Image registration in high dimensional feature space*

Huzefa Neemuchwala[b] and Alfred Hero[a]

[a]Dept EECS, University of Michigan, Ann Arbor, MI, 48109-2122
[b]Dept BME, University of Michigan, Ann Arbor, MI, 48109-2122

ABSTRACT

Image registration is a difficult task especially when spurrious image intensity differences and spatial variations between the two images are present. To robustify image registration algorithms to such spurrious variations it can be useful to employ an image registration matching criteria on higher dimensional feature spaces. This paper will present an overview of our recent work on image registration using high dimensional image features and entropic graph matching criteria. New entropic graph estimates of information divergence measures will be presented. We will demonstrate the advantage of our approach for ultrasound breast image registration.

Keywords: pattern matching, k-nearest neighbor graphs, information divergence estimation, multimodality image registration.

1. INTRODUCTION

Image registration methods select a sequence of intensity preserving transformations to maximize an image similarity measure between a reference image and a target, or secondary, image. The accuracy of the registration algorithm critically depends on two factors: the selection of a highly discriminating image feature space and the choice of similarity measure to match these image features. These factors are especially important when some of the intensity differences are due to the sensor itself, as arises in registration with different types of imaging sensors or registration of speckle-limited images. In such cases, it is well known that the standard linear cross correlation is a poor similarity measure. This has motivated the development of other measures that are robust to intensity distortions caused by the sensor modality including: optical flow matching[1]; level set matching[2]; Jensen difference minimization[3]; and mutual information (MI) maximization.[4]

The last two aforementioned methods can be called "entropic methods" since they use a matching criterion based on different similarity measures defined as relative entropies between the feature densities. Entropic methods have been shown to be virtually unbeatable for some medical imaging image registration applications.[5,6] Several properties of entropic methods have contributed to their popularity for image registration: 1) because they are statistically based measures they easily accommodate combinations of texture based and edge based registration features; 2) relative entropies are easily defined that are invariant to invertible intensity transformations on the feature space; 3) they are simple to compute and the number of local maxima can be controlled by suitably constraining the set of image transformations.

Entropic methods for registration have been largely based on density estimation techniques which are exceedingly difficult as feature dimension becomes high. Thus entropic registration methods have been limited to low dimensional feature spaces, such as pixel intensity levels, for which an estimate of feature density is feasible. This paper gives an overview of a recent class of extensions of entropic similarity measures that break this computational bottleneck for high dimensional features. We also present a comprehensive comparison between these extended entropic measures and the standard density estimation based methods for speckle limited ultrasound breast image registration.

*This work was supported in part by NIH grant 1P01CA87634 and by ARO contract DAAD19-02-1-0262.
Further author information (Send correspondence to A.H.): E-mail: A.H., H.N.: hero@umich.edu, hneemuch@umich.edu

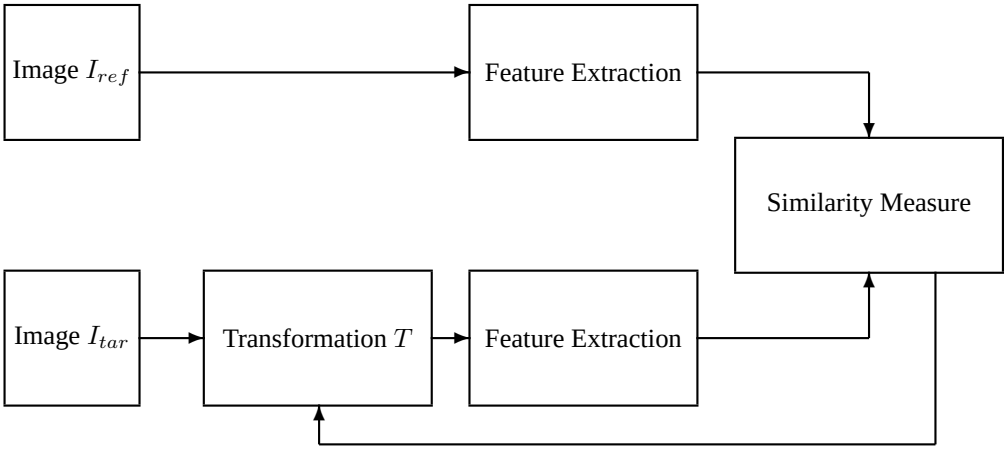


Figure 1. Block diagram of an image registration system

The guiding principle behind our extensions is the use of continuous quasi-additive power weighted graphs, such as the minimal spanning tree (MST) and k-Nearest Neighbor graph (kNNG), to estimate entropic similarity measures. These are discussed in Section 3. Different graph length functionals will allow us to approximate a wide variety of entropic matching criteria without the need to explicitly estimate densities or histograms. Building on our previous work,[7,8] in Sections 4 and 5 we will show how a kNNG can be used to approximate entropic similarity measures like the α-mutual information, α-Jensen divergence, and Geometric-Arithmetic mean affinity. Finally, in Section 6 will demonstrate how the combination of high dimensional wavelet features and kNNG similarity measures can lead to significant registration benefits in ultrasound breast imaging. More details on the methods presented here, along with other imaging applications, e.g., geo-registration and tracking, can be found in our recent book chapter.[9]

2. BACKGROUND

The three chief components of an image registration system (Figure 1) are: (1) definition and extraction of features that discriminate between different image poses I_{ref} and I_{tar}; (2) adaptation of a matching criterion that quantifies feature similarity, is capable of resolving important differences between images, yet is robust to image artifacts; (3) implementation of optimization techniques which allow fast search over possible transformations T. In this paper we shall be principally concerned with the second component of the system: the choice of matching criterion, also called a similarity or dissimilarity measure.

2.1. Mutual Information Image Registration

The mutual information (MI) can be interpreted as a similarity measure between the reference and target pixel intensities or as a dissimilarity measure between the joint density and the product of the marginals of these intensities. The MI was originally introduced for gray scale image registration.[4] Let X_0 be a reference image and consider a transformation of the target image X_1, defined as $X_T = T(X_1)$. We assume that the images are sampled on a grid of $M \times N$ pixels. Let (z_{0k}, z_{Tk}) be the pair of (scalar) gray levels extracted from identical (k-th) pixel locations in the reference and target images, respectively.

The basic assumption underlying MI image matching is that $\{(z_{0k}, z_{Tk})\}_{k=1}^{MN}$ are independent identically distributed (i.i.d.) realizations of a pair (Z_0, Z_T), $Z_T = T(Z_1)$, of random variables having joint density $f_{0,1}(z_0, z_T)$.

If the reference and the target images were perfectly correlated, e.g., identical images, then Z_0 and Z_T would be dependent random variables. On the other hand, if the two images were statistically independent, the joint density of Z_0 and Z_T would factor into the product of the marginals $f_{0,1}(z_0, z_T) = f_0(z_0)f_1(z_T)$. The (Shannon) MI measures the dissimilarity between the joint density and the product of the marginals

$$\mathrm{MI} = \int f_{0,1}(z_0, z_T) \log \left(\frac{f_{0,1}(z_0, z_T)}{f_0(z_0) f_1(z_T)} \right) dz_0 dz_T = H(f_0) + H(f_1) - H(f_{0,1}), \tag{1}$$

where $H(g) = -\int g \ln g$ denotes the Shannon entropy of density g.

For registering two discrete $M \times N$ images, one searches over a set of transformations of the target image to find the one that maximizes the MI (1) between the reference and the transformed target. We call this the "single pixel MI". In Viola and Wells[4] the authors empirically approximated the single pixel MI (1) by "histogram plug-in" estimates, which when extended to the αMI gives the estimate (neglecting unimportant normalization constants)

$$\widehat{\mathrm{MI}} \stackrel{\mathrm{def}}{=} \frac{1}{\alpha - 1} \log \sum_{z_0, z_T = 0}^{255} \hat{f}_{0,1}(z_0, z_T) \log \left(\frac{\hat{f}_{0,1}(z_0, z_T)}{\hat{f}_0(z_0) \hat{f}_1(z_T)} \right). \tag{2}$$

In (2) we assume 8-bit gray level, $\hat{f}_{0,1}$ denotes the joint intensity level "coincidence histogram"

$$\hat{f}_{0,1}(z_0, z_T) = \frac{1}{MN} \sum_{k=1}^{MN} I_{z_{0k}, z_{Tk}}(z_0, z_T), \tag{3}$$

and $I_{z_{0k}, z_{Tk}}(z_0, z_T)$ is the indicator function equal to one when $(z_{0k}, z_{Tk}) = (z_0, z_T)$ and equal to zero otherwise. Variants of this basic procedure have been applied to image registration by many authors.[5, 10] Other feature definitions have been proposed including gray level differences[11] and pixel pairs.[12]

To illustrate the MI registration procedure, the coincidence histogram is shown in Fig. 2 for the case of two ultrasound breast images X_0, X_1 (Fig. 3). Fig. 2 shows two cases. At top left is the coincidence histogram when the reference and secondary images are taken from the same two-dimensional slice of the US breast volume and are in perfect alignment ($X_0 = X_1$). At bottom left is the same histogram when the secondary image is rotated by 8°. The top right and bottom right panels in Fig. 2 are analogous except that the secondary images is extracted from a different two-dimensional slice separated from the reference by 2mm. In both cases the entropy $H(\hat{f}_{0,1})$ (dispersion) of the histogram is greater for the bottom panels (out of alignment) than for the top panels (in alignment) of the figure. Therefore, the MI can discriminate between the degrees of alignment.

3. GENERAL ENTROPIC DISSIMILARITY MEASURES

Let Z be a d-dimensional random vector and let $f(z)$ and $g(z)$ denote two possible densities for Z. Here Z will be a feature vector constructed from the reference image and the target image to be registered and f and g will be the feature densities. When the features are discrete valued the densities f and g should be interpreted as probability mass functions.

3.1. Measures Related to the Rényi Divergence

The Rényi α-divergence, also called the Rényi α-relative entropy, between f and g of fractional order $\alpha \in (0, 1)$[13] :

$$D_\alpha(f \| g) \quad = \quad \frac{1}{\alpha - 1} \log \int \; g(z) \left(\frac{f(z)}{g(z)} \right)^\alpha dz = \frac{1}{\alpha - 1} \log \int f^\alpha(z) g^{1-\alpha}(z) dz. \tag{4}$$

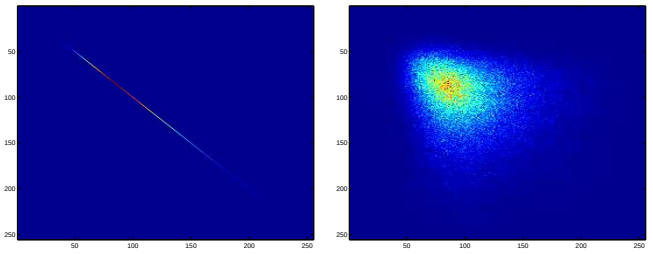


Figure 2. Joint coincidence histograms for single-pixel gray level features. Both horizontal and vertical axes of each panel are indexed over the gray level range of 0 to 255. (a): joint histogram scatter plot for the case that reference image (X_i) and secondary image (X_j) are the same slice of the US image volume (Case 142) at perfect 0° alignment ($X_j = X_i$). (c): same as (a) except that reference and secondary are misaligned by 8° relative rotation as in Fig. 3. (b): same as (a) except that the reference and secondary images are from adjacent (2mm separation) slices of the image volume. (d): same as (c) except that images are misaligned by 8° relative rotation.

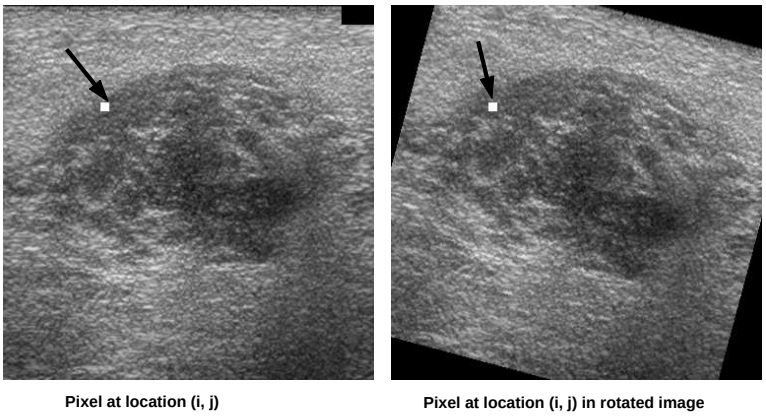


Figure 3. Single-pixel gray level coincidences are recorded by counting number of co-occurrences of a pair of gray level in the reference (a) and in the secondary (b) images at a pair of homologous pixel locations. Here the secondary image (b) is rotated by 15° relative to the reference image (a).

When the density f is supported on $[0,1]^d$ and g is uniform over this domain the (negative) α-divergence reduces to the Rényi α-entropy of f:

$$H_\alpha(f) = \frac{1}{1-\alpha}\log\int f^\alpha(z)dz. \tag{5}$$

When specialized to various values of α the α-divergence can be related to other well known divergence and affinity measures. Two of the most important examples are the Hellinger dissimilarity $-2\log\int\sqrt{f(z)g(z)}dz$ obtained when $\alpha = 1/2$, which is related to the Hellinger-Battacharya distance squared,

$$D_{Hellinger}(f\|g) \quad = \quad \int\left(\sqrt{f(z)}-\sqrt{g(z)}\right)^2 dz = 2\left(1-\exp\left(\tfrac{1}{2}D_{\frac{1}{2}}(f\|g)\right)\right), \tag{6}$$

and the Kullback-Liebler (KL) divergence obtained in the limit as $\alpha\to 1$ of (4),

$$\lim_{\alpha\to 1} D_\alpha(f\|g) = \int g(z)\log\frac{g(z)}{f(z)}dz. \tag{7}$$

Another divergence measure arises as a special cases of the Rényi α-divergence: the α-geometric-arithmetic mean divergence (α-GA)[14]

$$\alpha D_{GA}(f,g) \quad = \quad D_\alpha(pf+qg\|f^pg^q) = \frac{1}{\alpha-1}\log\int(pf(z)+qg(z))^\alpha(f^p(z)g^q(z))^{1-\alpha}dz, \tag{8}$$

where the weights p and $q=1-p$ are selected in the interval $(0,1)$. To our knowledge this measure has never been applied to image registration.

Finally, when the dissimilarity between a joint density $f(x,y)$ and the product of its marginals $g(x,y)=f(x)f(y)$ is of interest, the α-mutual information (αMI) can be defined from the α-divergence:

$$\alpha\text{MI} \quad = \quad D_\alpha(f\|g) = \frac{1}{\alpha-1}\log\int f^\alpha(x,y)f^{1-\alpha}(x)f^{1-\alpha}(y)dxdy. \tag{9}$$

In the limit as $\alpha\to 1$ this measure converges to the Shannon mutual information (MI).

3.2. Other Entropic Similarity Measures

Another divergence measure was introduced by Henze and Penrose[15] as the limit of the Friedman-Rafsky multivariate run-length statistic[16] and we shall call it the Henze-Penrose (HP) divergence

$$D_{HP}(f\|g) = \int\frac{p^2f^2(z)+q^2g^2(z)}{pf(z)+qg(z)}dz, \tag{10}$$

with respect to weights p and $q=1-p, p\in[0,1]$. To our knowledge this measure has not been applied to image registration.

An alternative entropic dissimilarity measure between two distributions is the α-Jensen difference[13]:

$$\Delta H_\alpha(p,f,g) = H_\alpha(pf+qg) - [pH_\alpha(f)+qH_\alpha(g)],, \tag{11}$$

with respect to weights p and $q=1-p, p\in[0,1]$. The α-Jensen difference has been applied to image registration.[3,17]

All of the above divergence measures can be obtained as special cases of the general class of f-divergences.[13] The α-Jensen difference shares the following properties with f-divergences: it depends on the features only through the feature density functions; it is a non-negative function and equal zero only when $f=g$; it is convex in f and g. On the other hand, unlike the divergences, the α-Jensen difference is not invariant to invertible transformations of the feature space Z. This means that the α-Jensen difference could depend on the feature parameterization, which is not desirable. We will see that this translates into reduced discrimination capability in image registration applications.

4. ENTROPIC GRAPH ESTIMATORS OF FEATURE SIMILARITY MEASURES

All of the similarity measures introduced in the previous section could be estimated by plugging in feature histogram or density estimates of the multivariate density f. This is the approach taken in virtually all previous image registration work. A deterrent to these approaches is the curse of dimensionality, which imposes prohibitive computational burden when attempting to construct histograms in large feature dimensions. An alternative approach, taken here, is to attempt to estimate the divergence directly without recourse to difficult density estimation. Such approaches have been developed for entropy estimation using the gap Vasicek estimator for one dimensional feature spaces[18] and entropic graph estimators have been developed for higher dimensions.[7, 19] As our previous work in entropic graph estimators forms the basis for approximating more general feature similarity metrics we will review it here.

4.1. Entropic Graphs for Entropy Estimation

Assume that an i.i.d. set of continuously valued feature vectors $\mathcal{Z}_n = \{z_1, \ldots, z_n\}$, $z \in \mathbf{R}^d$, have been collected from an image and that it is desired to estimate the entropy of the underlying feature density $f(z)$. An entropic graph estimator of entropy is constructed as follows. Considering the n points in $\mathcal{Z}_n$ as vertices, construct a a certain kind of minimal graph that spans these vertices. Assume that the total edge length of the graph satisfies the continuous and quasi additive property,[20] which is satisfied by graph constructions such as the minimal spanning tree, the traveling salesman tour solving the traveling salesman problem (TSP), the steiner tree, the Delaunay triangulation, and the k nearest neighbor graph[†] Then the total edge length function converges (a.s.) to a monotone function of the Rényi α-entropy of f as $n \to \infty$.

More specifically, define the length functional of such a minimal graph as

$$L_\gamma(\mathcal{Z}_n) = \min_{E \in \Omega} \sum_{e \in E} e^\gamma(\mathcal{Z}_n) = \sum_i e_i^\gamma,$$

where Ω is a set of graphs with specified properties, e.g., the class of acyclic or spanning graphs (leading to the MST), e is the euclidean length of an edge in Ω, γ is called the edge exponent or the power weighting constant, and $0 < \gamma < d$. The sum $\sum_i e_i^\gamma$ is an equivalent notation this length functional, where the $\{e_i\}_i$ are the lengths of the edges in the minimal graph. The determination of L_γ usually requires a combinatorial optimization over the set Ω but in some cases, e.g., the kNNG, this can be done in $O(n \log n)$ time.

The celebrated Beardwood, Halton and Hammersley (BHH) Theorem asserts that[20]

$$\lim_{n \to \infty} L_\gamma(\mathcal{Z}_n)/n^\alpha = \beta_{d,\gamma} \int f^\alpha(z) dz, \quad (a.s.) \tag{12}$$

where $\alpha = (d - \gamma)/d$ and $\beta_{d,\gamma}$ is a constant independent of f - it only depends on the type of graph construction (MST, kNNG, etc). Comparing this to the expression (5) for the Rényi entropy it is obvious that an entropy estimator can be constructed from the relation $(1 - \alpha)^{-1} \log (L_\gamma(\mathcal{Z}_n)/n^\alpha) = \hat{H}_\alpha(f) + c$, where $c = (1 - \alpha)^{-1} \log \beta_{d,\gamma}$ is a removable bias. Furthermore, it is seen that one can estimate entropy for different values of $\alpha \in [0, 1]$ by adjusting γ. For many minimal graph constructions the topology of the minimal graph is independent of γ and only a single combinatorial optimization is required to estimate H_α for all α.

4.2. Entropic Graph Estimate of α-Jensen Difference

The results of the last section can be applied to estimating the α-Jensen difference between feature densities of two images. Assume two sets of feature vectors $\mathcal{O}_{n_0} = \{o_i\}_{i=1}^{n_0}$ and $\mathcal{X}_{n_1} = \{x_i\}_{i=1}^{n_1}$ are extracted from images X_0 and X_1 and are i.i.d. realizations of random variables O and X having multivariate densities f_o and f_x, respectively. The

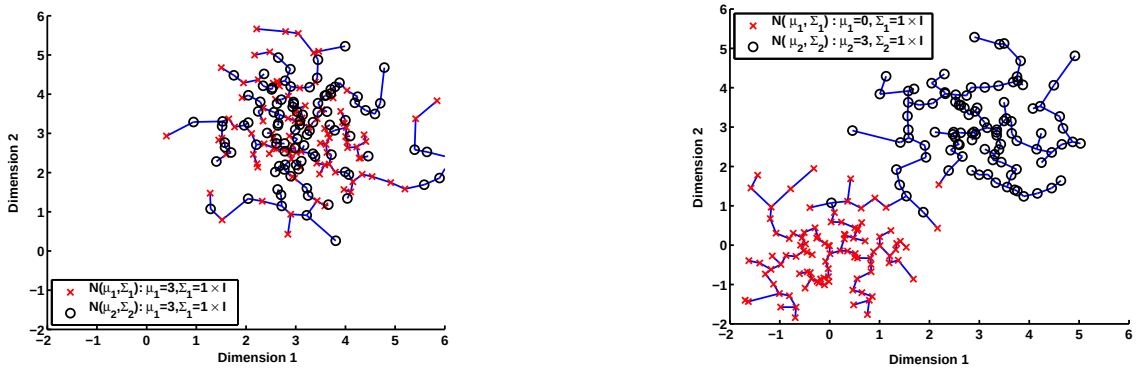
[†]Roughly speaking, continuous quasi additive functionals can be approximated closely by the sum of the weight functionals of minimal graphs constructed on a uniform partition of $[0, 1]^d$.

case of equal numbers $n_0 = n_1$ of features from X_0 and X_1 is the typical case in image registration when features are extracted at each pixel location. Define the set union $\mathcal{Z}_m = \mathcal{O}_{n_0} \cup \mathcal{X}_{n_1}$ containing $m = n_0 + n_1$ unordered feature vectors z_i. If n_0, n_1 increase at constant rate as a function of n then any consistent entropy estimator constructed from the vectors $\{z_i\}_{i=1}^{n_0+n_1}$ will converge to $H_\alpha(pf_0 + qf_1)$ as $n_0, n_1 \to \infty$ where $p = \lim_{n_0,n_1\to\infty} n_0/(n_0+n_1)$ and $q = 1-p$. This motivates the following finite sample entropic graph estimator of α-Jensen difference

$$\Delta\widehat{H}_\alpha(p, f_0, f_1) = \widehat{H}_\alpha(\mathcal{O}_{n_0} \cup \mathcal{X}_{n_1}) - \left[p\widehat{H}_\alpha(\mathcal{O}_{n_0}) + q\widehat{H}_\alpha(\mathcal{X}_{n_1})\right], \tag{13}$$

where $\widehat{H}_\alpha(\mathcal{Z}_0 \cup \mathcal{Z}_1)$ is the entropy estimator obtained from a MST or kNNG length functional constructed on the n point union of both sets of feature vectors, and the marginal entropy estimates $\widehat{H}_\alpha(\mathcal{O}_{n_0})$, $\widehat{H}_\alpha(\mathcal{X}_{n_1})$ are constructed on the individual sets of n_0 and n_1 feature vectors, respectively. We can similarly define a density-based estimator of α-Jensen difference. Observe that for rigid image registration problems (without cropping errors) the marginal entropies $\{H_\alpha(f_i)\}_{i=1}^K$ over the set of image transformations will be identical, obviating the need to compute estimates of the marginal α-entropies.

For illustration we show how the entropic graph αJensen difference estimator applies to a synthetic sample from two 2D feature distributions. The two densities are Gaussian bivariate densities with different means but identical (spherical) covariances. A sample from these two densities and the MST are shown in Fig. 4 for two different values of the mean parameter. Figure 5 shows the kNNG for the same realization. Note that the discriminating power of the kNN and MST reside in the sensitivity of the total edge length of these graphs to the difference in the means of the densities.



(a) MST $\mu_1 = \mu_2$ and $\Sigma_1 = \Sigma_2$ (b) MST $\mu_1 = \mu_2 - 3$ and $\Sigma_1 = \Sigma_2$

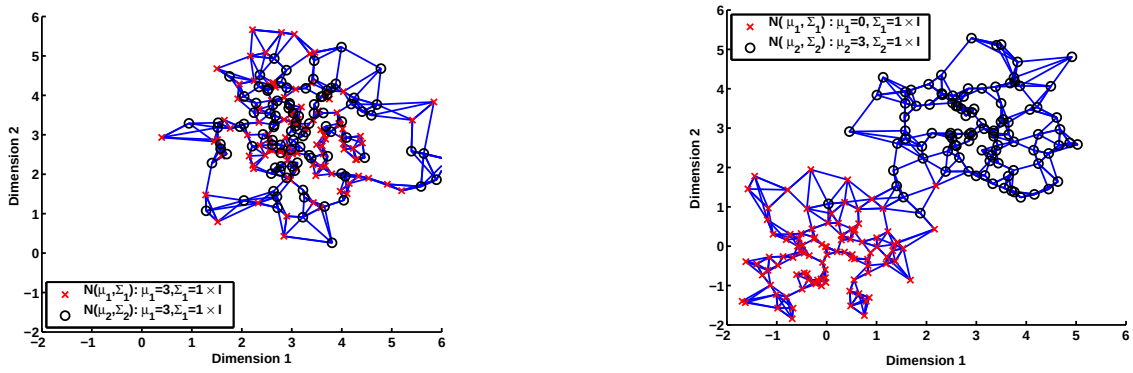
Figure 4. Illustration of MST estimate of α-Jensen difference for Gaussian case. Two bivariate normal distributions $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_1, \Sigma_1)$ are used. The 'x' labeled points are n_1 samples from $f_x = \mathcal{N}(\mu_1, \Sigma_1)$, whereas the 'o' labeled points are n_0 samples from $f_o = \mathcal{N}(\mu_2, \Sigma_2)$. (left) $\mu_1 = \mu_2$ and $\Sigma_1 = \Sigma_2$ and (right) $\mu_1 = \mu_2 - 3$ while $\Sigma_1 = \Sigma_2$. When normalized by $(n_0+n_1)^\alpha$ the sum of all edge lengths converges to $\Delta H_\alpha(p, f_x, f_o)$ (within a constant factor).

4.3. Entropic Graph Estimate of Henze-Penrose Affinity

Friedman and Rafsky[16] presented a multivariate generalization of the Wald-Wolfowitz runs statistic for the two sample problem. The Wald-Wolfowitz test statistic is used to decide between the following hypotheses on a pair of scalar random variables $X, O \in \mathbf{R}^d$ with densities f_x, f_o respectively:

$$H_0\colon f_x = f_o, \qquad\qquad H_1\colon f_x \neq f_o, \tag{14}$$

The test statistic is applied to an i.i.d. random sample $\{x_i\}_{i=1}^{n_1}, \{o_i\}_{i=1}^{n_0}$ from f_x and f_o. In the univariate Wald Wolfowitz test ($d = 1$), the $n_0 + n_1$ scalar observations $\{z_i\}_i = \{x_i\}_i, \{o_i\}_i$ are ranked in ascending order. Each

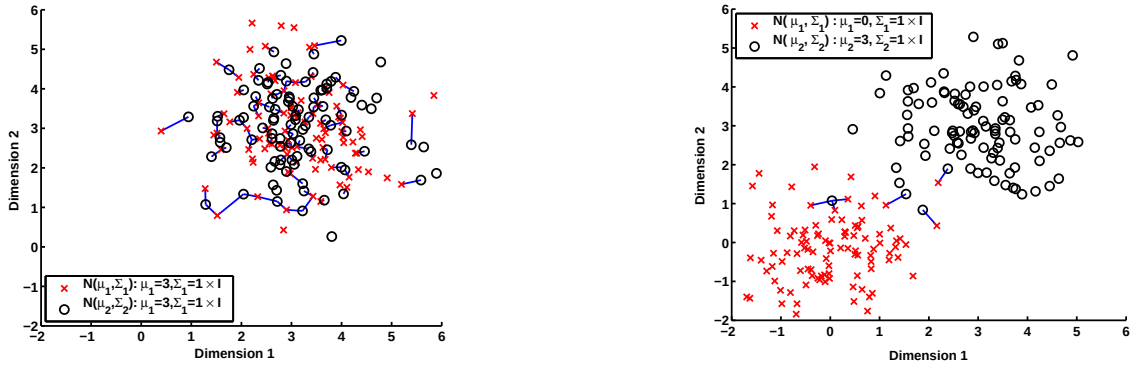


(a) kNN $\mu_1 = \mu_2$ and $\Sigma_1 = \Sigma_2$ (b) kNN $\mu_1 = \mu_2 + 3$ and $\Sigma_1 = \Sigma_2$

Figure 5. Illustration of kNN estimate of α-Jensen difference for Gaussian case illustrated in Fig. 4. When normalized by $(n_0 + n_1)^\alpha$ the sum of all edge lengths converges to $\Delta H_\alpha(p, f_o, f_x)$ (within a constant factor).

observation is then replaced by a class label X or O depending upon the sample to which it originally belonged, resulting in a rank ordered sequence. The Wald-Wolfowitz test statistic is the total number of runs (run-length) R_ℓ of X's or O's in the label sequence. As in run-length coding, R_ℓ, is the length of consecutive sequences of length ℓ of identical labels.

The Friedman-Rafsky (FR) test[16] generalizes the Wald-Wolfowitz test to d dimensions by clever use of the MST. The FR test proceeds as follows: 1) construct the MST on the pooled multivariate sample points $\{x_i\} \bigcup \{o_i\}$; 2) retain only those edges that connect an X labeled vertex to an O labeled vertex; 3) The FR test statistic, N, is defined as the number of edges retained. The hypothesis H_1 in (14) is accepted for smaller values of the FR test statistic. As shown by Henze and Penrose,[15] when normalized by the total number $n_0 + n_1$ of points, the FR test statistic N converges to 1 minus the Henze-Penrose divergence (10) between the distributions f_x and f_o. The FR test is illustrated in Fig. 6.



(a) FR $\mu_1 = \mu_2$ and $\Sigma_1 = \Sigma_2$ (b) FR $\mu_2 = \mu_1 + 3$ and $\Sigma_1 = \Sigma_2$

Figure 6. Illustration of Friedman and Rafsky's (FR) MST estimate of the Henze-Penrose divergence for Gaussian case illustrated in Fig. 4. The proportion of MST edges that connect feature vectors from different classes is a consistent estimate of $1 - D_{HP}(f_o \| f_x)$.

5. ENTROPIC GRAPH ESTIMATORS OF α-GA AND αMI

Assume for simplicity that the target and reference feature sets $\mathcal{O}_{n_0} = \{o_i\}_i$ and $\mathcal{X}_{n_1} = \{x_i\}_i$ have the same cardinality $n_0 = n_1 = n$. The estimators of α-GA and αMI are based on a kNNG-Voronoi partitioning heuristic, described below. While Voronoi and nearest neighbor approaches to entropy estimation have been proposed by Miller[21] and Kozachenko and Leonenko,[22] respectively, to our knowledge the heuristic below is new and is applicable to both entropy and divergence estimation.

5.1. kNNG-Voronoi Partitioning Heuristic

First consider the general problem of estimating a functional $\overline{\psi} = E[\psi(f)] = \int \psi(f(z))f(z)dz$ of a multivariate density $f(z)$, $z \in [0,1]^d$, based on a set of i.i.d. samples $\mathcal{Z}_n = \{z_1, \ldots, z_n\}$ from f. If we had a consistent estimator $\hat{f}_n$ of the density f, obtained independently from another set of samples $\mathcal{Z}_n^{'}$, then a consistent estimator, $\widehat{\psi}$, of $\overline{\psi}$ would be the "plug-in"

$$\widehat{\psi} = \frac{1}{n}\sum_{i=1}^{n} \psi(\hat{f}(z_i)). \tag{15}$$

Of course, actual implementation of a plug-in estimator is exactly what we are trying to avoid due to the inherent high complexity of density estimation in high dimensions. However, by judicious choice of an "unimplementable" plug-in density estimator followed by an approximation we will obtain an implementable estimator of $\overline{\psi}$.

Divide the total number n of samples into a disjoint training sample $\mathcal{Z}_{train}$, containing n_{train} points, and a test sample $\mathcal{Z}_{test}$, containing n_{test} points, $n = n_{train} + n_{test}$. We consider a partition density estimator of $f(z)$ built from the training sample using data dependent Voronoi tesselation of the domain of f. First, using the sample $\mathcal{Z}_{train}$ generate a K cell Voronoi partition Π using a Linde-Buzo-Gray (LBG) or K-means algorithm[23] on $[0,1]^d$. Then, in the notation of,[24] define the Voronoi partition density estimator

$$\hat{f}(z) = \frac{\mu(\Pi(z))}{\lambda(\Pi(z))} \tag{16}$$

where $\Pi(z)$ is the cell of the data-dependent Voronoi partition in $[0,1]^d$ containing the point z, μ is the empirical distribution of $\mathcal{Z}_{train}$ and λ is the Lebesgue measure. More specifically, for any set $\Pi \in [0,1]^d$, $\mu(\Pi)$ is the number of points of $\mathcal{Z}_{train}$ falling into Π divided by the total number n_{train} of points, and $\lambda(\Pi)$ is the volume of Π. The Voronoi partition density estimator (16) will be asymptotically consistent as long as $n_{train}, K \to \infty$ and $K/n_{train} \to 0$.[24] Therefore, as long as we chose the number K of cells in such a way that these conditions are satisfied, the following estimator of the functional (15):

$$\widehat{\psi} = \frac{1}{n_{test}} \sum_{z_i \in \mathcal{Z}_{test}} \psi(\hat{f}(z_i)) \tag{17}$$

will be asymptotically unbiased with variance going to zero.

Now, for the heuristic. Note that the density estimator (16) depends on the Voronoi partition only through the volume of the cell and its number of "counts" from $\mathcal{Z}_{train}$. Consider the form of the estimator (17) when we set $n_{train} = n_{test} = n$, $\mathcal{Z}_{train} = \mathcal{Z}_{test} = \mathcal{Z}_n$, and $K = n$. Note that this case violates the conditions for convergence stated above. Then each of the n points will be at the center of its own Voronoi cell and therefore $\mu(\Pi(z)) = 1/n$. If each cell is approximately spherical (similar assumption as Gersho's conjecture for asymptotic VQ) we can make the following kNN approximation of the cell volume:

$$\lambda(\Pi(z_i)) \approx c e_i^d, \tag{18}$$

where $e_i(\mathcal{Z}_n) = \min_j \|z_i - z_j\|$ is the distance from the point z_i to its nearest neighbor in the set $\mathcal{Z}_n$ and c is a constant. Substitution of this heuristic approximation into (15) gives the estimator

$$\widehat{\psi} = \frac{1}{n}\sum_{i=1}^{n} \psi((nce_i^d)^{-1}), \tag{19}$$

and $e_i = e_i(\mathcal{Z}_n)$.

Observe that in the special case of α-entropy $\psi(u) = u^{\alpha-1}$, $\alpha = (d-\gamma)/\gamma$ and the above estimator (19) reduces to

$$\widehat{\psi} = \frac{c^{\alpha-1}}{n^{\alpha}}\sum_{i=1}^{n} e_i^{\gamma},$$

which, up to a constant factor $c^{\alpha-1}$ is identical to the kNNG estimator of entropy introduced in Section 4. By the BHH Theorem we know that this estimator converges (a.s.) to the integral $c^{\alpha-1}\beta\int f^{\alpha}(z)dz$ which is identical, up to a scale factor, to $E[\psi]$. Thus, even though the heuristic was derived under some very questionable assumptions, which certainly invalidate consistency of the density estimator, we nonetheless preserved consistency of the entropy estimate.

5.2. kNNG Estimator of αGA

Assume an equal number of feature vectors $\mathcal{O}_n = \{o_i\}_{i=1}^n$ and $\mathcal{X}_n = \{x_i\}_{i=1}^n$ are acquired from images 1 and 2, where o_i and x_i are i.i.d. random variables distributed with densities f_o and f_x, respectively. Here we apply the heuristic approximation (19) to estimate $\alpha D_{GA}(f_o, f_x) = (\alpha-1)^{-1}\log I_{GA}(f_o, f_x)$, where $I_{GA}(f_o, f_x)$ is the integral in (8):

$$I_{GA}(f_o, f_x) = \int h^{\alpha}(z)(f_o^p(z)f_x^q(z))^{1-\alpha}dz = \int \left(\frac{f_o^p(z)f_x^q(z)}{h(z)}\right)^{1-\alpha} h(z)dz, \tag{20}$$

and $h(z) = pf_o(z) + qf_x(z)$. To convert this expression into an empirical estimate of the form (15) observe that h is the density function of the pooled sample $\mathcal{Z}_n = \{o_i, x_i\}_{i=1}^n$ with $p = q = 1/2$. Re-index (in no particular order) these $2n$ samples as $\{z_i\}_{i=1}^{2n}$. If the consistent partition density estimation procedure, discussed in the previous subsection, is used to estimate f_o, f_x and h from $\mathcal{O}_n$, $\mathcal{X}_n$ and $\mathcal{Z}_n$, respectively, we know that

$$\widehat{I_{GA}} = \frac{1}{2n}\sum_{i=1}^{2n}\left(\frac{\hat{f}_o^p(z_i)\hat{f}_x^q(z_i)}{\hat{h}(z_i)}\right)^{1-\alpha}, \tag{21}$$

is a consistent estimator of αGA divergence. We assume for simplicity that the support sets of f_o and f_x are contained in $[0,1]^d$. There is no loss of generality if actual support sets are bounded regions $\mathcal{S} \subset \mathbf{R}^d$ as they can be mapped inside the unit cube through coordinate transformation.

Next invoke the kNN-Voronoi heuristic and make the partition density estimator approximations

$$\hat{h}(z_i) = \frac{\mu(\Pi_z(z_i))}{\lambda(\Pi_z(z_i))} \approx \frac{c/n}{\min\{e_i^d(\mathcal{O}_n), e_i^d(\mathcal{X}_n)\}}, \qquad \hat{f}_o(z_i) = \frac{\mu(\Pi_o(z_i))}{\lambda(\Pi_o(z_i))} \approx \frac{c/n}{e_i^d(\mathcal{O}_n)}, \qquad \hat{f}_x(z_i) = \frac{\mu(\Pi_x(z_i))}{\lambda(\Pi_x(z_i))} \approx \frac{c/n}{e_i^d(\mathcal{X}_n)}.$$

Substitution of these approximations into (21) yields the entropic graph approximation to the α-GA mean divergence (8):

$$\widehat{\alpha D_{GA}} = \frac{1}{\alpha-1}\log\frac{1}{2n}\sum_{i=1}^{2n}\min\left\{\left(\frac{e_i(\mathcal{O}_n)}{e_i(\mathcal{X}_n)}\right)^{\gamma/2}, \left(\frac{e_i(\mathcal{X}_n)}{e_i(\mathcal{O}_n)}\right)^{\gamma/2}\right\}, \tag{22}$$

where unimportant constants have been omitted.

5.3. kNNG Estimator of αMI

Similar to Section 2.1 we assume that n vectors of paired features $z_i = (o_i, x_i) \in \mathbf{R}^{2d}$ are acquired from the two images, i.e., $\mathcal{Z}_n = \{z_i\}_{i=1}^n$ is the coincidence scatterplot of these features. Define $f_{ox}(z)$ the joint feature density and f_o and f_x the marginal densities of $o_i \in \mathbf{R}^d$ and $x_i \in \mathbf{R}^d$, respectively, and define the integral expression I_{MI}

$$I_{MI} = \int f^{\alpha}(ox)(u,v) f_o^{1-\alpha}(u) f_x^{1-\alpha}(v) du dv$$

appearing in the expression for the αMI (9), i.e., $\alpha\text{MI} = \frac{1}{\alpha-1}\log I_{MI}$. If a consistent partition density estimate of procedure, discussed in the previous subsection, is used to estimate f_{ox}, f_o and f_x, then it is easily seen that

$$\widehat{I_{MI}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{f}_o(o_i)\hat{f}_x(x_i)}{\hat{f}_{ox}(o_i,x_i)}\right)^{1-\alpha}, \tag{23}$$

is a consistent estimator of $I_M I$.

Application of the heuristic (18) yields

$$\hat{f}_{ox}(z_i) \approx \frac{c/n}{e_i^{2d}(\mathcal{Z}_n)}, \qquad \hat{f}_o(u_i) \approx \frac{c/n}{e_i^{d}(\mathcal{O}_n)}, \qquad \hat{f}_x(v_i) \approx \frac{c/n}{e_i^{d}(\mathcal{X}_n)}.$$

which when substituted into (23) gives the entropic graph approximation to the αMI

$$\widehat{\alpha MI} = \frac{1}{\alpha-1}\log\frac{1}{n^{\alpha}}\sum_{i=1}^{n}\left(\frac{e_i(\mathcal{Z}_n)}{\sqrt{e_i(\mathcal{O}_n)e_i(\mathcal{X}_n)}}\right)^{2\gamma}, \tag{24}$$

where $e_i(\mathcal{Z}_n)$ is the distance from the point $z_i = (o_i, x_i) \in \mathbf{R}^{2d}$ to its nearest neighbor in $\{Z_j\}$ and $e_i(\mathcal{O}_n)$ $(e_i(\mathcal{X}_n))$ is the distance from the point $o_i \in \mathbf{R}^d$, $(x_i \in \mathbf{R}^d)$ to its nearest neighbor in $\mathcal{O}_n$ $(\mathcal{X}_n)$. Again, unimportant constant factors have been omitted from (24).

6. APPLICATION TO ULTRASOUND BREAST IMAGING

Ultrasound (US) imaging is an important medical imaging modality for whole breast imaging that can aid discrimination of malignant from benign lesions, can be used to detect multi-focal secondary masses, and can quantify response to chemotherapy or radiation therapy. In Fig. 7 a set of twenty 2D slices extracted from a 3D volumetric US breast scanner is shown for twenty different patients (cases) receiving chemotherapy. The women were imaged on their backs with the transducer placed so as to image through the breast toward the chest wall. Some of the cases clearly exhibit tumors (delineated masses with shadows), others exhibit significant connective tissue structure (bright thin lines or edges), and all have significant speckle noise and distortions.

In registering ultrasound images of the breast, the reference and secondary images have genuine differences from each other due to biological changes and differences in imaging, such as positioning of the tissues during compression and angle dependence of scattering from tissue boundaries. The tissues are distorted out of a given image plane as well as within it. Speckle noise, elastic deformations and shadows further complicate the registration process thus making ultrasound breast images notoriously difficult to register. It is for this reason that conventional registration methods tend to have problems with US breast images. Here we will illustrate the advantages of matching on high dimensional feature spaces implemented with entropic similarity metrics.

6.1. Ultrasound Breast Database

To benchmark the various registration methods studied we evaluated the mean squared registration error for registering a slice of US breast image volume to an adjacent slice in the same image volume (case). For each case we added

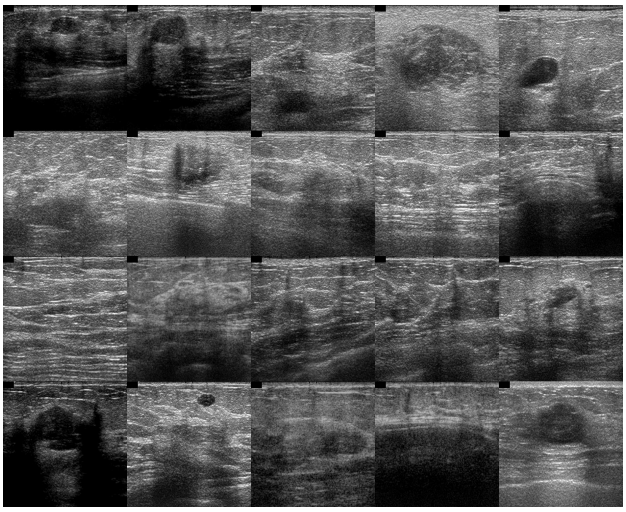


Figure 7. Ultrasound (US) breast scans from twenty volume scans of patients undergoing chemotherapy.

differing amounts of spatially homogeneous and independent random noise to both slices in order evaluate algorithm robustness. A training database of volumetric scans of 6 patients and a test database of 15 patient scans were created. Feature selection was performed using the training database and registration performance was evaluated over the test database. These databases were drawn from a larger database of 3D scans of the left or right breast of female subjects, aged 21-49 years, undergoing chemotherapy or going to biopsy for possible breast cancer. Each volumetric scan has a field of view of about 4cm^3 (voxel dimensions 0.1mm$^2 \times 0.5$mm) and encompasses the tumor, cyst or other structure of interest. The scans were acquired at 1cm depth resolution yielding 90 cross-sectional images at 0.4cm horizontal resolution. The patient data was collected with the intention to monitor therapy progress in the patients. Tumor/Cyst dimensions vary and can range from 5mm^3 to 1cm^3 or higher. As the aim of this study is to quantitatively compare different feature selection and registration methods we restricted our investigation to rotation transformations over $\pm 16^\circ$.

6.2. Feature Space

We have experimented with a large number of vector valued features including, Meyer 2D wavelet coefficents, grey level tag features, and curvelet features. Here we present results for vector valued features constructed by projecting image patches onto a basis for the patch derived from independent component analysis (ICA). The ICA basis is especially well suited for our purposes since it aims to obtain vector features which have statistically independent elements and can therefore facilitate estimation of αMI and other entropic measures.

Specifically, in ICA an optimal basis is found from a training set which decomposes images X_i in the training set into a small number of approximately statistically independent components $\{S_j\}$ each supported on an 8×8 pixel block (we choose an 8 by 8 block only for concreteness):

$$X_i = \sum_{j=1}^{p} a_{ij} S_j. \tag{25}$$

We select basis elements $\{S_j\}$ from an over-complete linearly dependent basis using randomized selection over the database. For image i the feature vectors z_i are defined as the coefficients $\{a_{ij}\}$ in (25) obtained by projecting each of its 8×8 sub-image blocks onto the basis.

Figure 6.2 illustrates the estimated 64 dimensional (8×8) ICA basis for the training database. The basis was ex-

tracted by training on over 100,000 randomly sampled 8×8 sub-images taken from the 6 volumetric breast ultrasound scans. The algorithm used for extraction was Hyvarinen and Oja's[25] `FastICA` ICA code (available from[26]) which uses a fixed-point algorithm to perform maximum likelihood estimation of the basis elements in the ICA data model (25). Given this ICA basis and a pair of to-be-registered image slices, coefficient vectors are extracted by projecting each 8×8 neighborhood in the images onto the basis set. Thus for αMI the coincidence scatter plot is in 128 dimensions; the number of dimensions of a coincidence feature extracted at a particular row-column coordinate in the pair of images. The feature space for the αJensen, αGA and Henze-Penrose registration criteria was constructed by pooling the two labeled sets of 64D feature vectors. Thus, the dimensionality of the feature space was 64D. MST or kNNG were constructed on the 64D feature spaces of the pooled sample. In either case these feature dimensions (128D or 64D) are too large for a histogram binning algorithm to be feasible, which prevented comparison to the full dimensional classical density plug-in MI registration criterion.

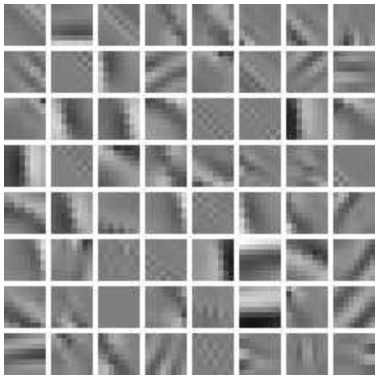


Figure 8. 8×8 ICA basis set obtained from training on randomly selected 8×8 blocks in the training database of breast scans.

6.3. Experimental Results

For each of the 15 scans in the test set 2 image slices were extracted in the depth direction perpendicular to the skin, such that they showed the cross-section of the tumor. These two slices have a seperation distance of about 5mm. At this distance, the speckle deccorelates but the underlying anatomy remains approximately unchanged. The first cross sectional slice was picked such that it intersected with the ellipsoidal-shaped tumor through its center. The second slice was picked closer to the edge of the tumor. These images thus show a natural decline in tumor size, as would be expected in time sampled scans of tumors responding to therapy. Since view direction changes from one image scan to the next for the same patient over time, rotational deformation is often deployed to correct these changes during registration. We simulated this effect by registering a rotationally deformed image with its unrotated slice-seperated counterpart, for each patient in the 15 test cases. Rotational deformation was in steps of 2 degrees such that the sequence of deformations was [-16 -8 -4 -2 0 (unchanged) 2 4 8 16] degrees. Further, the images were offset (relatively translated) by 0.5mm (5 pixels) laterally to remove any residual noise correlation since it can bias the registration results. Since some displacement can be expected from the handheld UL imaging process and the relative tissue motion of the compressible breast tissue, this is not unreasonable. For each deformation angle, divergence measures were calculated, where the 'registered state' is the one with 0 degrees of relative deformation.

For each extracted image slice we created 250 noisy replicates by adding truncated Gaussian noise. 8×8 neighborhoods of the ultrasound image replicates were projected onto the 64 dimensional ICA basis. The rms registration error is illustrated for six different algorithms in Fig. 9 as a function of the rms (truncated) Gaussian noise. Registration error was determined as the rms difference between the location of the peak in the matching criterion and the true rotation angle. Note from the figure that, except for the α-Jensen difference, the standard single pixel MI underperformes relative to the other methods. This is due to the superiority of the high dimensional ICA features used by these other methods. The α Jensen difference implemented with kNN vs MST give identical performance. Unlike the other

metrics, the α Jensen difference is not invariant to reparameterization, which explains its relatively poor performance for large rms noise. Finally, we remark that the runtime complexity of the kNN-based methods (off-the-shelf kdb-tree implemention) is lower than the MST-based methods (off-the-shelf Kruskal algorithm).

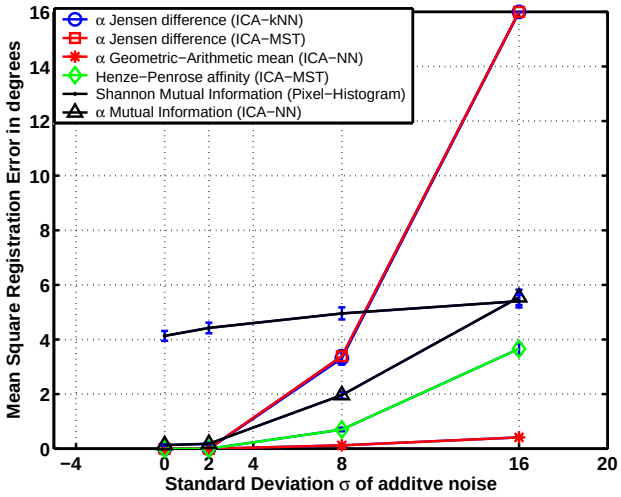


Figure 9. Rotational root mean squared error obtained from registration of ultrasound breast images using six different image similarity/dissimilarity criteria. Standard error bars are as indicated. These plots were obtained by averaging 15 cases, each with 250 Monte Carlo trials adding noise to the images prior to registration, corresponding to a total of 3750 registration experiments.

7. CONCLUSION

In this paper we have presented several extensions of our previous work on entropy estimation for image registration. These extensions include new kNN estimators of the mutual information (αMI) and geometric-arithmetic mean divergence (αGA). As compared to previous work in which estimated Jensen differences were used for registration, these divergence measures have the advantage of invariance to reparameterization of the feature space. While we do not yet have any convergence results for the kNN divergence estimators, there is circumstantial theoretical evidence that they do converge. Furthermore, our numerical evaluations show that these divergence estimators outperform previous approaches to image registration. We also introduced the Friedman-Rafsky (FR) multivariate run test, which is an estimator of Henze-Penrose divergence, as a new matching criterion for image registration. Our numerical experiments showed that the FR, αGA, and αMI significantly outperform previous approaches in terms of registration mean squared error. Of course, as compared to our kNNG divergence estimators, the FR method has the advantage of proven theoretical convergence but has the disadvantage of higher runtime complexity.

REFERENCES

1. M. Lefébure and L. Cohen, "Image registration, optical flow and local rigidity," *J. Mathematical Imaging and Vision*, vol. 14, no. 2, pp. 131–147, 2001.
2. Eric Debreuve, Michel Barlaud, Ivan Laurette, Gilles Aubert, and Jacques Darcourt, "Nonparametric and non-rigid registration method applied to myocardial-gated spect," *Proc. of IEEE Nuclear Science Symposium*, vol. 49, no. 3, 2002.
3. Y. He, A. Ben-Hamza, and H. Krim, "An information divergence measure for ISAR image registration," *Signal Processing*, Submitted, 2001.
4. P. Viola and W.M. Wells, "Alignment by maximization of mutual information," in *Proc. of 5th Int. Conf. on Computer Vision, MIT*, 1995, vol. 1, pp. 16–23.

5. C. R. Meyer, J. L. Boes, B. Kim, P. H. Bland, K. R. Zasadny, P. V. Kison, K. F. Koral, K. A. Frey, and R. L. Wahl, "Demonstration of accuracy and clinical versatility of mutual information for automatic multimodality image fusion using affine and thin-plate spline warped geometric deformations," *Medical Image Analysis*, vol. 1, no. 3, pp. 195–206, Apr. 1997.
6. D. Hill, P Batchelor, M. Holden, and D. Hawkes, "Medical image registration," *Phys. Med. Biol.*, vol. 26, pp. R1–R45, 2001.
7. A.O. Hero, B. Ma, O. Michel, and J. Gorman, "Applications of entropic spanning graphs," *IEEE Signal Processing Magazine*, vol. 19, no. 5, pp. 85–95, Sept. 2002, `www.eecs.umich.edu/~hero/imag_proc.html`.
8. H. Neemuchwala, A. O. Hero, and P. Carson, "Image matching using alpha-entropy measures and entropic graphs," *European Journal of Signal Processing*, To appear 2002.
9. H. Neemuchwala and A. O. Hero, *Image Fusion*, chapter Entropic Graphs for Registration, Marcel-Dekker and CRC Press, 2005.
10. A. Rangarajan, I.-T. Hsiao, and G. Gindi, "Integrating anatomical priors in ect reconstruction via joint mixtures and mutual information," in *IEEE Medical Imaging Conference and Symposium on Nuclear Science*, Oct. 1998, vol. III.
11. T. Butz and J. Thiran, "Affine registration with feature space mututal information," in *Lecture Notes in Computer Science 2208: MICCAI 2001*, Springer-Verlag berlin Heidelberg 2001, 2001, pp. 549–556.
12. D. Rueckert, M. Clarkson, D. Hill, and D. Hawkes, "Non-rigid registration using higher order mutual information," in *Proc. SPIE*, 2000, vol. 3979, pp. 438–447.
13. M. Basseville, "Distance measures for signal processing and pattern recognition," *Signal Processing*, vol. 18, pp. 349–369, 1989.
14. I. J. Taneja, "New developments in generalized information measures," *Advances in Imaging and Electron Physics*, vol. 91, pp. 37–135, 1995.
15. N. Henze and M. Penrose, "On the multivariate runs test," *Annals of Statistics*, vol. 27, pp. 290–298, 1999.
16. Jerome H. Friedman and Lawrence C. Rafsky, "Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests," *Annals of Statistics*, vol. 7, no. 4, pp. 697–717, 1979.
17. A. O. Hero, B. Ma, and O. Michel, "Imaging applications of stochastic minimal graphs," in *IEEE Int. Conf. on Image Processing*, Thessaloniki, Greece, Oct. 2001.
18. E. Miller and J. Fisher, "ICA using spacing estimates of entropy," in *Proc. Fourth International Symposium on Independent Component Analysis and Blind Signal Separation*, Nara, Japan, Apr. 2003, pp. pp. 1047–1052.
19. A.O. Hero and O. Michel, "Robust entropy estimation strategies based on edge weighted random graphs," in *Proc. of Meeting of Intl. Soc. for Optical Engin. (SPIE)*, San Diego, CA, July 1998, vol. 3459, pp. 250–261.
20. C. Redmond and J. E. Yukich, "Asymptotics for Euclidean functionals with power weighted edges," *Stochastic Processes and their Applications*, vol. 6, pp. 289–304, 1996.
21. E. Miller, "A new class of entropy estimators for multi-dimensional densities," in *Proc. IEEE Int. Conf. Acoust., Speech, and Sig. Proc.*, 2003, pp. 297–300.
22. L. F. Kozachenko and N. N. Leonenko, "Sample estimate of entropy of a random vector," *Problems of Information Transmission*, vol. 23, no. 1, pp. 95–101, 1987.
23. R. M. Gray, *Source Coding Theory*, Kluwer Academic, Norwell MA, 1990.
24. G. Lugosi and A. Nobel, "Consistency of data-driven histogram methods for density estimation and classification," *Annals of Statistics*, vol. 24, no. 2, pp. 687–706, 1996.
25. A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 1999.
26. A. Hyvärinen, "Fast ICA Code," `www.cis.hut.fi/projects/ica/fastica/`.