Forward Affine Point Set Matching Under Variational Bayesian Framework

QU Han-Bing[1, 3] CHEN Xi[2, 3] WANG Song-Tao[1] YU Ming[2, 3]

Abstract In this work, the affine point set matching is formulated under a variational Bayesian framework and the model points are projected forward into the scene space by a linear transformation. A directed acyclic graph is presented to represent the relationship between the parameters, latent variables, model and scene point sets and an iterative approximate algorithm is proposed for the estimation of the posterior distributions over parameters. Furthermore, the anisotropic covariance is assumed on the transition variable and one Gaussian component is provided for the inference of outlier points. Experimental results demonstrate that the proposed algorithm achieves good performance in terms of both robustness and accuracy.

Key words Point set matching, variational Bayesian, affine transformation, graphical model

Citation Qu Han-Bing, Chen Xi, Wang Song-Tao, Yu Ming. Forward affine point set matching under variational Bayesian framework. *Acta Automatica Sinica*, 2015, **41**(8): 1482−1494

DOI 10.16383/j.aas.2015.e130204

Point set matching is an important and well-studied problem in the fields of computer vision and arises in a variety of applications in other fields such as pattern recognition and computer graphics. The feature points in an image can be used for 2D-3D shape representation, alignment, registration and recognition[1−4]. The point set matching methods usually fall into two categories of matching strategies: 1) matching with prior correspondences between points[5−8], and 2) estimating the mapping function and point correspondences simultaneously[9−10]. In this paper, we focus on the study of affine point set matching without the prior point-to-point correspondence under probabilistic framework.

Iterative closest point (ICP) algorithm is one of most popular methods for point set matching problems, which obtains the rigid transforming matrix $\mathcal{T}$ by minimizing the mean-squares objective function between two point sets[11]. Although ICP is attractive for its efficiency, it can be easily trapped into local minima due to the strict selection of the best point-to-point assignment and is sensitive to initialization of transformation and choice of threshold which is needed to accept or reject a match[12]. Typical soft assignment registration is kernel correlation (KC) algorithm, which models each individual point sets by a kernel density function and then quantifies the (dis)similarity between them using an entropy measure[9]. The performance of KC method may be undermined under the circumstance of large-scale outliers for the absence of explicit noise and outlier models.

In recent years, probabilistic approach attracts more attention for solving point set matching problems. Chui and Rangarajan proposed mixture point matching (MPM) for point registration by use of a Gaussian mixture model with isotropic covariance, and an expectation maximization (EM) algorithm is proposed for the solution to maximum a posteriori (MAP) estimation of model parameters[13]. Coherent point drift (CPD) is another well-known EM-like point registration method which also assumes one same isotropic covariance for all the mixture components[14]. Instead of the isotropic assumption, Horaud et al. introduced an expectation conditional maximization for point registration algorithm for the issue of matching rigid and articulated shapes, where anisotropic covariance is assumed on each individual Gaussian components instead of the isotropic covariance[15]. Without the use of EM-like updates, Jian and Vemuri directly minimized KL-divergence between the distributions of model and scene point sets[12]. This formulation leads to a non-linear optimization problem under nonconvex rigidity constraints. Ma et al. made use of L_2E estimator for point set registration base on the assumption of one isotropic covariance and a coarse-to-fine deterministic annealing optimization[16]. All of the aforementioned algorithms utilize an extra uniform distribution to alleviate the influence of outliers. Recently, student′s-t mixture model was proposed for the non-rigid registration to overcome the vulnerability of Gaussian mixture model (GMM) to outliers and data longer than normal tails[17]. Though student distribution provides better performance in terms of outliers points than Gaussian distribution, the absence of outlier model would impair the generalization capability of this mixture model, especially in the case of multiple clusters of outliers.

As the Bayesian framework for point set matching, Gu and Kanade presented a multi-level generative model for the alignment of face images, in which the prior distribution of shape is modeled as Gaussian mixtures with constrained isotropic covariance and the deformation and transformation parameters are estimated by an EM updates under an MAP objective function[4]. Rangarajan et al. proposed Bayesian network for the relational shape matching in the absence of outlier model, and the pairwise correspondences between the template graphs and the data are estimated by Bethe free energy approach[18]. Zhou et al. presented a Bayesian mixture model to describe shape distribution and feature point visibility, where an EM algorithm is proposed to find the MAP estimation of the model parameters[19]. These aforementioned approaches provide the point estimations of transformation and correspondence, they are thus less prone to local optima for the absence of uncertainty modeling for the transformation parameters and outliers.

In order to provide a full Bayesian inference for model, Green and Mardia proposed a hierarchical graph models to match unlabeled or partial labeling points[20]. Their ap-

Manuscript received August 30, 2013; accepted February 9, 2015
Supported by Innovation Group Plan of Beijing Academy of Science and Technology (IG201506N), Youth Core Plan of Beijing Academy of Science and Technology (2014-30), Tianjin Science and Technology Projects (14RCGFGX00846)
Recommended by Associate Editor ZHANG Chang-Shui
1. Key Laboratory of Pattern Recognition, Beijing Academy of Science and Technology, Beijing 100094, China 2. School of Electronic and Information Engineering, Hebei University of Technology, Tianjin 300401, China 3. School of Computer Science and Engineering, Hebei University of Technology, Tianjin 300401, China

proach is based on Poisson process for hidden true point locations and Markov chain Monte Carlo (MCMC) is applied to the simultaneous estimation of matching and transforming. Czogiel et al. presented a Bayesian modeling of the predicted field overlap between pairs of point sets, and the posterior inference of molecular alignment is carried out using MCMC stochastic simulation as well[21]. Unlike stochastic approximation, variational inference is a kind of deterministic techniques for posterior approximation[22] and provides a monotonously increased lower bound for the true posteriors. Klami presented a point set matching solution to feature points matching based on Bayesian canonical correlation analysis (CCA), which disregards the transformation between two data sets and explicitly finds the correspondence between two point sets[23]. Simpson et al. propose a variational Bayesian registration framework to infer the level of regularization in the non-rigid registration and make a linear approximation to the transformation by taking a first order Taylor series expansion[18]. In their model, the target image data is assumed as being generated from a source image and the correspondences between points does not need not to be estimated during the registration process.

In this paper, we propose a probabilistic approach for the estimation of affine matching between two point sets under variational Bayesian framework. The contribution of our work can be summarized as follows. 1) The matching process is divided into two sub-phases, i.e., regression and clustering, where the regression phase estimates the affine transformation between two point sets and the clustering phase establishes the correspondences between point pairs. 2) Anisotropic covariance is assumed on each individual transformed model points to model the uncertainty of scattered scene points and thus the distribution of scene set is model by a Gaussian mixture model. 3) An extra Gaussian component is introduced to model the outliers and the outlier distribution is also estimated during the EM-like approximation process. Furthermore, comparison studies are made on synthetic and real data sets between our method and other state-of-the-art matching algorithms and the proposed approach demonstrates comparable performance in terms of both robustness and accuracy.

1 Problem formulation

The point set matching problem can be formulated as follows. Given a shape model set $\mathcal{M}$ (in the form of a vector of landmarks) consisting of K points, and a collection of points in scene $\mathcal{S}$, consisting of N points. The task of point set matching or point set registration is to find the correspondences between the model and scene points with a transformation $\mathcal{T}$ from $\mathcal{M}$ to $\mathcal{S}$ under certain objective function $\mathcal{F}$, while allowing outliers. The point set matching problem can be formulated under probabilistic integration as follows[24]:

$$p(\mathcal{S}|\mathcal{M}) = \int p(\mathcal{S}, \mathcal{T}|\mathcal{M}) p(\mathcal{T}) \mathrm{d}\mathcal{T} \tag{1}$$

Here $\mathcal{M} = \{\boldsymbol{m}_k\}_{k=1}^K$ is the model point set containing K feature components and $\mathcal{S} = \{\boldsymbol{s}_n\}_{n=1}^N$ is the scene point set containing N data points, where $\boldsymbol{m}_k$ and $\boldsymbol{s}_n$ are $D \times 1$ vectors, D is the dimension of individual point of model and scene, $\mathcal{T}$ is the transformation function. As affine point set matching is concerned, $\mathcal{T} = \{\boldsymbol{A}, \boldsymbol{b}\}$, where $\boldsymbol{A}$ is a $D \times D$ affine transformation matrix and $\boldsymbol{b}$ is a $D \times 1$ translation vector.

As the point numbers of model and scene may be different, we introduce a latent random variable $\mathcal{X} = \{\boldsymbol{x}_k\}_{k=1}^K$, which is called transition variable in this work, such that $\boldsymbol{x}_k = \boldsymbol{A}\boldsymbol{m}_k + \boldsymbol{b}$. It is obvious that the connection between model and scene sets can be established via the transition variable $\mathcal{X}$. If the mapping uncertainty is assumed to be on each transition variable $\boldsymbol{x}_k$, for example, Gaussian white noise with precision matrix $\boldsymbol{\Psi}$, the distribution of $p(\mathcal{X}|\mathcal{M})$ can be written as a typical statistical form of regression:

$$p(\mathcal{X}|\boldsymbol{A}, \boldsymbol{b}, \mathcal{M}, \boldsymbol{\Psi}) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{x}_k|\boldsymbol{A}\boldsymbol{m}_k + \boldsymbol{b}, \boldsymbol{\Psi}^{-1}) \tag{2}$$

After the transformation of model, one Gaussian mixture model with K components can be obtained in the scene space. Therefore, the scene data points can be seen as being generated from the Gaussian mixture model whose centers provide an internal structure according to the affine transformed model points.

For each scene observation, a 1-of-$(K+1)$ binary vector $\boldsymbol{z}_n$ with elements $\{z_{nk},\ k = 0, 1, \cdots, K\}$ is assumed on $\boldsymbol{s}_n$ to indicate which mixture the scene point is generated from. In order to model the outliers in the scene, an extra Gaussian component is added into the mixtures instead of the often-used uniform distribution. The subscript of outlier component is represented by 0 for simplicity, and $\boldsymbol{x}_0$ and $\boldsymbol{\Lambda}_0$ are the mean and covariance of outlier component, respectively. The random variable $\boldsymbol{z}_n$ indicates which component of mixture model data $\boldsymbol{s}_n$ belongs to. Let $\mathcal{Z} = \{\boldsymbol{z}_n,\ n = 1, \cdots, N\}$ denotes the collection of indicator variables. Given $\mathcal{X}$, $\mathcal{Z}$ and $\boldsymbol{\Lambda}$, the conditional probability of $\mathcal{S}$ can be written as follows:

$$p(\mathcal{S}|\mathcal{Z}, \mathcal{X}, \boldsymbol{\Lambda}) = \prod_{n=1}^{N} \prod_{k=0}^{K} \mathcal{N}(\boldsymbol{s}_n|\boldsymbol{x}_k, \boldsymbol{\Lambda}_k^{-1})^{z_{nk}} \tag{3}$$

where $\{\boldsymbol{\Lambda}_k\}_{k=0}^K$ are the precision matrices for the Gaussian components of mixture model. The coverage of outlier component is controlled by the prior scale matrix and the scattering of outlier points in the scene set.

As one can see from (3), the "data collapses" will occur in the case of $N \approx K$ under the maximum likelihood framework, because the mixture components can easily 'collapse' on some specific data points and there will present singularities in the estimation of parameters[22]. On the contrary, the Bayesian framework naturally avoids this kind of singularities for the reason that the prior is introduced into the matching of points.

The indicator variable $\mathcal{Z}$ is controlled by the mixing coefficient $\boldsymbol{\pi}$, which is the proportion of points in terms of each component to the whole size of scene. The conditional probability of indicator variable given $\boldsymbol{\pi}$ can be written as follows:

$$p(\mathcal{Z}|\boldsymbol{\pi}) = \prod_{n=1}^{N} \prod_{k=0}^{K} \pi_k^{z_{nk}} \tag{4}$$

π_0 is the mixing proportion of the number of outliers to the total number of scene points. The prior over mixing proportions $\boldsymbol{\pi}$ is a Dirichlet distribution with parameter $\boldsymbol{\alpha}^0$, which is the prior sample size for each component, and has the following probability formulation:

$$p(\boldsymbol{\pi}|\boldsymbol{\alpha}^0) = C(\boldsymbol{\alpha}^0) \prod_{k=0}^{K} {\pi_k}^{\alpha_k^0 - 1} \tag{5}$$

where $C(\boldsymbol{\alpha}^0)$ is normalization constant.

Following the idea of MacKay[25] and Neal[26], hierarchical prior is assumed to be on affine transformation $\tilde{\boldsymbol{A}} = [\boldsymbol{A}\ \boldsymbol{b}]$ in order to perform automatic relevance determination over the entries of $\tilde{\boldsymbol{A}}$. Each column of the $\tilde{\boldsymbol{A}}$ has a Gaussian prior with zero mean and a different precision parameter $\boldsymbol{\nu}$ which is drawn from a Gamma distribution with parameters $\boldsymbol{c}^0$ and $\boldsymbol{d}^0$:

$$p(\tilde{\boldsymbol{A}}|\boldsymbol{\nu}) = \prod_{j=1}^{D+1} p(\tilde{A}_{\cdot j}|\nu_j) = \prod_{j=1}^{D+1} \mathcal{N}(\tilde{A}_{\cdot j}|\boldsymbol{0}, \frac{\boldsymbol{I}}{\nu_j}) \tag{6}$$

and

$$p(\boldsymbol{\nu}|\boldsymbol{c}^0, \boldsymbol{d}^0) = \prod_{j=1}^{D+1} \text{Gamma}(\nu_j|c_j^0, d_j^0) \tag{7}$$

where $\boldsymbol{I}$ is a $D \times D$ identity matrix.

The noise precision matrix $\boldsymbol{\Psi}$ is assumed to be Wishart distribution with prior degree of freedom λ and prior scale matrix $\boldsymbol{U}^0$, which has the following distribution form:

$$p(\boldsymbol{\Psi}|\lambda^0, \boldsymbol{U}^0) = \frac{1}{C(\lambda^0, \boldsymbol{U}^0)} |\boldsymbol{\Psi}|^{\frac{\lambda^0 - D - 1}{2}} \mathrm{e}^{-\frac{1}{2}\mathrm{tr}(\boldsymbol{U}^{0\,-1}\boldsymbol{\Psi})} \tag{8}$$

where $|\cdot|$ is the determinant of matrix, $\mathrm{tr}(\cdot)$ is the trace function, and $C(\lambda^0, \boldsymbol{U}^0)$ is normalization constant.

Similar to the noise precision $\boldsymbol{\Psi}$, precision $\boldsymbol{\Lambda}_k$ of each individual mixture components is also assumed to be under Wishart distribution with prior degree of freedom γ_k^0 and prior scale matrix $\boldsymbol{W}_k^0$, which has the following probability:

$$p(\boldsymbol{\Lambda}|\boldsymbol{\gamma}^0, \boldsymbol{W}^0) = \prod_{k=0}^{K} \frac{1}{C(\gamma_k^0, \boldsymbol{W}_k^0)} |\boldsymbol{\Lambda}_k|^{\frac{\gamma_k^0 - D - 1}{2}} \mathrm{e}^{-\frac{1}{2}\mathrm{tr}(\boldsymbol{W}_k^{0\,-1}\boldsymbol{\Lambda}_k)} \tag{9}$$

where $C(\gamma_k^0, \boldsymbol{W}_k^0)$ is normalization constant for Wishart distribution as well. Generally, γ_0^0 and $\boldsymbol{W}_0^0$ in terms of outlier mixture are preset according to the prior information of outliers. A lower prior scale matrix $\boldsymbol{W}_0^0$ produces a broader covariance of the posterior of outlier mixture and endows equal probabilities over the scattered outliers.

There are seven random variables governing the probabilistic point set matching model: 1) transition variable $\mathcal{X}$, which dominates the centers of Gaussian mixtures, 2) transformation matrix $\tilde{\boldsymbol{A}} = [\boldsymbol{A}\ \boldsymbol{b}]$, 3) precision matrix $\boldsymbol{\Psi}$ with regard to the D-dimensional zero-mean regression noise, 4) hierarchical prior $\boldsymbol{\nu}$, which performs automatic relevance determination over $\boldsymbol{A}$ and $\boldsymbol{b}$, 5) precision matrix $\boldsymbol{\Lambda}$ of transition and outlier mixtures, 6) indicator latent variable $\mathcal{Z}$, and 7) mixing coefficient $\boldsymbol{\pi}$ determining the indicator variable.

There are four groups of prior parameters need to be preset over the latent variables $\boldsymbol{\Psi}$, $\boldsymbol{\nu}$, $\boldsymbol{\pi}$ and $\boldsymbol{\Lambda}$: 1) $\boldsymbol{c}^0$ and $\boldsymbol{d}^0$ are prior shape and inverse-scale parameters of $\boldsymbol{\nu}$, 2) λ^0 and $\boldsymbol{U}^0$ are prior degree of freedom and scale matrix of noise $\boldsymbol{\Psi}$, 3) $\boldsymbol{\gamma}^0$ and $\boldsymbol{W}^0$ are prior degree of freedom and scale matrix of component precision $\boldsymbol{\Lambda}$, and 4) $\boldsymbol{\alpha}^0$ is the prior sample size of mixing coefficients $\boldsymbol{\pi}$.

2 Bayesian point set matching

In this work, the affine point set matching is formulated under a full probabilistic framework, and all of the parameters are assumed as random variables. Furthermore, conjugate prior distributions are presumed on these variables in order that the posteriors will keep the same form as the priors. The affine matching process can be intuitively illustrated as a graphical model, which is shown in Fig. 1. According to (2) and (3), affine point set matching is divided into two statistical subphases: regression and clustering, and these two subphases are connected by the transition variable $\mathcal{X}$, which can also be observed in the graphical model. This division is important for Bayesian point set matching problem because it avoids nonlinear optimization over complex objective functions and a lot of statistical tools can be directly applied to the estimation of parameters.

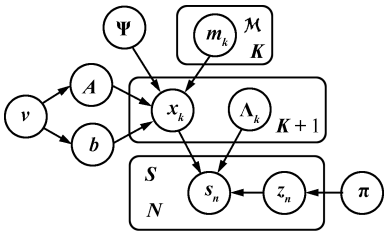


Fig. 1 Directed acyclic graph representing the point set matching (The top plate notation denotes the kth component of model $\mathcal{M}$ and the bottom plate notation denotes repetitions over the nth scene $\mathcal{S}$ point. Circled nodes are random variables and the shading ones are observations in the graph.)

According to this directed graph illustrated in Fig. 1, the conditional probability $p(\mathcal{S}|\mathcal{M})$ can be written as follows

$$\begin{aligned} p(\mathcal{S}|\mathcal{M}) = \int & p(\mathcal{S}|\mathcal{X}, \mathcal{Z}, \boldsymbol{\Lambda}) p(\mathcal{X}|\tilde{\boldsymbol{A}}, \mathcal{M}, \boldsymbol{\Psi}) p(\mathcal{Z}|\boldsymbol{\pi}) \times \\ & p(\boldsymbol{\pi}|\boldsymbol{\alpha}^0) p(\boldsymbol{\Lambda}|\boldsymbol{W}^0, \boldsymbol{\gamma}^0) p(\tilde{\boldsymbol{A}}|\boldsymbol{\nu}) p(\boldsymbol{\nu}|\boldsymbol{c}^0, \boldsymbol{d}^0) \times \\ & p(\boldsymbol{\Psi}|\boldsymbol{U}^0, \lambda^0) \mathrm{d}\mathcal{X} d\mathcal{Z} \mathrm{d}\tilde{\boldsymbol{A}} \mathrm{d}\boldsymbol{\Lambda} \mathrm{d}\boldsymbol{\Psi} \boldsymbol{\nu} \mathrm{d}\boldsymbol{\pi} \end{aligned} \tag{10}$$

The log-marginal likelihood $\log p(\mathcal{S}|\mathcal{M})$ is the objective function for the variational approximation, and it is difficult to optimize this function directly because the integral is in the logarithm function. However, the log-marginal likelihood can be decomposed into the summation of two items as $\log p(\mathcal{S}|\mathcal{M}) = \mathcal{F}(q) + \mathrm{KL}(q||p)$, where $\mathcal{F}(q)$ is a functional of approximate posterior $q(x)$ and $\mathrm{KL}(q||p)$ is the KL-divergence between true posterior $p(x)$ and approximate posterior $q(x)$ over random variable x. $\mathcal{F}(q)$ is called negative free energy and produces a lower bound to the $\log p(\mathcal{S}|\mathcal{M})$, which can be written as follows:

$$\begin{aligned} \mathcal{F}(q) = & \left\langle \log p(\mathcal{S}|\mathcal{X}, \mathcal{Z}, \boldsymbol{\Lambda}) \right\rangle_{q(\mathcal{X})q(\mathcal{Z})q(\boldsymbol{\Lambda})} + \\ & \left\langle \log p(\mathcal{X}|\tilde{\boldsymbol{A}}, \mathcal{M}, \boldsymbol{\Psi}) \right\rangle_{q(\mathcal{X})q(\tilde{\boldsymbol{A}})q(\boldsymbol{\Psi})} + \mathcal{H}[q(\mathcal{X})] + \\ & \left\langle \log p(\mathcal{Z}|\boldsymbol{\pi}) \right\rangle_{q(\mathcal{Z})q(\boldsymbol{\pi})} + \mathcal{H}[q(\mathcal{Z})] + \\ & \left\langle \log p(\boldsymbol{\pi}|\boldsymbol{\alpha}^0) \right\rangle_{q(\boldsymbol{\pi})} + \mathcal{H}[q(\boldsymbol{\pi})] + \\ & \left\langle \log p(\boldsymbol{\Lambda}|\boldsymbol{W}^0, \boldsymbol{\gamma}^0) \right\rangle_{q(\boldsymbol{\Lambda})} + \mathcal{H}[q(\boldsymbol{\Lambda})] + \\ & \left\langle \log p(\tilde{\boldsymbol{A}}|\boldsymbol{\nu}) \right\rangle_{q(\tilde{\boldsymbol{A}})q(\boldsymbol{\nu})} + \mathcal{H}[q(\tilde{\boldsymbol{A}})] + \\ & \left\langle \log p(\boldsymbol{\nu}|\boldsymbol{c}^0, \boldsymbol{d}^0) \right\rangle_{q(\boldsymbol{\nu})} + \mathcal{H}[q(\boldsymbol{\nu})] + \\ & \left\langle \log p(\boldsymbol{\Psi}|\boldsymbol{U}^0, \lambda^0) \right\rangle_{q(\boldsymbol{\Psi})} + \mathcal{H}[q(\boldsymbol{\Psi})] \end{aligned} \tag{11}$$

where $\mathcal{H}[p(x)] = -\int p(x) \log p(x) \mathrm{d}x$ is the entropy of random variable x and $\langle \cdot \rangle$ is the expectation operator. By taking the variational optimization of $\mathcal{F}(q)$ in terms of each in-

dividual approximate posteriors, one obtains a series of approximate posteriors with respect to the random variables, and the KL-divergence between the approximate posterior and the true posterior is also minimized during the approximate process.

It should be mentioned that our model is substantially different with other GMM-like point set matching models, such as MPM[13], CPD[14] and GMM-L2[12], in the following respects: 1) The full probabilistic model explicitly accounts for the matching uncertainty between two point sets and all the parameters are assumed with prior distributions for statistical inference. 2) Anisotropic covariance is presumed on the noise, outlier and each individual mixture components of the transition variable. 3) Bayesian framework provides our model with the ability to draw inference from variables, noise and outliers, and avoids the singularities while one component collapses on one data point, which arises frequently in the point-to-point matching.

Our approach is also fundamentally different with the other variational matching methods such as the variational Bayesian matching[23] and the probabilistic inference of image registration[8] in terms of the model setting. In fact, the variational Bayesian matching generally focuses on the natural language processing, for example the document alignment, sentence alignment or word alignment[23]. This approach directly infers the unknown co-occurrence between observations and has less consideration about the transformation between two data sets. This method cannot be directly applied to the alignment of two point sets with coordinates. On the contrary, the probabilistic image registration presumes the correspondences between two images are preset in the form of pixel-to-pixel correspondences, so there is no correspondence finding step in the model. These two methods may provide intuition for the point set matching problems in the estimation of transformation and correspondence.

3 Variational posterior distribution

According to the direct acyclic graph, the posterior approximate distribution $q(x)$ can be factorized into a product of a series of distributions, thus the priors and posteriors preserve the same distribution form under the conjugate-exponential structure. The variational approximate posteriors over latent variables and parameters can be written as follows:

1) Transition variable $\mathcal{X}$: The posterior of transition variable $\mathcal{X}$ is also a product of independent Gaussian distributions in terms of $\boldsymbol{x}_k$ with mean $\boldsymbol{\mu}_{\boldsymbol{x}_k}$ and covariance $\boldsymbol{\Sigma}_{\boldsymbol{x}_k}$, which is expressed as follows:

$$q(\mathcal{X}) = \prod_{k=1}^{K} q(\boldsymbol{x}_k) = \prod_{k=1}^{K} \mathcal{N}(\boldsymbol{x}_k|\boldsymbol{\mu}_{\boldsymbol{x}_k}, \boldsymbol{\Sigma}_{\boldsymbol{x}_k}) \tag{12}$$

where

$$\boldsymbol{\Sigma}_{\boldsymbol{x}_k}^{-1} = \sum_{n=1}^{N} \left\langle z_{nk}\boldsymbol{\Lambda}_k \right\rangle_{q(\mathcal{Z})q(\boldsymbol{\Lambda})} + \left\langle \boldsymbol{\Psi} \right\rangle_{q(\boldsymbol{\Psi})} \tag{13}$$

$$\boldsymbol{\mu}_{\boldsymbol{x}_k} = \boldsymbol{\Sigma}_{\boldsymbol{x}_k}\left[\sum_{n=1}^{N} \left\langle z_{nk}\boldsymbol{\Lambda}_k\boldsymbol{s}_n \right\rangle_{q(\mathcal{Z})q(\boldsymbol{\Lambda})} + \left\langle \boldsymbol{\Psi}(\boldsymbol{A}\boldsymbol{m}_k + \boldsymbol{b}) \right\rangle_{q(\boldsymbol{\Psi})q(\tilde{\boldsymbol{A}})}\right]$$

The mean and covariance of latent variable $\mathcal{X}$ are a combination of model and scene information. The transition variable $\mathcal{X}$ plays a core role in our Bayesian point set matching because it divides and combines the regression phase and the clustering phase during the approximation process.

2) Affine transformation $\boldsymbol{A}$ and $\boldsymbol{b}$: Following the idea of Beal and Ghahramani[27], the posterior over transformation is factorized over the rows of $\tilde{\boldsymbol{A}} = [\boldsymbol{A}\ \boldsymbol{b}]$ and the precision parameter $\boldsymbol{\nu}$ is placed over each column of $\tilde{\boldsymbol{A}}$. The posterior of $\tilde{\boldsymbol{A}}$ is a product of D Gaussian distributions, which can be written as follows:

$$q(\tilde{\boldsymbol{A}}) = \prod_{q=1}^{D} q(\tilde{\boldsymbol{A}}_q) = \prod_{q=1}^{D} \mathcal{N}(\tilde{\boldsymbol{A}}_q|\boldsymbol{\mu}_{\tilde{\boldsymbol{A}}_q}, \boldsymbol{\Gamma}_{\tilde{\boldsymbol{A}}_q}) \tag{14}$$

where $\boldsymbol{\mu}_{\tilde{\boldsymbol{A}}_q}$ and $\boldsymbol{\Gamma}_{\tilde{\boldsymbol{A}}_q}$ denote the posterior mean and covariance over the qth row of $\tilde{\boldsymbol{A}}$, respectively. The variational posterior updates are provided as follows:

$$\begin{aligned} \boldsymbol{\Gamma}_{\tilde{\boldsymbol{A}}_q}^{-1} &= \begin{bmatrix} \boldsymbol{\Upsilon}_{\boldsymbol{AA}}^{q} & \boldsymbol{\Upsilon}_{\boldsymbol{Ab}}^{q} \\ \boldsymbol{\Upsilon}_{\boldsymbol{bA}}^{q} & \boldsymbol{\Upsilon}_{\boldsymbol{bb}}^{q} \end{bmatrix} \\ \boldsymbol{\mu}_{\tilde{\boldsymbol{A}}_q} &= \begin{bmatrix} \boldsymbol{\mu}_{\boldsymbol{A}_q} & \boldsymbol{\mu}_{\boldsymbol{b}_q} \end{bmatrix}^{\mathrm{T}} \end{aligned} \tag{15}$$

with

$$\begin{aligned} \boldsymbol{\Upsilon}_{\boldsymbol{AA}}^{q} &= \mathrm{diag}\{\langle \boldsymbol{\nu}_{\boldsymbol{A}} \rangle_{q(\boldsymbol{\nu})}\} + \langle \boldsymbol{\Psi}_{qq} \rangle_{q(\boldsymbol{\Psi})} \sum_{k=1}^{K} \boldsymbol{m}_k\boldsymbol{m}_k^{\mathrm{T}} \\ \boldsymbol{\Upsilon}_{\boldsymbol{bb}}^{q} &= \langle \nu_{\boldsymbol{b}} \rangle_{q(\boldsymbol{\nu})} + K\langle \boldsymbol{\Psi}_{qq} \rangle_{q(\boldsymbol{\Psi})} \\ \boldsymbol{\Upsilon}_{\boldsymbol{Ab}}^{q} &= \langle \boldsymbol{\Psi}_{qq} \rangle_{q(\boldsymbol{\Psi})} \sum_{k=1}^{K} \boldsymbol{m}_k = \boldsymbol{\Upsilon}_{\boldsymbol{bA}}^{q\ \mathrm{T}} \\ \boldsymbol{\mu}_{\boldsymbol{A}_q} &= [\boldsymbol{\Gamma}_{q\cdot}]_{\boldsymbol{AA}} \left[\langle \boldsymbol{\Psi}_{qq} \rangle_{q(\boldsymbol{\Psi})} \sum_{k=1}^{K} \langle \boldsymbol{x}_{k,q}\boldsymbol{m}_k^{\mathrm{T}} \rangle_{q(\mathcal{X})} \right] \\ \boldsymbol{\mu}_{\boldsymbol{b}_q} &= [\boldsymbol{\Gamma}_q]_{\boldsymbol{bb}} \left[\langle \boldsymbol{\Psi}_{qq} \rangle_{q(\boldsymbol{\Psi})} \sum_{k=1}^{K} \langle \boldsymbol{x}_{k,q} \rangle_{q(\mathcal{X})} \right] \end{aligned} \tag{16}$$

The posterior distribution over $\tilde{\boldsymbol{A}}$ has a block diagonal covariance structure, which is a $D \times (D+1) \times (D+1)$ matrix.

3) Affine transformation precision $\boldsymbol{\nu}$: The posterior of precision parameter $\boldsymbol{\nu}$ in terms of the lth column of $\tilde{\boldsymbol{A}}$ is also a Gamma distribution and the updating parameters are expressed as follows:

$$\begin{aligned} c_l &= c_l^0 + \frac{D}{2} \\ d_l &= d_l^0 + \frac{1}{2}\sum_{q=1}^{D} \langle \tilde{A}_{ql}^2 \rangle \end{aligned} \tag{17}$$

4) Indicator variables $\mathcal{Z}$: The approximate distribution of $\mathcal{Z}$ takes the same functional form as the prior $p(\mathcal{Z}|\boldsymbol{\pi})$, which is expressed as follows:

$$q(\mathcal{Z}) = \prod_{n=1}^{N}\prod_{k=1}^{K} q(z_{nk}) = \prod_{n=1}^{N}\prod_{k=1}^{K} \hat{r}_{nk}^{z_{nk}} \tag{18}$$

where

$$\begin{aligned} \hat{r}_{nk} = \frac{1}{\mathcal{Z}_{nk}} \exp\Bigg\{ &\langle \log \pi_k \rangle_{q(\boldsymbol{\pi})} + \frac{1}{2}\langle \log|\boldsymbol{\Lambda}_k| \rangle_{q(\boldsymbol{\Lambda})} - \\ &\frac{1}{2}\left[\left\langle (\boldsymbol{s}_n - \boldsymbol{x}_k)^{\mathrm{T}}\boldsymbol{\Lambda}_k(\boldsymbol{s}_n - \boldsymbol{x}_k) \right\rangle_{q(\boldsymbol{\Lambda})q(\mathcal{X})} \right] \Bigg\} \end{aligned} \tag{19}$$

where $\mathcal{Z}_{nk}$ is a normalization constant for each data point, such that $\sum_{k=1}^{K} \hat{r}_{nk} = 1$. The two expectations in terms of the responsibility $\hat{r}_{nk}$ are evaluated as follows:

$$\left\langle \log \pi_k \right\rangle_{q(\boldsymbol{\pi})} = \psi(\alpha_k) - \psi(\hat{\alpha})$$

$$\left\langle \log|\boldsymbol{\Lambda}_k| \right\rangle_{q(\boldsymbol{\Lambda})} = \sum_{q=1}^{D} \psi\left(\frac{\gamma_k + 1 - q}{2}\right) + D\log 2 + \log|\boldsymbol{W}_k|$$

where $\psi(\cdot)$ is the digamma function, $|\cdot|$ is the determinant of matrix, and γ_k is the posterior update of the degree of freedom for $\boldsymbol{\Lambda}_k$.

5) Mixing coefficient $\boldsymbol{\pi}$: The posterior of kth component of $\boldsymbol{\pi}$ is updated according to α_k, which is expressed as follows

$$\alpha_k = \alpha_k^0 + \sum_{n=1}^{N} \hat{r}_{nk} \tag{20}$$

6) Noise precision matrix $\boldsymbol{\Psi}$: The variational posterior for the noise precision $\boldsymbol{\Psi}$ is also a Wishart distribution and the updates of degree of freedom and scale matrix are expressed as follows:

$$\begin{aligned} \lambda_k &= \lambda^0 + K \\ \boldsymbol{U}^{-1} &= \mathrm{diag}\Bigg\{\boldsymbol{U}^{0^{-1}} + \sum_{k=1}^{K} \big\langle (\boldsymbol{x}_k - \boldsymbol{A}\boldsymbol{m}_k - \boldsymbol{b}) \times \\ &\qquad (\boldsymbol{x}_k - \boldsymbol{A}\boldsymbol{m}_k - \boldsymbol{b})^{\mathrm{T}} \big\rangle_{q(\mathcal{X})q(\tilde{\boldsymbol{A}})} \Bigg\} \end{aligned} \tag{21}$$

The diag operator sets the off-diagonal terms to zero.

7) Precision matrix $\boldsymbol{\Lambda}$ of the Gaussian mixtures: The posterior of the component precision $\boldsymbol{\Lambda}$ is a Wishart distribution as well. The updates of degree of freedom and precision matrix in terms of each individual components can be expressed as follows

$$\begin{aligned} \gamma_k &= \gamma_k^0 + \sum_{n=1}^{N} \big\langle z_{nk} \big\rangle_{q(\mathcal{Z})} \\ \boldsymbol{W}_k^{-1} &= \boldsymbol{W}_k^{0^{-1}} + \sum_{n=1}^{N} \big\langle z_{nk}(\boldsymbol{s}_n - \boldsymbol{x}_k)(\boldsymbol{s}_n - \boldsymbol{x}_k)^{\mathrm{T}} \big\rangle_{q(\mathcal{Z})q(\mathcal{X})} \end{aligned} \tag{22}$$

8) Outlier mean $\boldsymbol{x}_0$ and precision $\boldsymbol{\Lambda}_0$: In order to simplify the posterior of outliers, the prior outlier mean is presumed to be on the original, and $\boldsymbol{x}_0$ and $\boldsymbol{\Lambda}_0$ are assumed to be mutually independent with other means and precisions of mixture components. Then the posterior updates of the outlier distribution can be written as follows:

$$\begin{aligned} \gamma_0 &= \gamma_0^0 + \sum_{n=1}^{N} \big\langle z_{n0} \big\rangle_{q(\mathcal{Z})} \\ \boldsymbol{x}_0 &= \frac{1}{\gamma_0} \sum_{n=1}^{N} \big\langle z_{n0}\boldsymbol{s}_n \big\rangle_{q(\mathcal{Z})} \\ \boldsymbol{W}_0^{-1} &= \boldsymbol{W}_0^{0^{-1}} + \sum_{n=1}^{N} \big\langle z_{n0}(\boldsymbol{s}_n - \boldsymbol{x}_0)(\boldsymbol{s}_n - \boldsymbol{x}_0)^{\mathrm{T}} \big\rangle_{q(\mathcal{Z})} \end{aligned} \tag{23}$$

Equations (12) to (23) are the posterior updates with respect to these random variables and parameters of our model. The EM-like updates produce a lower bound to approximate the log-marginal likelihood and the KL-divergence between approximate posterior $q(x)$ and true posterior $p(x)$ is also minimized during iterative updates.

4 Numerical solutions

The update equations (12) ∼ (23) are coupled and therefore must be solved iteratively. This is achieved by starting with initial guess of the parameters and cycling through the updating equations in terms of the moments until convergence. The learning steps of our method may be conveniently implemented according to Algorithm 1, which is illustrated in a pseudo-code form.

Algorithm 1. Variational Bayesian for forward affine point set matching (VBPSM)

1. **Input:**

$$\mathcal{M} = \{\boldsymbol{m}_k\}_{k=1}^{K}, \mathcal{S} = \{\boldsymbol{s}_n\}_{n=1}^{N}, \varepsilon_{\mathcal{F}}$$

2. **Initialization:**
hyperparameters: $\boldsymbol{c}^0$, $\boldsymbol{d}^0$, $\boldsymbol{\alpha}^0$, $\boldsymbol{U}^0$, λ^0, $\boldsymbol{W}^0$, $\boldsymbol{\gamma}^0$;
transformations: $\boldsymbol{A}$ and $\boldsymbol{b}$.
3. **Reconstruct:** $\mathcal{X} = \{\boldsymbol{x}_k\}_{k=1}^{K}$.
4. **repeat**
5. Infer the latent variable $\mathcal{Z}$ using (19).
6. Update the mixing coefficient $\boldsymbol{\alpha}$ using (20).
7. Update the degree of freedom and scale matrix $\boldsymbol{\gamma}$ and $\boldsymbol{W}$ using (22).
8. Update the prior shape and inverse-scale $\boldsymbol{c}$ and $\boldsymbol{d}$ using (17).
9. Infer the transformation parameters $\boldsymbol{A}$ and $\boldsymbol{b}$ using (15) and (16).
10. Infer the transition variable $\mathcal{X}$ using (13).
11. Update the noise degree of freedom and scale matrix λ and $\boldsymbol{U}$ using (21).
12. Update the outlier degree of freedom, mean and scale matrix $\boldsymbol{\gamma}_0$, $\boldsymbol{x}_0$ and $\boldsymbol{W}_0$ using (23).
13. Evaluate the negative free energy $\mathcal{F}$ using (11).
14. **until** $|\mathcal{F}(t+1) - \mathcal{F}(t)| < \varepsilon_{\mathcal{F}}$.

Fig. 2 illustrates the evolution of variational approximation process of $q(\mathcal{X})$ and negative free energy. $q(\mathcal{X})$ approximates to the distribution of scene data points as the negative free energy increasing approached to the lower bound of log-marginal likelihood. The stopping condition for the iterative process is either when $\epsilon_{\mathbf{F}}$ drops below 1e−8 or iteration numbers reaches 200. On average the algorithm converges in $1 \sim 3$ seconds and requires about $80 \sim 130$ iterations on 2D data sets.

Table 1 The comparison of computational time

Dataset	Points	GMM-L2 (s)	CPD (s)	VFC-L2E (s)	VBPSM (s)
Chinese	105	1.48	1.13	8.49	3.75
road	277	10.11	19.03	5.61	10.32

For a single iteration, the computational cost of updating the posteriors are approximately $\mathrm{O}(D^3KN)$ and $\mathrm{O}(D^2(N+D)K)$ by examining (19) and (22), respectively. In general, the coordinate dimension D is 2 or 3, so the computation complexity of our method is approximately $\mathrm{O}(NK)$. It is obvious that the Bayesian estimation of full posterior probability increases the computational complexity than the point estimation methods, such as CPD and VFC-L2E. Especially, one needs to compute the inverse matrix for the estimation of covariance matrix, which increases about D^3 times than the point estimation method.

A down-sampling of model points will be helpful for large-scale registration problems in order to construct sparse centers. A simple computational comparison is shown in Table 1 in terms of GMM-L2, CPD, VFC-L2E and our method. It should be noted that there are still lots of works left for future computational optimization, including prior parameters optimization, algorithm optimization for large-scale point set registration and variable inference during the learning process.

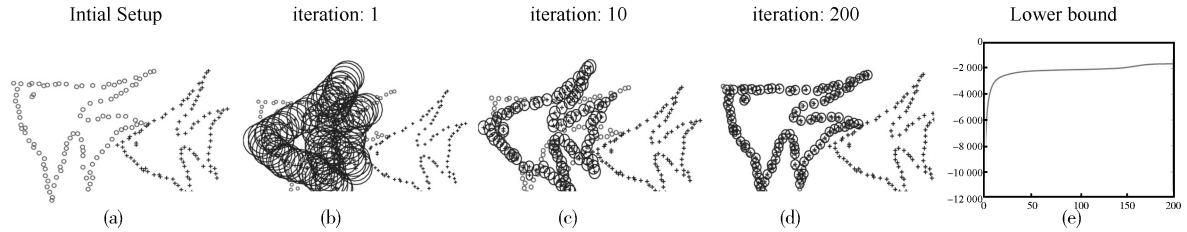


Fig. 2 (a) ∼ (d) Evolution of VBPSM with respect to 2D-fish data. Model and scene are represented by black '+' and gray 'o', respectively. The mixture components of $q(\mathcal{X})$ are illustrated by the ellipses with transformed model points as centers. (e) The evolution of negative free energy

Under variational Bayesian framework, the prior plays an important role in the point set matching process. In general, the hyperparameters of prior distribution are preset according to the prior knowledge. In our model, the elements of $\boldsymbol{c}^0$ and $\boldsymbol{d}^0$ are set to be 0.001 and 0.001, respectively, which means non-informative prior on affine transformation $\tilde{\boldsymbol{A}}$[28]. The element in $\boldsymbol{\alpha}^0$ is set to be 1, which indicates that point-to-point correspondence. As the terms of noise are concerned, λ^0 and $\boldsymbol{U}^0$ are set to be $D+1$ and $10\times\boldsymbol{I}$, respectively, which means the posterior is determined predominantly by the regression error and the prior has less influence on the approximate posterior.

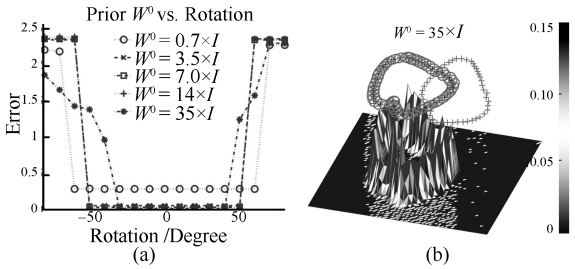


Fig. 3 The influence of prior scale matrix $\boldsymbol{W}^0$ on the matching performance ((a) The matching error and the range of rotation convergence in terms of different prior scale matrix; (b) The matching result of VBPSM with rotation angle 30°. The model, scene and transition mixtures $\mathcal{X}$ are illustrated on the top layer. The approximate distribution of $p(\mathcal{S}|\mathcal{M})$ is shown under the data.)

Precision $\boldsymbol{\Lambda}$ controls the matching accuracy and robustness in our approach and it also takes advantage of Baysian method to overcome the 'data collapses' by introducing prior into the point-to-point matching. Through the observation of (22), the posterior $\boldsymbol{W}$ is composed of two terms: the prior $\boldsymbol{W}^0$ and the data $\boldsymbol{s}_n$. Therefore, if just one data point coincides with the transition center, the prior $\boldsymbol{W}^0$ will take effect to avoid the singularity. Furthermore, the prior scale matrix $\boldsymbol{W}^0$ has an effect for overcoming local optima and achieving higher matching accuracy during the matching process. For example, a smaller $\boldsymbol{W}^0$ leads to a smaller posterior precision $\boldsymbol{\Lambda}$ and a wider mixture covariance. Although wider covariance causes more inaccurate point-to-point alignment, the larger uncertainty provides the model with the ability to be less prone to local minima. On the contrary, fine point-to-point matching (narrow posterior covariance) is at the expense of being apt to be trapped into local minima.

Fig. 3 illustrates the influence of prior scale matrix $\boldsymbol{W}^0$ on matching performance and Fig. 3 (b) gives the variational approximate in the case of $\boldsymbol{W}^0 = 35\times I$. The scene points are obtained by rotating model from $-80°$ to $80°$. In the case of lower prior $\boldsymbol{W}^0$, the range of convergence is wider with lower matching accuracy. On the contrary, higher prior produces more accurate matching result with more difficulty in convergence. Therefore, an empirical coarse-to-fine strategy is proposed for the variational approximate point set matching, that is, firstly taking advantage of small value of prior scale matrix $\boldsymbol{W}^0$ to grasp the main tendency of scene points, and then using the higher prior value to obtain more precise matching results. This strategy is similar to tuning annealing temperature T in VFC-L_2E[11] and MPM[14], and the coarse-to-fine correlation matching[29]. In this paper, prior $\boldsymbol{W}^0$ is increased from $0.1\times I$ to $20\times I$ for the subphases of approximation and the γ_k^0 is set to $2\times(D+1)$ according to the work of Fraley and Raftery[30].

5 Experimental results

In this section, we present some results from the application of our method to both synthetic and real data sets including 2D and 3D point sets. Quantitative experiments are also provided for the performance comparison with other competing point set matching algorithms. Our method is referred to as VBPSM for these affine point set matching experiments. The algorithm is implemented in Matlab v7.12 and tested on an Intel Core2 CPU 2.67 GHz with 4 GB RAM.

5.1 2D affine point set matching

Firstly, VBPSM is applied to matching various clean 2D point sets in order to test validity of our approach[1]. Moderate deformations of scene data sets are obtained by the following affine transformations from model set: 1) translating model with a displacement $[T_x\ T_y]=[-3.6\ 0.5]$, 2) scaling the point set with $S_x=1.1$, $S_y=1.3$ with respect to x and y axes, respectively, 3) rotating the point set with an angle between $-50°\sim 50°$, and 4) finally, shearing the shape with $sh_x=0.4$ and $sh_y=-0.15$. Fig. 4 shows the matching results from VBPSM on five data sets. As one can see from the results, VBPSM obtains good matching performance on these clean data sets and it can also provide an anisotropic covariance over each individual transition components.

In the second quantitative experiment, we test the convergence performance of VBPSM on five clean 2D data sets in comparison with KC[9], GMM-L2[13] and CPD[15] methods. Each data set is rotated from $-\pi$ to π and the four algorithms are performed on the alignment of the original point set with the rotated one. The range of convergence angles is reported in Table 2 with respect to these four

[1]The fish-1 data set was obtained from https://sites.google.com/site/myronenko[31]. The road dataset was obtained from http://www.cs.cmu.edu/~ytsin/KCReg/[9] and the other datasets were obtained from http://www.cise.ufl.edu/~anand/students/chui/research.html[13].

methods on five data sets. The comparison shows that VBPSM algorithm achieves comparable performance to the other three approaches in terms of rotation, except that the KC method achieve the best convergence range than the others on the road data.

In the third quantitative experiment, two quantitative comparisons are conducted on four affine point set matching algorithms. These experimental results are implemented on 2D fish-2 and Chinese character data in Fig. 2. As in previous experiments, the scene data sets are generated by the following affine moderate transformations of original model point set: 1) translating model with a displacement $[T_x\ T_y] = [-0.50\ 0.50]$, 2) scaling the point set with $S_x = 1.20$, $S_y = 1.15$ with respect to x and y axes, respectively, 3) rotating the point set with angle 30°, and 4) finally, shearing the shape with $sh_x = 0.10$ and $sh_y = 0.15$. Since the true affine parameters are known in advance, the Euclidean norm between the true transformation and the estimated affine parameters is used as the metric of matching accuracy with respect to KC, GMM-L2, CPD and our method. For each level of noise, we conduct 20 runs and take the mean of these 20 results as measurement to compare the performance of four algorithms.

To test the ability of these algorithms to occlusion, we delete a certain number of points from at head part of one fish and points from at the tail part of another. The deleted points is increased from 1 to 21. In the robust-to-outlier test, we increase the ratio of the number of outliers to the number of scene data points from 0.2 to 2.0, and the clean model set is transformed to match the corrupted scene. The outliers are generated from uniform distributions with respect to each outliers-to-data level.

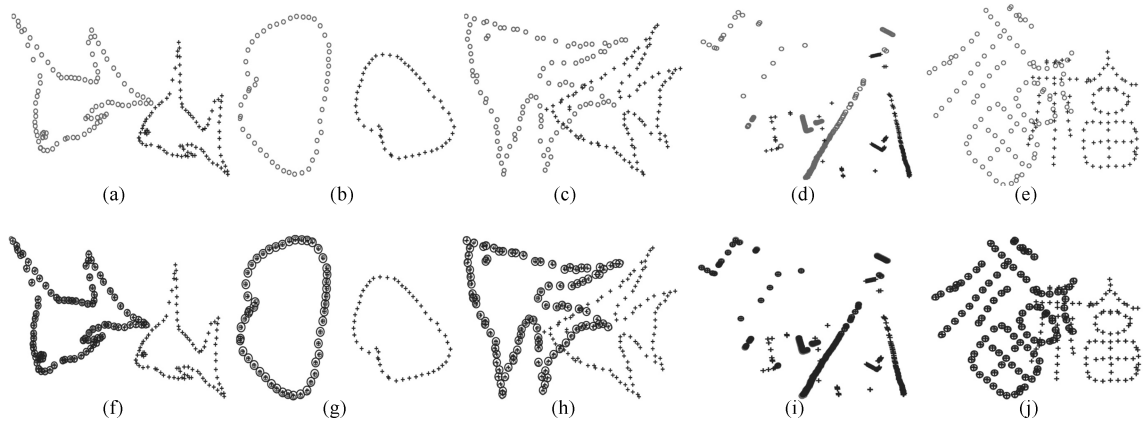


Fig. 4 The application of VBPSM on various 2D data sets ((a) ∼ (e) Initial setup of model (black '+') and scene(gray 'o') in terms of (from left to right) fish-1, contour, fish-2, road and Chinese datasets; (f) ∼ (j) Matching results. The transition variable $\mathcal{X}$ is illustrated with mean (gray '+') and covariance (black ellipse) around model points.)

Table 2 Convergence range of KC[9], GMM-L2[12], CPD[14] and VBPSM methods in terms of rotation angles (in degree) on five 2D datasets

Dataset	VBPSM	CPD	GMM-L2	KC
fish-1	[−70° 70°]	[−67° 67°]	[−65° 66°]	[−65° 78°]
fish-2	[−93° 94°]	[−69° 69°]	[−70° 70°]	[−77° 70°]
contour	[−96° 96°]	[−87° 86°]	[−91° 91°]	[−87° 86°]
road	[−113° 110°]	[−115° 53°]	[−103° 104°]	[−115° 127°]
Chinese	[−91° 91°]	[−73° 72°]	[−62° 62°]	[−88° 90°]

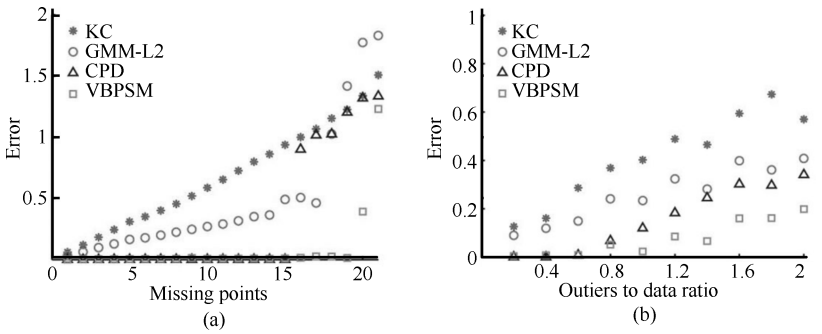


Fig. 5 Performance comparison of KC[9], GMM-L2[12], CPD[14] and VBPSM on (a) fish-2 data with different missing points and (b) Chinese character with different outliers to data ratio (x-axis is the ratio of the number of outliers to the number of clean scene points.)

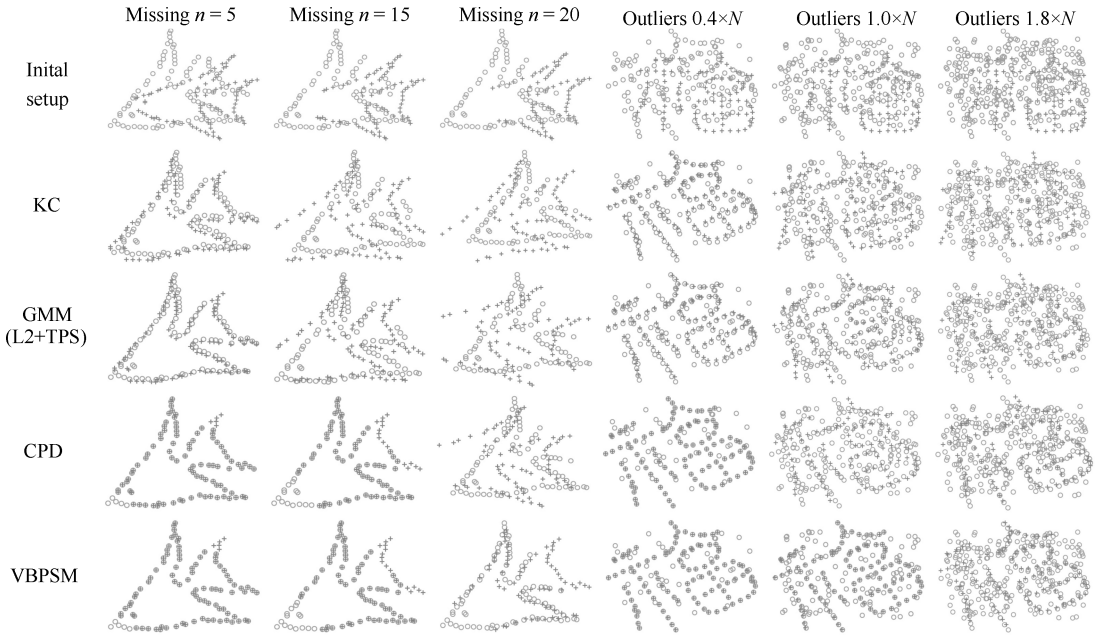


Fig. 6 Performance comparisons of KC[9], GMM-L2[12], CPD[14] and VBPSM methods on data with missing points and outliers (Top row is initialization with model (black '+') and scene (gray 'o'). Row 2 $\sim$ 5 are matching results obtained from four methods (black '+' is the transformed center from model).)

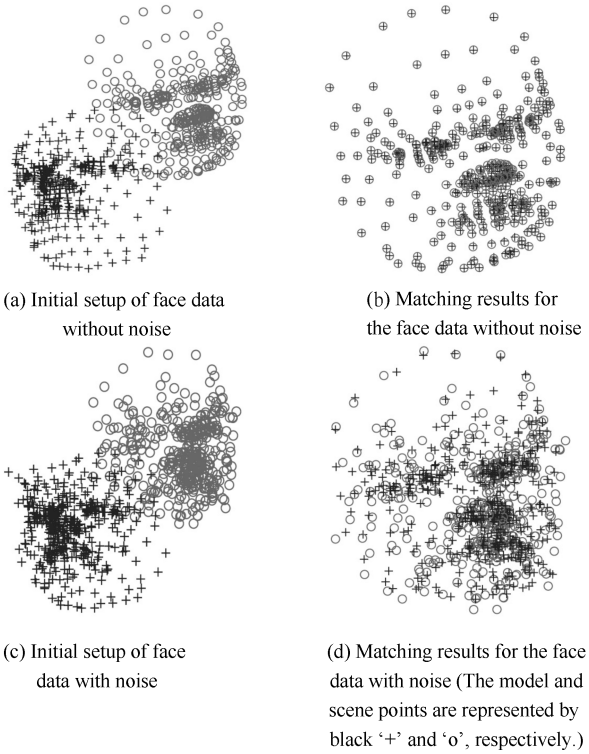
Fig. 5 illustrates these two quantitative comparison results. In the missing points experiment in Fig. 5 (a), we observe that when the number of deleted points is larger than 20, none of these algorithms is able to achieve a successful alignment. VBPSM obtains successful matching results even when the maximum number of missing points reaches 20, whereas the maximum numbers of CPD and GMM-L2 are 15 and 17, respectively. As far as the outlier test is concerned, Fig. 5 (b) demonstrates that VBPSM has better matching performance than KC, GMM-L2 and CPD with respect to every ratio level. It is evident that VBPSM achieves more accurate matching performance compared to KC, GMM-L2 and CPD, especially in the presence of outliers and missing data.

As is the case of most other matching algorithms, the proposed algorithm iteratively performs local updates and is thus prone to local optima as well. However, VBPSM performs well in practice as reported in these experiments, because the variational framework explicitly accounts for matching uncertainty and is thus less prone to local optima. Fig. 6 gives the matching results from the four algorithms with respect to missing data and outliers at different noise level.

In fact, VBPSM provides not only the means of the transformed points but also their covariances with respect to these ones. The uncertainty of transition variable $\mathcal{X}$ will be a variate according to the relative locations of scene points to the transformed model point set. As far as other matching methods are concerned, the covariances are all presumed to be same so as to simplify the matching algorithm. The anisotropic covariance in terms of individual mixture endows VBPSM the flexibility to model the matching uncertainty of each data points in essence.

5.2 3D affine point set matching

In the 3D point set matching experiment, VBPSM is applied to 3D face matching experiment and the results are demonstrated in Fig. 7. There are $N = 392$ points of model and scene, respectively. In this experiment, we firstly perform our method on clean data and then test the validity of our method on noised data. In the noise experiment setup, we first delete 64 data points around the left eye in scene. After that, SNR $= 20$ Gaussian white noise is added to each point of model and scene, respectively. Finally, $0.2\times N$ uniformly distributed outliers are independently being appended to the model and scene.



(a) Initial setup of face data without noise

(b) Matching results for the face data without noise

(c) Initial setup of face data with noise

(d) Matching results for the face data with noise (The model and scene points are represented by black '+' and 'o', respectively.)

Fig. 7 The matching results of VBPSM on 3D face data with and without noise

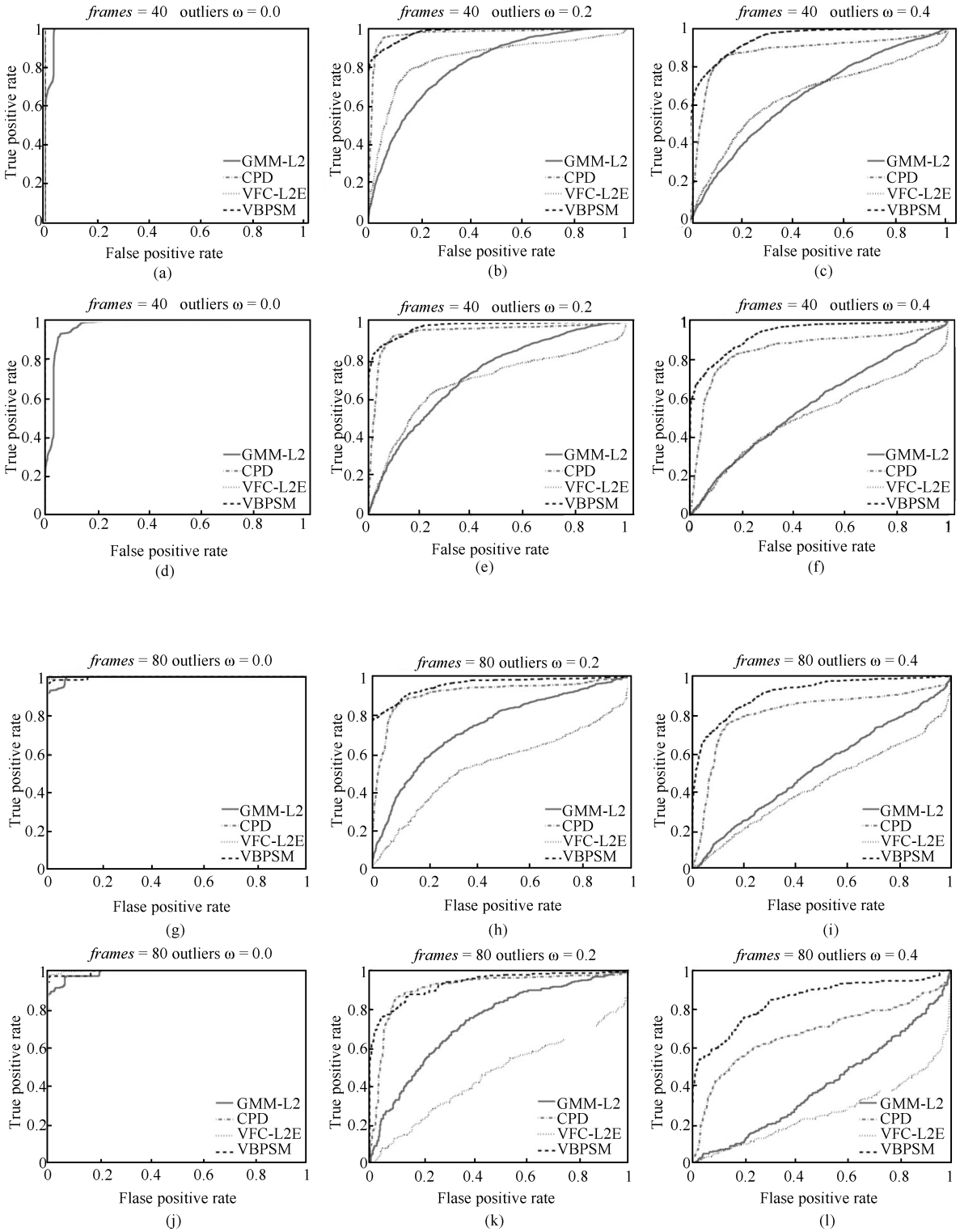
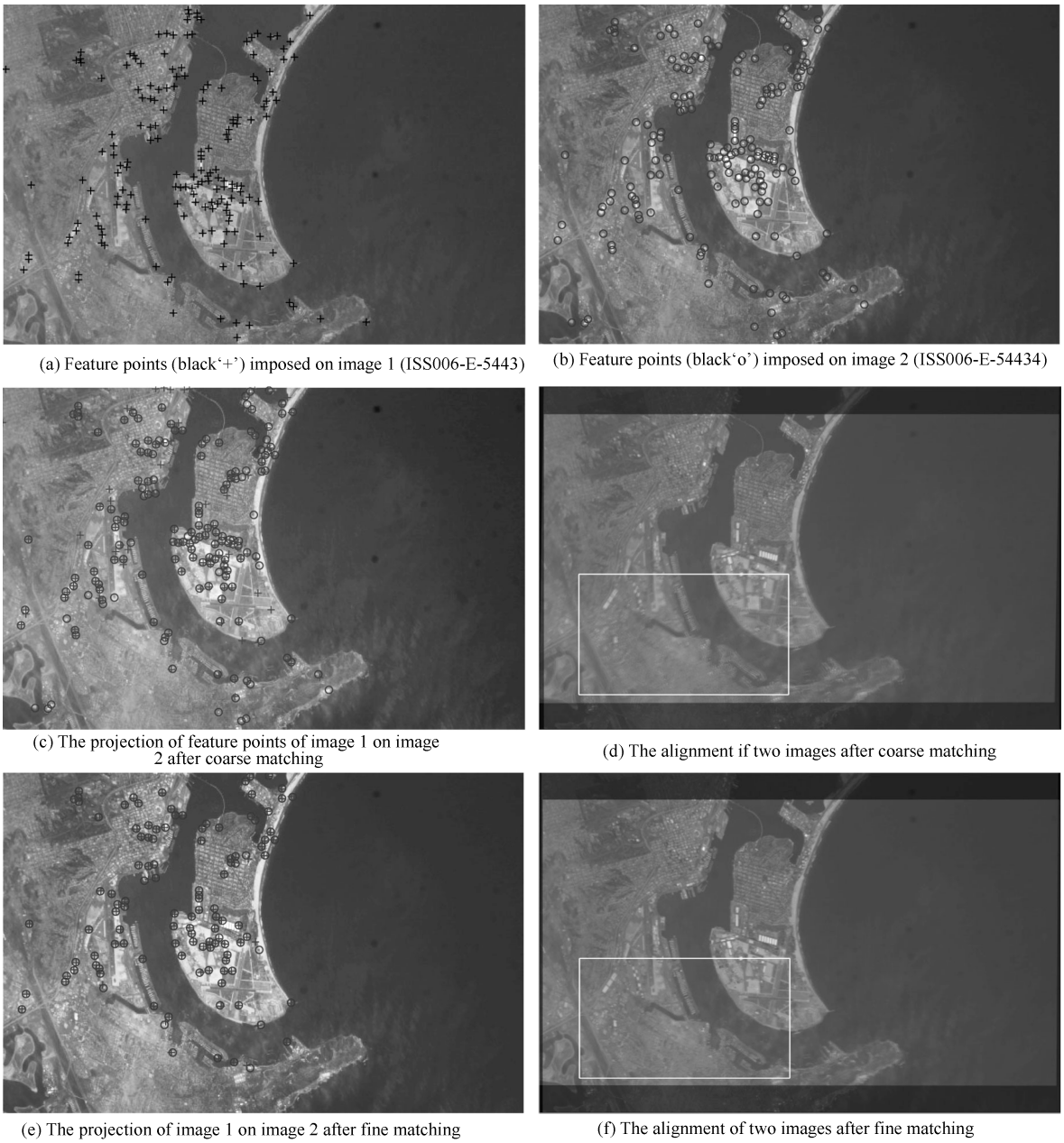


Fig. 8 Performance comparison of GMM-L2[12], CPD[14], VFC-L2E[11] and VBPSM on *CMU* house sequence dataset in terms of different outlier-to-data ratios w, which are 0.0, 0.2 and 0.4, respectively. (a) ∼ (c): ROC curves obtained from data with separation frames 40; (d) ∼ (f): ROC curves obtained from data with separation frames 60; (g) ∼ (i): ROC curves obtained from data with separation frames 80; (j) ∼ (l): ROC curves obtained from data with separation frames 100

Fig. 7 (a) and Fig. 7 (c) demonstrate the initial setups for clean and the noised cases. Fig. 7 (b) and Fig. 7 (d) provide the matching results obtained using VBPSM. One can see from the results that VBPSM achieves good performance on 3D face data not only in the clean case but also in the presence of Gaussian noise, outliers and missing points.



(a) Feature points (black'+') imposed on image 1 (ISS006-E-5443)

(b) Feature points (black'o') imposed on image 2 (ISS006-E-54434)

(c) The projection of feature points of image 1 on image 2 after coarse matching

(d) The alignment if two images after coarse matching

(e) The projection of image 1 on image 2 after fine matching

(f) The alignment of two images after fine matching

Fig. 9 Alignment of two astronaut images

5.3 Quantitative comparison

In the quantitative comparison, quantitative experiments are conducted on CMU house sequence[2]. There are totally 111 toy house images from different viewpoints, and 30 landmark points are manually marked as known correspondence. Uniform outliers are added to the model and scene simultaneously, and a set of outlier points are appended to both sets with the same outlier-to-data ratio w. Image pairs spaced by 40, 60, 80 and 100 frames are selected for the comparison of matching accuracy, and four point set matching algorithms are applied to these data sets in terms of outlier-to-data ratios of $w=0.0$, $w=0.2$ and $w=0.4$, respectively. The parameters are tuned for the best at the outlier-to-data ratio $w=0.0$ and separation frames $=40$, and they are kept unchanged in the other cases for these four algorithms.

Fig. 8 illustrates the receiver operating characteristic (ROC) curves obtained by these algorithms. For outlier-to-data ratio $w=0.0$, CPD, VFC and VBPSM achieve almost the same matching performance, whereas the area under ROC curve (AUC) is almost equal to 1. As the outlier-to-data ratio w increases, our method achieves better performance than the other three algorithms in all of these cases. The full probabilistic modeling of uncertainty endows VBPSM robustness to outliers and the ability of being less prone to local minima. Though the other three algorithms achieve good performance in the data without outliers, their matching ability may be ruined greatly in

[2]http://vasc.ri.cmu.edu/idb/html/motion/house/index.html

the presence of spurious outliers.

5.4 Real applications

In this experiment, we perform image registration of two astronaut photos of earth by VBPSM. The images are downloaded from the "The Gateway to Astronaut Photography of Earth" program of Image Science and Analysis Laboratory, NASA-Johnson Space Center. Two photographes, ISS006-E-54433 and ISS006-E-54434, are camera captures of San Diego bay, North Island of California from the international space station in the same day of May 2003. These photographes are converted to gray images of size 639×647 from their original color JPG format for convenience sake. There are 172 and 163 corners extracted from two images respectively by Harris corner detector, as shown in Fig. 9 (a) and Fig. 9 (b). In our experiment, no putative matches are assumed for forward image alignment, and we conduct the point set matching directly on the detected feature points in two images.

In this paper, we propose a coarse-to-fine strategy to align the two images by VBPSM algorithm. In the coarse phase, we perform VBPSM to match the two feature point sets and then a responsibility matrix between two point sets is obtained, as is shown in Fig. 10 (a). There are outliers in both feature point sets after the first step and these outliers would deteriorate alignment. Fig. 9 (c) provides the projection of feature points of image 1 on image 2, and Fig. 9 (d) gives the alignment result obtained using the affine parameters estimated by VBPSM.

It is evident that there are a number of mismatches between these two point sets after coarse matching and these outliers will blur the composite image, as is shown in the white window in left bottom of Fig. 9 (d). As the responsibility matrix between two feature points has been obtained after the coarse estimation of affine transformation, one can find these corresponding point pairs by the observation of response amplitude $\hat{r}_{nk}$. The corresponding point pairs are obtained by the following steps: 1) Finding the most probable corresponding point in image 1 with respect to each point in image 2 by maximizing the responsibility matrix along the column, 2) Finding the most probable corresponding point in image 2 with respect to each point in image 1 by maximizing the responsibility matrix along the row, 3) If the two corresponding positions are same and the response $\hat{r}_{nk}$ at this position is greater than a threshold, then these two feature points are assumed to be corresponding pair. The threshold is empirically set to be 0.2 in this experiment.

After the elimination of spurious points, VBPSM algorithm is re-run on these two new point sets. Fig. 10 (b) shows the re-estimated responsibility matrix. Fig. 9 (e) provides the projection of corresponding point pairs, and Fig. 9 (f) demonstrates the merging result of two images after the the fine matching. As one can see from the result, most outliers are excluded by this method, and the detected inlying point pairs are adequate to achieve a good estimation of the affine transformation between two images. The blurring effect is eliminated after the fine step, which can be seen from the content in the white window of Fig. 9 (f) by comparing with that of Fig. 9 (d).

6 Conclusion

In this work, we propose a variational approximation algorithm for affine point set matching under Bayesian framework. Matching two point sets is considered as a probabilistic inference process, where the transformed model points are represented by Gaussian mixtures and the scene set is treated as data points. Furthermore, the affine transformation parameters are assumed to be uncertain as well, and thus the probabilistic distribution is placed over each individual parameters. Under the variational Bayesian framework, an iteratively updating algorithm is formulated to obtain the approximate posteriors of parameters, and the negative free energy approaches to the lower bound of log marginal probability.

The contributions of this work can be summarized as the follows. A graphical model is presented for the affine point set matching and a fully probabilistic approximation is derived for the inference of the matching uncertainty, which has not been exactly derived in previous work to our knowledge. A closed form solution is formulated to approximate the posterior distributions of latent variables under the probabilistic framework. To be mentioned, the approximate distribution is not restricted to Gaussian mixtures, other probabilistic model, such as Student mixtures which is robust to outliers, can be conveniently introduced under this framework. The proposed algorithm demonstrates robust and accurate performance in comparison with other state-of-the-art point registration algorithms on various synthetic and real datasets.

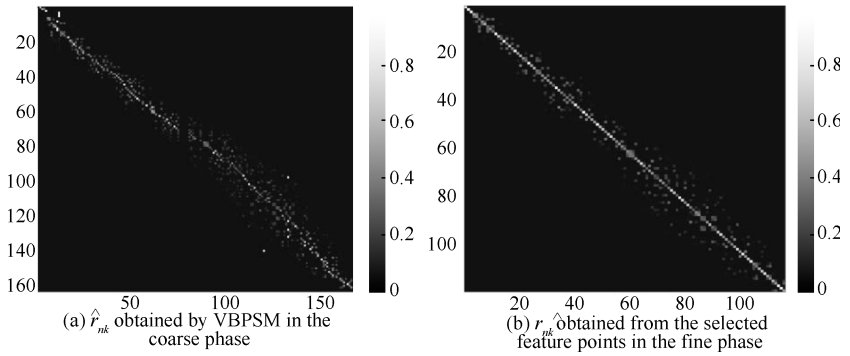


Fig. 10 Responsibility matrix $\hat{r}_{nk}$ between feature points in two images

References

1 Caetano T S, Caelli T, Schuurmans D, Barone D A C. Graphical models and point pattern matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2006, **28**(10): 1646−1663

2 Cootes T F, Taylor C J, Cooper D H, Graham J. Active shape models — their training and applications. *Computer*

Vision and Image Understanding, 1995, **61**(1): 38−59

3 Chen T Z, Li Y. A high performance edge point feature match method of SAR images. *Acta Automatica Sinica*, 2014, **39**(12): 2051−2063

4 Gu L, Kanade T. A generative shape regularization model for robust face alignment. In: Proceedings of the 10th European Conference on Computer Vision. Marseille, France: Springer, 2008. 413−426

5 Pilet J, Lepetit V, Fua P. Real-time nonrigid surface detection. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, USA: IEEE, 2005. 822−828

6 Pizarro D, Bartoli A. Feature-based deformable surface detection with self-occlusion reasoning. *International Journal of Computer Vision*, 2012, **97**(1): 54−70

7 Zhao J, Ma J Y, Tian J W, Ma J, Zhang D Z. A robust method for vector field learning with application to mismatch removing. In: Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition. Providence, RI: IEEE, 2011. 2977−2984

8 Simpson I J A, Schnabel J A, Groves A R, Andersson J L R, Woolrich M W. Probabilistic inference of regularisation in non-rigid registration. *NeuroImage*, 2012, **59**(3): 2438−2451

9 Tsin T, Kanade T. A correlation-based approach to robust point set registration. In: Proceedings of the 8th European Conference on Computer Vision. Prague, Czech: Springer, 2004. 558−569

10 Chui H, Rangarajan A. A new point matching algorithm for non-rigid registration. *Computer Vision and Image Understanding*, 2003, **89**(2−3): 114−141

11 Zhang Z Y. Iterative point matching for registration of free-form curves and surfaces. *International Journal of Computer Vision*, 1994, **13**(2): 119−152

12 Jian B, Vemuri C C. Robust point set registration using Gaussian mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011, **33**(8): 1633−1645

13 Chui H, Rangarajan A. A feature registration framework using mixture models. In: Proceedings of the 2000 IEEE Workshop on Mathematical Methods in Biomedical Image Analysis. Hilton Head Island, USA: IEEE, 2000. 190−197

14 Myronenko A, Song X B. Point set registration: coherent point drift. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010, **32**(12): 2262−2275

15 Horaud R P, Forbes F, Yguel M, Dewaele G D, Zhang J. Rigid and articulated point registration with expectation conditional maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011, **33**(3): 587−602

16 Ma J Y, Zhao J, Tian J W, Tu Z W, Yuille A. Robust estimation of nonrigid transformation for point set registration. In: Proceedings of the 2003 IEEE Conference on Computer Vision and Pattern Recognition. Portland, OR: IEEE, 2013. 2147−2154

17 Zhou Zhi-Yong, Li Li-Hua, Zheng Jian, Kuai Dou-Jie, Hu Su, Zhang Tao. Point sets non-rigid registration using student′s-t mixture model with spatial constraints. *Acta Automatica Sinica*, 2014, **40**(4): 683−696 (in Chinese)

18 Rangarajan A, Coughlan J, Yuille A L. A Bayesian network framework for relational shape matching. In: Proceedings of the 9th IEEE International Conference on Computer Vision. Nice, France: IEEE, 2003. 671−678

19 Zhou Y, Zhang W, Tang X O, Shum H. A Bayesian mixture model for multi-view face alignment. In: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego, CA, USA: IEEE, 2005. 741−746

20 Green P J, Mardia K V. Bayesian alignment using hierarchical models, with applications in protein bioinformatics. *Biometrika*, 2006, **93**(2): 235−254

21 Czogiel I, Dryden I L, Brignell C J. Bayesian matching of unlabeled marked point sets using random fields, with application to molecular alignment. *The Annals of Applied Statistics*, 2011, **5**(4): 2603−2629

22 Bishop C M. *Pattern Recognition and Machine Learning*. New York: Springer, 2006

23 Klami A. Variational Bayesian matching. In: Proceedings of the 4th Asian Conference on Machine Learning. Singapore, 2012. 205−220

24 Govindu V M, Werman M. On using priors in affine matching. *Image and Vision Computing*, 2004, **22**(14): 1157−1164

25 MacKay D J C. Bayesian non-linear modeling for the energy prediction competition. *ASHRAE Transactions*, 1994, **100**: 1053−1062

26 Neal R M. *Bayesian Learning for Neural Networks*. New York: Spring, 1996

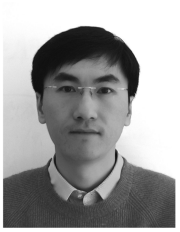
27 Beal M, Ghahramani Z. Variational Bayesian learning of directed graphical models with hidden variables. *Bayesian Analysis*, 2006, **1**(4): 793−832

28 Beal M J. Variational Algorithms for Approximate Bayesian Inference [Ph. D. dissertation], University of Cambridge, Cambridge, UK, 2003.

29 Lu C P, Mjolsness E. Two-dimensional object localization by coarse-to-fine correlation matching. In: Proceedings of the 1994 Advances in Neural Information Processing Systems 6. San Francisco, CA: Morgan Kaufmann, 1994. 985−992

30 Fraley C, Raftery A E. How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, 1998, **41**(8): 578−588

31 Myronenko A, Song X B, Carreira-Perpiñán M Á. Non-rigid point set registration: coherent point drift. In: Proceedings of the 2006 Advances in Neural Information Processing Systems 19. Vancouver, British Columbia, Canada, 2006. 1009−1016



QU Han-Bing Received the M. S. and Ph. D. degrees from Harbin Institute of Technology (HIT) and Institute of Automation, Chinese Academy of Sciences (CASIA) in 2003 and 2007, respectively. Currently, He is an associate professor in Beijing Institute of New Technology Applications and is the director of Key Laboratory of Pattern Recognition, Beijing Academy of Science and Technology (BJAST). He is also a committee member of Intelligent Automation Committee of Chinese Association of Automation (IACAA). His research interest covers biometrics, machine learning, pattern recognition and computer vision. Corresponding author of this paper.
E-mail: quhanbing@gmail.com

CHEN Xi Received the M. S. degree from School of Computer Science and Engineering, Hebei University of Technology (HEBUT) in 2013. Currently, He is a Ph. D. candidate in School of Electronic and Information Engineering, HEBUT, and a research assistant in Beijing Institute of New Technology Applications (BIONTA) and the Key Laboratory of Pattern Recognition, Beijing Academy of Science and Technology (BJAST). His research interest covers image processing, machine learning, pattern recognition and computer vision.
E-mail: csxichen@outlook.com

WANG Song-Tao Received the M. S. degree from Harbin University Of Science and Technology (HUST) in 2009, and is currently a Ph. D. candidate in the Higher Educational Key Laboratory for Measuring & Control Technology and Instrumentations of Heilongjiang Province, HUST. He is also a research assistant in Beijing Institute of New Technology Applications (BIONTA) and the Key Laboratory of Pattern Recognition, Beijing Academy of Science and Technology (BJAST). His research interest covers pattern recognition and computer vision, especially the visual saliency detection in surveillance scenarios.
E-mail: wangsongtao1983@163.com

YU Ming Received the B. S. degree in electrical and information engineering from Beijing University of Posts and Telecommunications in 1986, and the M. S. degree in electronic systems from the 54th Research Institute of China Electronics Technology Group Corporation (Shijiazhuang) in 1989. He received the Ph. D. degree in communication and information systems from Beijing Institute of Technology, Beijing, China, in 1999. Since 2000, he has been a full professor at School of Computer Science and Engineering of Hebei University of Technology, Tianjin, China. His research interest covers image processing and pattern recognition, image and video compression, visual computing and application, and biometrics.
E-mail: yuming@hebut.edu.cn