Y. Luo, Z. Wei, Non-rigid point set registration using dual-feature finite mixture model and global-local structural preservation, Pattern Recognit. 80 (2018) 183–195.
[61] A. Dempster, N. Laird, D.B. Rubin, Maximum likelihood from incomplete data via the em algorithm, J. R. Statist. Soc. Series B 39 (1) (1977) 1–38.
[62] C.M. Bishop, Pattern Recognition and Machine Learning, Springer-Verlag, New York, NY, USA, 2006.
[63] K. Nigam, A.K. McCallum, S. Thrun, T. Mitchell, Text classification from labeled and unlabeled documents using em, Mach. Learn. 39 (2–3) (2000) 103–134.

[64] J. Ma, X. Jiang, J. Jiang, J. Zhao and X. Guo, LMR: Learning a two-class classifier for mismatch removal, IEEE Trans. Image Process. (to be published, doi:10.1109/TIP.2019.2906490).
[65] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.
[66] J. Ma, H. Zhou, J. Zhao, Y. Gao, J. Jiang, J. Tian, Robust feature matching for remote sensing image registration via locally linear transforming, IEEE Trans. Geosci. Remote Sens. 53 (12) (2015) 6469–6481.
[67] J. Ma, J. Zhao, J. Tian, Z. Tu, A. Yuille, Robust estimation of nonrigid transformation for point set registration, in: Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2013, pp. 2147–2154.
[68] S. Umeyama, Least-squares estimation of transformation parameters between two point patterns, IEEE Trans. Pattern Anal. Mach. Intell. 13 (4) (1991) 376–380.
[69] A. Myronenko, X. Song, On the closed-form solution of the rotation matrix arising in computer vision problems, arXiv:0904.1613 (2009).
[70] C.A. Micchelli, M. Pontil, On learning vector-valued functions, Neural Comput. 17 (1) (2005) 177–204.
[71] R. Rifkin, G. Yeo, T. Poggio, Regularized least-squares classification, Advances in Learning Theory: Methods, Model and Applications, MIT Press, Cambridge, MA, USA, 2003.
[72] J. Ma, J. Zhao, J. Tian, X. Bai, Z. Tu, Regularized vector field learning with sparse approximation for mismatch removal, Pattern Recognit. 46 (12) (2013) 3519–3532.
[73] J.L. Bentley, Multidimensional binary search trees used for associative searching, Commun. ACM 18 (9) (1975) 509–517.
[74] A. Vedaldi, B. Fulkerson, VLFeat - an open and portable library of computer vision algorithms, in: Proc. ACM Int. Conf. Multimedia, 2010, pp. 1469–1472.
[75] G. Wang, Z. Wang, Y. Chen, W. Zhao, Robust point matching method for multimodal retinal image registration, Biomed. Signal Process. Control 19 (2015) 68–76.
[76] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, L. van Gool, A comparison of affine region detectors, Int. J. Comput. Vis. 65 (1) (2005) 43–72.
[77] A. Jinda-Apiraksa, V. Vonikakis, S. Winkler, California-ND: An annotated dataset for near-duplicate detection in personal photo collections, in: Proc. Int. Workshop on Quality of Multimedia Experience, 2013, pp. 142–147.
[78] X. Bai, X. Yang, L.J. Latecki, W. Liu, Z. Tu, Learning context-sensitive shape similarity by graph transduction, IEEE Trans. Pattern Anal. Mach. Intell. 32 (5) (2010) 861–874.
[79] C. Carmeli, E. De Vito, A. Toigo, Vector valued reproducing kernel hilbert spaces of integrable functions and mercer theorem, Anal. Appl. 4 (2006) 377–408.