ROBUST IMAGE MATCHING VIA FEATURE GUIDED GAUSSIAN MIXTURE MODEL

Jiayi Ma[1]*, Junjun Jiang*[2]*, Yuan Gao*[3]*, Jun Chen*[4]*, and Chengyin Liu*[1]

[1]Electronic Information School, Wuhan University, Wuhan 430072, China
[2]School of Computer Science, China University of Geosciences, Wuhan 430074, China
[3]Department of Electronic Engineering, City University of Hong Kong, Hong Kong
[4]School of Automation, China University of Geosciences, Wuhan 430074, China

ABSTRACT

In this paper, we propose a novel feature guided Gaussian mixture model (FG-GMM) for image matching, which typically requires matching two sets of feature points extracted from the given images. We formulate the problem as estimation of a feature guided mixture of densities: a GMM is fitted to one point set, such that both the centers and local features of the Gaussian densities are constrained to coincide with the other point set. The problem is solved under a unified maximum-likelihood framework together with an iterative semi-supervised Expectation-Maximization (EM) algorithm initialized by the confident feature correspondences. The image transformation is specified in a reproducing kernel Hilbert space and a sparse approximation is adopted to achieve a fast implementation. Extensive experiments on various real images show the robustness of our approach, which consistently outperforms other state-of-the-art methods.

Index Terms— Image matching, feature guided, GMM, semi-supervised EM

1. INTRODUCTION

Establishing reliable correspondence between two images is a fundamental problem in computer vision and multimedia, and it is a critical prerequisite in a wide range of applications including 3D reconstruction, tracking, super-resolution, content based image retrieval [1, 2, 3, 4, 5, 6]. In this paper, we formulate it as a matching problem between two sets of discrete points where each point is an image feature, extracted by a feature detector, and has a local image descriptor, such as Scale Invariant Feature Transform (SIFT) [7].

During the last decades, a variety of methods have been introduced to address the matching problem. A popular strategy is to first construct a set of putative point correspondences according to a *similarity constraint* which requires that points can only match points with similar descriptors, and then remove the false correspondences and estimate the transformation parameters (either rigid or non-rigid) robustly based on a *geometric constraint* which requires that the matches satisfy an underlying geometrical requirement [3, 8]. Examples of this strategy include the hypothesize-and-verify RANSAC and analogous algorithms [8, 9, 10] which are based on parametric models, and smooth motion field interpolation methods [3, 11, 12, 13] based on non-parametric models. However, the putative set in the first step typically contains only a small part of the whole existing true correspondences [14], and it will be even less for low-quality or small overlapping images, which may lead to inadequate correspondences for computing the transformation parameters in the second step. Therefore, it is of particular advantage to develop a technique that is able to preserve most of the existing true matches.

Rather than compute the point correspondence and spatial transformation separately, another popular strategy is to estimate these two variables jointly [15, 16]. These methods typically involve an iteration process which alternates between the correspondence and the transformation estimation. The Iterated Closest Point (ICP) algorithm [17] is one of the best known point matching approaches. It uses nearest-neighbor relationships to assign a binary correspondence and then uses the estimated correspondences to refine the transformation. Chui and Rangarajan established a general framework for non-rigid matching called TPS-RPM [15], which replaces the nearest point strategy of ICP with soft assignments within a continuous optimization framework involving deterministic annealing. In the recent past, the point matching is typically solved by probabilistic methods [16, 18, 19, 20, 21]. These methods formulate matching as the estimation of a mixture of densities using Gaussian mixture models (GMMs), and the problem is solved within the framework of maximum likelihood and the EM algorithm. The methods mentioned above generate a correspondence matrix between the original two feature point sets, and hence do not suffer from missing true matches. However, the feature points in these methods are treated as pure spatial coordinates, that is to say, the feature descriptors are entirely discarded, which may easily lead to suboptimal solution in case of severely degraded data such as

The authors gratefully acknowledge the financial supports from the National Natural Science Foundation of China under Grant nos. 61503288 and 61501413, and the China Postdoctoral Science Foundation under Grant no.2015M570665 (e-mail: jyma2010@gmail.com).

large outlier percentage. Therefore, it is necessary to incorporate the local appearance information of feature points in the formulation and helps to establish better point correspondences.

In this paper, we propose a novel feature guided Gaussian mixture model (FG-GMM) to address the problem of robust image matching. The new formulation possesses the advantages of incorporating local feature information as well as preserving most of the existing true matches. More precisely, we formulate point matching as the estimation of a feature guided mixture of densities: a GMM is fitted to one point set, such that both the centers and local features of the Gaussian densities are constrained to coincide with the other point set. The problem is solved under a unified maximum-likelihood framework together with a semi-supervised EM algorithm initialized by the confident feature correspondences. The spatial transformation is modeled in a functional space, called the reproducing kernel Hilbert space (RKHS) [22], in which the transformation function has an explicit kernel representation. In addition, we provide a fast implementation based on sparse approximation to improve the computational efficiency.

2. METHOD

This section describes the proposed matching algorithm. We start by introducing the feature guided GMM formulation for matching feature sets with associated descriptors, and then give the optimization method based on semi-supervised EM. Finally, we provide some implementation details.

2.1. Feature Guided Gaussian Mixture Model

Suppose we obtain two sets of features extracted respectively from two given images, e.g. $\{\mathcal{X}, \mathcal{S}_\mathbf{x}\}$ and $\{\mathcal{Y}, \mathcal{S}_\mathbf{y}\}$, where $\mathcal{X} = \{\mathbf{x}_n\}_{n=1}^N$ and $\mathcal{Y} = \{\mathbf{y}_m\}_{m=1}^M$ are 2D column vectors indicating 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 vectors. We call the two feature sets the model feature set and the target feature set, respectively. The goal is to establish accurate correspondences between the two feature sets and estimate a spatial transformation $\mathcal{T}$ which warps the model features to the target features.

Without considering the associated feature descriptors, the point matching can be formulated as the estimation of a mixture of densities: A Gaussian mixture model (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 $\mathcal{T}(\mathcal{X})$ [16, 18, 19]. Let $\mathcal{Z} = \{z_m \in \mathbb{N}_{N+1} : m \in \mathbb{N}_M\}$ be a set of latent variables, with each variable z_m assigning a target point $\mathbf{y}_m$ to a GMM centroid $\mathcal{T}(\mathbf{x}_n)$, if $z_m = n$, $1 \leq n \leq N$, or to an additional outlier class, if $z_m = N+1$. The GMM probability density function then can be defined as

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

In this paper, we generalize the formulation to register feature sets with associated descriptors. More specifically, let π_{mn} be the membership probability of the GMM, which is typically assumed to be equal for all GMM components in the original formulation, i.e., $\pi_{mn} = \frac{1}{N}, \forall m \in \mathbb{N}_M, n \in \mathbb{N}_N$ [16, 19]; instead, we assign its value based on the associated feature descriptor vectors $\mathcal{S}_\mathbf{x}$ and $\mathcal{S}_\mathbf{y}$. To this end, we first match $\mathcal{S}_\mathbf{x}$ and $\mathcal{S}_\mathbf{y}$ according to a descriptor similarity constraint, for example, comparing the distance of the closest neighbor to that of the second-closest neighbor (we call it distance ratio) and matching them if the distance ratio is below a predefined threshold t [7]. Then we assign $\pi_{mn} = \tau$ if $S(\mathbf{x}_n)$ is matched to $S(\mathbf{y}_m)$, where parameter τ, $0 \leq \tau \leq 1$, could be considered as the confidence of a feature correspondence. For the rest elements of $\{\pi_{mn}\}_{m=1,n=1}^{M,N}$, we set them to either $(1-\tau)/(N-1)$ or $1/N$, so that it satisfies $0 \leq \pi_{mn} \leq 1$ together with $\forall\, m$, $\sum_{n=1}^N \pi_{mn} = 1$. Note that the matched correspondences may be contaminated by some false correspondences and typically contain only a small part of the true correspondences [14].

For point matching, a popular assumption is the equal isotropic covariances $\sigma^2\mathbf{I}$ on all GMM components and the uniform distribution $1/a$ for the outliers [16]. We denote by $\boldsymbol{\theta} = \{\mathcal{T}, \sigma^2, \gamma\}$ the set of unknown parameters, where $\gamma \in [0, 1]$ is the percentage of outliers. The mixture model in Eq. (1) then takes the form

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

The parameters $\boldsymbol{\theta}$ can be estimated by maximizing the likelihood, or equivalently by minimizing the negative log-likelihood function

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

where we have made the i.i.d. data assumption. The correspondence probability between 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 $\mathcal{T}$ will be obtained from the optimal solution $\boldsymbol{\theta}^*$.

2.2. The Semi-Supervised EM Algorithm

The EM algorithm is a technique for learning and inference in the context of latent variables. It alternates between two steps: an expectation step (E-step) and a maximization step (M-step). We follow standard notation [23] and omit some terms that are independent of $\boldsymbol{\theta}$. Considering the negative log-likelihood function, i.e., Eq. (3), the complete-data log-

likelihood is then given by

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

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

E-Step: It aims 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 using the current estimated parameters $\boldsymbol{\theta}^{\text{old}}$. As we have part confident feature correspondences obtained based on the associated descriptors, we consider the semi-supervised EM [24] rather than the original EM. More specifically, we compute p_{mn} according to the following two rules:

(i) For the target features $\{\mathbf{y}_m\}$ with knowing correspondences, we expect them to play a role of anchors leading the EM iteration to avoid or alleviate getting trapped into local minima. Thus we set

$$p_{mn} = \pi_{mn}, \quad 1 \leq n \leq N. \quad (5)$$

(ii) For the target features $\{\mathbf{y}_m\}$ without knowing correspondences, the posterior distribution can be computed by applying Bayes rule:

$$p_{mn} = \frac{\pi_{mn} e^{-\frac{\|\mathbf{y}_m - \mathcal{T}(\mathbf{x}_n)\|^2}{2\sigma^2}}}{\sum_{k=1}^{N} \pi_{mk} e^{-\frac{\|\mathbf{y}_m - \mathcal{T}(\mathbf{x}_k)\|^2}{2\sigma^2}} + \frac{2\gamma\pi\sigma^2}{(1-\gamma)a}}. \quad (6)$$

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

$$\gamma = 1 - M_{\mathbf{P}}/M, \quad (7)$$

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

The estimation of $\mathcal{T}$ is a complicated procedure, which will be discussed later. Once the semi-supervised EM converges, we obtain the estimated spatial transformation $\mathcal{T}$. Besides, the feature correspondences can be computed based on the posterior distribution $\{p_{mn}\}_{m=1,n=1}^{M,N}$. With a predefined threshold η, we obtain the correspondence set $\mathcal{I}$:

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

In this process, we update p_{mn} associated with the knowing correspondences one more time by using Eq. (6) rather than Eq. (5), which would be beneficial if the knowing correspondences contain false matches.

2.3. Estimation of Transformation

We consider the terms of $\mathcal{Q}(\boldsymbol{\theta})$ that are related to $\mathcal{T}$, it is estimated by minimizing a weighted empirical error $\mathcal{Q}(\mathcal{T}) = \frac{1}{2\sigma^2} \sum_{m=1}^{M} \sum_{n=1}^{N} p_{mn} \|\mathbf{y}_m - \mathcal{T}(\mathbf{x}_n)\|^2$. This is not tractable since the feature sets typically suffer from noise and outliers, and the problem will be even big in the non-rigid case as the solution of $\mathcal{T}$ is not unique. Here we consider the slow-and-smooth model [3], where a smoothness functional $\phi(\mathcal{T})$ is imposed on the transformation to ensure well-posedness. Thus we obtain an energy function

$$\mathcal{E}(\mathcal{T}) = \frac{1}{2\sigma^2} \sum_{m=1}^{M} \sum_{n=1}^{N} p_{mn} \|\mathbf{y}_m - \mathcal{T}(\mathbf{x}_n)\|^2 + \lambda \phi(\mathcal{T}), \quad (10)$$

with $\lambda > 0$ controlling the trade-off between the two terms.

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 reproducing kernel Hilbert space (RKHS) [25] (associated with a particular kernel). The smoothness functional can then be defined as the square norm, i.e., $\phi(\mathcal{T}) = \phi(\mathbf{f}) = \|\mathbf{f}\|_{\mathcal{H}}^2$. 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 [16, 3].

Theorem 1. *The optimal solution of Eq.* (10) *is given by*

$$\mathcal{T}(\mathbf{x}) = \mathbf{x} + \mathbf{f}(\mathbf{x}) = \mathbf{x} + \sum_{n=1}^{N} \Gamma(\mathbf{x}, \mathbf{x}_n) \mathbf{c}_n, \quad (11)$$

with the coefficient set $\{\mathbf{c}_n\}$ *determined by a linear system*

$$(d(\mathbf{P}^{\mathrm{T}}\mathbf{1})\boldsymbol{\Gamma} + 2\lambda\sigma^2\mathbf{I})\mathbf{C} = \mathbf{P}^{\mathrm{T}}\mathbf{Y} - d(\mathbf{P}^{\mathrm{T}}\mathbf{1})\mathbf{X}, \quad (12)$$

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

The proof of Theorem 1 is similar to that in [16, 3], which is given in the supplementary material.

Fast Implementation. The algorithm requires at least $O(N^3)$ complexity due to the requirement of solving the linear system (12), which may cause significant computational problem in case of large scale feature sets. Consequently, we adopt a sparse approximation, and randomly pick only a subset of size L input points $\{\tilde{\mathbf{x}}_l\}_{l=1}^{L}$ to have nonzero coefficients in the expansion of the solution (i.e. Eq. (11)). This follows [3] who found that this approximation works well and that simply selecting a random subset of the input points in this manner, performs no worse than more sophisticated and time-consuming methods. Thus we seek a solution of form

$$\mathbf{f}(\mathbf{x}) = \sum_{l=1}^{L} \Gamma(\mathbf{x}, \tilde{\mathbf{x}}_l) \mathbf{c}_l. \quad (13)$$

The chosen point set $\{\tilde{\mathbf{x}}_l\}_{l=1}^{L}$ is somewhat analogous to control points. By using the sparse approximation, the linear

Algorithm 1: The FG-GMM algorithm

Input: Image pair, parameters t, τ, λ, η, β, L
Output: Correspondence set $\mathcal{I}$

Extract two feature sets using SIFT: $\{\mathcal{X}, \mathcal{S}_{\mathbf{x}}\}$, $\{\mathcal{Y}, \mathcal{S}_{\mathbf{y}}\}$;
Match the two feature sets using a descriptor similarity constraint together with a distance ratio threshold t;
Assign the membership probability π_{mn};
Set a to the volume of the output space;
Construct matrix $\mathbf{\Gamma}$ or $\mathbf{E}$ using definition of Γ;
Initialize $\mathbf{C} = \mathbf{0}$, γ, $p_{mn} = \pi_{mn}$, σ^2 (using Eq. (8));
repeat
  E-step:
    Update $\mathbf{P}$ by Eqs. (5) and (6);
  M-step:
    Update $\mathbf{C}$ based on linear system (12) or (14);
    Update σ^2 and γ by Eqs. (8) and (7);
until $\mathcal{Q}$ *converges*;
Correspondence set $\mathcal{I}$ is determined by Eq. (9).

system (12) becomes

$$(\mathbf{U}^{\mathrm{T}}\mathrm{d}(\mathbf{P}^{\mathrm{T}}\mathbf{1})\mathbf{U} + 2\lambda\sigma^2\mathbf{\Gamma}^s)\mathbf{C}^s = \mathbf{U}^{\mathrm{T}}\mathbf{P}^{\mathrm{T}}\mathbf{Y} - \mathbf{U}^{\mathrm{T}}\mathrm{d}(\mathbf{P}^{\mathrm{T}}\mathbf{1})\mathbf{X}, \tag{14}$$

where the coefficient matrix $\mathbf{C}^s = (\mathbf{c}_1, \cdots, \mathbf{c}_L)^{\mathrm{T}} \in \mathbb{R}^{L\times 2}$, $\mathbf{\Gamma} \in \mathbb{R}^{L\times L}$ with $\mathbf{\Gamma}_{ij} = \kappa(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j) = e^{-\beta\|\tilde{\mathbf{x}}_i - \tilde{\mathbf{x}}_j\|^2}$, and $\mathbf{U} \in \mathbb{R}^{N\times L}$ with $\mathbf{U}_{ij} = \kappa(\mathbf{x}_i, \tilde{\mathbf{x}}_j) = e^{-\beta\|\mathbf{x}_i - \tilde{\mathbf{x}}_j\|^2}$. The derivation of Eq. (14) is similar as that of Eq. (12), please see the supplementary material for further details. By using this sparse approximation, the time complexity for solving the linear system is reduced from $O(N^3)$ to $O(L^2N)$.

We summarize our matching algorithm in Algorithm 1.

2.4. Implementation Details

The performance of feature matching algorithms depends, typically, on the coordinate system in which feature points are expressed. We use data normalization to control for this. More specifically, we perform a linear re-scaling so that the spatial positions of the two feature point sets both have zero mean and unit variance. Note that the constant a of the uniform distribution in Eq. (2) is the area of the second image (i.e., the range of $\mathbf{y}_m$), and it should be set according to the data normalization.

Parameter setting. There are mainly seven parameters in our method: t, τ, λ, η, γ, β and L. Parameter t is a distance ratio threshold used to establish the initial correspondences based on feature descriptors. Parameter τ is used to assign the membership probability π_{mn} which is the confidence of a knowing correspondence. Parameter λ controls the influence of the local geometrical constraint on the transformation $\mathcal{T}$. Parameter η is a threshold, which is used for deciding the correctness of a correspondence. Parameter γ reflects our initial assumption on the amount of inliers in the correspondence sets. Parameters β determines how wide the range of interaction between feature points. Parameter L is the required number of control points for sparse approximation. We set $t = 0.8$, $\tau = 0.9$, $\lambda = 3$, $\eta = 0.5$, $\gamma = 0.9$, $\beta = 0.1$ and $L = 15$, throughout our experiments.

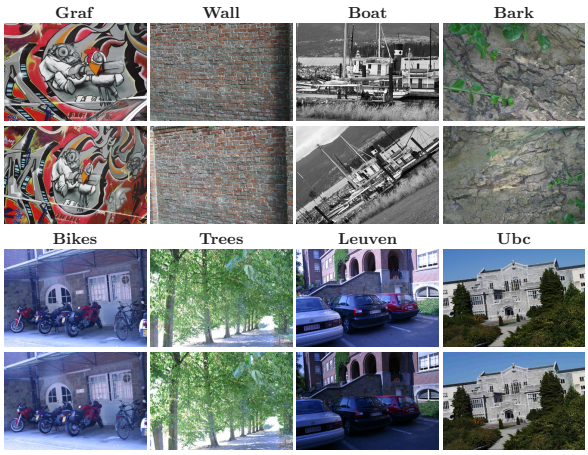

Fig. 1. Examples of images in the dataset [26].

3. EXPERIMENTAL RESULTS

We test the performance of our proposed algorithm on real images. The experiments are performed on a laptop with 2.5-GHz Intel Core CPU, 8-GB memory, and MATLAB code.

3.1. Datasets and Settings

To test the capability of handling non-rigid deformation, we conduct experiments on several image pairs involving deformable objects, which is frequently encountered in image retrieval. We further test our method on the dataset of Mikolajczyk *et al.* [26], which contains 40 image pairs either of planar scenes or taken by camera in a fixed position during acquisition. The images, therefore, 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 as well as JPEG compression. Some examples are given in Fig. 1. To determine the match correctness on this dataset, we use the same overlap error criterion as in [3].

The open source VLFEAT toolbox [27] is used to determine the putative correspondences of SIFT [7]. The experimental results are evaluated by precision and the number of identified correct matches, where the precision is defined as the ratio of the identified correct matches number and the preserved correspondence number. We compare our FG-GMM algorithm with other four state-of-the art matching methods, such as RANSAC [9], ICF [11], VFC [3] and CPD [16]. We implement ICF and tune all parameters accordingly to find

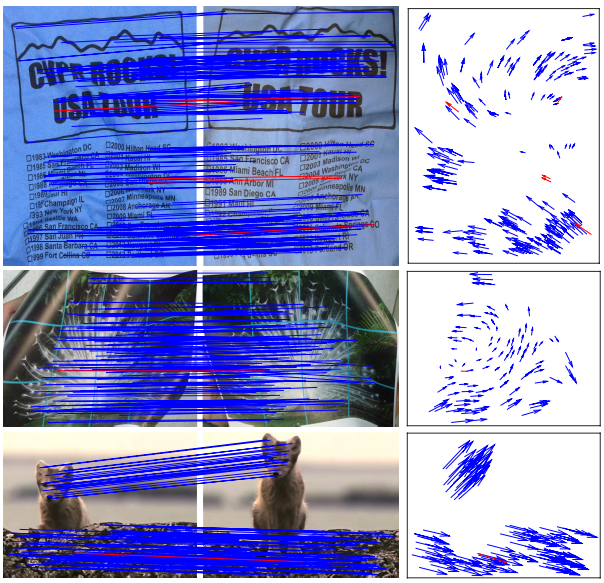

Fig. 2. Results of our FG-GMM on three typical image pairs (e.g., *T-shirt*, *Peacock* and *Fox*) involving deformable objects. The precisions and identified correct match numbers are $(97.85\%, 288)$, $(99.07\%, 107)$, and $(99.29\%, 139)$. Blue and red lines/arrows indicate correct and false matches, 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.

optimal settings. The other three methods are implemented by using publicly available codes. Throughout all the experiments, the parameters of five methods are all fixed.

3.2. Results on Non-Rigid Images

We first give some intuitive performance of our proposed FG-GMM on three typical image pairs involving deformable objects, as shown in Fig. 2. The first pair consists of scenes of two different deformations with illumination changes of a T-shirt. In the second pair, we first add a regular grid on it, and then warp it and take two views with different deformations. The third pair is two frames extracted from a video. Such matching problem is frequently encountered in near-duplicate image retrieval. From the results, we see that our FG-GMM is able to establish accurate feature matching, and the precisions are 97.85%, 99.07% and 99.29%, respectively. The motion fields related to the three image pairs are provided in the right column of Fig. 2, we see that the degree of the non-rigid deformation is quite large, where different parts of the scenes have different motion manners. However, the variation of the motion field is slow-and-smooth, which guarantees our method working well in such case.

To demonstrate the advantages of our method, we report the results of other four state-of-the-art methods, as shown in Table 1. Clearly, our FG-GMM has consistently better precisions and can identify much more true matches. RANSAC has satisfying precisions, as it can identify a majority of

Table 1. Comparison of precisions and preserved correct match numbers on image pairs involving deformable objects.

	T-shirt	*Peacock*	*Fox*
RANSAC [9]	$(89.21\%, 124)$	$(96.61\%, 57)$	$(97.33\%, 73)$
ICF [11]	$(95.00\%, 76)$	$(97.67\%, 42)$	$(98.70\%, 76)$
VFC [3]	$(96.18\%, 126)$	$(98.44\%, 63)$	$(98.94\%, 93)$
CPD [16]	$(90.00\%, 45)$	$(96.92\%, 63)$	$(95.31\%, 61)$
FG-GMM	$(\mathbf{97.85\%}, \mathbf{288})$	$(\mathbf{99.07\%}, \mathbf{107})$	$(\mathbf{99.29\%}, \mathbf{139})$

the putative correspondences which 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 well if the deformation is complex. This can be seen from the *T-shirt* pair with larger degree of deformation, in which RANSAC has much lower precision. By contrast, the precisions of the two non-parametric based methods ICF and VFC are better. Nevertheless, these three methods operate on a set of putative correspondences, which suffer from missing true correspondences and hence, the numbers of identified correct matches are much smaller compared to our FG-GMM[1]. For the CPD method, we found in our evaluation that it completely failed on all the three pairs (here we omit the detail results for clarity), which can be attributed to the large outlier percentages in the feature set. Note that CPD only uses the spatial positions of feature points; its poor results demonstrate the significance of using local appearance information during image matching. We also test CPD on two feature sets obtained from the putative sets as those in RANSAC, ICF and VFC, and report the results in Table 1. We see that the results are still not that good compared to VFC; here we give an explanation as follows: VFC only needs to remove false matches from a putative set, it has initial correspondence information compared to CPD, which is beneficial for solving the matching problem.

3.3. Results on An Image Dataset

We next conduct experiments on the dataset of Mikolajczyk *et al*. [26]. The statistics of the precision and identified correct match number for RANSAC, ICF, VFC and our FG-GMM are given in Fig. 3. Here we do not report the results of CPD, as it again fails on most of the image pairs. From the results, we see that our FG-GMM has the best average precision (i.e., 95.84%) and largest average identified correct match number (i.e., 920.15), followed by VFC and RANSAC. Note that RANSAC works well on this dataset, since the image transformation satisfies a parametric model such as homography.

We also test the fast version of our FG-GMM on this dataset. The average number of extracted SIFT features for

[1]To keep more true matches, a possible solution is to enlarge the size of the putative set. But this will rapidly reduce the correct match percentage in the putative set, and hence badly degrades the matching performance [14].

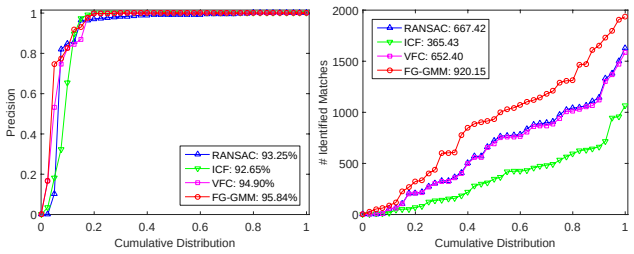

Fig. 3. Precision (left) and identified correct match number (right) of RANSAC [9], ICF [11], VFC [3] and our FG-GMM with respect to the cumulative distribution on the dataset of Mikolajczyk *et al*. [26].

an image is about 2630, which should be a large scale problem for image matching, and hence it is desirable to seek an efficient implementation. The average precision and identified correct match number of our fast FG-GMM are about 95.21% and 922.34, which are similar to the results of the original FG-GMM. The average run times of the original and fast FG-GMM are about 34.23s and 16.35s per image pair. We see that the fast implementation can significantly reduce the computational complexity without sacrifice in accuracy.

4. CONCLUSION

In this paper, we presented a feature guided Gaussian mixture model (FG-GMM) for robust image matching. A key characteristic of our approach is that it can preserve much more true feature matches and can incorporate local feature information during matching. The semi-supervised EM algorithm is introduced to solve the problem which is formulated as a maximum-likelihood estimation. We also provide an efficient implementation of our method to reduce the computational complexity without significantly reducing the quality of the matching. Experiments on public available dataset demonstrate that our approach yields superior results to those of the state-of-the-art methods, and it will be beneficial for multimedia applications such as content based vedio/image retrieval.

5. REFERENCES

[1] L. Li, Z. Wu, Z.-J. Zha, S. Jiang, and Q. Huang, "Matching content-based saliency regions for partial-duplicate image retrieval," in *ICME*, 2011, pp. 1–6.

[2] Y. Uchida and S. Sakazawa, "Accurate feature matching and scoring for re-ranking image retrieval results," in *ICME*, 2013, pp. 1–6.

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

[4] J. Zhao, J. Ma, J. Tian, J. Ma, and D. Zhang, "A robust method for vector field learning with application to mismatch removing," in *CVPR*, Jun. 2011, pp. 2977–2984.

[5] J. Jiang, X. Ma, Z. Cai, and R. Hu, "Sparse support regression for image super-resolution," *IEEE Photonics Journal*, vol. 7, no. 5, pp. 1–11, 2015.

[6] J. Jiang, R. Hu, Z. Han, and T. Lu, "Efficient single image super-resolution via graph-constrained least squares regression," *Multimedia Tools and Applications*, vol. 72, no. 3, pp. 2573–2596, 2014.

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

[8] P. H. S. Torr and A. Zisserman, "MLESAC: A new robust estimator with application to estimating image geometry," *Computer Vision and Image Understanding*, vol. 78, no. 1, pp. 138–156, 2000.

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

[10] O. Chum and J. Matas, "Matching with PROSAC - progressive sample consensus," in *CVPR*, 2005, pp. 220–226.

[11] X. Li and Z. Hu, "Rejecting mismatches by correspondence function," *International Journal of Computer Vision*, vol. 89, no. 1, pp. 1–17, 2010.

[12] J. Ma, J. Zhao, J. Tian, Z. Tu, and A. Yuille, "Robust estimation of nonrigid transformation for point set registration," in *CVPR*, 2013, pp. 2147–2154.

[13] J. Ma, J. Zhao, Y. Ma, and J. Tian, "Non-rigid visible and infrared face registration via regularized gaussian fields criterion," *Pattern Recognition*, vol. 48, no. 3, pp. 772–784, 2015.

[14] C. Wang, L. Wang, and L. Liu, "Progressive mode-seeking on graphs for sparse feature matching," in *ECCV*, pp. 788–802. 2014.

[15] H. Chui and A. Rangarajan, "A new point matching algorithm for non-rigid registration," *Computer Vision and Image Understanding*, vol. 89, pp. 114–141, 2003.

[16] A. Myronenko and X. Song, "Point set registration: Coherent point drift," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 12, pp. 2262–2275, 2010.

[17] P. J. Besl and N. D. McKay, "A method for registration of 3-d shapes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 2, pp. 239–256, 1992.

[18] B. Jian and B. C. Vemuri, "Robust point set registration using gaussian mixture models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 8, pp. 1633–1645, 2011.

[19] R. Horaud, F. Forbes, M. Yguel, G. Dewaele, and J. Zhang, "Rigid and articulated point registration with expectation conditional maximization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 587–602, 2011.

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

[21] Y. Gao, J. Ma, J. Zhao, J. Tian, and D. Zhang, "A robust and outlier-adaptive method for non-rigid point registration," *Pattern Analysis and Applications*, vol. 17, no. 2, pp. 379–388, 2014.

[22] N. Aronszajn, "Theory of reproducing kernels," *Transactions of the American Mathematical Society*, vol. 68, no. 3, pp. 337–404, 1950.

[23] C. M Bishop, *Pattern Recognition and Machine Learning*, Springer-Verlag, New York, NY, USA, 2006.

[24] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell, "Text classification from labeled and unlabeled documents using em," *Machine Learning*, vol. 39, no. 2-3, pp. 103–134, 2000.

[25] C. A. Micchelli and M. Pontil, "On learning vector-valued functions," *Neural Computation*, vol. 17, no. 1, pp. 177–204, 2005.

[26] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. van Gool, "A comparison of affine region detectors," *International Journal of Computer Vision*, vol. 65, no. 1, pp. 43–72, 2005.

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