Group-Wise Point-Set Registration based on Rényi's Second Order Entropy

Luis G. Sanchez Giraldo
University of Miami, FL, USA
lgsanchez@cs.miami.edu

Erion Hasanbelliu, Murali Rao, Jose C. Principe
University of Florida, FL, USA
ehasanbelliu@gmail.com, mrao@ufl.edu, principe@cnel.ufl.edu

Abstract

In this paper, we describe a set of robust algorithms for group-wise registration using both rigid and non-rigid transformations of multiple unlabelled point-sets with no bias toward a given set. These methods mitigate the need to establish a correspondence among the point-sets by representing them as probability density functions where the registration is treated as a multiple distribution alignment. Holder's and Jensen's inequalities provide a notion of similarity/distance among point-sets and Rényi's second order entropy yields a closed-form solution to the cost function and update equations. We also show that the methods can be improved by normalizing the entropy with a scale factor. These provide simple, fast and accurate algorithms to compute the spatial transformation function needed to register multiple point-sets. The algorithms are compared against two well-known methods for group-wise point-set registration. The results show an improvement in both accuracy and computational complexity.

1. Introduction

Point-set registration is a common problem in computer vision, pattern recognition, medical imaging, robotics, and many other fields. Tasks such as image registration, face recognition, object tracking, image stitching or 3D object fusion all require registration of features/point-sets. Features representing an object's contour or other distinguishing characteristics are extracted from an image and compared against those of other objects or templates, where first, a correspondence between the point-sets is established, and then, the spatial transformation that aligns them is retrieved. A similarity measure is used to compare the point-sets, where the choice depends on the object features and the problem.

The research community has developed many techniques to solve point-set registration. The iterative closest point (ICP) algorithm [1] is the most popular method, which uses the nearest neighbour relationship to assign binary correspondence and then determines the least squares transformation relating the point-sets. The method is very simple, but it exhibits local convergence due to a non-differentiable cost function. In addition, it suffers in the presence of outliers and is not suitable for non-rigid transformations. Robust point matching (RPM) [4] improves upon ICP by employing a global to local search and soft assignment for the correspondence. The method jointly determines the correspondence and the transformation between the two point-sets via deterministic annealing and soft-assignment. However, the method is not stable in the presence of outliers, and due to the soft-assignment approach, the complexity of the method is high.

To eliminate the point correspondence requirement, a global approach is taken where the point-sets are represented as probability density functions (PDF). Tsin and Kanade [21] were the first to propose such a method where they modelled the two point-sets as kernel density functions and evaluated their similarity and transformation updates using the kernel correlation between the two density estimates. Glaunes et al. [8] matched the two point-sets by representing them as weighted sums of Dirac delta functions, where Gaussian functions were used to 'soften' the Dirac delta functions and diffeomorphic transformations were used to minimize the distance between the two distributions. Jian and Vemuri [12], [13] extended this approach by representing the densities as Gaussian mixture models (GMM). They derive a closed-form expression to compute the L_2 distance between the two Gaussian mixtures and the update method to align the two point-sets.

All these methods are limited to registering only a pair of point-sets. In certain applications, such as medical image registration, there is a need to simultaneously register a group of point-sets. None of these methods is directly extendible to group-wise alignment of multiple point-sets. In addition, these methods are all biased (unidirectional update) where one point-set acts as the target and the other is transformed to align with it. In applications where group-wise registration is required, there may not be an actual template to match against (e.g. medical images). Therefore, there is a need to determine a middle, common orientation/position to align the point-sets. Estimating a meaning-

1063-6919/17 $31.00 © 2017 IEEE
DOI 10.1109/CVPR.2017.263

ful average shape from a set of unlabelled shapes is a key challenge in deformable shape modelling.

Chui et al. [6] presented a joint clustering and matching algorithm that finds a mean shape from multiple shapes represented by unlabelled point-sets. Their process follows a similar approach to their previous work in [5] where explicit correspondence needs to be determined first. In addition, the method is not robust to outliers, so stability is not always guaranteed.

Wang et al. [22] proposed a method for group-wise registration where the point-sets are represented as density functions. Based on the same principles as the PDF registration methods above, their algorithm simultaneously registers the point-sets and determines the mean set without solving for correspondences or selecting any specific point-set as a reference. Their approach minimizes the Jensen-Shannon divergence among cumulative distribution functions (CDFs). They shifted from PDFs because CDFs are more immune to noise and are also well defined since CDF is an integral measure. However, the CDF estimation is computationally very expensive and there are no closed-form solutions to their updates.

Chen et al. [3] developed another group-wise registration method based on the Havrda-Charvát divergence for CDFs. Similar to [22], they use the cumulative residual entropy to represent the CDFs. Their method, CDF-HC, generalizes the CDF-JS but it is much simpler to implement and computationally more efficient.

In this paper, we introduce a set of methods for group-wise registration based on Rényi's quadratic entropy. The major improvement of our methods is that they provide a closed-form solution for the updates, which makes them much simpler and faster to compute than both CDF-based methods above, with no loss in accuracy. We compare our methods against CDF-JS and CDF-HC on various data sets.

2. Background

In [9] and [10], a density-based point-set registration method is introduced. The similarity between the two PDFs is measured using an information theoretic measure, the Cauchy-Schwarz (CS) divergence [15], derived from the Cauchy-Schwarz inequality [18]:

$$\int f(x)g(x)dx \leq \sqrt{\int f^2(x)dx \int g^2(x)dx}, \qquad (1)$$

where in the case of PDFs, the equality holds if and only if $f(x) = Cg(x)$ with $C = 1$. The Cauchy-Schwarz divergence is then defined as:

$$\mathcal{D}_{\mathcal{CS}}(f\|g) = -\log \frac{\left(\int f(x)g(x)dx\right)^2}{\int f^2(x)dx \int g^2(x)dx}. \qquad (2)$$

The divergence directly compares PDF similarity and is expressed in terms of inner products of the two PDFs which essentially estimate a normalized Rényi's quadratic cross-entropy. Suppose we have two d-dimensional point-sets $\mathbf{X}_f = \{\mathbf{x}_1^{(f)}, \ldots, \mathbf{x}_{M_f}^{(f)}\}$ and $\mathbf{X}_g = \{\mathbf{x}_1^{(g)}, \ldots, \mathbf{x}_{M_g}^{(g)}\}$. It is assumed that $\mathbf{x}_i^{(f)} \in \mathbf{X_f}$ are drawn from f and $\mathbf{x}_j^{(g)} \in \mathbf{X}_g$ drawn from g. The CS divergence between point-sets $\mathbf{X_f}$ and $\mathbf{X_g}$ is computed by plugging in the Parzen density estimates $\hat{f}$ and $\hat{g}$ of the PDFs f and g into (2). Namely, the Parzen density estimate for f is given by [14]:

$$\hat{f}(\mathbf{x}; \mathbf{X}_f) = \frac{1}{M_f} \sum_{i=1}^{M_f} \kappa\left(\frac{\mathbf{x} - \mathbf{x}_i^{(f)}}{\sigma}\right), \qquad (3)$$

where $\kappa(\cdot)$ is a valid kernel (window) function and σ its bandwidth parameter. The estimate $\hat{g}$ of g is obtained in a similar way. The Gaussian function

$$G_\sigma(x, x_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\|x - x_i\|^2}{2\sigma^2}\right), \qquad (4)$$

is considered as the kernel of choice for its properties: symmetry, positive definitiness, and exponential decay controlled via the kernel bandwidth. By using the Gaussian kernel, the plug in estimator of Cauchy-Schwarz divergence has a closed-form expression in terms of convolutions operators making it simple to estimate. An important term in (2) is called the *cross-information potential* [15, 11], which corresponds to the inner product between f and g. The empirical estimator of the cross-information potential has the following form:

$$\begin{aligned} CIP(\mathbf{X}_f, \mathbf{X}_g) &= \int \hat{f}(x)\hat{g}(x)dx \qquad (5) \\ &= \frac{1}{M_f M_g} \sum_{i=1}^{M_f} \sum_{j=1}^{M_g} G_{\sqrt{2}\sigma}\left(\mathbf{x}_i^{(f)}, \mathbf{x}_j^{(g)}\right). \end{aligned}$$

This term is crucial in determining the similarity between the two point-sets. When data points are interpreted as particles, the information potential measures the interaction of the field created by one set of particles on the locations specified by the other set. The Cauchy-Schwarz information potential field exerts information forces on samples of the second point-set forcing them to move toward a path that will provide the most similarity between the two PDFs. The CS-divergence is restricted to pairwise comparisons. To account for group-wise comparisons, we propose two extensions of the above idea that allow the comparison of more than two density functions simultaneously. The first proposed method uses Hölder's divergence as a direct extension of the Cauchy-Schwarz divergence. The second extension, which is the main focus of this work, is obtained based on Jensen's inequality on the information potential.

3. Group-wise registration based on Hölder's divergence

A generalized version of Hölder's inequality to N functions $\{f_k\}$ is given by:

$$\left(\int \prod_{k=1}^{N} f_k(x)dx\right)^N \leq \prod_{k=1}^{N} \int {f_k}^N(x)dx. \quad (6)$$

Based on (6), Hölder's divergence is thus defined as:

$$\mathcal{D}_{\mathcal{H}}(\{f_k\}) = -\log \frac{\left(\int \prod_{k=1}^{N} f_k(x)dx\right)^N}{\prod_{k=1}^{N} \int {f_k}^N(x)dx}, \quad (7)$$

which can be written as a difference of logarithms,

$$\begin{aligned}\mathcal{D}_{\mathcal{H}}(\{f_k\}) = &-N\log\int \prod_{k=1}^{N} f_k(x)dx \\ &+\sum_{k=1}^{N}\log\int {f_k}^N(x)dx.\end{aligned} \quad (8)$$

Suppose we have N point-sets $\mathbf{X}_k$, $k \in 1,\ldots,N$. Each $\mathbf{X}_k = \{\mathbf{x}_1^{(k)},\ldots,\mathbf{x}_{M_k}^{(k)}\}$, where $\mathbf{x}_i^{(k)} \in \mathbb{R}^d$. Similar to (5), an empirical estimator of Hölder's divergence, with closed-form solution, can be obtained using Parzen density approximation with Gaussian kernel,

$$\begin{aligned}\mathcal{D}_{\mathcal{H}}(\{\hat{f}_k\}) = &-N\log\sum_{i_1=1}^{M_1}\cdots\sum_{i_N=1}^{M_N}\prod_{p=1}^{N-1}\prod_{q=p+1}^{N} G_{p,q} \\ &+\sum_{k=1}^{N}\log\sum_{i_{1,k}=1}^{M_k}\cdots\sum_{i_{N,k}=1}^{M_k}\prod_{p=1}^{N-1}\prod_{q=p+1}^{N} G_{k,k},\end{aligned} \quad (9)$$

where $\gamma = \sqrt{N}\sigma$, $G_{p,q} = G_\gamma(\mathbf{x}_{i_p}^{(p)} - \mathbf{x}_{i_q}^{(q)})$, and $G_{k,k} = G_\gamma(\mathbf{x}_{i_{p,k}}^{(k)} - \mathbf{x}_{i_{q,k}}^{(k)})$. Although having a closed-form solution for Hölder's divergence estimator is appealing at first glance, it also reveals its complexity. A direct implementation of Hölder's divergence (9) would require $\mathcal{O}(M_1\cdots M_N N^2 d)$ operations which become prohibitively large as the number of shapes and data points increase. Clever manipulations of the terms in (9) can reduce the number of operations to $\mathcal{O}\left(\left(\max_i\{M_i\}\right)^2 N^2 d\right)$. In the following, we propose an alternative objective function with $\mathcal{O}\left(\left(\max_i\{M_i\}\right)^2 N^2 d\right)$ complexity.

4. Group-wise registration based on Rényi's second order entropy

To reach a closed-form but also computationally efficient solution, we propose an objective function based on Jensen's inequality. However, unlike [22] where a group-wise registration method was proposed using the Jensen-Shannon (JS) divergence, our method is based on the direct comparison of information potentials derived from Rényi's second order entropy.

4.1. Rényi's second order entropy and the information potential

Rényi's entropy is a generalization of Shannon's entropy [19] for which the logarithm operation lies outside the expectation operator. Rényi's entropy of a continuous random variable X taking values in $\mathcal{X}$, and with probability density function f, is defined as:

$$H_\alpha(X) = -\frac{1}{1-\alpha}\log\int_{\mathcal{X}} f^\alpha(x)\mathrm{d}x. \quad (10)$$

In the limit $\alpha \to 1$, (10) approximates Shannon's differential entropy [17]. For $\alpha = 2$, an empirical estimator $\hat{H}_2$ of H_2, which has closed form solution and it is also differentiable, can be obtained using Parzen windows. Let $\mathbf{X}_p$ be an *i.i.d* sample of size M drawn from f. The expression,

$$\hat{H}_2(\mathbf{X}_p) = -\log\frac{1}{M^2}\sum_{i=1}^{M}\sum_{j=1}^{M} G_{\sqrt{2}\sigma}(\mathbf{x}_j - \mathbf{x}_i), \quad (11)$$

is an estimator of Rényi's second order entropy. The quantity inside of the logarithm in (11) is known as the information potential, which as we will see below, allows the formulation of a well behaved cost function for group-wise registration.

4.2. Group-wise registration based on information potentials

Let $\mathbf{X} = \{\mathbf{X}_1,\ldots,\mathbf{X}_N\}$ be a collection of N shapes to be aligned. Each $\mathbf{X}_k$ is an array of M_k points representing a shape. Here, a shape is considered to be a set of *i.i.d* samples drawn from a common random variable that have undergone an unknown transformation T_k. Our goal is to find a set of transformations $S_k = T_k^{-1}$ that map the collection of shapes $\mathbf{X}$ to a collection $\tilde{\mathbf{X}} = \{\tilde{\mathbf{X}}_k = T_k^{-1}(\mathbf{X}_k)\}$ regarded as *i.i.d* samples drawn from the same underlying random variable $\tilde{X}$. Since the underlying random variable that the shapes are assumed to be sampled from is unknown, it is necessary to impose a set of regularity conditions on the transformation mappings S_k and the estimated matching distributions from each $\tilde{\mathbf{X}}_i$.

The information potential $IP(f) = \int f^2(x)\mathrm{d}x$ is a convex function since it corresponds to the squared L^2 norm of a density function[1]. From the convexity of the information

[1]Here, we restrict to the space of probability density functions with bounded L^2 norm.

potential, the following inequality holds:

$$IP\left(\sum_{k=1}^{N}\Pi_k\hat{f}_k\right) \leq \sum_{k=1}^{N}\Pi_k IP(\hat{f}_k), \tag{12}$$

where $\hat{f}_k$ is the Parzen density estimate for the kth point-set, and

$$\Pi_k = \frac{M_k}{\sum_{k=1}^{N} M_k}. \tag{13}$$

This indicates that the weighted sum of the information potentials of each of the point-sets $\tilde{\mathbf{X}}_k$ is greater than the information potential of their union $\tilde{\mathbf{X}}$, and equality is valid if and only if all the point-sets are the same. From this observation, we propose as a cost function, the difference between the two terms in the inequality (12). The proposed cost function, expressed in terms of the point-sets $\tilde{\mathbf{X}}_k$, can be written as:

$$\mathcal{J} = \sum_{k=1}^{N}\Pi_k IP(\tilde{\mathbf{X}}_k) - IP\left(\bigcup_{k=1}^{N}\tilde{\mathbf{X}}_k\right). \tag{14}$$

Multi-shape alignment can be achieved by minimizing the above difference. As will will show below, transforming the samples $\mathbf{X}_k$ to $\tilde{\mathbf{X}}_k$, so that their weighted average information potential is equivalent to the information potential of their union $\tilde{\mathbf{X}}$, yields the desired alignment.

5. Point-set registration

As we mentioned above, a collection of N point-sets $\mathbf{X}_k$, assumed to be transformed versions of samples drawn from a common random variable $\tilde{X}$, can be aligned to a common shape by finding a set of transformations S_k such that the estimated distributions of $S_k(\mathbf{X}_k)$ minimize the objective function (14). Nevertheless, this is an ill-posed problem because the distribution of $\tilde{X}$ is unknown. To select a particular solution, we impose constraints on the set of estimated transformations as well as the target distribution. For each $\mathbf{X}_k$, the desired transformation can be broken into two parts: 1) *affine* transformation, and 2) *non-rigid* transformation. The affine transformation has a global effect on the shape and is mainly composed of linear transformations such as rotation, scaling, shear, and a translation. The non-rigid transformation accounts for local deformations that cannot be expressed by the affine transformation. In addition, we assume that the function is *smooth* in the sense that two similar inputs correspond to two similar outputs. This assumption is crucial when dealing with the non-rigid part of the transformation. Smoothness of the solution is enforced by introducing a regularization term into our problem.

5.1. Affine transformation

Let $\mathbf{X}_k$ denote a matrix of size $M_k \times d$, where each row vector is a point in $\mathbb{R}^d$. An affine transformation is obtained by a linear transformation $\mathbf{A}^{\mathrm{T}}$ followed by the addition of the vector $\mathbf{t} \in \mathbb{R}^d$. For a point $\mathbf{x} \in \mathbf{R}^d$, the affine transformation S_{aff} is expressed in terms of $\mathbf{A}$ and $\mathbf{t}$, as follows:

$$S_{\text{aff}}(\mathbf{x}|\mathbf{A},\mathbf{t}) = \mathbf{A}^{\mathrm{T}}\mathbf{x} + \mathbf{t}. \tag{15}$$

A more compact form of (15) can be obtained by considering homogeneous coordinates, where the original vector points are extended to $(d+1)$ dimensions with the last dimension as 1,

$$S_{\text{aff}}(\mathbf{x}|\mathbf{A},\mathbf{t}) = \left[\mathbf{A}^{\mathrm{T}}|\mathbf{t}\right]\begin{bmatrix}\mathbf{x}\\1\end{bmatrix}. \tag{16}$$

From here on, it is assumed that the points $\mathbf{x}_i^{(k)}$ have been extended to deal with translations implicitly as shown in (16).

5.2. Non-rigid transformation

Non-rigid transformations model local deformations, which is a non-linear operation on the coordinate points of the shape. In this work, we employ radial basis function (RBF) expansions to compute the non-rigid transformation. In the RBF expansion, the transformation is defined as a linear combination of M basis functions $\phi_i(\mathbf{x})$,

$$S_{\text{nr}}(\mathbf{x}|\mathbf{W},\boldsymbol{\phi}) = \sum_{i=1}^{M}\mathbf{w}_i\phi_i(\mathbf{x}) = \sum_{i=1}^{M}\mathbf{w}_i\phi\left(\|\mathbf{x}-\mathbf{x}_i\|\right). \tag{17}$$

Each of the basis elements ϕ_i is a nonlinear function of the Euclidean distance between the point $\mathbf{x}$, at which the function is evaluated, and the corresponding center point $\mathbf{x}_i$. The most common RBFs used for non-rigid transformations are thin plate splines (TPS) [2] and Gaussian RBFs. The key advantage of both of these functions is that their approximation errors approach to zero asymptotically. However, the non parametric nature of the RBF expansion where all data points are also control points, if not handled properly, can have a significant impact on the computational complexity. Although this is an important problem on its own right, it is out of the scope of this paper, and will not be addressed any further.

The transformation S_k is the linear combination of the affine and non-rigid components as follows:

$$\begin{aligned} S_k(\mathbf{x}|\mathbf{A}_k,\mathbf{t}_k,\mathbf{W}_k,\boldsymbol{\phi}_k) &= S_{\text{aff}}(\mathbf{x}|\mathbf{A}_k,\mathbf{t}_k) + S_{\text{nr}}(\mathbf{x}|\mathbf{W}_k,\boldsymbol{\phi}_k) \\ &= \left[\mathbf{A}_k^{\mathrm{T}}|\mathbf{t}_k\right]\begin{bmatrix}\mathbf{x}\\1\end{bmatrix} + \mathbf{W}_k^{\mathrm{T}}\boldsymbol{\phi}_k(\mathbf{x}). \end{aligned} \tag{18}$$

The mapping $\boldsymbol{\phi}_k$ is defined by the set of basis functions centred at the points of $\mathbf{X}_k$, which yields $\mathbf{W}_k$ of size $M_k \times d$. To simplify notation, we will denote the affine mapping by

$\mathbf{A}$. It should be understood that $\mathbf{x}$ is the augmented vector and $\mathbf{t}$ is contained in $\mathbf{A}$. Equation (18) is then written as:

$$S_k(\mathbf{x}|\mathbf{A}_k, \mathbf{W}_k, \boldsymbol{\phi}_k) = \mathbf{A}_k^{\mathrm{T}}\mathbf{x} + \mathbf{W}_k^{\mathrm{T}}\boldsymbol{\phi}_k(\mathbf{x}). \tag{19}$$

For each shape $\mathbf{X}_k$, the transformed shape $\tilde{\mathbf{X}}_k$ can be compactly written in matrix form,

$$\tilde{\mathbf{X}}_k = [\mathbf{X}_k|\mathbf{1}]\,\mathbf{A}_k + \boldsymbol{\Phi}_k\mathbf{W}, \tag{20}$$

where $\boldsymbol{\Phi}_k$ is the matrix of all pairwise evaluations of ϕ, that is, $(\boldsymbol{\Phi}_k)_{ij} = \phi\left(\|\mathbf{x}_i^{(k)} - \mathbf{x}_i^{(k)}\|\right)$. However, since the columns of the augmented matrix $[\mathbf{X}_k|\mathbf{1}]$ could also belong to the span of the columns of $\boldsymbol{\Phi}_k$, part of the affine transformation could be implicitly contained on the second term of the *r.h.s.* of (20). It is possible to decouple the affine and non-rigid components by projecting $\boldsymbol{\Phi}_k$ to the kernel of the span of the columns of $\mathbf{X}_k$. The projection can be obtained based on the QR factorization of $[\mathbf{X}_k|\mathbf{1}]$. Let $\mathbf{Q}_k\mathbf{R}_k = [\mathbf{X}_k|\mathbf{1}]$, where $\mathbf{Q}_k = \left[\mathbf{Q}_k^{\mathcal{Y}}|\mathbf{Q}_k^{\perp}\right]$. A decoupled version of (20) can be obtained by replacing $\boldsymbol{\Phi}_k$ with

$$\bar{\boldsymbol{\Phi}}_k = \left(\mathbf{Q}_k^{\perp}{\mathbf{Q}_k^{\perp}}^{\mathrm{T}}\right)\boldsymbol{\Phi}_k\left(\mathbf{Q}_k^{\perp}{\mathbf{Q}_k^{\perp}}^{\mathrm{T}}\right). \tag{21}$$

Out of sample extensions of the solution are given by:

$$S_k(\mathbf{x}) = \mathbf{A}_k^{\mathrm{T}}\mathbf{x} + \mathbf{W}_k^{\mathrm{T}}\bar{\boldsymbol{\Phi}}_k{\boldsymbol{\Phi}_k}^{\dagger}\boldsymbol{\phi}_k(\mathbf{x}), \tag{22}$$

where ${\boldsymbol{\Phi}_k}^{\dagger}$ denotes the Moore-Penrose pseudo-inverse of $\boldsymbol{\Phi}_k$. From this point on, to ease notation, $\boldsymbol{\Phi}_k$ will be understood as the decoupled version obtained in (21), and $\mathbf{X}_k$ as the augmented matrix that accounts for the shift parameter in the affine transformation.

5.3. Regularization

Incorporating a *stabilizer* into the solution is an important part of shape registration because it enforces *smoothness* on the transformation function. Since we are dealing with a finite set of points, there can be an infinite number of transformations that match the corresponding points, but have very different behavior for unseen portions of the shape. Choosing the smoothest solution through a regularization term provides a unique tractable solution.

The regularization theory originates from the work of Tikhonov [20] where the existing optimization problem is augmented with a regularization term. In our case, we have:

$$\underset{\{S_k\}}{\text{minimize}}\mathcal{J}(\{S_k\}) + \lambda\Omega(\{S_k\}), \tag{23}$$

where $\Omega(\{S_k\})$ is the regularization term that constraints the *smoothness* of set of functions $\{S_k\}$, and λ is the free parameter that trades-off between the alignment $\mathcal{J}(\{S_k\})$ objective and the smoothness of the solution. For the non-rigid transformation, the regularization term is given by the pseudo-norm,

$$\Omega(\{S_k\}) = \mathrm{tr}\left(\mathbf{W}_k^{\mathrm{T}}\boldsymbol{\Phi}_k\mathbf{W}_k\right). \tag{24}$$

5.4. Solving for transformation matrices A and W

The cost function $\mathcal{J}$ can be expressed in terms of the cross-information potential (5). First, we have that for each transformed shape $\tilde{\mathbf{X}}_k$ with M_k points,

$$IP(\tilde{\mathbf{X}}_k) = CIP(\tilde{\mathbf{X}}_k, \tilde{\mathbf{X}}_k). \tag{25}$$

The information potential of the collection of transformed shapes $\{\tilde{\mathbf{X}}_k\}$ is given by:

$$IP\left(\bigcup_{k=1}^{N}\tilde{\mathbf{X}}_k\right) = \frac{1}{\left(\sum_{k=1}^{N}M_k\right)^2}\sum_{k,\ell=1}^{N}M_kM_\ell CIP(\tilde{\mathbf{X}}_k, \tilde{\mathbf{X}}_\ell). \tag{26}$$

The cross-information potential can be expressed in compact form using Gram matrices. For a given kernel, in particular for the Gaussian kernel of fixed width, let the matrix $\mathbf{G}(\tilde{\mathbf{X}}_k, \tilde{\mathbf{X}}_\ell)$ contain all pairwise evaluations of points in $\tilde{\mathbf{X}}_k$ and $\tilde{\mathbf{X}}_\ell$. For short, we will denote $\mathbf{G}(\tilde{\mathbf{X}}_k, \tilde{\mathbf{X}}_\ell)$ as $\mathbf{G}_{k\ell}$ and its entries by

$$G_{ij}^{(k\ell)} = G_{\sqrt{2}\sigma}\left(\tilde{\mathbf{x}}_i^{(k)}, \tilde{\mathbf{x}}_j^{(\ell)}\right). \tag{27}$$

Then, the cross-information potential can be written as:

$$CIP(\tilde{\mathbf{X}}_k, \tilde{\mathbf{X}}_\ell) = \mathbf{1}_{M_k}^{\mathrm{T}}\mathbf{G}_{k\ell}\mathbf{1}_{M_\ell}, \tag{28}$$

where $\mathbf{1}_M$ denotes a vector of length M and equal entries $1/M$. The partial derivatives of the cost function can be easily computed based on the partial derivatives of the cross-information potential, which are shown below:

$$\begin{aligned}\frac{\partial CIP(\tilde{\mathbf{X}}_k, \tilde{\mathbf{X}}_\ell)}{\partial \mathbf{A}_k} &\propto {\mathbf{X}_k}^{\mathrm{T}}\mathbf{G}_{k\ell}\tilde{\mathbf{X}}_\ell - {\mathbf{X}_k}^{\mathrm{T}}\mathrm{dg}(M_\ell\mathbf{G}_{k\ell}\mathbf{1}_{M_\ell})\tilde{\mathbf{X}}_k,\\ \frac{\partial CIP(\tilde{\mathbf{X}}_k, \tilde{\mathbf{X}}_\ell)}{\partial \mathbf{W}_k} &\propto {\boldsymbol{\Phi}_k}^{\mathrm{T}}\mathbf{G}_{k\ell}\tilde{\mathbf{X}}_\ell - {\boldsymbol{\Phi}_k}^{\mathrm{T}}\mathrm{dg}(M_\ell\mathbf{G}_{k\ell}\mathbf{1}_{M_\ell})\tilde{\mathbf{X}}_k,\end{aligned} \tag{29}$$

where $\mathrm{dg}(\mathbf{v})$ denotes a diagonal matrix with the values of $\mathbf{v}$ on its main diagonal.

5.5. Normalized information potential

Likewise differential entropy, the information potential is not scale invariant, so using the cost function as presented above to align the point-sets can end up at a saddle point on the performance surface. A global minimum could be reached by simply collapsing all the point-sets into to a single point. To ensure that such a collapse does not occur, we introduce a normalized form of the cost function that confines the solution space. Here, we rely on the fact that the entropy of a random variable depends on its standard deviation. To prevent unwanted, scaled-down solutions, we modify the cost function in (14) dividing the information

potentials by the squared root of the total variance, which corresponds to the trace of the covariance matrix. The cost function corresponds to the following difference of normalized information potentials:

$$\mathcal{J} = \sum_{k=1}^{N} \Pi_k \frac{IP(\tilde{\mathbf{X}}_k)}{\sqrt{\mathrm{tr}(\tilde{\mathbf{C}}_k)}} - \frac{IP\left(\bigcup \tilde{\mathbf{X}}_k\right)}{\sqrt{\mathrm{tr}(\tilde{\mathbf{C}}_\cup)}}, \qquad (30)$$

where $\tilde{\mathbf{C}}_k$ are the covariance matrices of each transformed shape $\tilde{\mathbf{X}}_k$, and $\tilde{\mathbf{C}}_\cup$ is the covariance of the union of all transformed shapes.

6. Experimental results

To illustrate the group-wise registration capabilities of the proposed method, Fig. 1 shows an example of multiple shapes of the same object rotated, translated, and scaled to different sizes. Fig. 1(A) shows the initial position of the objects. The rest of the subfigures show steps through the alignment process. Note that there are no jumps occurring during the registration process. This is not the case for the two methods based on CDFs that we compare against. The resulting alignment is an average of rotation and scaling of all the individual shapes. It does not coincide with any of the original shapes, but it reflects a compromise in orientation and size of the different point-sets.

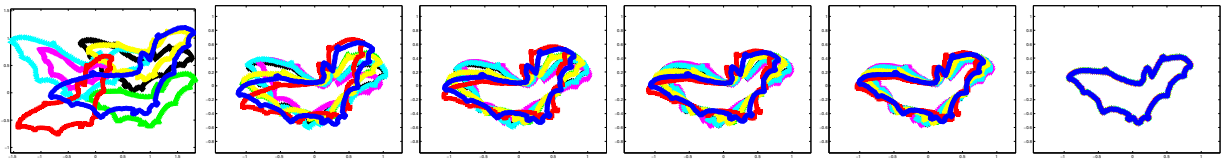
6.1. Groupwise registration for atlas construction

The following example shows group-wise registration using both affine and non-rigid transformations on a dataset borrowed from [3]. The dataset contains points extracted from the outer contours of the corpus callosum (CC) of seven subjects. In this experiment, we demonstrate the ability of our algorithm for *unbiased* 2D atlas construction. In addition to the information potential and normalized information potential differences, we also provide results based on Hölder's divergence. Fig. 2 shows the registration results for the two algorithms and also for CDF-HC [3]. In our experiments, we modified the optimization routine used for CDF-HC with the *minimize* routine that uses conjugate gradients and approximate line searches based on polynomial interpolation with Wolfe-Powel conditions from [16]. The first seven images 2(a)-2(g) show the deformation of each point-set to the atlas generated by one of the three methods. The color scheme used in the first seven images is as follows: the initial point-set is denoted with blue '+', the deformed points sets are denoted with circles which are in color green, black, red and magenta corresponding to CDF-HC, Holder's, IP, and normalized IP algorithms. Image 2(h) shows the superimposed point-sets before the registration. Images 2(i)-2(l) show the superimposed point-sets after registration for each method. Notice that CDF-HC follows more closely point-set 7, Figure 2(g), and its final registration is smaller than the rest of the point-sets. Holder's final registration has a close alignment with the initial positions of each individual point-set and is quite similar to the unconstrained IP solution. However, both get scaled down in the y-axis. This is due to the scaling problem previously discussed. Image 2(l) shows the results of the normalized IP with the constraint term in (30). This demonstrates the importance of the constraint when we minimize the cost. The results demonstrate that our methods not only provide a final registration that resembles the average shape more closely, but they also provide a better fit of the final point-sets.

To compare the registration capabilities of the methods against noise, we add a few outlier points in one of the point-sets, namely, point-set 7. Fig. 3(a) shows the point-sets before registration plus the outlier samples which are shown in dark stars. Sub-figures 3(b), 3(c), and 3(e) show the final registration for CDF-HC, Holder's, and normalized information potential, respectively. In addition to the seven point-sets after registration, each sub-figure also contains point-set 7 from the registration without the outlier samples to show the deviations that may have occurred in the final registration due to the outliers. The final registration shows that our method is more robust to outliers than CDF-HC. When compared against point-set 7 of the registration without outliers, our method has very little deviation, whereas CDF-HC shows a more pronounced convergence of the point-sets towards the outliers.

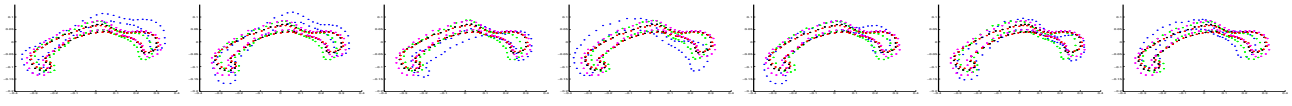
We need to point out that even though our methods perform very well, we have two free parameters in the cases when TPS are used; namely, the kernel size σ required to estimate the information potential, and the regularization parameter λ that controls the 'smoothness' of the function. In addition, when Gaussian RBFs are employed, a kernel bandwidth β that determines the locality influence of control points must be defined, as well. Determining these parameters and establishing an optimal annealing rate are problem dependent. In our examples above using the corpus callosum data, we use TPSs as basis functions, $\lambda = 0.1$ for the information potential and normalized information potential algorithms and 0.01 for Holder's algorithm. For all three proposed algorithms, we initialize $\sigma = 0.1$ and apply 98% annealing rate. To observe the effect of different regularization parameter values, Fig. 5 shows four different values for λ while initializing $\sigma = 0.2$ with 98% annealing rate.

Figure 5 depicts the behavior of the alignment when λ is changed by one order of magnitude. It can be seen how lower values of λ allow the points to freely move and provide an almost perfect overlap. Nevertheless, applying the learned transformation from the noisy subsampled shapes to the original shapes exposes the overfitting phenomena that

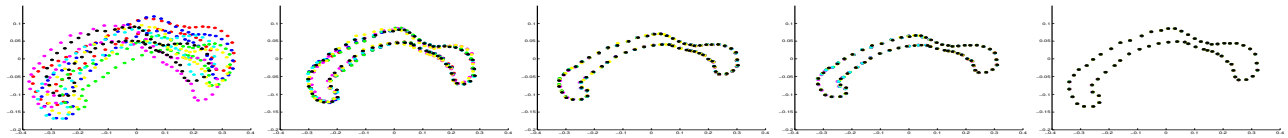


(a) initial position (b) step 5 (c) step 20 (d) step 50 (e) step 80 (f) final alignment

Figure 1. Example of multiple shapes aligned using the Information potential cost function.

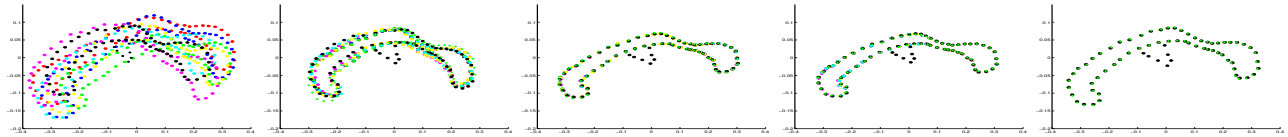


(a) point-set 1 (b) point-set 2 (c) point-set 3 (d) point-set 4 (e) point-set 5 (f) point-set 6 (g) point-set 7



(h) Before Registr. (i) CDF-HC (j) Holder's (k) IP difference (l) Normalized IP difference

Figure 2. Example of unbiased group-wise non-rigid registration on real CC data sets. Performance comparison using CDF-HC, estimate of Holder's, IP, and normalized IP methods.



(a) Before Regist. (b) CDF-HC (c) Holder's (d) IP difference (e) Normalized IP difference

Figure 3. Example of unbiased groupwise no-rigid registration on outlier noise. Compare the three methods: CDF-HC, estimate of Holder's, IP difference and normalized IP difference.

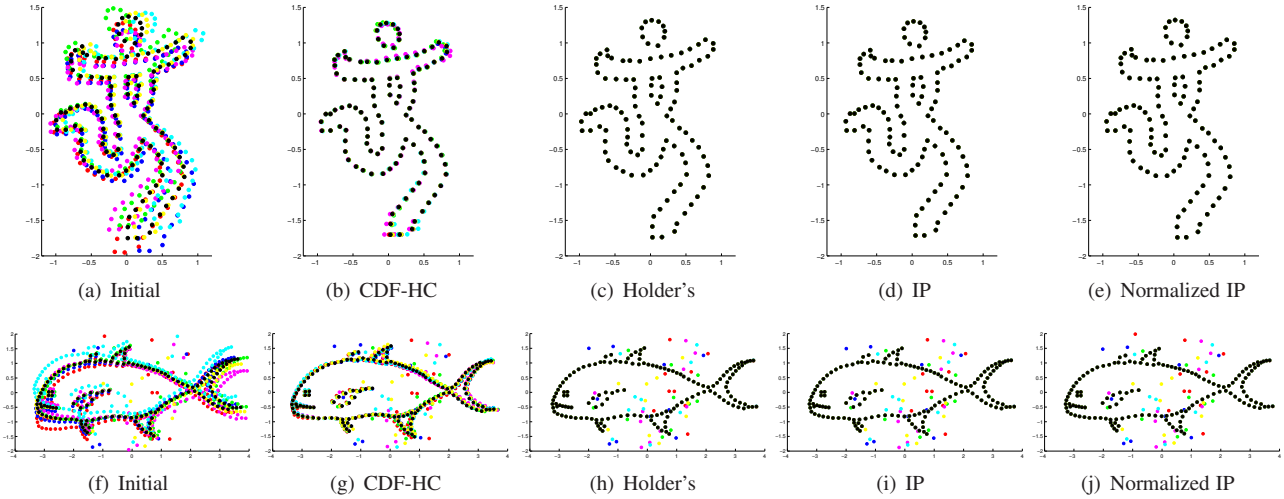
can result when the non-rigid transformation is not properly regularized.

6.2. Groupwise registration of biased datasets

To demonstrate the accuracy and robustness against noise, we borrowed two data-sets from [3]: olympic logo and noisy fish. These datasets were synthetically generated to be biased. The first point-set is the original set and the other six were generated by passing the point-set through various non-rigid transformations using thin-plate splines. For the second dataset, the fish, in addition to the transformation, ten randomly generated jitter points were also inserted into each point-set. Notice, this is a different kind of perturbation from the one employed in the regularization experiments (5), where the point locations were altered by adding Gaussian noise. The initial point-set positions and the final registration results for CDF-HC, Holder's, IP, and normalized IP are shown in Fig. 4. For the olympic and fish data $\lambda = 0.01$, $\sigma = 0.5$ with 99% annealing rate for all three algorithms.

To analytically measure the accuracy of the registration - the similarity between the final point-sets, we compute the Kolmogorov-Smirnov (KS) statistic [7] between the ground truth point-set, the first set, and the final registered point-sets. The average KS-statistic results for the corpus callosum (CC7), CC7 with outliers CC7(+out), CC7 with outliers registration but only considering the original points for the KS statistic CC7(-out), olympic logo, and the noisy fish datasets are shown in Table 1. We also include KS-statistic results of the CDF-JS [22] for the olympic and fish datasets that are listed in [3]. The results show that our methods perform better than CDF-JS and CDF-HC. It is important to highlight the difference between the KS statistic on the CC7 with outliers data set when the outliers are removed from the set. When the outliers are removed after registration to compute the KS statistic, the normalized IP algorithm not only exposes the best performance both also the largest decrease in KS statistic in agreement with the visual inspection from Figure 3.

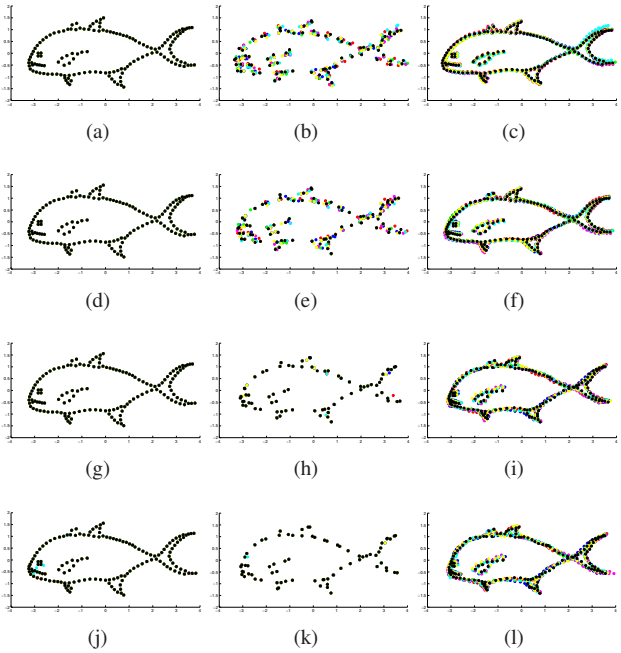
The normalized information potential algorithm performs well on any transformation type and noise level. In addition, in comparison with CDF-HC, the information potential algorithm is faster to compute. Table 2 shows the average computation time of CDF-HC, Hölder's, information potential, and normalized information potential on three different sets of data, where all three methods are set to run



(a) Initial (b) CDF-HC (c) Holder's (d) IP (e) Normalized IP

(f) Initial (g) CDF-HC (h) Holder's (i) IP (j) Normalized IP

Figure 4. Example of biased group-wise non-rigid registration. Compare the four methods: Holder's, information potential, normalized information potential, and CDF-HC on the olympic logo and noisy fish datasets.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Figure 5. Registration results for different regularization levels λ. The first row displays the resulting registration for $\lambda = 0.1$ using the original set of shapes 5(a), a sub-sampled set of points which are corrupted with zero mean and 0.05 standard deviation *i.i.d.* Gaussian noise 5(b), and the original set of shapes transformed using the parameters learned based on the sub-sampled noisy shapes 5(c). Similarly, row 2 (5(d), 5(e), and 5(f)) corresponds to $\lambda = 0.01$, and rows 3 (5(g), 5(h), and 5(i)) and 4 (5(j), 5(k), and 5(l)) to $\lambda = 0.001$ and $\lambda = 0.0001$, respectively.

for 300 epochs. The results show that the proposed methods, Hölder's, information potential, and normalized information potential, are computationally less demanding. Our methods perform much faster than CDF-HC and CDF-JS.

Table 1. KS statistic

	CDF-JS	CDF-HC	Holder	IP	Norm-IP
CC7	N/A	0.0661	0.0555	0.0503	**0.0317**
CC7(+out)	N/A	0.0577	0.0635	0.0539	**0.0405**
CC7(-out)	N/A	0.0556	0.0556	0.0529	**0.0317**
olympic	0.1103	0.0295	0.0206	**0.0177**	**0.0177**
fish(out)	0.1314	0.0462	0.0387	**0.0383**	**0.0383**

Table 2. Run time

	CDF-HC	Holder	IP	Norm IP
CC7	252s	28s	30s	31s
CC7(outliers)	270s	29s	30s	31s
olympic	746s	74s	57s	60s
fish(outliers)	1946s	131s	111s	117s

7. Conclusions

In this paper, we presented a robust algorithm to simultaneously register multiple unlabelled point-sets represented as density functions. We used the argument of the logarithm in Rényi's second order entropy and Jensen's inequality to yield a cost function with closed-form solution that allows gradient based parameter updates. We observed that the alignment based on the poposed cost function focuses on high density regions making it robust to the presence of outliers. Furthermore, the proposed normalized cost function avoids the necessity of impossing constraints on the rigid transformation, which can be difficult to set, and the regularized non-rigid transformation can handle cases where points-sets are noisy. These poperties arise from the averaging behaviour of the group-wise registration, where no point-set is employed as reference.

References

[1] Besl, P.J., McKay, N.D.: A method for registration of 3-d shapes. IEEE Trans. Pattern Anal. Mach. Intell. **14**(2), 239–256 (1992) 1

[2] Bookstein, F.L.: Principal warps: Thin-plate splines and the decomposition of deformations. IEEE Trans. Pattern Anal. Mach. Intell. **11**(6), 567–585 (1989) 4

[3] Chen, T., Vemuri, B.C., Rangarajan, A., Eisenschenk, S.J.: Group-wise point-set registration using a novel cdf-based havrda-charvat divergence. Int. J. Comput. Vision **86**(1), 111–124 (2010) 2, 6, 7

[4] Chui, H., Rangarajan, A.: A new algorithm for non-rigid point matching. In: Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on, vol. 2, pp. 44 –51 (2000) 1

[5] Chui, H., Rangarajan, A.: A new point matching algorithm for non-rigid registration. Comput. Vis. Image Underst. **89**(2-3), 114–141 (2003) 2

[6] Chui, H., Rangarajan, A., Zhang, J., Leonard, C.: Unsupervised learning of an atlas from unlabeled point-sets. Pattern Analysis and Machine Intelligence, IEEE Transactions on **26**(2), 160 –172 (2004) 2

[7] Feller, W.: On the kolmogorov-smirnov limit theorems for empirical distributions. The Annals of Mathematical Statistics **12**(2), 177–189 (1948) 7

[8] Glaunes, J., Trouve, A., Younes, L.: Diffeomorphic matching of distributions: a new approach for unlabelled point-sets and sub-manifolds matching. In: Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, vol. 2 (2004) 1

[9] Hasanbelliu, E., Sanchez Giraldo, L., Principe, J.: A robust point matching algorithm for non-rigid registration using the cauchy-schwarz divergence. In: Machine Learning for Signal Processing (MLSP), 2011 IEEE International Workshop on (2011) 2

[10] Hasanbelliu, E., Sanchez Giraldo, L., Principe, J.: Information theoretic shape matching. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(125), 2436–2451 (2014) 2

[11] Jenssen, R., Principe, J.C., Erdogmus, D., Eltoft, T.: The cauchy-schwarz divergence and parzen windowing: Connections to graph theory and mercer kernels. Journal of the Franklin Institute **343**(6), 614 – 629 (2006) 2

[12] Jian, B., Vemuri, B.: A robust algorithm for point set registration using mixture of gaussians. In: Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, vol. 2, pp. 1246 – 1251 (2005) 1

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

[14] Parzen, E.: On estimation of a probability density function and mode. The Annals of Mathematical Statistics **33**(3), 1065–1076 (1962) 2

[15] Principe, J.C.: Information Theoretic Learning, Renyi's Entropy and Kernel Perspectives. Springer, New York, NY (2010) 2

[16] Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. The MIT Press, Cambridge, Massachusetts (2006) 6

[17] Renyi, A.: Selected papers of Alfred Renyi. Akademia Kiado, Budapest (1976) 3

[18] Rudin, W.: Principles of Mathematical Analysis. McGraw-Hill Publishing Co., New York (1976) 2

[19] Shannon, C., Weaver, W.: The Mathematical Theory of Communication. University of Illinois Press, Urbana (1949) 3

[20] Tikhonov, A.N., Arsenin, V.I.: Solutions of Ill-Posed Problems. Winston and Sons, Washington (1977) 5

[21] Tsin, Y., Kanade, T.: A correlation-based approach to robust point set registration. In: Computer Vision, 2004. ECCV 2004. Eighth European Conference on, vol. 3023, pp. 558 –569 (2004) 1

[22] Wang, F., Vemuri, B., Rangarajan, A.: Groupwise point pattern registration using a novel cdf-based jensen-shannon divergence. In: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on (2006) 2, 3, 7