Image registration methods in high-dimensional space

Huzefa Neemuchwala[a,c], Alfred Hero[a,b] and Paul Carson[a,c]
Dept Biomedical Engineering[a], University of Michigan, Ann Arbor MI 48109
Dept EECS[b], University of Michigan, Ann Arbor, MI, 48109-2122
Dept Radiology[c], University of Michigan, Ann Arbor MI 48109-0533

July 9, 2006

Abstract

Quantitative evaluation of similarity between feature densities of images is an important step in several computer vision and data-mining applications such as registration of two or more images, retrieval and clustering of images. Previously we had introduced a new class of similarity measures based on entropic graphs to estimate Rènyi's α-entropy, α-Jensen difference divergence, α-mutual information and other divergence measures for image registration. Entropic graphs such as the minimum spanning tree (MST) and k-Nearest neighbor (kNN) graph allow the estimation of such similarity measures in higher dimensional feature spaces. A major drawback of histogram-based estimates of such measures is that they cannot be reliably constructed in higher dimensional feature spaces.

In this paper, we shall briefly extrapolate upon the use of entropic graph based divergence measures mentioned above. Additionally, we shall present estimates of other divergence viz the Geometric-Arithmetic mean divergence and Henze-Penrose affinity. We shall present the application of these measures for pairwise image registration using features derived from independent component analysis of the images. An extension of pairwise image registration is to simultaneously register multiple images, a challenging problem that arises while constructing atlases of organs in medical imaging. Using entropic graph methods we show the feasibility of such simultaneous registration using graph based higher dimensional estimates of entropy measures. Finally we present a new non-linear correlation measure that is invariant to non-linear transformations of the underlying feature space and can be reliably constructed in higher dimensions. We present an image clustering experiment to demonstrate the robustness of this measure to non-linear transformations and contrast it with the clustering performance of the linear correlation coefficient.

Keywords: image registration, divergence estimation, k-nearest neighbor graphs.

[1] Further author information: This work was performed when the first author was a graduate student in Biomedical Engineering at University of Michigan, Ann Arbor. (Send correspondence to A.H.): E-mail: A.H., H.N. P.C.: `hero@umich.edu, hneemuch@umich.edu, pcarson@umich.edu. This work was supported in part by NIH grant 1P01CA87634 and by ARO contract DAAD19-02-1-0262.`

1 Introduction

The accuracy of image matching algorithms critically depend 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 of speckle-limited images or when images of objects exhibit non-linear intensity relationship. In such cases, it is well known that the standard linear cross correlation is a poor similarity measure.

To overcome limitations of linear correlation, Viola and Wells [1] and Maes et. al. [2] devised a similarity measure based on the Kullback-Liebler [3] information divergence between the joint feature density and the product of the marginal densities. This is the mutual information (MI) measure and it quantifies the non-linear correlation between images as the amount of statistical dependency in the underlying joint probability distribution functions (pdf); where the pdf is estimated using pixel intensity histograms. Although the pixel-histogram method overcomes the nonlinear correlation problem, drawbacks abound due to the use of histogram density estimators. Histograms are efficient density estimators in low dimensions, but cannot be reliably constructed in higher dimensional feature spaces (> 4) thus limiting themselves to applications where dimensionality of feature space is very low. Several applications such as in multi-image and multi-sensor registration require the higher dimensional feature descriptors to effectively capture signal properties. Unfortunately, the pixel-histogram method cannot be directly extended to address these problems.

Ma and Hero [4] proposed the use of entropic-graph methods for image registration. As contrasted to the previous approaches, entropic graphs estimate an information divergence without the need to compute histogram density estimates. Our approach is based on the entropic graph based estimate of Rényi's α-entropy introduced by [5, 6, 7] and developed by Ma for image registration [4]. An entropic graph is any graph whose normalized total weight (sum of the edge lengths) is a consistent estimator of α-entropy. An example of an entropic graph is the k-nearest neighbor graph and due to its low computational complexity it is an attractive entropic graph algorithm. This graph estimator can be viewed as a multidimensional generalization of the Vasicek-Shannon entropy estimator for one dimensional features [8, 9]. Graph methods sidestep the issue of density estimation and have asymptotic convergence to the Rényi α-entropy of the feature distribution.

This paper extends our previous work with regards to using entropic graphs for registration. Here we present the applications of entropic graphs for robust pairwise image registration and extensions to multi-image registration. We also introduce a new measure of non-linear correlation that can be estimated using entropic graphs and is shown to be more robust to non-linear transformations than the linear correlation coefficient (CC). Previously [10], we had demonstrated the advantages of cross modality image registration algorithms that used divergence measures calculated on higher dimensional feature spaces using entropic graph methods such as the minimum spanning tree and k-Nearest neighbor graphs. Divergence was estimated using the α-Jensen difference that is a generalization of the Shannon-Jensen divergence. In [11, 12] we presented entropic graph based estimation of Henze-Penrose affinity, α-MI and α-Geometric arithmetic mean divergence. An overview of our previous work is presented in some detail here to ease understanding of concepts related to entropic graph based estimation of entropy and divergence.

This paper is arranged as follows: Section 2 briefly introduces different divergence measures based on Rényi's generalized divergence. 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 [6, 10, 11, 12], in Sections 3 and 4 we will show how a kNNG can be used to estimate $alpha$-entropy, Henze-Penrose affinity α-mutual information and Geometric-Arithmetic mean divergence. Section 5 introduces a new non-linear correlation method based on entropic graphs. Section 6 will demonstrate how the combination of high dimensional ICA features and kNNG similarity measures can lead to significant registration benefits in ultrasound breast imaging. In section 7 we explain the utility of higher dimensional matching toward simultaneous registration of three images. Lastly, section 8 presents a clustering example to contrast the performance of the NLCC versus the CC in the face of image corruption due to non-linear distortion.

2 General Entropic Dissimilarity Measures

$\mathcal{Z}$ is a d-dimensional random vector and $f(z)$ and $g(z)$ denote two possible densities for $\mathcal{Z}$. Here $\mathcal{Z}$ 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.

2.1 Measures Related to the Rényi Divergence

The basis for entropic methods of image fusion is a measure of dissimilarity between densities f and g. The Rényi α-divergence, also called the Rényi α-relative entropy, between f and g of fractional order $\alpha \in (0,1)$

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

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:

$$H_\alpha(f) = \frac{1}{1-\alpha}\log\int f^\alpha(z)dz. \tag{2}$$

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 Hellinger-Battacharya distance squared,

$$D_{Hellinger}(f\|g) \quad = \quad \int\left(\sqrt{f(z)}-\sqrt{g(z)}\right)^2 dz = 2\left(1-\exp\left(\tfrac{1}{2}D_{\frac{1}{2}}(f\|g)\right)\right), \tag{3}$$

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

$$\lim_{\alpha\to 1} D_\alpha(f\|g) = \int g(z)\log\frac{g(z)}{f(z)}dz. \tag{4}$$

Another divergence measure arises as a special cases of the Rényi α-divergence: the α-geometric-arithmetic mean divergence (α-GA) [13]

$$\alpha D_{GA}(f,g) \quad = \quad D_\alpha(pf+qg\|f^pg^q) = \frac{1}{\alpha-1}\log\int(pf(z)+qg(z))^\alpha(f^p(z)g^q(z))^{1-\alpha}dz, \tag{5}$$

where the weights p and $q = 1-p$ are selected in the interval $(0,1)$. The α-GA divergence is a measure of the discrepancy between the arithmetic mean and the geometric mean of f and g, respectively, with respect to weights p and $q = 1-p, p \in [0,1]$. The α-GA divergence can thus be interpreted as the dissimilarity between the weighted arithmetic mean $pf(x)+qg(x)$ and the weighted geometric mean $f^p(x)g^q(x)$. Similarly to the α-Jensen difference (10), the α-GA divergence is equal to zero if and only if $f = g$ (a.e.) and is otherwise greater than zero. 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:

$$\alpha\mathrm{MI} \;=\; D_\alpha(f\|g) = \frac{1}{\alpha-1}\log\int f^\alpha(x,y)f^{1-\alpha}(x)f^{1-\alpha}(y)dxdy. \tag{6}$$

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

$$\mathrm{MI} = \int f_{0,1}(z_0,z_T)\log\left(\frac{f_{0,1}(z_0,z_T)}{f_0(z_0)f_1(z_T)}\right)dz_0dz_T = H(f_0)+H(f_1)-H(f_{0,1}), \tag{7}$$

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 (7) between the reference and the transformed target. We call this the “single pixel MI”. In Viola and Wells [14] the authors empirically approximated the single pixel MI (7) by “histogram plug-in” estimates, which when extended to the αMI gives the estimate (neglecting unimportant normalization constants)

$$\widehat{\mathrm{MI}} \overset{\mathrm{def}}{=} \frac{1}{\alpha-1}\log\sum_{z_0,z_T=0}^{255}\hat{f}_{0,1}(z_0,z_T)\log\left(\frac{\hat{f}_{0,1}(z_0,z_T)}{\hat{f}_0(z_0)\hat{f}_1(z_T)}\right). \tag{8}$$

2.2 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

$$D_{HP}(f\|g) = \int \frac{p^2f^2(z)+q^2g^2(z)}{pf(z)+qg(z)}dz, \tag{9}$$

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 [17]:

$$\Delta H_\alpha(p, f, g) = H_\alpha(pf + qg) - [pH_\alpha(f) + qH_\alpha(g)], , \tag{10}$$

with respect to weights p and $q = 1 - p, p \in [0, 1]$. The α-Jensen difference has been applied to image registration [18, 19]. For detailed discussion on this divergence measure please refer to [10, 11, 12].

All of the above divergence measures can be obtained as special cases of the general class of f-divergences [17]. The 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.

3 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 [20] and entropic graph entropic graph estimators have been developed for higher dimensions [6, 21]. As our previous work in entropic graph estimators forms the basis for approximating more general feature similarity metrics we will review it here.

3.1 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 [22], 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[2] 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 [22]

$$\lim_{n \to \infty} L_\gamma(\mathcal{Z}_n)/n^\alpha = \beta_{d,\gamma} \int f^\alpha(z) dz, \quad (a.s.) \tag{11}$$

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 (2) for the Rényi entropy it is obvious that an entropy estimator can be constructed from the relation $(1-\alpha)^{-1} \log\left(L_\gamma(\mathcal{Z}_n)/n^\alpha\right) = \hat{H}_\alpha(f) + c$, 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.

[2] 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 $[0,1]^d$.

3.2 Entropic Graph Estimate of Henze-Penrose Affinity

Friedman and Rafsky [16] presented a multivariate generalization of the Wald-Wolfowitz 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:

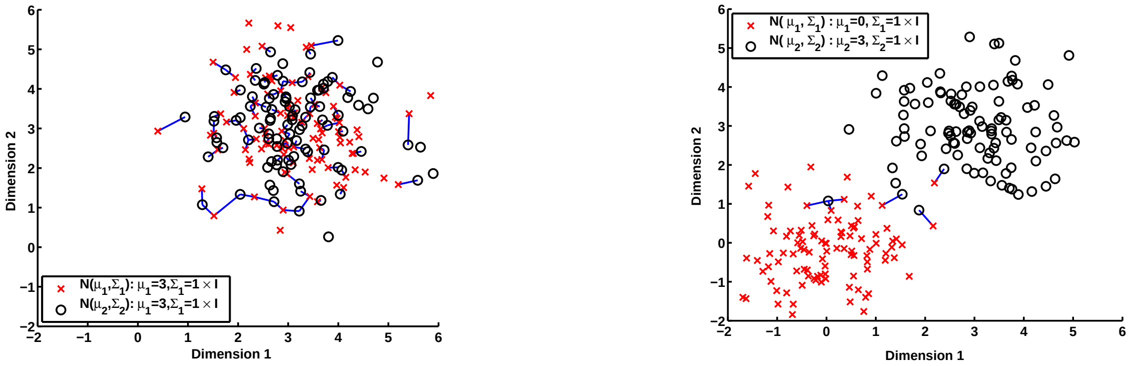
$$H_0\colon f_x = f_o, \qquad\qquad H_1\colon f_x \neq f_o, \tag{12}$$

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 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 (12) 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 (9) between the distributions f_x and f_o. The FR test is illustrated in Fig. 1.

4 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 [23] and Kozachenko and Leonenko [24], respectively, to our knowledge the heuristic below is new and is applicable to both entropy and divergence estimation.



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

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

Figure 1: Illustration of Friedman and Rafsky's (FR) MST estimate of the Henze-Penrose divergence for the case of two Gaussian densities. (a) The two densities have the mean and variance parameters. (b) the mean of one distribution is now shifted so that the densities diverge. The proportion of MST edges that connect feature vectors from different classes is a consistent estimate of $1 - D_{HP}(f_o \| f_x)$.

4.1 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 kNNG-Voronoi partitioning heuristic approximation from [25]. This heuristic allows us to approximate the volume of cellular Voronoi partitions on the feature density using kNN graph edge lengths. 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 (5):

$$I_{GA}(f_o, f_x) = \int h^\alpha(z)(f_o^p(z) f_x^q(z))^{1-\alpha} dz = \int \left(\frac{f_o^p(z) f_x^q(z)}{h(z)} \right)^{1-\alpha} h(z) dz, \tag{13}$$

and $h(z) = pf_o(z) + qf_x(z)$. Finally, 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 kNNG-Voronoi partition density estimation procedure discussed in [25], 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

$$\widehat{I_{GA}} = \frac{1}{2n} \sum_{i=1}^{2n} \left(\frac{\hat{f}_o^p(z_i) \hat{f}_x^q(z_i)}{\hat{h}(z_i)} \right)^{1-\alpha}, \tag{14}$$

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

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

Substitution of these approximations into (14) yields the entropic graph approximation to the α-GA mean divergence (5):

$$\widehat{\alpha D_{GA}} = \frac{1}{\alpha - 1} \log \frac{1}{2n} \sum_{i=1}^{2n} \min\left\{ \left(\frac{e_i(\mathcal{O}_n)}{e_i(\mathcal{X}_n)}\right)^{\gamma/2}, \left(\frac{e_i(\mathcal{X}_n)}{e_i(\mathcal{O}_n)}\right)^{\gamma/2} \right\}, \tag{15}$$

where unimportant constants have been omitted.

4.2 kNNG Estimator of αMI

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 scatter-plot 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 (6), 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

$$\widehat{I_{MI}} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\hat{f}_o(o_i)\hat{f}_x(x_i)}{\hat{f}_{ox}(o_i, x_i)} \right)^{1-\alpha}, \tag{16}$$

is a consistent estimator of $I_M I$. Here, we note that according to the definition of a consistent estimator, a consistent estimator of $I_M I$ is one that converges in probability to $I_M I$ as the sample size grows.

Application of the kNNG-Voronoi partitioning heuristic ([25]) 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 (16) gives the entropic graph approximation to the αMI

$$\widehat{\alpha MI} = \frac{1}{\alpha - 1} \log \frac{1}{n^\alpha} \sum_{i=1}^{n} \left(\frac{e_i(\mathcal{Z}_n)}{\sqrt{e_i(\mathcal{O}_n) e_i(\mathcal{X}_n)}} \right)^{2\gamma}, \tag{17}$$

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 (17).

4.3 Implementation Issue

The stable computation of the α-MI estimator (Equation 17) requires that $e_i(o)$ and $e_i(x)$ be non-zero whenever $e_i(o \times x)$ is non-zero (Figure 2). If either $e_i(o)$ or $e_i(x)$ is zero, α-MI cannot be calculated due to division-by-zero problems. For continuously distributed features $\{O_i\}$ and $\{\mathcal{X}_i\}$ the probability of stable computation is one, since the probability that any two feature components be exactly equal is zero. However, for practical applications where the feature space is quantized to finite precision arithmetic, the probability of stable computation is strictly less than one. In fact, it can be shown that the probability of stable computation of the α-MI estimator rapidly goes to zero as the number of feature realizations gets large.

A remedy for this is randomization. To avoid zero values of $e_i(o)$ and $e_i(x)$, a small amount of uniform noise may be added to the feature coefficient. This randomization disperses points uniformly in an area around their discretized value. This process is consistent with the assumption that local distribution of continuously valued feature vectors is uniform around their discretized values. In simulations with discretized 8-bit pixel intensity features, univariate uniform noise with a variance $\sigma^2 = 0.02$ was added to each pixel intensity. This ensured that no two intensities were exactly the same and thus enabling stable computation of αMI. Another approach is to replace $e_i(o)$ and $e_i(x)$ with $\max(e_i(o), \epsilon)$ and $\max(e_i(x), \epsilon)$, where $\epsilon << 1$ [26].

5 A non-linear correlation measure

The simple form of Equation 17 is suggestive of a non-linear correlation measure between the features $\{O_i\}$ and $\{\mathcal{X}_i\}$ that eliminates the implementation issue discussed above. Indeed, if "e_i" in Equation 17 is redefined as the statistical expectation "E", then the α-MI estimator takes the appearance of a linear correlation coefficient between $\{O_i\}$ and $\{\mathcal{X}_i\}$. However, as explained above, the ratio $e_i(o \times x)/\sqrt{e_i(o)e_i(x)}$ is not bounded between 0 and 1, rather it can take values that are arbitrarily large. The following modification of Equation 17 can be used to ensure that the non-linear correlation measure lie between 0 and 1. This new measure is called the non-linear correlation coefficient (NLCC).

Let $e_i(o \times x)$ be the distance from i-th feature pair (o_i, e_i) to its nearest neighbor as before. Instead of $e_i(o)$ and $e_i(x)$ being the coordinate-wise nearest neighbor distances along the feature coordinate axes $\mathcal{X}$ and O (See Figure 2) we define $\tilde{e}_i(o)$ and $\tilde{e}_i(x)$ the associated nearest neighbor distances in the plane (see Figure 3). The quantity $\tilde{e}_i(o \times x)/\sqrt{\tilde{e}_i(o)\tilde{e}_i(x)}$ is now bounded between 0 and 1. In particular, it is equal to one when the nearest neighbor to (o_i, x_i) is also the coordinate-wise nearest neighbor to (o_i, x_i) along the coordinate axes O and $\mathcal{X}$.

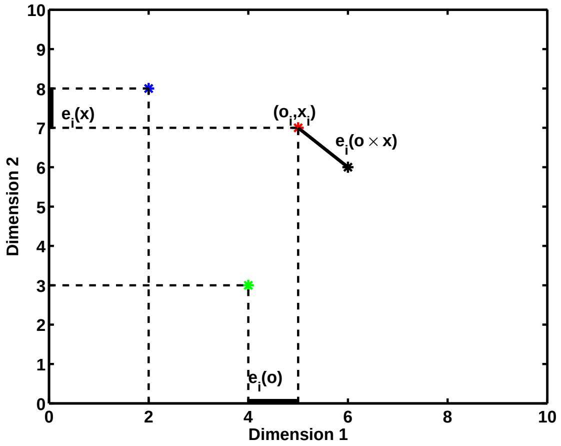


Figure 2: Illustration of the distances $e_i(o \times x)$, $e_i(o)$ and $e_i(x)$ used in the α-MI estimator (Equation 17)

In particular the quantity

$$\hat{\rho} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\tilde{e}_i(o \times x)}{\sqrt{\tilde{e}_i(o)\tilde{e}_i(x)}} \right) \tag{18}$$

is equal to one when the nearest neighbor graph is monotone (increasing or decreasing) piecewise linear curve in the plane 4. Thus if the features are realizations of the random vector $(O, \mathcal{X})$ which obeys the monotone model:

$$\Theta = g(\mathcal{X}), \tag{19}$$

where $g(\cdot)$ is a monotonic increasing function, the NLCC $\hat{\rho}$ will equal 1 with probability one. This motivates

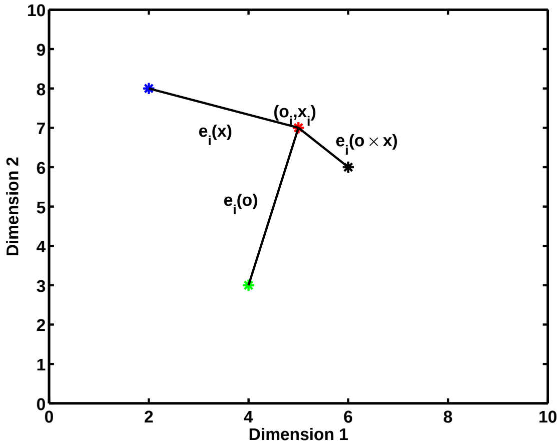


Figure 3: Illustration of modified distances $e_i(x)$ and $e_i(o)$ used to stabilize the estimator (Equation 17), defining the non-linear correlation coefficient (NLCC)

the use of ρ as a measure of information between Θ and $\mathcal{X}$. Unfortunately, if the actual model is

$$\Theta = g(\mathcal{X}) + w \tag{20}$$

where w is additive noise, $\hat{\rho}$ will converge to zero as $n \to \infty$ for any continuous random variable w. It can be shown that the rate of convergence in this case is $n^{\frac{-\gamma}{2d}}$. This motivates the modification of the NLCC to:

$$\hat{\rho}_{NLCC} = \frac{1}{n^{1-\gamma/2d}} \sum_{i=1}^{n} \left(\frac{\tilde{e}_i(o \times x)}{\sqrt{\tilde{e}_i(o)\tilde{e}_i(x)}} \right). \tag{21}$$

This modified correlation now takes values between 0 and ∞. A normalized version can be defined as:

$$\hat{\rho} = \frac{\hat{\rho}_{NLCC}}{1 + \hat{\rho}_{NLCC}} \tag{22}$$

that is between zero and one.

We illustrate the NLCC by comparing it to the linear correlation coefficient 23 for two simple models. The linear correlation coefficient is defined as:

$$\hat{\rho}_{CC} = \frac{\frac{1}{n}\sum_{i=1}^{n}(o_i - \bar{o})(x_i - \bar{x})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(o_i - \bar{o})^2 \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}} \tag{23}$$

where $\bar{o} = 1/n \sum_{i=1}^{n} o_i$ and $\bar{x} = 1/n \sum_{i=1}^{n} x_i$ are sample means.

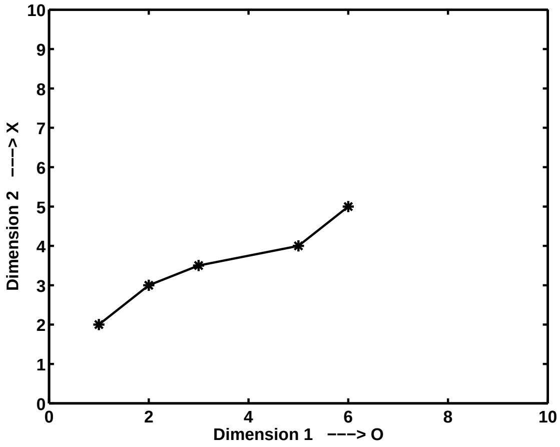


Figure 4: The Nearest Neighbor Graph over the realizations $\{(o_i \times x_i)\}_{i=1}^{N}$ of the paired features describes a monotone function in the plane. For this case, the NLCC $\hat{\rho} = 1$

5.1 Numerical experiments with NLCC

Consider the linear model $\Theta = a\mathcal{X} + w$, where $a^2 = \rho_{CC}^2/(\rho_{CC}^2 + 1)$. Figure 5 shows a plot of the linear (Equation 23) and nonlinear (Equation 21) correlation coefficients, $\hat{\rho}_{CC}$ and $\hat{\rho}_{NLCC}$ for this model as functions of the number of points N for various values of a. As a increases, the linear correlation increases but does not reach one due to the presence of additive noise w. In the limit as $N \to \infty$ the non-linear correlation coefficient converges to a constant.

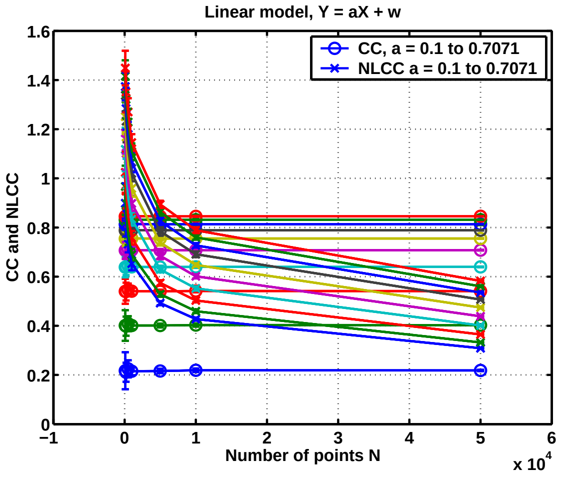


Figure 5: Comparison of Linear and non-linear correlation coefficient for a linear model

Now consider the nonlinear model given by $\Theta = ag(\mathcal{X}) + w$; $g(\mathcal{X}) = b\mathcal{X}^3$. As shown in Figure 6, the linear correlation coefficient remains unchanged at the value corresponding to the relation between Θ and $\mathcal{X}$. The non-linear correlation, however increases with a, showing that it responds to changes in the non-linear relation between Θ and $\mathcal{X}$.

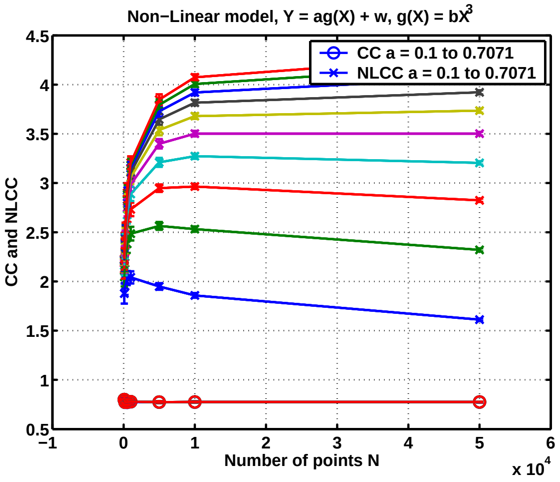


Figure 6: Comparison of Linear and non-linear correlation coefficient for a nonlinear model

Figure 7 confirms these findings. It illustrates the relation between the linear and non-linear correlation coefficients for both linear and non-linear models. The values are plotted for $N = 50000$ and a increases from 0.1 to 0.7071.

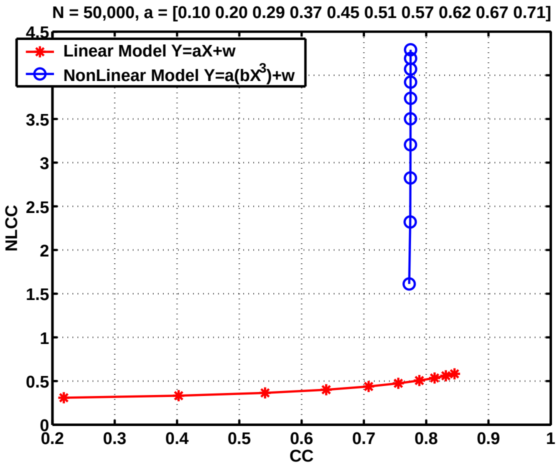


Figure 7: Plot of CC v/s NLCC for N = 50000 and a = 0.1 to 0.7071

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. 8 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 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

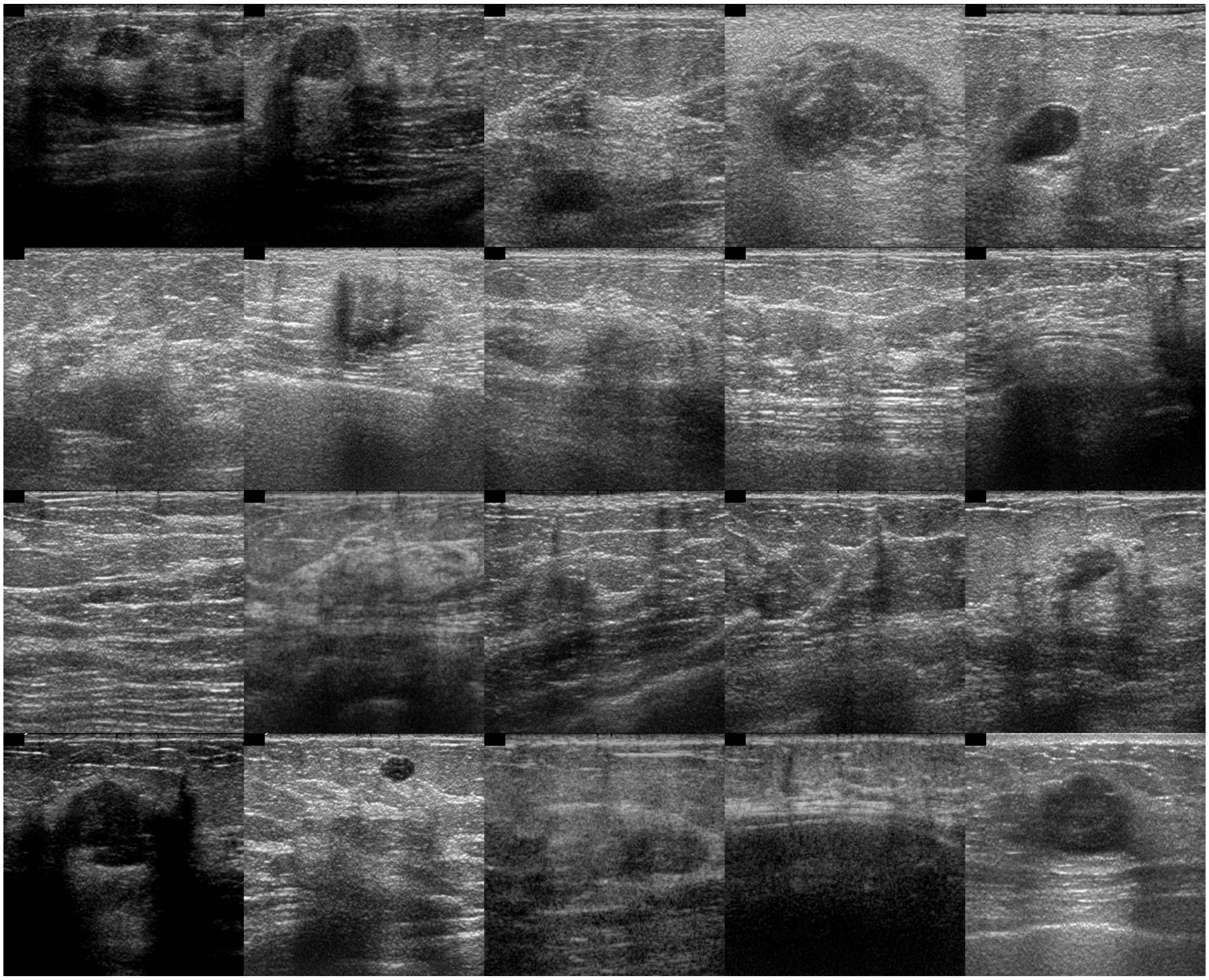


Figure 8: Ultrasound (US) breast scans from twenty volume scans of patients undergoing chemotherapy.

about 4cm^3 (voxel dimensions $0.1\text{mm}^2 \times 0.5\text{mm}$) 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 $\pm 16°$.

6.2 Feature Space

We have experimented with a large number of vector valued features including, Meyer 2D wavelet coefficients, 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):

$$X_i = \sum_{j=1}^{p} a_{ij} S_j. \tag{24}$$

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 (24) 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 extracted 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 [27] `FastICA` ICA code (available from [28]) which uses a fixed-point algorithm to perform maximum likelihood estimation of the basis elements in the ICA data model (24). Note that no pruning is performed on the ICA basis vectors. The 64D ICA is a full decomposition of the 8×8 patch of image. 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.

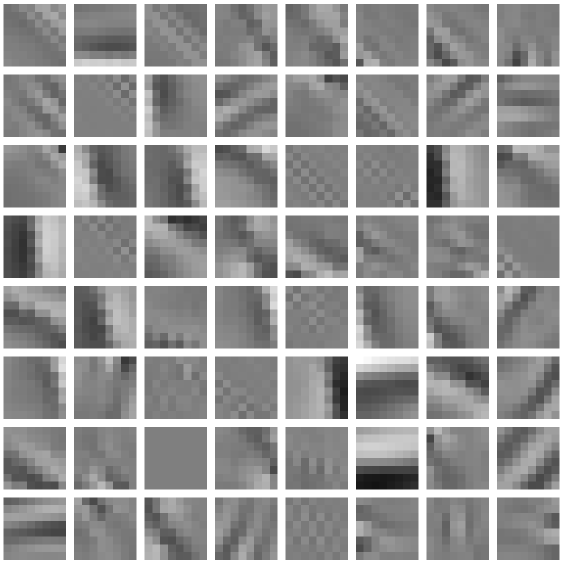


Figure 9: 8×8 ICA basis set obtained from training on randomly selected 8×8 blocks in the training database of breast scans.

Recently, Kybic [26] used the kNN graph to estimate MI by randomly grouping higher dimensional feature vectors. Divergence was calculated as the mean divergence over m such groupings of n points each. In our approach, all extracted feature vectors are used to estimate divergence. In experiments where feature vectors were partitioned (e.g. using k-Means clustering) before building the NN tree over the centroids of these partitions, we noticed a drop in registration accuracy. Kybic reports that divergence estimation bias decreased for $m > 50$ and registration error was lower than histogram estimates of divergence.

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 separation 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-separated 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.

Figure 11 shows average objective function plots for the registration experiment discussed above. Thirty different noise realizations were added to the fifteen test images at every angle of rotational deformation to give $N = 400$ different images for calculation of the matching functions. In the figure, each graph plots the sample mean, $\hat{\mu}_\theta$, calculated over the N measurements at each angle, θ. The standard deviation of $\hat{\mu}_\theta$, also called the standard error of the measurements, is given by $\sigma_{M_\theta} = \sigma_\theta/\sqrt{N}$ for $\theta \in \{-16^\circ, \ldots, +16^\circ\}$, where σ_θ is the standard deviation of the N measurements made at each rotational deformation. To normalize the images it is important to discount for the relative scaling between the matching functions. Hence, $\hat{\mu}_\theta$ of each matching function is normalized such that $\max(\sigma_{M_\theta})$ is unity. This restricts arbitrary scaling and also discounts for any scaling inherent in the computation of the matching function. In each row, the extent on the search space is identical. This facilitates comparison of two divergence estimates and also allows for comparison of a particular divergence as noise increases. It can readily be seen from the trends that at low

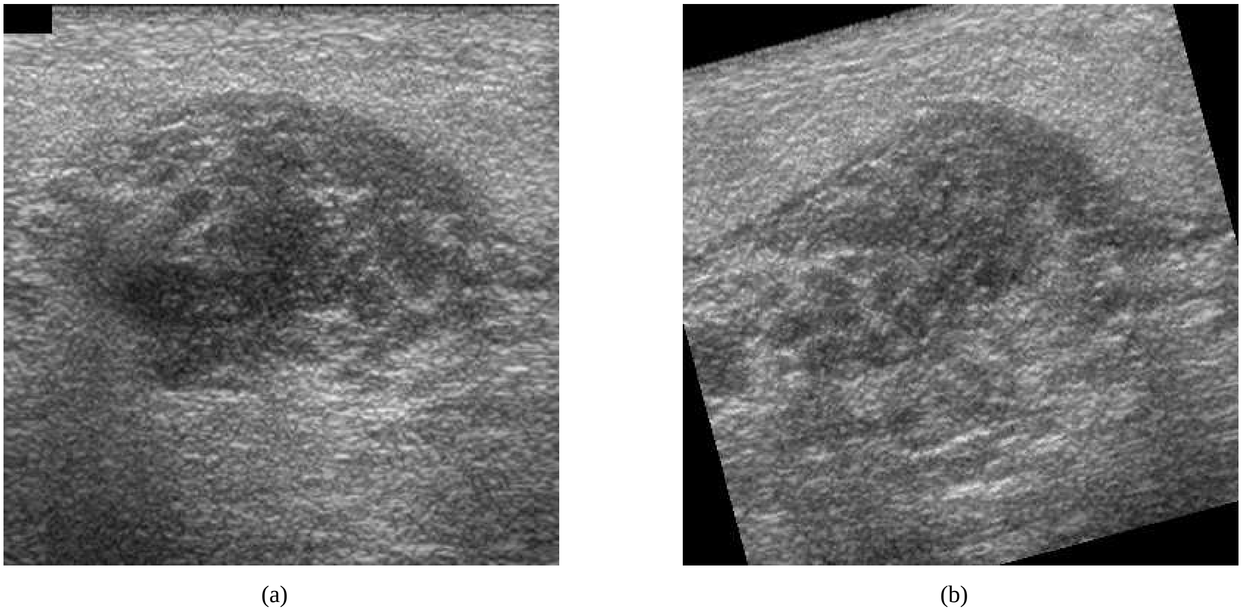


Figure 10: UL Images of the breast separated and rotationally deformed. (a) Cross-sectional image through center of tumor. (b) Rotated cross-sectional image acquired at a distance 5mm away from Image in (a).

levels of noise, all feature based estimates have sharper peaks than the Shannon MI estimate using pixel features. Further, as noise increases some divergence estimates, notably α GA and αMI divergence between the ICA features of the images, maintain sensitivity to rotational 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. 12 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 re-parameterization, 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 implementation) is lower than the MST-based methods (off-the-shelf Kruskal algorithm).

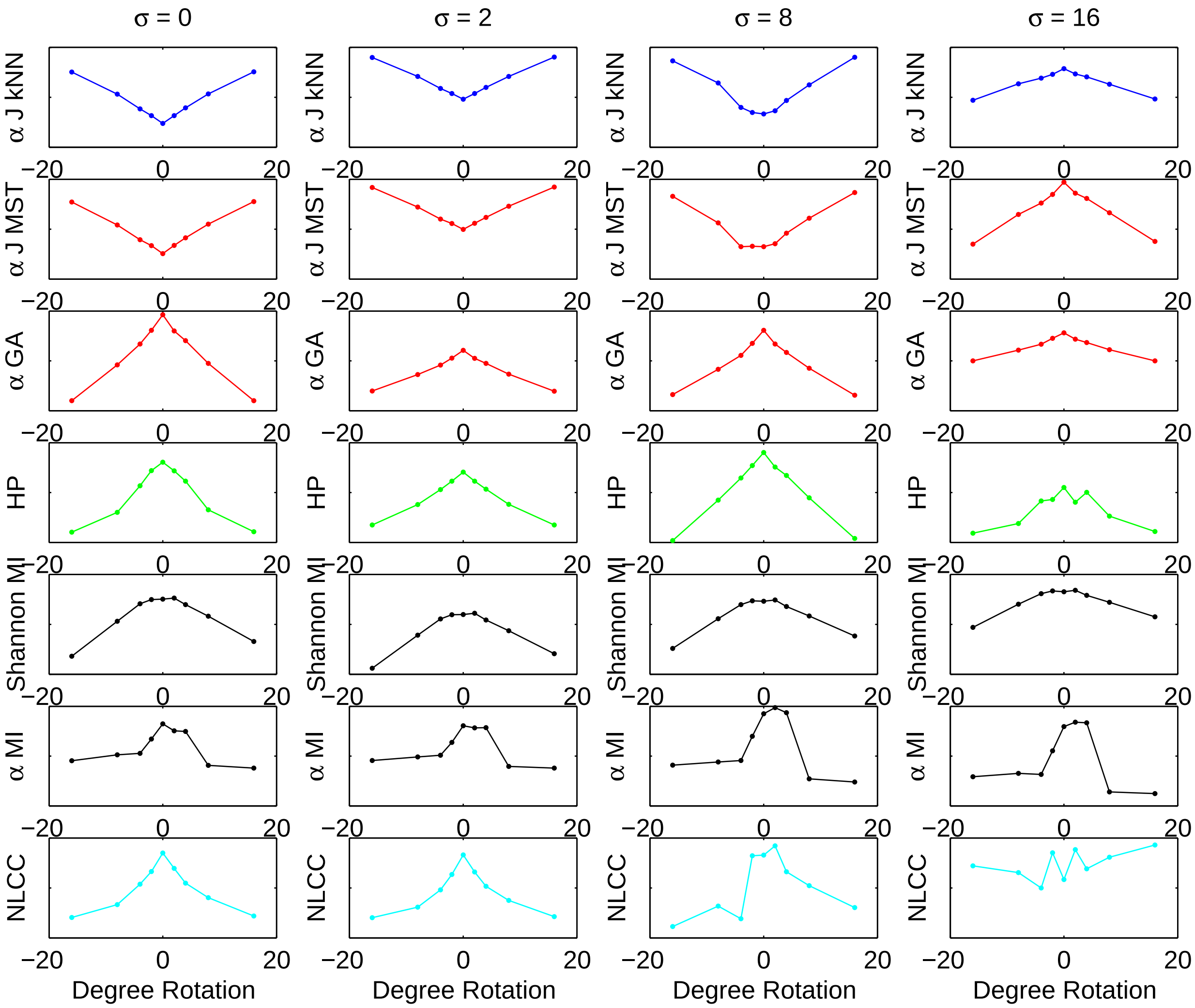


Figure 11: Normalized average profiles of image matching criteria for registration of UL breast images taken from two slices of the image volume database under decreasing SNR. All plots are normalized with respect to the maximum variance in the sampled observations.(row 1) kNN-based estimate of α-Jensen difference divergence between ICA features of the two images, (row 2) MST-based estimate of α-Jensen difference divergence between ICA features of the two images, (row 3) NN estimate of α Geometric-Arithmetic mean affinity between ICA features, (row 4) MST based estimate of Henze-Penrose affinity between ICA features, (row 5) Shannon Mutual Information estimated using pixel feature histogram method, (row 6) α Mutual Information estimated using NN graphs on ICA features and lastly, (row 7) NN estimate of the Non-linear correlation coefficient between the ICA feature vectors. Columns represent objective function under increasing additive noise. Column 1-4 represent additive truncated Gaussian noise of standard deviation, $\sigma =$ 0, 2, 8 and 16. Rotational deformations were confined to ± 16 degrees.

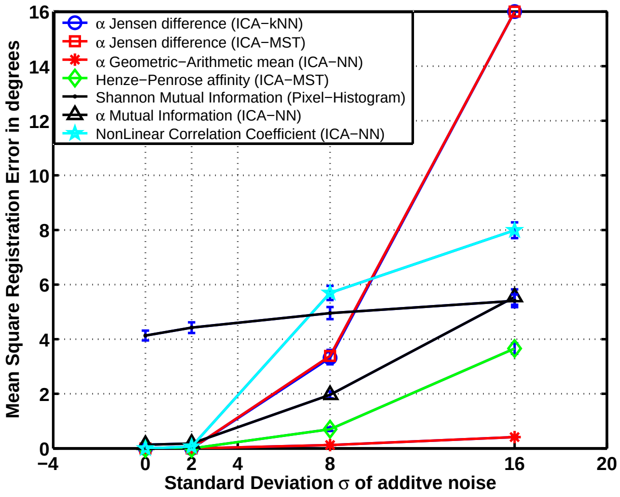


Figure 12: Rotational root mean squared error obtained from registration of ultrasound breast images using six different image similarity/dissimilarity criteria. Standard error bars are as indicated. These plots were obtained by averaging 15 cases, each with 250 Monte Carlo trials adding noise to the images prior to registration, corresponding to a total of 3750 registration experiments.

7 Simultaneous multi-image registration

Multi-image registration deals with the problem of registering three or more images simultaneously. In breast cancer therapy patient progress is monitored by periodic UL scans of the breast. Radiologists often register breast images of a patient collected at periodic intervals to monitor tumor growth or recession. One approach is to sequentially register pairs of images from time A to time B, time B to time C and so on. Besides being cumbersome and expensive, this process may lead to the accumulation of registration errors. A less expensive solution that may be able to avoid error accumulation is to register all the sequential scans (A,B,C,...) simultaneously. This section demonstrates the utility of entropic graph methods to simultaneously register three or more images.

7.1 Divergence estimation for multi-image registration

Evaluation of divergence for multiple images is straightforward. The α-MI between d-dimensional features $\{\mathcal{X}_i\}_{i=1}^N$, $\{\mathcal{O}_i\}_{i=1}^N$, $\{\mathcal{Y}_i\}_{i=1}^N$ extracted from three images, I_1, I_2, I_3, respectively is an extension of Equation

17 as follows:

$$\widehat{\alpha MI} = \frac{1}{\alpha - 1} \log \frac{1}{n^\alpha} \sum_{i=1}^{n} \left(\frac{e_i(x \times o \times y)}{\sqrt{e_i(x) e_i(o) e_i(y)}} \right)^{3\gamma}, \tag{25}$$

where $e_i(x \times o \times y)$ is the distance from the point $z_i = [x_i, o_i, y_i] \in \mathbb{R}^{3d}$ to its nearest neighbor in $\{Z_j\}_{j \neq i}$ and $e_i(x)$ $(e_i(o))$ $(e_i(y))$ is the distance from the point $x_i \in \mathbb{R}^d$, $(o_i \in \mathbb{R}^d)$, $(y_i \in \mathbb{R}^d)$ to its nearest neighbor in $\{X_j\}_{j \neq i}(\{\mathcal{O}_j\}_{j \neq i})\{Y_j\}_{j \neq i}$ respectively.

Similarly, building on Equation 15 α-GA can be estimated between one reference and two target images as follows:

$$\begin{aligned} \widehat{\alpha D_{GA}} &= \frac{1}{\alpha - 1} \log \frac{1}{3n} \sum_{i=1}^{3n} \min\{r_j\}_{j=1}^{3} \\ r_1 &= \min\left\{ \left(\frac{e_i(o)}{e_i(x)} \right)^{\gamma/2}, \left(\frac{e_i(x)}{e_i(o)} \right)^{\gamma/2} \right\}, \\ r_2 &= \min\left\{ \left(\frac{e_i(x)}{e_i(y)} \right)^{\gamma/2}, \left(\frac{e_i(y)}{e_i(x)} \right)^{\gamma/2} \right\}, \\ r_3 &= \min\left\{ \left(\frac{e_i(y)}{e_i(o)} \right)^{\gamma/2}, \left(\frac{e_i(o)}{e_i(y)} \right)^{\gamma/2} \right\}, \end{aligned} \tag{26}$$

where $e_i(x)$, $e_i(o)$ and $e_i(y)$ are the distances from a point $z_i \in \{\{x_i\}^i, \{o_i\}^i, \{y_i\}^i\} \in \mathbb{R}^d$ to its nearest neighbor in $\{\mathcal{X}_i\}_i$, $\{\mathcal{O}_i\}_i$ and $\{\mathcal{Y}_i\}_i$, respectively. Here, as above $\alpha = (d - \gamma)/d$.

Shannon MI can be estimated using pixel features by extending Equation 8 to histogram estimates of the joint pdf in three dimensional space as follows:

$$\widehat{\alpha MI} \overset{\text{def}}{=} \frac{1}{\alpha - 1} \log \sum_{x,o,y=0}^{255} \hat{f}_{0,1}^{\alpha}(x, o, y) \left(\hat{f}_x(x) \hat{f}_o(o) \hat{f}_y(y) \right)^{1-\alpha}. \tag{27}$$

In (27) we assume 8-bit gray level, $\hat{f}_{x,o,y}$ denotes the joint intensity level "coincidence histogram"

$$\hat{f}_{x,o,y}(x, o, y) = \frac{1}{MN} \sum_{k=1}^{MN} I_{x_k, o_k y_k}(x, o, y), \tag{28}$$

and $I_{x_k, o_k y_k}(x, o, y)$ is the indicator function equal to one when $(x_k, o_k, y_k) = (x, o, y)$ and equal to zero otherwise.

Equation 28 requires building a histogram in the three dimensional joint space of the three images. Generalizing to N images, it can easily be seen that a N-dimensional histogram would be required to estimate Shannon MI using the histogram plug-in method. As discussed earlier, the curse of dimensionality restricts the estimation of Shannon MI in higher dimensions. On comparison with Equations 25 and 27 it is seen that estimation of α-MI and α-GA do not suffer from this curse-of-dimensionality since the complexity of the kNN graph grows only linearly in the dimension.

In the following section, the performance of entropic graph based divergence estimates of α-MI and α-GA is compared with traditional histogram estimation techniques of Shannon MI.

7.2 Quantitative performance evaluation in multi-image registration

The methods used to evaluate performance of divergence estimates for the two-image case are extended to three images. The database of UL images is divided, as before, into training and testing sets. 64D ICA are estimated on the training set and used as features for registration. Test images are extracted from each volumetric scan in the test dataset. A ±5mm depth directional distance separates the reference image I_{ref} from the two target images I_{tar_1} and I_{tar_2}. ICA basis coefficient features are extracted from the reference and target images using the standard sub-block projection technique, as before. Registration performance is evaluated over rotational deformation within the range $\pm 16^\circ$. Figure 13 shows an example registration scenario where the reference images is shown to be sandwiched between two target images that are rotated.

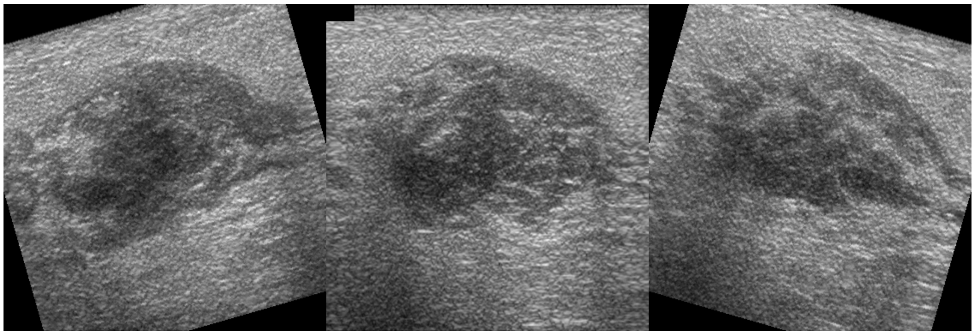


Figure 13: Multi-image registration scenario illustrated using three UL images of the breast where the reference image is sandwiched between two target images that are rotated $\pm 16^\circ$ respectively.

In Figure 14 shows the registration performance of the 16 test image sets. Mis-registration error is measured as the sum of mean-squared misregistration errors along each of the target images, and can hence vary from 0° to 32°. The SNR in all the images is progressively decreased by adding truncated uncorrelated Gaussian noise. Mean misregistration error is obtained by Monte-Carlo simulations over 30 different noise realizations on each of the 16 image. Thus, every point in the graph is the mean error over 480 measurements. Standard error bars are as shown.

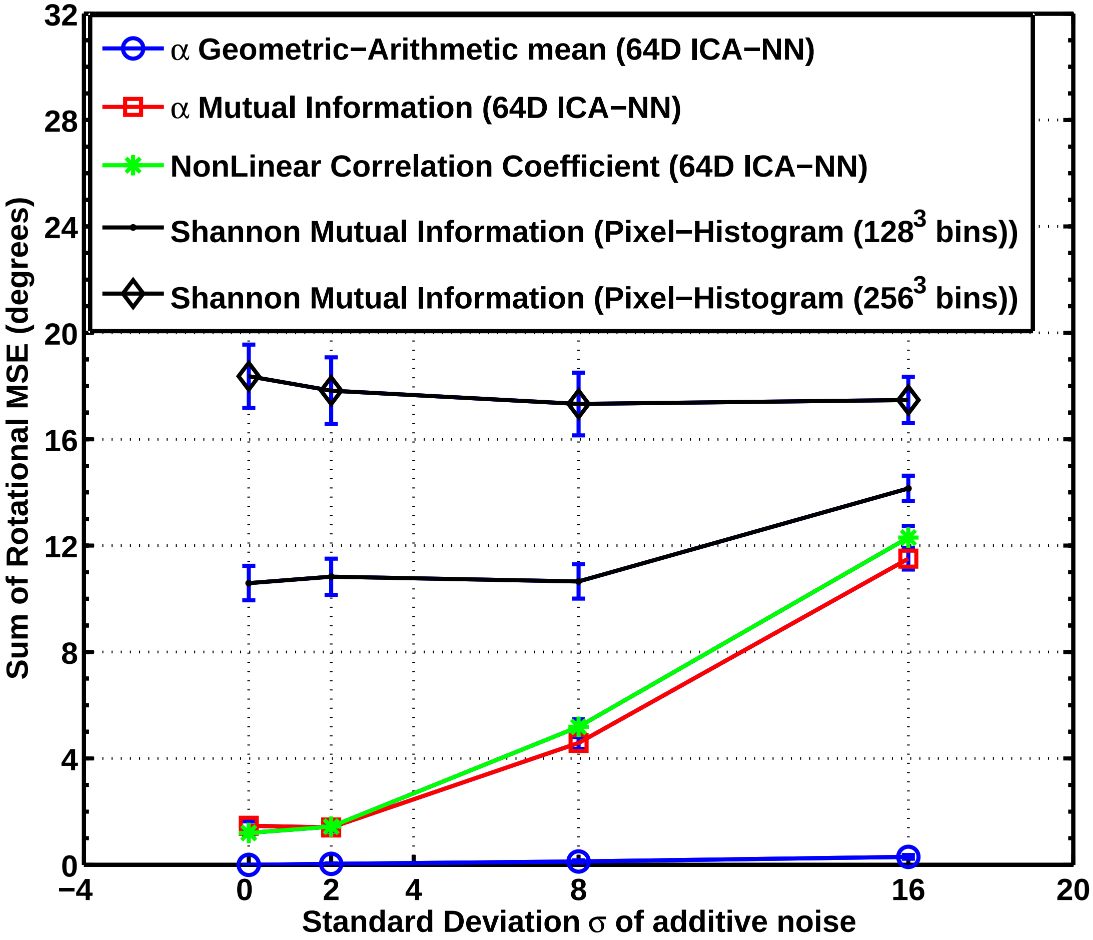


Figure 14: Multi-image registration scenario illustrated using three UL images of the breast where the reference image is sandwiched between two target images that are rotated ±16° respectively.

8 Image Clustering

Non-linear transformations often creep into the image acquisition process; common sources being lens distortion in cameras, changes in light patterns or, sensor specific nonlinearities such as magnetic field inhomogeneities in magnetic resonance imaging. Since the non-linear transformations cannot be recovered by a linear measure of similarity such as the linear correlation coefficient, in such situations, the use of a measure invariant to non-linear transformations could be justified as being more robust approach. While the mutual information measure, as used by Viola and Wells [1], is invariant to non-linear transformations it is a joint statistic that requires one-to-one feature correspondence and is difficult to calculate directly in higher-dimensional spaces. The NLCC that we introduced in section 5 does not suffer from these drawbacks since it is not a joint statistic and can also be reliably calculated in higher-dimensional spaces using the graph-based methods we describe.

In this section, we attempt to use this invariance property of NLCC for an image clustering application. In this preliminary example we demonstrated clustering of images sampled from the Corel image database. 15 labeled images were randomly picked and resized to 100×100 pixels using bilinear interpolation. Six different non-linear transformation functions, including quadratic, cubic, parabolic, sigmoid, inverse sigmoid and reverse video were applied to the images in the intensity space.

Here is a quick mathematical description of the non-linear transformations. Also see Figure 15 for a graphical illustration of the transform. Let x correspond to the set of intensity features extracted from the original image. Let y correspond to the set of intensity features generated by applying a non-linear transformation $T(x)$.

Quadratic transformation

$$y = T(x) = a * x^2 \tag{29}$$

Cubic Transformation

$$y = T(x) = a * x^3 \tag{30}$$

Parabolic Transformation

$$y = T(x) = a(x - x_c)^2 + y_c \tag{31}$$

Third-Order Polynomial Transformation

$$y = T(x) = a * x^2 + b * x^3 \tag{32}$$

Sigmoidal Transformation

$$y = T(x) = \frac{1}{1 + \exp -a * (x - b)} \tag{33}$$

Inverse-Sigmoid Transformation

$$y = T(x) = \frac{-1}{a} * \log\left(\frac{1 - x}{x}\right) + b \tag{34}$$

Reverse Video Transformation

$$y = T(x) = \max(x) - x \tag{35}$$

Images of the 15 objects used in this clustering study were transformed non-linearly using the formulations described above. Further, reverse-video versions of each image were also transformed and added to the dataset. Finally, using different values of the parameters a and b, every image in the dataset has 21 additional transformed counterparts to create a dataset of 330 unique images. The LCC and NLCC were then estimated between all images of the databased picked 2 at a time. There are $\binom{330}{2}/2$ such combinations. The linear and non-linear CC were then calculated for all such image pairs. To visualize the resultant cloud of relative positions of these images where distance is measured using the similarity measure, we project them onto a 2D space using a variant of the multidimensional scaling algorithm as used in the Pajek [29] software package. The relative estimates provided by MDS algorithms are accurate up to a rotation of the co-ordinate positions of the vertices. The resultant mappings can be seen in Figures 16 and 17. The performance of the clustering result is measured using a clustering figure-of-merit called the Dunn's validity index [30] defined as:

$$D_{n_c} = \min_{i=1,\ldots,n_c} \left\{ \min_{j=i,\ldots,n_c} \left(\frac{d(c_i, c_j)}{\max_{k=1,\ldots,n_c} diam(c_k)} \right) \right\}, \tag{36}$$

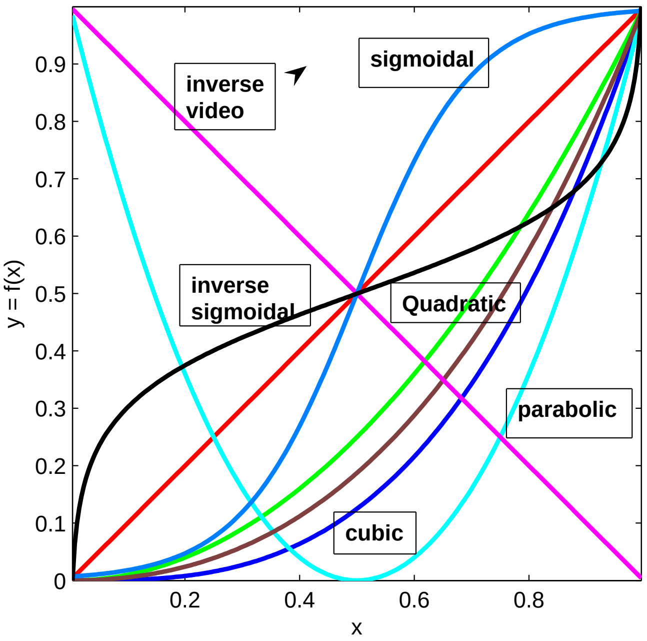


Figure 15: Non-linear transformations applied to images from the Corel database.

where $d(c_i, c_j)$ is the dissimilarity function between two clusters c_i and c_j and is defined as $d(c_i, c_j) = \min(x \in c_i, y \in c_j)d(x, y)$ and $diam(c_i)$ is the diameter of the cluster representing its dispersion and given as $diam(c_i) = \max(x, y \in c_i)d(x, y)$. Dunn's technique is well-suited to illustrate clustering performance since it attempts to identify clusters that are compact and well separated. In this experiment the number of classes are known apriori (15 image classes) and the validity index is used to measure the performance of the clustering algorithm. A higher value of D_{n_c} thus implies that the algorithm can cluster the data into 15 partitions with better separation between classes and more compactness within each class.

Figures 16 and 17 above show clustering performance of the LCC and NLCC respectively. The vertices represent the images in the lower dimensional space. The bi-directional links between images each have an associated weight w_{ij}^{LCC} and $w_{i,j}^{NLCC}$, where i and j index over images and the super-script signifies the distance measure of the link. By thresholding on w we can visualize only the strong links. Thus absent links imply that the link weights were low and the images were not perceived to be similar. In the first figure we can see that the LCC has a highly disperse cluster with a great amount of inter-mingling between

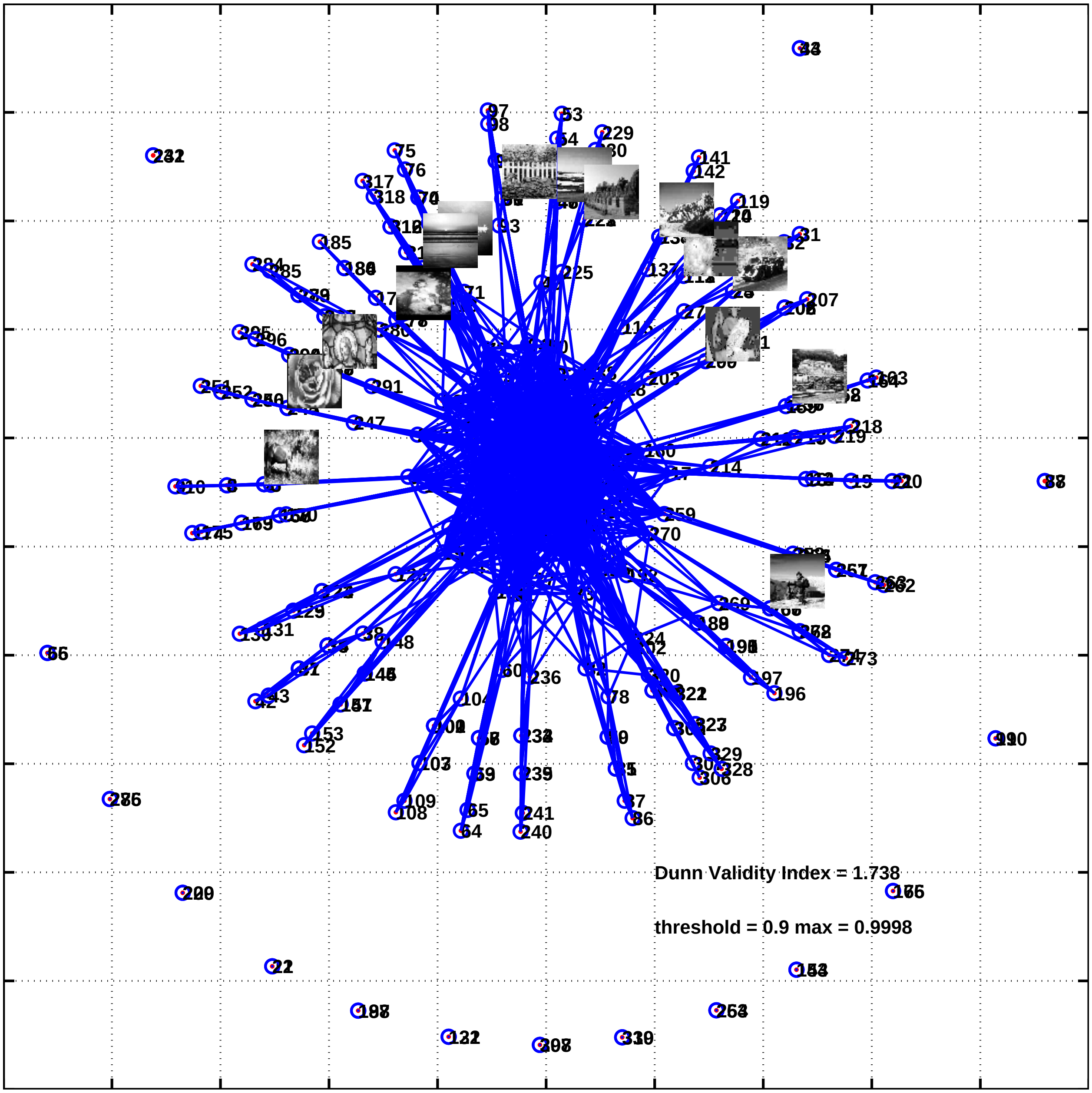


Figure 16: Demonstration of image clustering using the linear correlation coefficient. Intensity images of 15 objects were each transformed using a non-linear function. Using the CC as a similarity function the images were projected onto a 2D scale using a MDS algorithm [29]. The nodes of the graph represent images while the edges represent similarity between images. For clarity, only edge weight greater than a particular threshold (0.9) of the CC are shown.

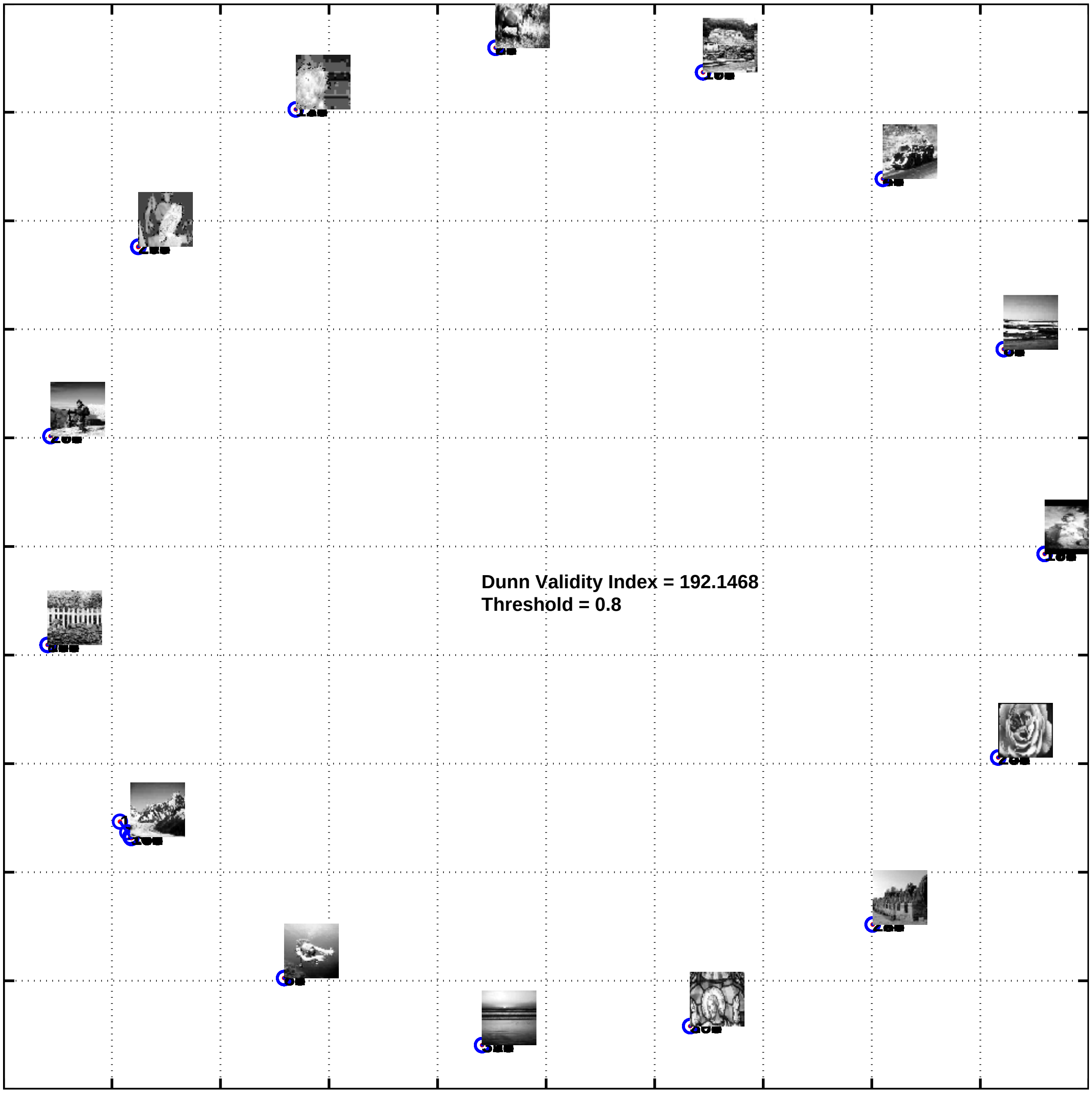


Figure 17: Demonstration of image clustering using the non-linear correlation coefficient. Intensity images of 15 objects were each transformed using a non-linear function. Using the NLCC as a similarity function the images were projected onto a 2D scale using a MDS algorithm [29]. The nodes of the graph represent images while the edges represent similarity between images. For clarity, only edge weight greater than a particular threshold (0.8) of the NLCC are shown.

classes. This is also reflected in the lower Dunn's validity index for the clustering of 15 classes. The NLCC, however, shows tight clustering and scores much higher on the Dunn's validity index. Earlier, in section 5 we saw that the NLCC is invariant to non-linear transformations of the underlying image intensity features. By definition, the linear CC is invariant only to linear transformations of image intensity features. Hence the clustering of objects under the influence of non-linear transformations on the feature space is much better behaved when the NLCC is used as a dissimilarity measure.

9 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) and a new measure of non-linear correlation. As compared to previous work in which estimated Jensen differences were used for registration, these divergence measures have the advantage of invariance to re-parameterization 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.

The new kNN estimators of the αMI and αGA have the advantage of invariance to re-parameterization of the feature space. While convergence results for the kNN divergence estimators were not provided there is circumstantial theoretical evidence that they do converge. Furthermore, the numerical evaluations show that these divergence estimators outperform previous approaches to image registration. This paper 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. 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.

The performance of αGA and Henze-Penrose have exceeded those of other divergence measures. We hypothesize that the combination of low-dimensional complexity through the exclusive use of marginal spaces and invariance to transformations has led to superior noise performance and robustness in these measures as compared to others. Unlike the other metrics, the αJensen difference is not invariant to re-parameterization, which explains its relatively poor performance for large RMS noise.

An exciting extension of this work is in registration of multiple images. Multiple images could be registered simultaneously to form an atlas. Multi-image registration could also be used to simultaneously register time-sampled imagery such as those acquired during periodic UL examination for cancer detection and management.

Lastly, we have introduced a new measure of non-linear correlation. Based on an extension of αGA and αMI measures, the NLCC is estimated using the kNN graph to adaptively partition space based on local density of samples. We contrast its performance to the linear CC and find this measure to be robust in the face of non-linear intensity transformations.

References

[1] P. Viola and W. M. Wells III, “Alignment by maximization of mutual information,” in *Proceedings of IEEE International Conference on Computer Vision*, Los Alamitos, CA, Jun. 1995, pp. 16–23.

[2] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, “Multimodality image registration by maximization of mutual information,” *IEEE Transactions on Medical Imaging*, vol. 16, no. 2, pp. 187–198, Apr. 1997.

[3] S. Kullback and R.A. Leibler, “On information and sufficiency,” *Ann. Math. Statist.*, vol. 22, pp. 79–86, 1951.

[4] B. Ma, *Parametric and non-parametric approaches for multisensor data fusion*, Ph.D. thesis, University of Michigan, Ann Arbor, MI 48109-2122, 2001, `www.eecs.umich.edu/˜hero/research.html`.

[5] A.O. Hero and O. Michel, "Asymptotic theory of greedy approximations to minimal k-point random graphs," *IEEE Trans. on Inform. Theory*, vol. IT-45, no. 6, pp. 1921–1939, Sept. 1999.

[6] 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`.

[7] A.O. Hero, J. Costa, and B. Ma, "Convergence rates of minimal graphs with random vertices," *IEEE Trans. on Inform. Theory*, vol. submitted, 2002, `www.eecs.umich.edu/˜hero/det_est.html`.

[8] O. Vasicek, "A test for normality based on sample entropy," *J. Royal Statistical Society, Ser. B*, vol. 38, pp. 54–59, 1976.

[9] J. Beirlant, E. J. Dudewicz, L. Györfi, and E.C. van der Meulen, "Nonparametric entropy estimation: an overview," *Intern. J. Math. Stat. Sci.*, vol. 6, no. 1, pp. 17–39, june 1997.

[10] H. Neemuchwala, A. O. Hero, and P. Carson, "Image registration using entropy measures and entropic graphs," *European Journal of Signal Processing, Special issue on content based information retrieval*, vol. 85, no. 2, pp. 277–296, Feb. 2005.

[11] H. Neemuchwala and A. O. Hero, *Multi-sensor Image Fusion and its applications*, chapter Entropic Graphs for Registration, Marcel-Dekker and CRC Press, 2005.

[12] H. Neemuchwala, *Entropic graphs for image registration*, Ph.D. thesis, University of Michigan, Ann Arbor, MI, 2005.

[13] I. J. Taneja, "New developments in generalized information measures," *Advances in Imaging and Electron Physics*, vol. 91, pp. 37–135, 1995.

[14] 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.

[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] M. Basseville, “Distance measures for signal processing and pattern recognition,” *Signal Processing*, vol. 18, pp. 349–369, 1989.

[18] 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.

[19] Y. He, A. Ben-Hamza, and H. Krim, “An information divergence measure for ISAR image registration,” *Signal Processing*, Submitted, 2001.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] H. Neemuchwala and A.O. Hero, “Image registration in higher dimensional feature space,” in *Proc. of the SPIE Conference on Electronic Imaging*, San Jose, CA, Jan. 2005.

[26] J. Kybic, “High-dimensional mutual information estimation for image registration,” in *Proc. of the IEEE Int. Conf. on Image Processing*, 2004, pp. 1779–1782.

[27] A. Hyvärinen and E. Oja, “Independent component analysis: algorithms and applications,” *Neural Networks*, vol. 13, no. 4-5, pp. 411–430, 1999.

[28] A. Hyvärinen, “Fast ICA Code,” `www.cis.hut.fi/projects/ica/fastica/`.

[29] Batagelj V. and Mrvar A., *Graph Drawing Software*, chapter Pajek - Analysis and visualization of large networks, Springer, 2003.

[30] J.C. Dunn, “Well seperated clusters and optimal fuzzy partitions,” *Journal of Cybernetics*, vol. 4, pp. 95–104, 1974.