Robust Variational Bayesian Point Set Registration

Jie Zhou [1,2,†] Xinke Ma [1,2,†] Li Liang [1,2] Yang Yang [1,2,*] Shijin Xu [1,2] Yuhe Liu [1,2] Sim Heng Ong [3]
[1] School of Information Science and Technology, Yunnan Normal University
[2]Laboratory of Pattern Recognition and Artificial Intelligence, Yunnan Normal University
[3] Department of Electrical and Computer Engineering, National University of Singapore
[†] Jie Zhou and Xinke Ma contributed equally to this work.
[*]Yang Yang is corresponding author.
[*] Email: yyang_ynu@163.com

Abstract

In this work, we propose a hierarchical Bayesian network based point set registration method to solve missing correspondences and various massive outliers. We construct this network first using the finite Student's t *latent mixture model (TLMM), in which distributions of latent variables are estimated by a tree-structured variational inference (VI) so that to obtain a tighter lower bound under the Bayesian framework. We then divide the TLMM into two different mixtures with isotropic and anisotropic covariances for correspondences recovering and outliers identification, respectively. Finally, the parameters of mixing proportion and covariances are both taken as latent variables, which benefits explaining of missing correspondences and heteroscedastic outliers. In addition, a cooling schedule is adopted to anneal prior on covariances and scale variables within designed two phases of transformation, it anneal priors on global and local variables to perform a coarse-to-fine registration. In experiments, our method outperforms five state-of-the-art methods in synthetic point set and realistic imaging registrations.*

1. Introduction

Point set registration is the process of finding one-to-one correspondence of two point sets (the model and scene) in ways of which supposed spatial transformations are exerted on the model point set. It is actually to recover the deformation while taking the noise, outlier and missing into account. Point set registration plays an indispensable role in numerous applications such as remote sensing image registration [12, 33, 37], medical image analysis [22], deformable motion tracking [9, 26], intermediate frames interpolation, *etc*. We summarize the existing problems of traditional point set registration algorithms in dealing with outliers, missing correspondences and objective function optimization.

There are three approaches to deal with outliers: (i) The first approach is to construct extra modeling for outliers. Robust Point Matching (RPM) [4] utilizes a correspondence matrix where extra entries are added to address outliers, and the number of outliers is reasonable restrained by conducting a linear assignment to the correspondence matrix. Mixture point matching (MPM) [3] and Coherent Point Drift (CPD) [20] employ an extra component of uniform distribution embedded into Gaussian mixture model to handle outliers, but the correspondences are badly biased because the extra fixed component can not capture distributions of positional Gaussian outliers. Han-Bing Qu *et al*. [27] proposes extra Gaussian mixtures to fit outliers, it alleviates the estimation bias arising from positional dense outliers even though it may lack robustness to the outliers presenting as non-Gaussianity. (ii) The second approach is to employ a more robust parameter estimator (*e.g*., L_2E [8, 14]) to minimize the difference between the two point sets which are taken as the means of two Gaussian mixture models (GMM). Tsin *et al*. in the kernel correlation (KC) [32] maximize the correlation of the kernel densities constructed by the two sets. Jian and Vemuri [8], Ma *et al*. [14] apply the L_2 estimator to maximize the correlation of two Gaussian densities. Their methods [8, 14, 32] are different from the Maximum Likelihood estimates (MLE) which will bring about a bias estimation if there is a fraction of outliers, because the penalty intensity of L_2E is more smooth than that of quadratic MLE in the truncation of their penalty curves [8, 14]. So even the remote point (containing the outliers that cause the biased estimates) can also be assigned with a low probability. Therefore, the distribution of outliers can be captured by the tails of all Gaussian components.

But the tuning for the tail of penalty curve relies on an artificial annealing to adjust covariance directly, and an identical covariance for all model and scene points can not free the uncertainty to vary in terms of the context of each point. In addition, the results of their method are usually unsatisfactory when outliers has the same pattern as inliers. (iii) The third approach aims to capture the tail which, in general, contains outliers which can lead to biased estimates. Hence, this approach is essential to select a heavy tailed distribution where the tail can be tuned to accommodate varying degrees of outliers. Besides, the empirical assumption for distribution inevitably brings departures to reality distributions. Thus, a family density with broader distributions is needed to obtain a better generality to the outliers representing as Gaussianity or non-Gaussianity. Recently, the Student's *t*-distribution is proposed to alleviate the vulnerability of Gaussian to outliers caused by its quadratic decaying tail. Student's *t*-distribution is widely used in medical image segmentation [23,24], data classification [29] and data clustering [30]. In addition, Gaussian can be regarded as a special case of Student's *t* when the degree of freedom (dof) increases to infinity. Peel and McLachlan [25] introduce an alternative form of Student's *t* to figure out analytical solutions arduously with an EM procedure. Gerogiannis *et al*. [5, 6] propose a SMM based registration algorithm that shows a better robustness than GMM based methods. However, the trade-off between efficiency and accuracy is still a problem during objective function optimization. (i) The first is to construct extra modeling for outliers. Robust Point Matching (RPM) [4] utilizes a correspondence matrix where extra entries are added to address outliers, and the number of outliers is reasonable restrained by conducting a linear assignment to the correspondence matrix. Mixture point matching (MPM) [3] and Coherent Point Drift (CPD) [20] employ an extra component of uniform distribution embedded into Gaussian mixture model to handle outliers, but the correspondences are badly biased because the extra fixed component can not capture distributions of positional Gaussian outliers. Han-Bing Qu *et al*. [27] propose extra Gaussian mixtures to fit outliers. It alleviates the estimation bias arising from positional dense outliers even though it may lack robustness to the outliers presenting as non-Gaussianity.

There are also three ways to deal with the problem of missing correspondence: (i) Missing correspondence can cause severe changes to the transformation invalidating the maintaining of point set structure. Existing methods mostly rely on more certain correspondences and constrained transformation (*e.g*., affine or even rigid) to address this problem [3,4,8,14–17,20,32,35]. CPD and its extensions [7,35,36] add spatial coherence constraints on transformation. Recently, Ma *et al*. [15–17] and Zhang *et al*. [35, 36] present the global-local point structure to obtain more certain correspondences. But their method do not face with this problem directly. (ii) MLE based methods [3, 4, 8, 14, 17, 20, 32, 35] are to minimize the Kullback-Leibler (KL) divergence of distributions [8] of the two sets where each point has the same covariance, but the KL divergence (also called the I-projection) shows a property that the model has a tendency to under-estimate the support of the distribution of the scene set [18] because the asymmetric KL divergence are not a metric. Therefore, the model point tends to lock onto one of the scene points, L_2E based methods [8, 14] use a symmetric divergence elegantly to solve the problem of mixing correspondence arising from the asymmetric KL divergence. However, their approach is still indirect, and the mixing proportion of the disappeared model should be reweighted to gain at least a small value. (iii) For the variational method, the optimization for parameter, hierarchically taken as random variable, is yielded from a prior governed functional space, it also acquires a whole predictive distribution rather than only one point estimate. Han-Bing Qu *et al*. [27] use the Dirichlet prior to re-weight mixing proportion and optimize objective function under the Variational Bayesian (VB) framework. The same idea also arises in DSMM [38,39], it discovers that the varying mixing proportion is beneficial to address missing correspondences.

We summarize three problems of objective function optimization: (i) The intuitive EM framework is used widely in numerous works [17, 35, 38, 39], and the fact that MLE is implemented by the expectation maximization (EM) algorithm [21], boosts greatly the development of registration. Nevertheless, an undesirable over-fitting problem will arise from the unbounded likelihood function of MLE, and those MLE based point estimators [8, 14, 17, 35, 38, 39] also tend to be trapped into poor local optima easily. (ii) From the variational perspective, the parameter taken as random variable is estimated to acquire a predictive distribution which is then bounded by a prior by Bayesian method, it alleviates significantly the over-fitting problem of MLE. However, the current VB based methods [6, 27, 38, 39] are all to follow the mean-field full factorization to approximate individually the marginal densities of model posterior, so they can not capture the correlations among variables and certainly lose the expressiveness of variational distribution. (iii) The ill-posed no-rigid deformation calls for a deterministic annealing to stimulate the algorithm to escape from poor local optima, the traditional annealing process [4, 8, 14, 20, 32, 32] update the covariance directly, therefore the effect of empirical parameter setting on algorithm stability is reduced. Continuing the discussion regarding VB methods, the optimization of objective ELBO requires a trade-off between variational distributions that could better express the underlying distributions and that with high entropy. VBPSM is essential to anneal the same prior on covariances to achieve the same traditional

annealing process [4, 8, 14, 20, 32].

2. Method

We propose a new point set registration method to deal with the problems involving outlier, missing correspondence and objective function optimization. In section.2.1.1, we apply a broader class distribution namely the heavy-tailed *t*-distribution which endows our model with nature robustness to outliers. Furthermore, its tail is adaptively tuned by using a hierarchical Bayesian network which, under the Variational Bayesian framework, is reformulated where both mixing proportion and covariances are taken as latent variables, so it gives our model the better interpretations of mixing correspondence and heteroscedasticity compared with the non-Bayesian approach. And then in section.2.1.2, we separate the mixture model into the two where the transition mixture with isotropic covariances is constrained by transformation, and the other with anisotropic covariances is used to find the outliers with multiple clusters. This separation is complement with the reformulated Student's *t* latent mixture model (TLMM) that further used to capture tails. For the optimization, in section.2.2, the tree-structured variational factorization (SVI) is employed to induce variational dependency for obtaining a higher-fidelity variational distributions under the two-steps iterative optimization, *i.e.*, the Variational Bayesian Expectation Maximization (VBEM) algorithm in section.2.2.1 and section.2.2.2. In addition, we deduce the transformation beyond the conjugate family in section.2.3, and the designed two phases of transformation based on the global and local cooling schedule are introduced in section.2.4.

2.1. Hierarchical Bayesian model for point set registration

Given a model set $\mathcal{M} = \{m_k\}_{k=1}^{K}$ with K points and a scene set $\mathcal{S} = \{s_n\}_{n=1}^{N}$ with N points, where each m_k and s_n has the same dimension D, the $\mathcal{T}$ is supposed as a latent variable that transforms the model onto the scene point sets and an auxiliary variable $\mathcal{T}(\mathcal{M})$ is introduced to denote the transition between the model and the scene.

2.1.1 Student's *t* mixture model

The Student's *t*, a family of distributions, is comprised by an infinite mixture of overlapped Gaussians with the same mean but different scaled covariances, so its tail can be tuned by varying the scaling of covariances, and using the one with heavier tail, at least in the beginning of iterations, is less tend to give biased estimates in response to outliers. For the convenience of conjugate inference, we use the alternative form called the Gaussian scale mixture which can be written as the convolution of a Gaussian with a Gamma distribution:

$$\mathcal{S}t(\mathbf{s}_n;\mu,\Upsilon,l) = \int \mathcal{N}\big(\mathbf{s}_n;\mu,(\varepsilon_n\Upsilon)^{-1}\big)\mathcal{G}(\varepsilon_n;\tfrac{l}{2},\tfrac{l}{2})\,d\varepsilon_n, \tag{1}$$

where $\mathcal{S}t$, $\mathcal{N}$ and $\mathcal{G}$ denote Student's *t*, Gaussian and Gamma probability density function, respectively. We find that the precision Υ is scaled in terms of variable ε that is governed by a Gamma distribution with the same prior parameters $\frac{l}{2}, \frac{l}{2}$. The finite Student's *t* mixture model (SMM) takes advantage of the superior statistical properties of Student's *t*-distribution. Furthermore, we divide SMM into two different mixtures for simultaneously accomplishing registration and outlier rejection. The transition mixture, with K components, is used to find correspondence, and the outlier mixture, with K_0 components, is used to process outliers. Hence, the SMM is written as:

$$p(\mathbf{s}_n;\varphi,\phi) = \sum_{k=1}^{K}\pi_k\mathcal{S}t\big(\mathbf{s}_n;\varphi_k\big)\sum_{k=K+1}^{K+K_0}\pi_k\mathcal{S}t(\mathbf{s}_n;\phi_k), \tag{2}$$

where π denotes the mixing proportion, *i.e.*, $\sum_{k=1}^{k=K+K_0}\pi_k = 1$. It is in effect the posterior of the fuzzy correspondence inferred from parameters φ_k and ϕ_k, so it plays an essential role as re-estimated responsibility of all components [1]. And regarding the missing correspondence can be interpreted as the disappeared mixing proportion since it does not generate scene point, so the parameterized mixing proportions are taken as random variables jointly governed by a Dirichlet prior which frees the mixing proportions to vary and also takes effect on re-weighting less for models which are more likely to miss correspondences, that is to say, it reduces the contributions of those models to promote transformation, and those models in turn tend to be dragged by other models which have more responsibility during the next iterations.

2.1.2 The latent variable model of SMM

The latent assignment variable $\mathcal{Z} = \{\mathbf{z}_n\}_{n=1}^{K+K_0}$ of mixture model is comprised by $K+K_0$ vectors, and each of which is drawn from a categorical distribution, that is, all its entries are zero except for a one at the position indicating correspondence. So the distribution of $\mathcal{Z}$ conditional on π has a multinomial density:

$$p(\mathbf{z}_n|\pi) = \prod_{k=1}^{K+K_0}\pi_k^{z_{nk}}. \tag{3}$$

The re-formulate mixture model can be written as:

$$p(\mathbf{s}_n|\mathcal{H},\Theta) = \prod_{k=1}^{K}\mathcal{N}\big(\mathbf{s}_n|\mathcal{T}(m_k),(u_{nk}\Lambda_k)^{-1}\big)^{z_{nk}} \prod_{k=K+1}^{K+K_0}\mathcal{N}(\mathbf{s}_n|x_k,(v_{nk}\Gamma_{nk})^{-1})^{z_{nk}}, \tag{4}$$

where x_k denotes the mean of one component in outlier mixture, and $\mathcal{U} = \{\mathbf{u}_n\}_{n=1}^N$ and $\mathcal{V} = \{\mathbf{v}_n\}_{n=1}^N$ are the latent scale variables of two mixtures respectively, and both of them are Gamma distributed:

$$\begin{aligned} p(\mathbf{u}_n|\mathbf{z}_n) &= \prod_{k=1}^{K} \mathcal{G}(u_{nk}; \frac{a_k}{2}, \frac{a_k}{2})^{z_{nk}}, \\ p(\mathbf{v}_n|\mathbf{z}_n) &= \prod_{k=K+1}^{K+K_0} \mathcal{G}(v_{nk}; \frac{b_k}{2}, \frac{b_k}{2})^{z_{nk}}. \end{aligned} \tag{5}$$

We define two collections named model structure and model parameter, *i.e.*, $\mathcal{H} = \{\mathcal{Z}, \mathcal{U}, \mathcal{V}\}$, $\Theta = \{\mathcal{T}, \Lambda, \mathcal{X}, \Gamma, \pi, a, b\}$, respectively. We can verify that taking integral of the product of Eq.3, Eq.5 and Eq.4 over $\mathcal{H}$ leads to the generative model showed in Eq.2. According to the conjugacy property that posterior is preserved in the same exponential family of prior, the prior imposed on mean, precision and mixing proportion is deduced in terms of the likelihood terms shown in Eq.14, Eq.5 and Eq.4. Specifically, the mixing proportions are jointly imposed on a prior of Dirichlet $\mathcal{D}(\pi; \alpha_0)$, the means and precisions of outlier mixtures are jointly governed by a Gaussian-Wishart prior $\mathcal{NW}(\mathcal{X}_k, \Gamma_k; \eta_0, \zeta_0, \varepsilon_0, \beta_0)$, and a Gamma $\mathcal{G}(\Lambda_k; \rho_0/2, \phi_0/2)$ is imposed on each transition precision Λ_k. But, we do not consider priors on degree of freedom a and b since they are no conjugate priors. So we collect all hyper-parameters, *i.e.*, $\vartheta = \{\rho, \phi, \varepsilon, \beta, \eta, \zeta\}$. Note that the transformation will be inferred directly by taking derivation to obtain a point estimate. The more detailed discussion can refer to section.2.3. So far, the completed TLMM is represented by a directed acyclic graph, as shown in Fig.1.

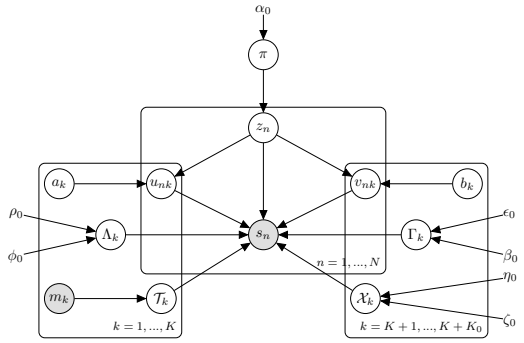


Figure 1. Directed acyclic graph of hierarchical probability model for point set registration

Furthermore, the joint distribution of all variables and parameters can be factorized by the chain rule of probability, as shown below:

$$\begin{aligned} p(\mathcal{S}, \mathcal{H}, \Theta) =& p(\mathcal{S}|\mathcal{H}, \Theta)p(\mathcal{U}|\mathcal{Z})p(\mathcal{V}|\mathcal{Z})p(\mathcal{Z}|\pi)p(\pi) \\ & p(\mathcal{X})p(\Gamma)p(\mathcal{T})p(\Lambda). \end{aligned} \tag{6}$$

2.2. Tree-structured variational inference

The idea of Bayesian is to make inferences of posterior involving an intractable integral over the ensemble of all unknown quantities contained in the presented probabilistic model. An alternative to that exact inference is to resort the approximation by using a parameterized and therefore simpler distribution, because this distribution can be tuned by the variational parameter to minimize the departure to true posterior through variational method. Thus, this problem is in fact turned into an optimization over variational parameters with variational inference (VI) to minimize the commonly used Kullback-Leibler (KL) divergence of variational distribution (also known as the proxy) and model posterior. It is worth noting that we can lower bound the evidence to embody the term of KL divergence, mathematically, the logarithmic evidence taking on a summation:

$$\ln p(\mathcal{S}) = \mathcal{L}(q(\mathcal{H}), q(\Theta), \mathcal{S}) + \mathbb{KL}(q(\mathcal{H}, \Theta)||p(\mathcal{H}, \Theta|\mathcal{S})), \tag{7}$$

where $\mathcal{L}$:

$$\begin{aligned} &\mathcal{L}(q(\mathcal{H}), q(\Theta), \mathcal{S}) \\ &= \int_{\Theta}\int_{\mathcal{H}} q(\mathcal{H}, \Theta)[\ln p(\mathcal{S}, \mathcal{H}, \Theta) - \ln q(\mathcal{H}, \Theta)]d\mathcal{H}d\Theta. \end{aligned} \tag{8}$$

$\mathcal{L}$ is known as the evidence lower bound (ELBO) and, as a result, this optimization is equivalent to maximize the ELBO, as it can be verified from Eq.7 that the ELBO is exact when the posterior $p(\mathcal{H}, \Theta|\mathcal{S})$ matches the variational distribution $q(\mathcal{H}, \Theta)$. Note that the item $p(\mathcal{S}, \mathcal{H}, \Theta)$ in Eq.7 joints all observations, variables and parameters. Following the fully factorial assumption which is the most common tractable form of VI, the proxy is factorized as a product of independent marginals, respectively named the model structure $\mathcal{H}$ and model parameter Θ, *i.e.*, $q(\mathcal{H}, \Theta) = q(\mathcal{H})q(\Theta)$. Then, the two steps iterative optimization of VBEM for non-convex ELBO can be concluded as follows:

$$\begin{aligned} q^{t+1}(\mathcal{H}) &= \arg\max_{q(\mathcal{H})} \quad \mathcal{L}(q(\mathcal{H}), q^t(\Theta), \mathcal{S}), \\ q^{t+1}(\Theta) &= \arg\max_{q(\Theta)} \quad \mathcal{L}(q^{t+1}(\mathcal{H}), q(\Theta), \mathcal{S}), \end{aligned} \tag{9}$$

where $q^t(\cdot)$ denotes the posterior that has already iterated t times. The advanced coordinate ascent algorithm VBEM approximates posterior for one variable, which requires taking expectation of its complete conditional (the product of the conditional densities of all variables which depend on the one variable in the directed graphical model shown in Fig.1) with respect to the remaining variables. Therefore, the approximated posterior for the one variable absorbs the mean values of those remaining variables locating in the one variable's Markov-blanket.

2.2.1 VB-E: Derivation of the latent variables

Experiential independency assumptions may be correct and valid among some variables in a specific problem. It is used to trade approximate accuracy for feasibility or efficiency. In this work we use a superior tree-structured factorization over the variational distributions so that it can be written as the following product:

$$q(\mathcal{U},\mathcal{V},\mathcal{Z}) = q(\mathcal{U}|\mathcal{Z})q(\mathcal{V}|\mathcal{Z})q(\mathcal{Z}). \tag{10}$$

The tree-structured factorization induces dependencies between $\mathcal{U}$ and $\mathcal{Z}$, and also between $\mathcal{V}$ and $\mathcal{Z}$, so it makes possible to obtain a tighter ELBO. In literature experiments [10, 30] the tree-structured factorization shows superiority over the full independent assumption as it yields more accurate approximations at the same cost of the scene set containing significant amount of outliers. The conditional density can be deduced by Lagrange multiplier. The expression for the updating of proxy $q(\mathbf{u}_n|z_{nk}=1)$ is given as follows:

$$q(\mathbf{u}_n|k) \propto \exp\langle \ln[p(\mathbf{s}_n|\mathbf{u}_n,k)p(\mathbf{u}_n|k)]\rangle_{q(\mathbf{u}_n|k)}, \tag{11}$$

where $\neg q(\mathbf{u},\mathbf{z})$ denotes the product in Eq.10 except for the factors containing u or z. According to the conjugate prior of the distribution employed, its posterior $q(u_n|k)$ should follow Gamma distribution, *i.e.*, $\mathcal{G}(u_n;\omega_{nk},\tau_{nk})$. Then its posterior parameters can be updated:

$$\omega_{nk} = \frac{a_{nk}+D}{2}; \tau_{nk} = \frac{a_{nk}+\tilde{\varphi}_{nk}}{2}, \tag{12}$$

where $\tilde{\varphi}_{nk}$ denotes the expectation of a Mahalanobis distance, *i.e.*, $\tilde{\varphi}_{nk} = \langle (s_n - \mathcal{T}(m_k))^T u_{nk} \Lambda_k (s_n - \mathcal{T}(m_k))\rangle$, and $\langle\cdot\rangle$ represents an expectation operator. In the same manner, we define $\tilde{\varepsilon}_{nk} = \langle [s_n - x_k]^T \Gamma_k [s_n - x_k]\rangle$, so the variational distribution $q(v_n|k)$ which follows the distribution $\mathcal{G}(u_n;\omega_{nk},\tau_{nk})$ can be updated as follows:

$$\gamma_{nk} = \frac{b_0+D}{2}; \delta_{nk} = \frac{b_0+\tilde{\varepsilon}_{nk}}{2}. \tag{13}$$

The joint posterior of the latent assignment and the scale variable can be obtained from the expectation of their complete conditional. Through marginalization and normalization we can acquire the expectation of z_{nk}:

$$\langle z_{nk}\rangle = \frac{q(z_{nk}=1)}{\sum_{k=1}^{K+K_0} q(z_{nk}=1)}. \tag{14}$$

2.2.2 VB-M: Derivation of the model parameters

The sample size α_k of mixing proportion π is updated:

$$\alpha_k = \alpha_0 + \sum_{k=1}^{K+K_0} \langle z_{zk}\rangle. \tag{15}$$

The posterior parameters of outlier mixtures are written as:

$$\begin{aligned}
\hat{\zeta} &= \zeta_0 + \sum_{n=1}^{N}\langle z_{nj}\rangle, \\
\hat{\varepsilon} &= \varepsilon_0 + \sum_{n=1}^{N}\langle z_{nj}\rangle, \\
\hat{\eta} &= \frac{1}{\hat{\zeta}}\big[\zeta_0\eta_0 + \sum_{n=1}^{N}\langle z_{nj}\rangle \mathbf{s}_n\big], \\
\hat{\beta}^{-1} &= \hat{\beta}_0^{-1} + \frac{1}{\hat{\zeta}}\big[\zeta_0\eta_0\eta_0^{\mathsf{T}} + \mathcal{S}\langle\Phi_k\rangle\mathcal{S}^{\mathsf{T}}\big],
\end{aligned} \tag{16}$$

where $\Phi_k = d([z_{1k}, z_{2k}, ..., z_{Nk}])$, and for $K+1 \le k \le K+K_0$, the function $d(\cdot)$ converts a vector into a diagonal matrix. The posterior parameters of transition mixtures are updated as follows:

$$\begin{aligned}
\rho_k &= \rho_0 + \frac{1}{2}\sum_{n=1}^{N}\langle z_{nk}\rangle, \\
\phi_k &= \phi_0 + \frac{1}{2}\sum_{n=1}^{N} C_{nk}\langle [s_n - \mathcal{T}(m_k)]^T [s_n - \mathcal{T}(m_k)]\rangle.
\end{aligned} \tag{17}$$

2.3. Non-conjugate derivation for transformation

We first consider the derivation for the non-rigid transformation which, according to the Riesz representation theorem [28], can be expressed as a sum form, *i.e.*, $\mathcal{T}(\mathcal{M}) = \mathcal{M} + \mathbf{GB}$, by constructing a Gaussian radial basis function (GRBF), *i.e.*, $G_{ij}(\mathcal{M}) = \exp(-1/2\kappa^2 \|m_i - m_j\|_2^2)$. Then, the problem is converted to solve coefficient matrix $\mathbf{B}$ for GRBF as a result of this representation. Secondly, we regularize the transformation by imposing a non-conjugate prior thereon as constraint so that the points in transition mixture can move coherently to preserve their global structure, *i.e.*, $\ln p(\mathcal{T}) = \lambda/2\mathrm{Tr}(\mathbf{B}^{\mathsf{T}}\mathbf{GB})$. Finally, we can obtain the posterior of $\mathcal{T}$ that is shown as below:

$$\ln q(\mathcal{T}) \propto -\frac{1}{2}\sum_{n=1}^{N}\sum_{k=1}^{K}\langle z_{nk}\rangle\tilde{\varphi}_{nk} - \frac{\lambda}{2}\ln p(\mathcal{T}), \tag{18}$$

where $\mathrm{Tr}(\cdot)$ denotes the trace, parameter κ (default set to 3) controls the spatial smoothness, and parameter λ (default set to 2) controls the regularization intensity. Note that the independent items of transformation are omitted. Besides, the posterior distribution of $\mathcal{T}$ is concave and symmetric, so its mode that is also the only existing extreme point, equals to its mean. Therefore, it is unnecessary to bring a specific parametric constraint to the functional family of the transformation, so the optimal solution of the matrix $\mathbf{B}$, can be obtained by taking partial derivative on the both sides of Eq.18:

$$\mathbf{B} = [d(\mathbf{C}\mathbf{1}_N)^{-1}\mathbf{C}\mathcal{S} - \mathcal{M}][\mathbf{G} + \lambda d(\mathbf{C}\mathbf{1}_N)^{-1}]^{-1}, \tag{19}$$

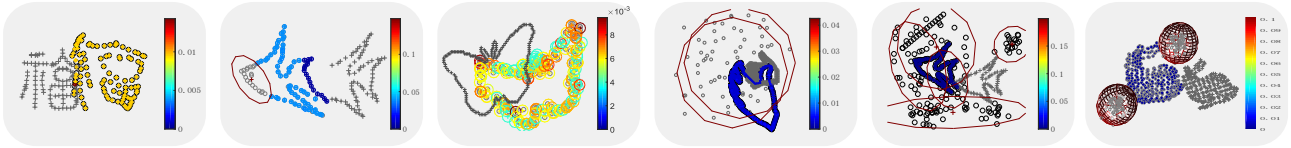


Figure 2. This figure shows the results of six tests. The initial position of model points and scene points are denoted by '+' and 'o', respectively. Transition and outlier mixtures are represented by circle and ellipse which are colored according to estimated mixing proportion $q(\pi)$.

where $\mathbf{1}_d$ denotes the d dimensional identity vector. And we define the matrix with entry $C_{nk} = \langle z_{nk} \rangle \langle u_{nk} \rangle$. As for the transformation confirmed to be rigid/affine, it is also easy to derive them in terms of [19], and their derivations are rarely related to our contributions, so the updating for them is omitted. Besides, it is worth noting that the matrix $\mathbf{C}$ is the equivalent of the probability matrix $\mathbf{P}$ given in [19].

2.4. Two phases transformation and prior annealing

The Variational Bayesian scheme with deterministic annealing technology achieves the penalization for low-entropy distributions in advance, and then cooling temperature through slowly relaxing the intensity of penalty is to re-weight the balance between them. An appropriate cooling schedule employs a series of stages for decreasing temperature realizes the refinement of transformation. Because the differential entropy is proportional to the value of covariance, this value can be a measurement of the uncertainty of correspondences. Specifically, in a temperature stage the prior matrixes on precisions, *i.e.*, ϕ_0, can promote the convergency for algorithm globally. In addition, the two-phases transformation is embedded into a cooling schedule, (*i.e.*, $\phi_0 := 100, 50, 30, 10, 1$), each value of ϕ_0 iterating 100 times. The priors on local variables have an impact on re-weighting all correspondences. Thus, the prior of each scale variable (u_0) is set in terms of local context of every possible correspondence so better express the existing local feature detector, *e.g.*, shape context [2, 11], sift [13] and C-NN based detectors [31], *etc*. Furthermore, to achieve the switch of the two-phases transformation, the rigid transformation is used to capture the global tendency in the first 200 iterations, and the feasible non-rigid transformation is subsequently used to speed up convergence and then obtains a more accurate result. We can summarize our algorithm using a pseudocode, as shown in 11.

Algorithm 1: Robust Variational Bayesian point set registration.

input : $\mathcal{M} = \{m_j\}_{j=1}^{J}, \mathcal{S} = \{s_n\}_{n=1}^{N}, \Delta$
initialise: $\rho_0, \phi_0, \varepsilon_0, \beta_0, \eta, \zeta_0, a_0, b_0$
repeat
  anneal ϕ_0, a_k and b_k;
  VB E-step:
  update $q(\mathcal{U}|\mathcal{Z})$ by Eq.12;
  update $q(\mathcal{V}|\mathcal{Z})$ by Eq.13;
  infer the assignment variable $\mathcal{Z}$ by Eq.14;
  VB M-step:
  update $q(\pi)$ by Eq.15;
  update $q(\mathcal{X})$ and $q(\Gamma)$ by Eq.16;
  calculate $\mathcal{T}$ by Eq.19;
until *the* ELBO *in* Eq.*8 increases less than* Δ;

3. Experimental Results

Both the point set and image registration examples are carried out to compare our method with five state-of-the-art methods [8, 14, 20, 27, 35]. Besides, the image stitching examples are provided to verify the validity of our method. The experimental datasets consist of five types: (a) the synthetic point sets are provided from the TPS-RPM [4] and CPD [20] for point set registration, (b) the 3D motion human tracking is consistent with [27], (c) the Graffiti set is provide by the Oxford data set, the remote sensing image and hyper-spectral and visible image are provided from RS dataset, (d) the transerse plain brain MRI and the retinal and the fingerprint images are provide by Zhang et al. [36], and (e) the image stitching data are downloaded from 720 cloud[1]. All experiments are implemented in MATLAB on a laptop with a 2.90GHz Intel Core CPU and a 16GB RAM.

3.1. Evaluation setting

Three general experimental evaluation methods demonstrate the performance of our method: (I) In the point set registration, we follow the same assessment criteria in [34]; (II) In the 3D motion tracking point set registration, we follow the same assessment criteria in [26]; (III) In the image registration, we follow the same assessment criteria in [14].

3.2. Point registration experiments

We firstly demonstrate the performance of our algorithm under six kinds of cases. As shown in Fig. 2, the first column is the registration example of non-rigid deformation. In the second column, both model and scene miss 21 points. The mixing proportion is re-weighted after learning, and then the missing components are correctly identified

[1]The 720 cloud platform is available at https://720yun.com/.

with small mixing proportion. Besides, anisotropic Gaussian components fit outliers distribution and thus weaken the interference of disordered outliers. In the third column, the registration example of noise shows that Gaussian components with identical covariances is not easily collapse for noised scene points. In the fourth column, the registration example of scaling and rotation with uniform outliers is tested. In the fifth column, the registration example of nonuniform outliers is demonstrated. In the last column, we test the performance of our method on the 3D rabbit which includes double nonuniform outliers.

The nine synthesized contour data sets, Chinese character, hand, fish1, line, horse, face, heart, fish2 and butterfly, each of which respectively contains 105, 302, 98, 60, 198, 317,96, 91 and 172 points are used. The average performances on deformation, missing, outlier, rotation, outlier+deformation 8, outlier+rotation $\pm 70°$, outlier+missing 0.5, missing+deformation 8, missing+rotation $\pm 70°$ and missing+outlier 1 are shown in Fig. 3. We follow

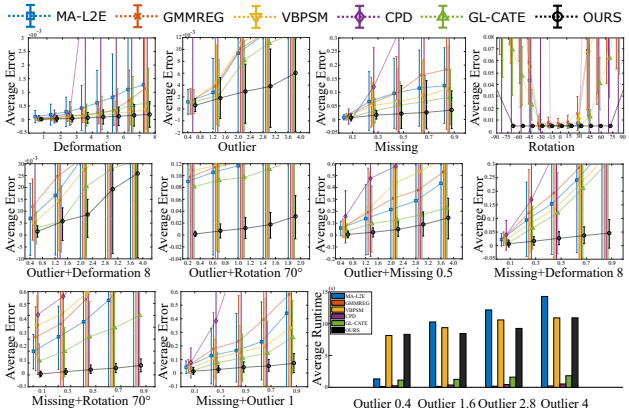


Figure 3. Comparison of our method against five state-of-the-art methods including MA-L2E, CPD, VBPSM,GMMREG and GL-CATE on nine point sets and showcase of the average runtime of our method for different outlier levels.

the experimental settings as in [34]. Meanwhile, We also test the average runtime of MA-L2E, GMMREG, VBPSM, CPD, GL-CATE and our method on Outlier 0.4, Outlier 1.6, Outlier 2.8 and Outlier 4 using fish 1 (98 points). In addition, the point set registration examples of deformation(D), rotation(R), missing(M), outliers(O), deformation and outliers(D,O), deformation and rotation(D,R), deformation and missing(D,M), outliers and rotation(O,R) and rotation and missing(R,M) are demonstrated in Fig. 4.

3.3. 3D human motion capture registration

We test the performance of our method on 3D human motion tracking compared with others. There are 450 frames and 42 markers on the body of experimenter, and their error analysis are provided in Fig. 5. Experimental

Table 1. The comparison of RMSE, MAE and SD for five methods in GID dataset. The percentage of each unit is averaged. The best results are identified in bold.

Method	LLT	VFC	CPD	Yang et al.	Ours
RMSE	43.27	54.51	87.33	11.85	**3.11**
MAE	49.36	61.63	99.34	13.92	**4.41**
SD	35.79	39.20	71.45	9.76	**1.39**

results demonstrate that our method achieves better performances than state-of-the-art methods.

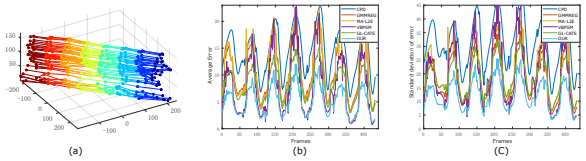


Figure 5. Registration performance of the 3D human motion capture. (a): original data, (b): comparison of mean absolute error and (c): comparison of the standard deviation.

3.4. Image registration experiments

Image registration examples including the graf image registration, the remote sensing image registration, the transerse plain brain MRI image registration, the hyperspectral and visible image registration, retinal image registration and fingerprint image registration of our method are provided in Fig. 6. Meanwhile, the root mean square error (RMSE), the max error (MAE) and the standard deviation (SD) are provided in Table 1. Our method achieves prominent registration results on images with small overlapping regions and large viewpoint changes. Image stitching examples including the mongolian yurt image stitching and the Leshan Buddha image stitching are provided in Fig. 7.

4. Conclusion

In this work, we employ TLMM to address the issues of mixing correspondences and outliers under the Variational Bayesian framework, and improve the optimization of a constructed Bayesian network by a tree-structured Variational inference which favors a tighter lower bound that the obtained variational distributions are more expressive. Experimental results demonstrate that our approach achieves favorable performances.

5. Acknowledgement

This work was supported by (i) National Nature Science Foundation of China [41661080, 41971392]; (ii) Yunnan Ten-thousand Talents Program.

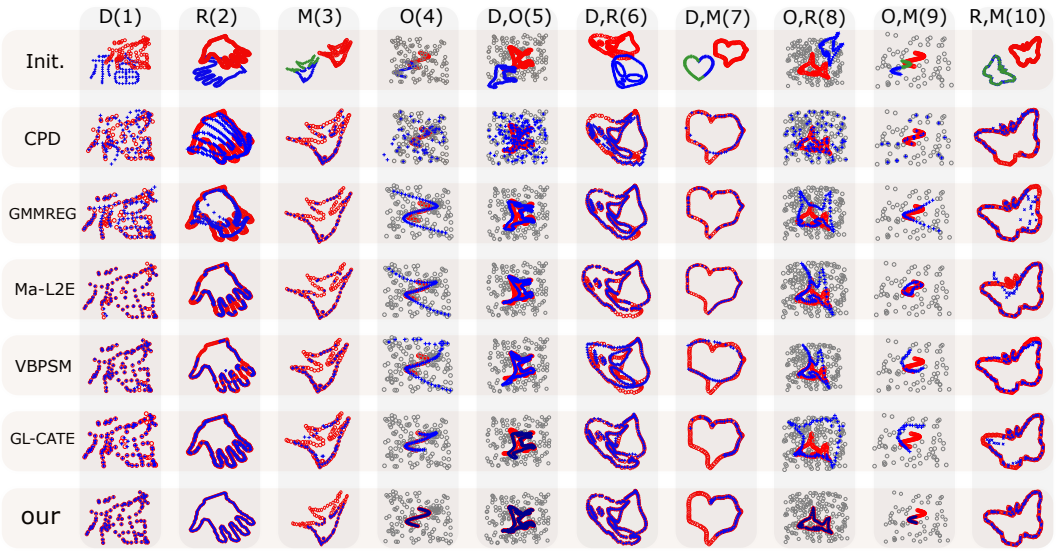
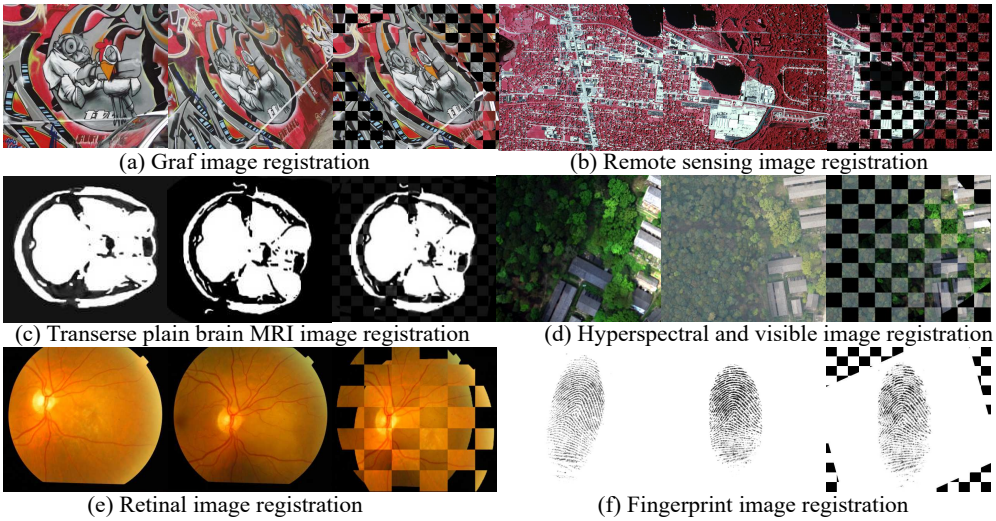


Figure 4. Registration examples of the ten scenarios are Chinese character, hand, fish 1, line, horse, face, heart, fish 2 and butterfly, respectively. From the first column to the fourth column: deformation(8), rotation(85°), missing(0.5) and outlier(3). From the fifth column to the last column: deformation and outlier(8+2), deformation and rotation(8+85°), deformation and missing(8+0.6), missing and rotation(0.8+85°) and outlier and missing(2+0.5).



(a) Graf image registration

(b) Remote sensing image registration

(c) Transerse plain brain MRI image registration

(d) Hyperspectral and visible image registration

(e) Retinal image registration

(f) Fingerprint image registration

Figure 6. Image registration examples. (a)-(f): From the first column to the third column, the sensed image, the reference image and the checkboard image.

Figure 7. Image stitching examples. **Top Row**: Mongolian yurt image stitching. **Bottom Row**: Leshan Buddha image stitching.

References

[1] C. Archambeau and M. Verleysen. Robust bayesian clustering. *Neural Networks the Official Journal of the International Neural Network Society*, 20(1):129–138, 2007. 3

[2] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE transactions on pattern analysis and machine intelligence*, 24(4):509–522, 2002. 6

[3] H. Chui and A. Rangarajan. A feature registration framework using mixture models. In *Proceedings IEEE Workshop on Mathematical Methods in Biomedical Image Analysis. MMBIA-2000 (Cat. No. PR00737)*, pages 190–197. IEEE, 2000. 1, 2

[4] H. Chui and A. Rangarajan. A new algorithm for non-rigid point matching. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, volume 2, pages 44–51. IEEE, 2000. 1, 2, 3, 6

[5] D. Gerogiannis, C. Nikou, and A. Likas. Robust image registration using mixtures of t-distributions. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007. 2

[6] D. Gerogiannis, C. Nikou, and A. Likas. The mixtures of student's t-distributions as a robust framework for rigid registration. *Image and Vision Computing*, 27(9):1285–1294, 2009. 2

[7] V. Golyanik, B. Taetz, G. Reis, and D. Stricker. Extended coherent point drift algorithm with correspondence priors and optimal subsampling. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016. 2

[8] B. Jian and B. Vemuri. Robust point set registration using gaussian mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:1633, 2011. 1, 2, 3, 6

[9] B. Jian and B. C. Vemuri. Robust point set registration using gaussian mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(8):1633–1645, 2011. 1

[10] S. Jianyong and K. Ata. A fast algorithm for robust mixtures in the presence of measurement errors. *IEEE Transactions on Neural Networks*, 21(8):1206–1220, 2010. 5

[11] M. Körtgen, G.-J. Park, M. Novotni, and R. Klein. 3d shape matching with 3d shape contexts. In *The 7th central European seminar on computer graphics*, volume 3, pages 5–17. Budmerice, 2003. 6

[12] H. Li, W. Ding, X. Cao, and C. Liu. Image registration and fusion of visible and infrared integrated camera for medium-altitude unmanned aerial vehicle remote sensing. *Remote Sensing*, 9(5):441, 2017. 1

[13] D. G. Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999. 6

[14] J. Ma, W. Qiu, Z. Ji, M. Yong, and Z. Tu. Robust l2e estimation of transformation for non-rigid registration. *IEEE Transactions on Signal Processing*, 63(5):1115–1129, 2015. 1, 2, 3, 6

[15] J. Ma, J. Wu, J. Zhao, J. Jiang, H. Zhou, and Q. Z. Sheng. Nonrigid point set registration with robust transformation learning under manifold regularization. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14, 2018. 2

[16] J. Ma, J. Zhao, J. Jiang, H. Zhou, and X. Guo. Locality preserving matching. *International Journal of Computer Vision*, 127(5):512–531, May 2019. 2

[17] J. Ma, J. Zhao, and A. L. Yuille. Non-rigid point set registration by preserving global and local structures. *IEEE Transactions on image Processing*, 25(1):53–64, 2016. 2

[18] K. P. Murphy. Machine learning: A probabilistic perspective (adaptive computation and machine learning series), 2018. 2

[19] A. Myronenko and X. Song. Point set registration: Coherent point drift. *IEEE transactions on pattern analysis and machine intelligence*, 32(12):2262–2275, 2010. 6

[20] A. Myronenko, X. Song, and M. A. Carreira-Perpinán. Non-rigid point set registration: Coherent point drift. In *Advances in neural information processing systems*, pages 1009–1016, 2007. 1, 2, 3, 6

[21] R. M. Neal and G. E. Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer, 1998. 2

[22] T. M. Nguyen and Q. J. Wu. Multiple kernel point set registration. *IEEE transactions on medical imaging*, 35(6):1381–1394, 2016. 1

[23] T. M. Nguyen and Q. M. Wu. Bounded asymmetrical student's-t mixture model. *IEEE Transactions on Cybernetics*, 44(6):857–869, 2017. 2

[24] T. M. Nguyen and Q. M. J. Wu. Robust student's-t mixture model with spatial constraints and its application in medical image segmentation. *IEEE Transactions on Medical Imaging*, 31(1):103–116, 2011. 2

[25] D. Peel and G. J. Mclachlan. Robust mixture modelling using the t distribution. *Statistics and Computing*, 10(4):339–348, 2000. 2

[26] H. B. Qu, J. Q. Wang, B. Li, and M. Yu. Probabilistic model for robust affine and non-rigid point set matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):371–384, 2016. 1, 6

[27] H.-B. Qu, J.-Q. Wang, B. Li, and M. Yu. Probabilistic model for robust affine and non-rigid point set matching. *IEEE transactions on pattern analysis and machine intelligence*, 39(2):371–384, 2017. 1, 2, 6

[28] B. Schölkopf, R. Herbrich, and A. J. Smola. A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer, 2001. 5

[29] G. Sfikas, C. Nikou, N. Galatsanos, and C. Heinrich. Mr brain tissue classification using an edge-preserving spatially variant bayesian mixture model. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 43–50. Springer, 2008. 2

[30] J. Sun, A. Zhou, S. Keates, and S. Liao. Simultaneous bayesian clustering and feature selection through student's t mixtures model. *IEEE Transactions on Neural Networks and Learning Systems*, 29(4):1187–1199, 2017. 2, 5

[31] Y. Tian, B. Fan, and F. Wu. L2-net: Deep learning of discriminative patch descriptor in euclidean space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 661–669, 2017. 6

[32] Y. Tsin and T. Kanade. A correlation-based approach to robust point set registration. In *European conference on computer vision*, pages 558–569. Springer, 2004. 1, 2, 3

[33] K. Yang, A. Pan, Y. Yang, S. Zhang, S. Ong, and H. Tang. Remote sensing image registration using multiple image features. *Remote Sensing*, 9(6):581, 2017. 1

[34] Y. Yang, S. Ong, and K. Foong. A robust global and local mixture distance based non-rigid point set registration. *Pattern Recognition*, 48:156–173, 2015. 6, 7

[35] S. Zhang, K. Yang, Y. Yang, Y. Luo, and Z. Wei. Non-rigid point set registration using dual-feature finite mixture model and global-local structural preservation. *Pattern Recognition*, 80:183–195, 2018. 2, 6

[36] S. Zhang, Y. Yang, K. Yang, Y. Luo, and S.-H. Ong. Point set registration with global-local correspondence and transformation estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2669–2677, 2017. 2, 6

[37] M. Zhao, B. An, Y. Wu, H. Van Luong, and A. Kaup. Rfvtm: a recovery and filtering vertex trichotomy matching for remote sensing image registration. *IEEE Transactions on Geoscience and Remote Sensing*, 55(1):375–391, 2017. 1

[38] Z. Zhou, B. Tong, G. Chen, J. Hu, J. Zheng, and Y. Dai. Direct point-based registration for precise non-rigid surface matching using students-t mixture model. *Biomedical Signal Processing and Control*, 33:10–18, 2017. 2

[39] Z. Zhou, J. Tu, C. Geng, J. Hu, B. Tong, J. Ji, and Y. Dai. Accurate and robust non-rigid point set registration using studentst mixture model with prior probability modeling. *Scientific reports*, 8(1):8742, 2018. 2