Entropic graphs for registration

Huzefa Neemuchwala and Alfred Hero

Abstract

In many applications, fusion of images acquired via two or more sensors requires image alignment to an identical pose, a process called image registration. Image registration methods select a sequence of transformations to maximize an image similarity measure. Recently a new class of entropic-graph similarity measures was introduced for image registration, feature clustering and classification. This chapter provides an overview of entropic graphs in image registration and demonstrates their performance advantages relative to conventional similarity measures. In this chapter we introduce : techniques to extend image registration to higher dimension feature spaces using Rényi's generalized α-entropy. The α-entropy is estimated directly through continuous quasi additive power weighted graphs such as the minimal spanning tree (MST) and k-Nearest Neighbor graph (kNN). Entropic graph methods are further used to approximate similarity measures like the α mutual information, α-Jensen divergence, Henze-Penrose affinity and Geometric-Arithmetic mean affinity. These similarity measures offer robust registration benefits in a multisensor environment. Higher dimensional features used for this work include basis functions like multidimensional wavelets and independent component analysis (ICA). Registration is performed on a database of multisensor satellite images. Lastly, we demonstrate the sensitivity of our approach by matching local image regions in a multimodal medical imaging example.

This manuscript will appear as a chapter in ”Multi-sensor image fusion and its applications,” Eds. R. S. Blum and Z. Liu, Marcel-Dekker, Inc 2004. This work was supported in part by NIH grant 1P01CA87634 and by ARO contract DAAD19-02-1-0262.

At University of Michigan Ann Arbor, Huzefa Neemuchwala is with the Departments of Biomedical Engineering, Electrical Engineering and Computer Science and Radiology. Alfred Hero is with Electrical Engineering and Computer Science, Biomedical Engineering and Statistics.

Entropic graphs for registration

I. INTRODUCTION

Given 2D or 3D images gathered via multiple sensors located at different positions, the multi-sensor image registration problem is to align the images so that have an identical pose in a common coordinate system (Figure 1). Image registration is becoming a challenging multi-sensor fusion problem due to the increased diversity of sensors capable of imaging objects and their intrinsic properties. In medical imaging, cross sectional anatomic images are routinely acquired by magnetic induction (Magnetic Resonance Imaging, MRI), absorption of accelerated energized photons (X-Ray Computed Tomography, CT) and ultra high frequency sound (Ultrasound) waves. Artifacts such as motion, occlusion, specular refraction, noise, inhomogeneities in the object and imperfections in the transducer compound the difficulty of image registration. Cost and other physical considerations canm constrain the spatial or spectral resolution and the signal to noise ratio (SNR). Despite these hindrances, image registration is now commonplace in medical imaging, satellite imaging and stereo vision. Image registration also finds widespread usage in other pattern recognition and computer vision applications such as image segmentation, tracking and motion compensation. A comprehensive survey of the image registration problem, its applications, and implementable algorithms can be found in [52], [51]. Image fusion is defined as task of extracting co-occurring information from multisensor images. Image registration is hence a precursor to fusion. Image fusion finds several applications in medical imaging where it is used to fuse anatomic and metabolic information [72], [53], [24], and build global anatomical atlases [80].

The three chief components of an effective image registration system (Figure 2) are: (1) definition of features that discriminate between different image poses; (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. In this chapter, we shall be principally concerned with the first two components of the system. In a departure from conventional pixel-intensity features, we present techniques that use higher dimensional features extracted from images. We adapt traditional pixel matching methods that rely on entropy estimates to include higher dimensional features. We propose a general class of information

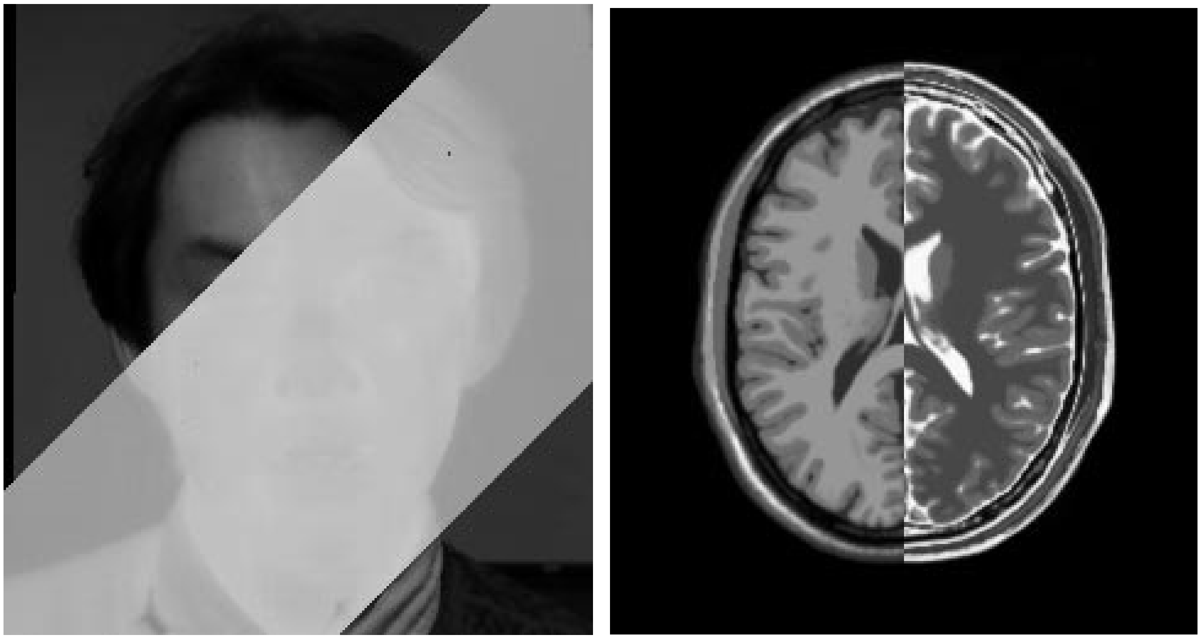


Fig. 1. Image fusion: (left) Co-registered images of the face acquired via visible light and longwave senors. (right) Registered brain images acquired by time-weighted responses . Face and brain images courtesy ([23]) and ([16]) respectively.

theoretic feature similarity measures that are based on entropy and divergence and can be empirically estimated using entropic graphs, such as the minimal spanning tree (MST) or k-Nearest Neighbor (kNN) graph, and do not require density estimation or histograms.

Traditional approaches to image registration have included single pixel gray level features and correlation type matching functions. The correlation coefficient is a poor choice for the matching function in multi-sensor fusion problems. Multi-sensor images typically have intensity maps that are unique to the sensors used to acquire them and a direct linear correlation between intensity maps may not exist (Fig 3). Several other matching functions have been suggested in the literature [37], [42], [66]. Some of the most widespread techniques are: histogram matching [39]; texture matching [2]; intensity cross correlation [52]; optical flow matching [47]; kernel-based classification methods [17]; boosting classification methods [19], [44]; information divergence minimization [81], [77], [76], [29]; and mutual information (MI) maximization [84], [28], [53], [11]. The last two methods can be called ”entropic methods” since both use a matching criterion defined as a relative entropy between the feature distributions. The main advantage

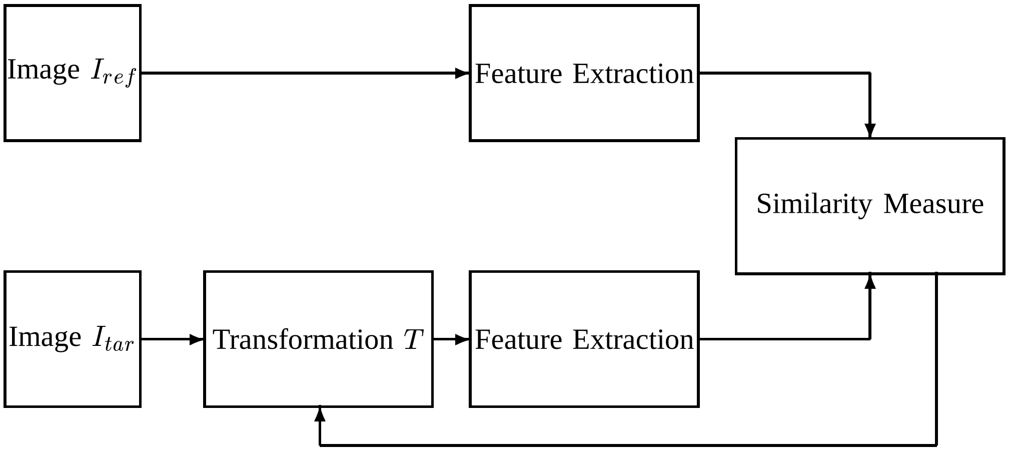


Fig. 2. Block diagram of an image registration system

of entropic methods is that they can capture non-linear relations between features in order to improve discrimination between poor and good image matches. When combined with a highly discriminatory feature set, and reliable prior information, entropic methods are very compelling and have been shown to be virtually unbeatable for some multimodality image registration applications [48], [53], [37]. However, due to the difficulty in estimating the relative entropy over high dimensional feature spaces, the application of entropic methods have been limited to one or two feature dimensions. The independent successes of relative entropy methods, e.g., MI image registration, and the use of high dimensional features, e.g., SVM's for handwriting recognition, suggest that an extension of entropic methods to high dimensions would be worthwhile. Encouraging initial studies on these methods have been conducted by these authors and can be found in [60], [58].

Here we describe several new techniques to extend methods of image registration to high dimensional feature spaces. Chief among the techniques is the introduction of entropic graphs to estimate a generalized α-entropy: Rényi's α-entropy. These entropic graph estimates can be computed via a host of combinatorial optimization methods including the MST and the k-Nearest neighbor graph (kNNG). The computation and storage complexity of the MST and kNNG-based estimates increase linearly in feature dimension as opposed to the exponential rates of histogram-based estimates of entropy. Furthermore, as will be shown,

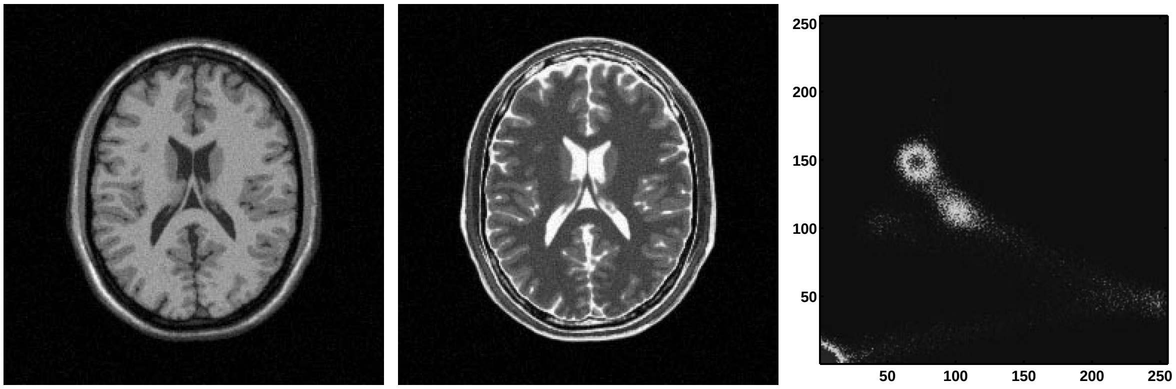


Fig. 3. MRI images of the brain, with additive noise. (left) T1 weighted I_1, (center) T2 weighted I_2. Images courtesy [16]. Although acquired by a single sensor, the time weighting renders different intensity maps to identical structures in the brain. (right) Joint gray-level pixel coincidence histogram is clustered and does not exhibit a linear correlation between intensities.

entropic graphs can also be used to estimate more general similarity measures. Specific examples include the α-mutual information (α-MI), α-Jensen difference divergence, the Henze-Penrose (HP) affinity, which is a multidimensional approximation to the Wald-Wolfowitz test [85], and the α-geometric-arithmetic (α-GA) mean divergence [79]. To our knowledge, the last two divergence measures have never been utilized in the context of image registration problems. We also explore variants of entropic graph methods that allow estimation with faster asymptotic convergence properties and reduced computational complexity.

The α-entropy of a multivariate distribution is a generalization of the better known Shannon entropy. Alfred Rényi introduced the α-entropy in a 1961 paper [71] and since then many important properties of α-entropy have been established [4]. From Rényi's α-entropy the Rényi α-divergence and the Rényi α-mutual information (α-MI) can be defined in a straightforward manner. For $\alpha = 1$ these quantities reduce to the standard (Shannon) entropy, (Kullback-Liebler) divergence, and (Shannon) MI, respectively. Another useful quantity that can be derived from the α-entropy is the α-Jensen difference, which is a generalization of the standard Jensen difference and has been used here in our extension of entropic pattern matching methods to high feature dimension. As we will show, this generalization allows us to define an image matching algorithm that benefits from a simple estimation procedure and an extra degree of freedom (α).

Some additional comments on relevant prior work by us and others is in order. Various forms of

α-entropy have been exploited by others for applications including: reconstruction and registration of interferometric synthetic aperture radar (ISAR) images [29], [26]; blind deconvolution [25]; and time-frequency analysis [3], [86]. Again, our innovation with respect to these works is the extension to high dimensional features via entropic graph estimation methods. On the other hand, the alpha-entropy approaches described here should not be confused with entropy-alpha classification in SAR processing [15] which has no relation whatsoever to our work. A tutorial introduction to the use of entropic graphs to estimate multivariate α-entropy and other entropy quantities was published by us in a recent survey article [35]. As introduced in [36] and studied in [35], [34] 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 minimal spanning tree 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 [83], [7].

We have developed experiments that allows the user to examine and compare our methods with other methods currently used for image fusion tasks. The applications presented in this chapter are primarily selected to illustrate the flexibility of our method, in terms of selecting high dimensional features. However, they help us compare and contrast multidimensional entropy estimation methods. In the first example we perform registration on images obtained via multi-band satellite sensors. Images acquired via these geostationary satellites serve in research related to heat dissipation from urban centers, climactic changes and other ecological projects. Thermal and visible light images captured for the Urban Heat Island [68] project form a part of the database used here. NASA's visible earth project [57] also provides images captured via different satellite sensors, and such multi-band images have been used here to provide a rich representative database of satellite images. Thermal and visible-light sensors image different bands in the electromagnetic spectrum and thus have different intensity maps, removing any possibility of using correlation-based registration methods.

As a second example we apply our methods to registering medical images of the human brain acquired under dual modality (T1,T2 weighted) magnetic resonance imaging. Simulated images of the brain under different time-echo responses to magnetic excitation are used. Different areas in the brain (neural tissue, fat and water) have distinct magnetic excitation properties. Hence, they express different levels of excitation when appropriately time-weighted. This example qualifies as a multisensor fusion example due to the disparate intensity maps generated by the imaging sequence, commonly referred to as the T1 and T2 time weighted MRI sequences. We demonstrate an image matching technique for MRI images sensitive to local perturbations in the image.

Higher dimensional features used for this work include those based on independent component analysis (ICA) and multidimensional wavelet image analysis. Local basis projection coefficients are implemented by projecting local 8 by 8 sub-images of the image onto the ICA basis for the local image matching example from medical imaging. Multi-resolution wavelet features are used for registration of satellite imagery. Local feature extraction via basis projection is a commonly used technique for image representation [74], [82]. Wavelet bases are commonly used for image registration as is evidenced in [87], [78], [43]. ICA features are somewhat less common but have been similarly applied by Olshausen, Hyvärinen and others [49], [41], [64]. The high dimensionality (= 64 for local basis projections) of these feature spaces precludes the application of standard entropy-based pattern matching methods and provides a good illustration of the power of our approach. The ability of the wavelet basis to capture spatial-frequency information in a hierarchical setting makes them an attractive choice for use in registration.

The paper is organized as follows: Section II introduces various entropy and α-entropy based similarity measures such as Rényi entropy and divergence, mutual information and α-Jensen difference divergence. Section III describes continuous Euclidean functionals such as the MST and the kNNG that asymptotic converge to the Rényi entropy. Section IV presents the Henze-Penrose test statistic as a divergence measure for image registration. Next, Section VI describes, in detail, the feature based matching techniques used in this work, different types of features used and the advantages of using such methods. Computational considerations involved in constructing graphs are discussed in Section VII. Finally, Sections VIII and IX present the experiments we conducted to compare and contrast our methods with other registration algorithms.

II. Entropic Feature Similarity/Dissimilarity Measures

In this section we review entropy, relative entropy, and divergence as measures of dissimilarity between probability distributions. Let Y be a q-dimensional random vector and let $f(y)$ and $g(y)$ denote two possible densities for Y. Here Y will be a feature vector constructed from the reference image and the target image to be registered and f and g will be multidimensional feature densities. For example, information divergence methods of image retrieval [76], [21], [82] specify f as the estimated density of the reference image features and g as the estimated density of the target image features. When the features are discrete valued the densities f and g are interpreted as probability mass functions.

A. Rényi Entropy and Divergence

The basis for entropic methods of image fusion is a measure of dissimilarity between densities f and g. A very general entropic dissimilarity measure is the Rényi α-divergence, also called the Rényi α-relative entropy, between f and g of fractional order $\alpha \in (0,1)$ [71], [18], [4] :

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

When the density f is supported on a compact domain and g is uniform over this domain the 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 $-2\log\int\sqrt{f(z)g(z)}dz$ obtained when $\alpha = 1/2$, which is related to the Hellinger-Battacharya distance squared,

$$D_{Hellinger}(f\|g) = \int\left(\sqrt{f(z)}-\sqrt{g(z)}\right)^2 dz \tag{3}$$

$$= 2\left(1-\exp\left(\tfrac{1}{2}D_{\frac{1}{2}}(f\|g)\right)\right), \tag{4}$$

and the Kullback-Liebler (KL) divergence [46], obtained in the limit as $\alpha \to 1$,

$$\lim_{\alpha\to 1} D_\alpha(f\|g) = \int g(z)\log\frac{g(z)}{f(z)}dz. \tag{5}$$

B. Mutual Information and α-Mutual Information

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 introduced for gray scale image registration [84] and has since been applied to a variety of image matching problems [28], [48], [53], [69]. 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 the k-th pixel location 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 $f_{0,1}(z_0, z_T) = f_0(z_0)f_1(z_T)$. This suggests using the α-divergence $D_\alpha(f_{0,1}(z_0, z_T)\|f_0(z_0)f_1(z_T))$ between $f_{0,1}(z_0, z_T)$ and $f_0(z_0)f_1(z_T)$ as a similarity measure. For $\alpha \in (0,1)$ we call this the α-mutual information (or α-MI) between Z_0 and Z_T and it has the form

$$\begin{aligned}\alpha MI &= D_\alpha(f_{0,1}(Z_0, Z_T) \parallel f_0(Z_0)f_1(Z_T)) && (6)\\ &= \frac{1}{\alpha-1}\log\int f_{0,1}^{\alpha}(z_0, z_T)f_0^{1-\alpha}(z_0)f_i^{1-\alpha}(z_T)dz_0dz_T. && (7)\end{aligned}$$

When $\alpha \to 1$ the α-MI converges to the standard (Shannon) MI

$$\text{MI} = \int f_{0,1}(z_0, z_T)\log\left(\frac{f_{0,1}(z_0, z_T)}{f_0(z_0)f_1(z_T)}\right)dz_0dz_T. \tag{8}$$

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 (8) between the reference and the transformed target. The MI is defined using features $(Z_0, Z_T) \in \{z_{0k}, z_{Tk}\}_{k=1}^{MN}$ equal to the discrete-valued intensity levels at common pixel locations (k, k) in the reference image and the rotated target image. We call this the “single pixel MI”. In [84], the authors empirically approximated the single pixel MI (8) by “histogram plug-in" estimates, which when extended to the α-MI gives the estimate

$$\widehat{\alpha MI} \stackrel{\text{def}}{=} \frac{1}{\alpha - 1} \log \sum_{z_0, z_T = 0}^{255} \hat{f}_{0,1}^{\alpha}(z_0, z_T) \left(\hat{f}_0(z_0) \hat{f}_1(z_T) \right)^{1-\alpha}. \tag{9}$$

In (9) we assume 8-bit gray level, $\hat{f}_{0,1}$ denotes the joint intensity level "coincidence histogram"

$$\hat{f}_{0,1}(z_0, z_T) = \frac{1}{MN} \sum_{k=1}^{MN} I_{z_{0k}, z_{Tk}}(z_0, z_T), \tag{10}$$

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. Other feature definitions have been proposed including gray level differences [11] and pixel pairs [73].

Figure 4 illustrates the MI alignment procedure through a multisensor remote sensing example. Aligned images acquired by visible and thermally sensitive satellite sensors, generate a joint gray level pixel coincidence histogram $f_{0,1}(z_0, z_1)$. Note, that the joint gray-level pixel coincidence histogram is not concentrated along the diagonal due to the multisensor acquisition of the images. When the thermal image is rotationally transformed, the corresponding joint gray-level pixel coincidence histogram $f_{0,1}(z_0, z_T)$ is dispersed, thus yielding a lower mutual information than before.

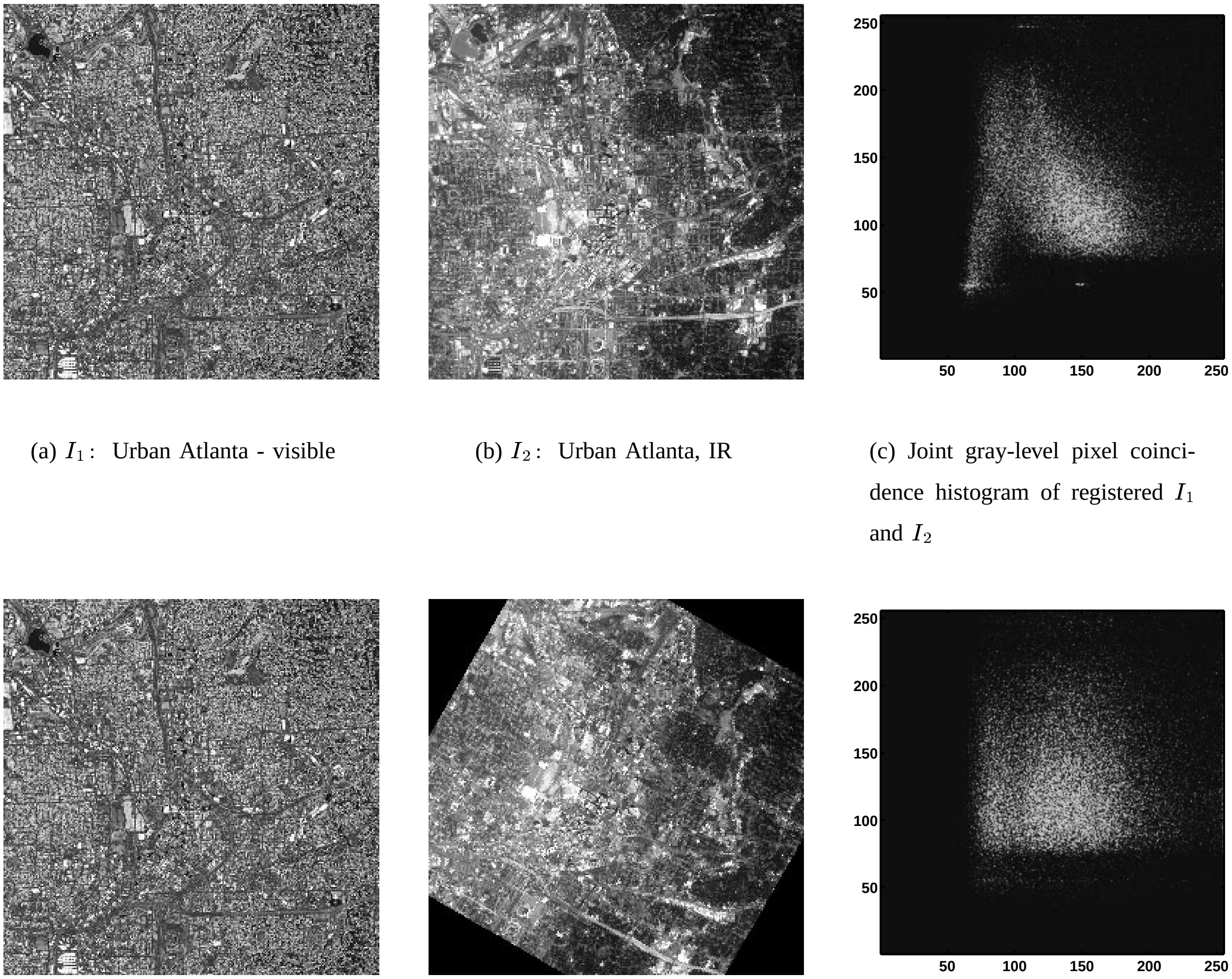
1) Relation of α-MI to Chernoff Bound: The α-MI (7) can be motivated as an appropriate registration function by large deviations theory through the Chernoff bound. Define the average probability of error $P_e(n)$ associated with a decision rule for deciding whether Z_T and Z_0 are independent (hypothesis H_0) or dependent (hypothesis H_1) random variables based on a set of i.i.d. samples $\{z_{0k}, z_{Tk}\}_{k=1}^{n}$, where $n = MN$. For any decision rule, this error probability has the representation:

$$P_e(n) = \beta(n) P(H_1) + \alpha(n) P(H_0), \tag{11}$$

where $\beta(n)$ and $\alpha(n)$ are the probabilities of Type II (say H_0 when H_1 true) and Type I (say H_1 when H_0 true) errors, respectively, of the decision rule and $P(H_1) = 1 - P(H_0)$ is the prior probability of H_1. When the decision rule is the optimal minimum probability of error test the Chernoff bound implies that [20]:

$$\lim_{n \to \infty} \frac{1}{n} \log P_e(n) = - \sup_{\alpha \in [0,1]} \left\{ (1 - \alpha) D_\alpha(f_{0,1}(z_0, z_T) \| f_0(z_0) f_1(z_T) \right\}. \tag{12}$$

Thus the mutual α-information gives the asymptotically optimal rate of exponential decay of the error probability for testing H_0 vs H_1 as a function of the number $n = MN$ of samples. In particular, this



(a) I_1: Urban Atlanta - visible

(b) I_2: Urban Atlanta, IR

(c) Joint gray-level pixel coincidence histogram of registered I_1 and I_2

(d) I_1

(e) $T(I_2)$

(f) Joint gray-level pixel coincidence histogram of I_1 and $T(I_2)$

Fig. 4. Mutual information based registration of multisensor, visible and thermal infrared, images of Atlanta acquired via satellite [68]. Top row (in-registration): (a) Visible light image I_1 (b) Thermal image I_2 (c) Joint gray-level pixel coincidence histogram $\hat{f}_{0,1}(z_0, z_1)$. Bottom row (out-of-registration): (d) Visible light image, unaltered I_1 (e) Rotationally transformed thermal image $T(I_2)$ (f) Joint gray-level pixel coincidence histogram shows wider dispersion $\hat{f}_{0,1}(z_0, z_T)$.

implies that the α-MI can be used to select optimal transformation T that maximizes the right side of (12). The appearance of the maximization over α implies the existence of an optimal parameter α ensuring the lowest possible registration error. When the optimal value α is not equal to 1 the MI criterion will be suboptimal in the sense of minimizing the asymptotic probability of error. For more discussion of the issue of optimal selection of α we refer the reader to [33].

C. α-Jensen Dissimilarity Measure

An alternative entropic dissimilarity measure between two distributions is the α-Jensen difference. This function was independently proposed by Ma [32] and He *et al* [29] for image registration problems. It was also used by Michel *et al* in [54] for characterizing complexity of time-frequency images. For two densities f and g the α-Jensen difference is defined as [4]

$$\Delta H_\alpha(p,f,g) = H_\alpha(pf+qg) - [pH_\alpha(f) + qH_\alpha(g)], \tag{13}$$

where $\alpha \in (0,1)$ and $p \in [0,1]$ and $q = 1-p$. As the α-entropy $H_\alpha(f)$ is strictly concave in f, Jensen's inequality implies that $\Delta H_\alpha(p,f,g) > 0$ when $f \neq g$ and $\Delta H_\alpha(p,f,g) = 0$ when $f = g$ (a.e.). Thus the α-Jensen difference is a bone fide measure of dissimilarity between f and g.

The α-Jensen difference can be applied as a surrogate optimization criterion in place of the α-divergence. One identifies $f = f_1(z_T)$ and $g = f_0(z_0)$ in (13). In this case an image match occurs when the α-Jensen difference is minimized over i. This is the approach taken by [29], [32] for image registration applications and discussed in more detail below.

D. α-Geometric-Arithmetic Mean Divergence

The α-geometric-arithmetic (α-GA) mean divergence [79] is another measure of dissimilarity between probability distributions. Given continuous distributions f and g, the α-GA :

$$\begin{aligned} \alpha D_{GA}(f,g) &= D_\alpha(pf+qg \| f^p g^q) && (14)\\ &= \frac{1}{\alpha-1}\log\int (pf(z)+qg(z))^\alpha (f^p(z)g^q(z))^{1-\alpha} dz && (15)\end{aligned}$$

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 (13), the α-GA divergence is equal to zero if and only if $f = g$ (a.e.) and is otherwise greater than zero.

E. Henze-Penrose Affinity

While divergence measures dissimilarity between distributions, similarity between distributions can be measured by affinity measures. One measure of affinity between probability distributions f and g is

$$A_{HP}(f,g) = 2pq \int \frac{f(z)g(z)}{pf(z) + qg(z)} dz, \tag{16}$$

with respect to weights p and $q = 1 - p, p \in [0, 1]$. This affinity measure was introduced by Henze and Penrose [30] as the limit of the Friedman-Rafsky statistic [27] and we shall call it the Henze-Penrose (HP) affinity. The HP affinity can be related to the divergence measure:

$$D_{HP}(f \| g) = 1 - A_{FR}(f,g) = \int \frac{p^2 f^2(z) + q^2 g^2(z)}{pf(z) + qg(z)} dz \tag{17}$$

All of the above divergence measures can be obtained as special cases of the general class of f-divergences, e.g., as defined in [18], [4]. In this article we focus on the cases for which we know how to implement entropic graph methods to estimate the divergence. For motivation consider the α-entropy (2) which could be estimated by plugging in feature histogram estimates of the multivariate density f. A deterrent to this approach is the curse of dimensionality, which imposes prohibitive computational burden when attempting to construct histograms in large feature dimensions. For a fixed resolution per coordinate dimension the number of histogram bins increases geometrically in feature vector dimension. For example, for a 32 dimensional feature space even a coarse 10 cells per dimension would require keeping track of 10^{32} bins in the histogram, an unworkable and impractically large burden for any envisionable digital computer. As high dimensional feature spaces can be more discriminatory this creates a barrier to performing robust high resolution histogram-based entropic registration. We circumvent this barrier by estimating the α-entropy via an entropic graph whose vertices are the locations of the feature vectors in feature space.

III. Continuous Quasi Additive Euclidean Functionals

A principal focus of this article is the use of minimal graphs over the feature vectors $\mathcal{Z}_n = \{z_1, \ldots, z_n\}$, and their associated minimal edge lengths, for estimation of entropy of the underlying feature density $f(z)$. For consistent estimates we require convergence of minimal graph length to a entropy related quantity. Such convergence issues have been studied for many years, beginning with Beardwood, Halton and Hammersley [6]. The monographs of Steele [75] and Yukich [88] cover the interesting developments in this area. In the general unified framework of Redmond and Yukich [70] a widely applicable convergence result can be invoked for graphs whose length functionals can be shown to Euclidean, continuous and quasi additive. This result can often be applied to minimal graphs constructed by minimizing a graph length function L_γ of the form:

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

where Ω is a set of graphs with specified properties, e.g., the class of acyclic or spanning graphs, e is an edge in Ω, $\|e\|$ is the Euclidean length of e, γ is called the edge exponent or the power weighting constant, and $0 < \gamma < d$. The determination of L_γ requires a combinatorial optimization over the set Ω.

If $\mathcal{Z}_n = \{z_1, \ldots, z_n\}$ is a random i.i.d. sample of d-dimensional vectors drawn from a Lebesgue multivariate density f and the length functional L_γ is continuous quasi additive then the following limit holds [70]

$$\lim_{n \to \infty} L_\gamma(\mathcal{Z}_n)/n^\alpha = \beta_{d,\gamma} \int f^\alpha(z) dz, \quad (a.s.) \tag{18}$$

where $\alpha = (d-\gamma)/d$ and $\beta_{d,\gamma}$ is a constant independent of f. Comparing this to the expression (2) for the Rényi entropy it is obvious that an entropy estimator can be constructed as $(1-\alpha)^{-1} \log\left(L_\gamma(\mathcal{Z}_n)/n^\alpha\right) = H_\alpha(f) + c$, 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. In many cases 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.

A few words are in order concerning the sufficient conditions for the limit (18). 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 $[0,1]^d$. Examples of graphs with continuous quasi additive length functionals are the Euclidean minimal spanning tree (MST), the traveling salesman tour solving the traveling salesman problem (TSP), the steiner tree, the Delaunay triangulation, and the k nearest neighbor graph (kNNG). An example of a graph that does not have a continuous quasi additive length functional is the k-point MST (kMST) discussed in [36].

Even though any continuous quasi additive functional could in principle be used to estimate entropy via relation (18), only those that can be simply computed will be of interest to us here. An uninteresting example is the TSP length functional $L_\gamma^{TSP}(\mathcal{Z}_n) = \min_{C\in c}\sum_{e\in C}\|e\|^\gamma$, where C is a cyclic graph that spans the points $\mathcal{Z}_n$ and visits each point exactly once. Construction of the TSP is NP hard and hence is not attractive for practical image fusion applications. The following sections describe, in detail, the MST and kNN graph functionals.

A. Minimal Spanning Tree for Entropy Estimation

A spanning tree is a connected acyclic graph which passes through all n feature vectors in $\mathcal{Z}_n$. The MST connect these points with $n-1$ edges, denoted $\{e_i\}$, in such a way as to minimize the total length:

$$L_\gamma(\mathcal{Z}_n) = \min_{e\in T}\sum_e \|e\|^\gamma, \tag{19}$$

where T denotes the class of acyclic graphs (trees) that span $\mathcal{Z}_n$. See Figures 5 and 6 for an illustration when $\mathcal{Z}_n$ are points in the unit square. We adopt $\gamma = 1$ for the following experiments.

The MST length $L_n = L(\mathcal{Z}_n)$ is plotted as a function of n in Figure 7 for the case of an i.i.d. uniform sample (right panel) and non-uniform sample (left panel) of $n = 100$ points in the plane. It is intuitive that the length of the MST spanning the more concentrated non-uniform set of points increases at a slower rate in n than does the MST spanning the uniformly distributed points. This observation has motivated the MST as a way to test for randomness in the plane [38]. As shown in [88], the MST length is a continuous quasi additive functional and satisfies the limit (18). More precisely, with $\alpha \overset{\text{def}}{=} (d-\gamma)/d$ the log of the length function normalized by n^α converges (a.s.) within a constant factor to the α-entropy.

$$\lim_{n\to\infty}\log\left(\frac{L_\gamma(\mathcal{Z}_n)}{n^\alpha}\right) = H_\alpha(f) + c_{MST}, \quad \text{(a.s.)}, \tag{20}$$

Thus we can identify the difference between the asymptotes shown on the left Figure 7 as the difference between the α-entropies of the uniform and non-uniform densities ($\alpha = 1/2$). Thus, if f is the underlying

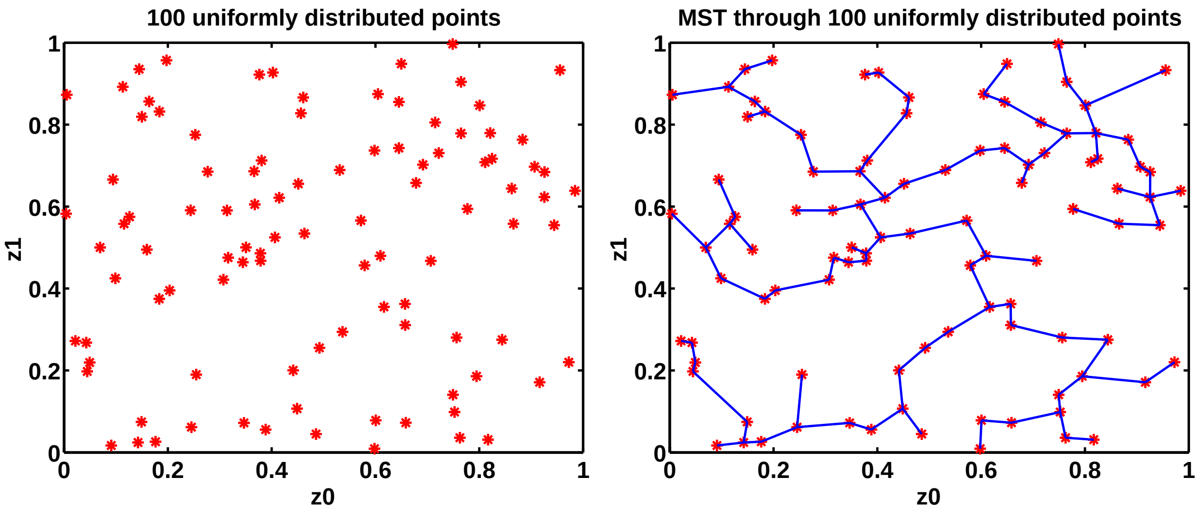


Fig. 5. A set of $n = 100$ uniformly distributed points $\{Z_i\}$ in the unit square in $\mathbb{R}^2$ (left) and the corresponding Minimal Spanning Tree (MST) (right).

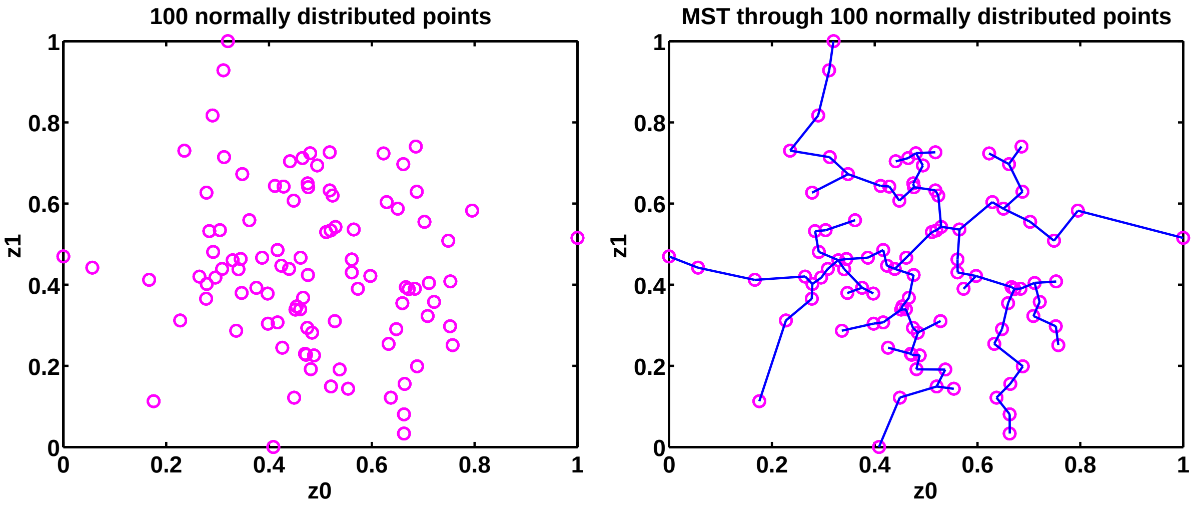


Fig. 6. A set of $n = 100$ normally distributed points $\{Z_i\}$ in the unit square in $\mathbb{R}^2$ (left) and the corresponding Minimal Spanning Tree (MST) (right).

density of $\mathcal{Z}_n$, the α-entropy estimator

$$\widehat{H}_\alpha(\mathcal{Z}_n) = 1/(1-\alpha)\left[\log L_\gamma(\mathcal{Z}_n)/n^\alpha - \log \beta_{d,\gamma}\right], \tag{21}$$

is an asymptotically unbiased and almost surely consistent estimator of the α-entropy of f where $\beta_{d,\gamma}$ is a constant which does not depend on the density f.

The constant $c_{MST} = (1-\alpha)^{-1}\log \beta_{d,\gamma})$ in (20) is a bias term that can be estimated offline. The constant $\beta_{d,\gamma,k}$ is the limit of $L_\gamma(\mathcal{Z}_n)/n^\alpha$ as $n \to \infty$ for a uniform distribution $f(z) = 1$ on the unit cube $[0,1]^d$. This constant can be approximated by Monte Carlo simulation of mean MST length for a large number of uniform d-dimensional random samples.

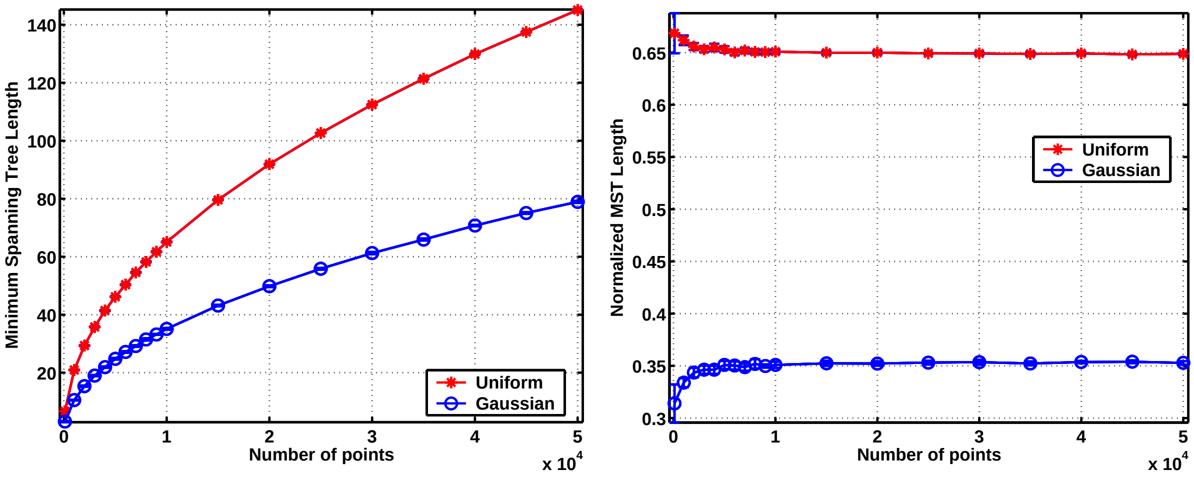


Fig. 7. Mean Length functions L_n of MST implemented with $\gamma = 1$ (left) and $L_n/\sqrt{n}$ (right) as a function of n for uniform and normal distributed points.

The MST approach to estimating the α-Jensen difference between the feature densities of two images can be implemented as follows. Assume two sets of feature vectors $\mathcal{Z}_0 = \{z_0^{(i)}\}_{i=1}^{n_0}$ and $\mathcal{Z}_1 = \{z_1^{(i)}\}_{i=1}^{n_1}$ are extracted from images X_0 and X_1 and are i.i.d. realizations from multivariate densities f_0 and f_1, respectively. In the applications explored in this paper $n_0 = n_1$ but it is worthwhile to maintain this level of generality. Define the set union $\mathcal{Z} = \mathcal{Z}_0 \cup \mathcal{Z}_1$ containing $n = n_0 + n_1$ unordered feature vectors. 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 $n \to \infty$ where $p = \lim_{n\to\infty} n_0/n$. This motivates the following finite sample entropic graph estimator of α-Jensen difference

$$\Delta\widehat{H}_\alpha(p, f_0, f_1) = \widehat{H}_\alpha(\mathcal{Z}_0 \cup \mathcal{Z}_1) - \left[p\widehat{H}_\alpha(\mathcal{Z}_0) + q\widehat{H}_\alpha(\mathcal{Z}_1)\right], \tag{22}$$

where $p = n_0/n$, $\widehat{H}_\alpha(\mathcal{Z}_0 \cup \mathcal{Z}_1)$ is the MST entropy estimator constructed on the n point union of both sets of feature vectors and the marginal entropies $\widehat{H}_\alpha(\mathcal{Z}_0)$, $\widehat{H}_\alpha(\mathcal{Z}_1)$ 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 affine image registration problems 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.

As contrasted with histogram or density plug-in estimator of entropy or Jensen difference, the MST-based estimator enjoys the following properties [33], [31], [36]: it can easily be implemented in high dimensions; it completely bypasses the complication of choosing and fine tuning parameters such as histogram bin size, density kernel width, complexity, and adaptation speed; as the topology of the MST does not depend on the edge weight parameter γ, the MST α-entropy estimator can be generated for the entire range $\alpha \in (0,1)$ once the MST for any given α is computed; the MST can be naturally robustified to outliers by methods of graph pruning. On the other hand the need for combinatorial optimization may be a bottleneck for a large number of feature samples for which accelerated MST algorithms are necessary.

B. Nearest Neighbor Graph Entropy Estimator

The k-nearest neighbor graph is a continuous quasi additive power weighted graph is a computationally attractive alternative to the MST. Given i.i.d vectors $\mathcal{Z}_n$ in $\mathbb{R}^d$, the 1-nearest neighbor of z_i in $\mathcal{Z}_n$ is given by

$$\arg \min_{z \in \mathcal{Z}_n \setminus \{z_i\}} \|z - z_i\|, \tag{23}$$

where $\|z - z_i\|$ is the usual Euclidean (L_2) distance in $\mathbb{R}^d$. For general integer $k \geq 1$, the k-nearest neighbor of a point is defined in a similar way [8], [12], [62]. The kNN graph puts a single edge between each point in $\mathcal{Z}_n$ and its k-nearest neighbors. Let $\mathcal{N}_{k,i} = \mathcal{N}_{k,i}(\mathcal{Z}_n)$ be the set of k-nearest neighbors of z_i in $\mathcal{Z}_n$. The kNN problem consists of finding the set $N_{k,i}$ for each point z_i in the set $\mathcal{Z}_n - \{z\}$.

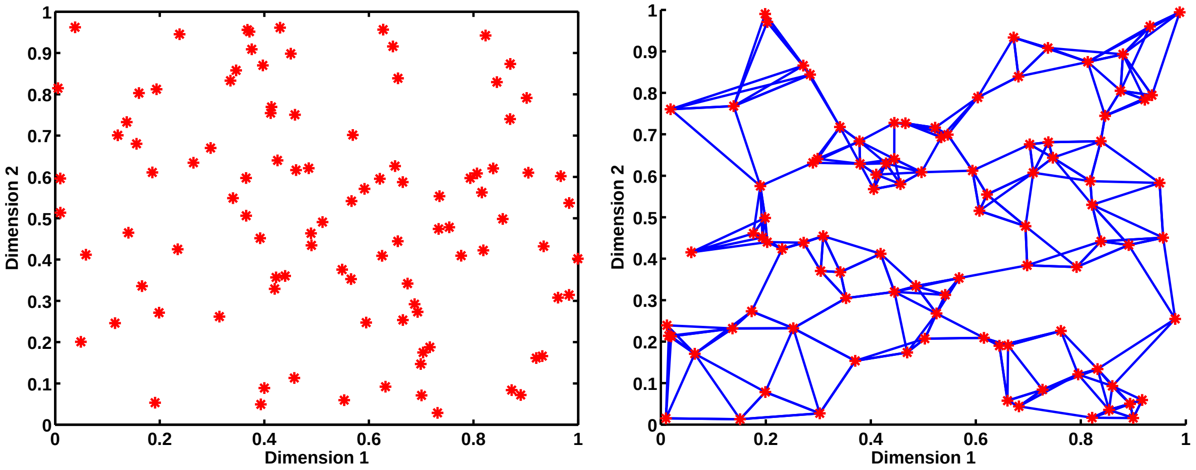


Fig. 8. A set of $n = 100$ uniformly distributed points $\{Z_i\}$ in the unit square in $\mathbf{R^2}$ (left) and the corresponding k-Nearest Neighbor graph $(k = 4)$ (right).

This problem has exact solutions which run in linear-log-linear time and the total graph length is:

$$L_{\gamma,k}(\mathcal{Z}_n) = \sum_{i=1}^{N} \sum_{e \in N_{k,i}} \|e\|^{\gamma}. \tag{24}$$

In general, the kNN graph will count edges at least once, but sometimes count edges more than once. If two points X_1 and X_2 are mutual k-nearest neighbors, then the same edge between X_1 and X_2 will be doubly counted.

Analogously to the MST, the log length of the kNN graph has limit

$$\lim_{n\to\infty} \log\left(\frac{L_{\gamma,k}(\mathcal{X}_n)}{n^{\alpha}}\right) = H_{\alpha}(f) + c_{kNNG}, \quad \text{(a.s.)}. \tag{25}$$

Once again this suggests an estimator of the Renyi α-entropy

$$\widehat{H}_{\alpha}(\mathcal{Z}_n) = 1/(1-\alpha)\left[\log L_{\gamma,k}(\mathcal{Z}_n)/n^{\alpha} - \log \beta_{d,\gamma,k}\right], \tag{26}$$

As in the MST estimate of entropy, the constant $c_{kNNG} = (1-\alpha)^{-1} \log \beta_{d,\gamma,k}$ can be estimated off-line by Monte Carlo simulation of the kNNG on random samples drawn from the unit cube. The complexity of the kNNG algorithm is dominated by the nearest neighbor search, which can be done in $O(n \log n)$ time for n sample points. This contrasts with the MST that requires a $O(n^2 \log n)$ implementation.

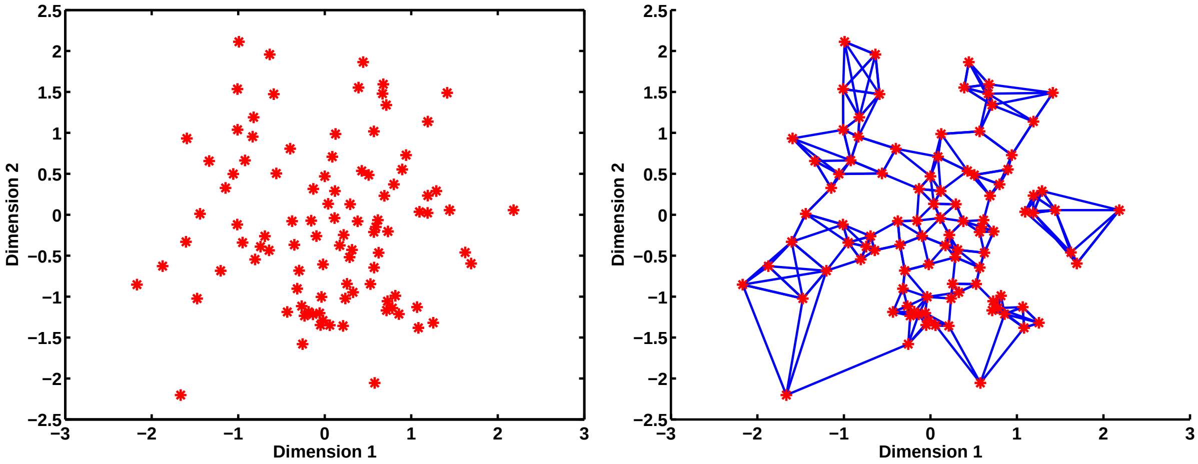


Fig. 9. A set of $n = 100$ normally distributed points $\{Z_i\}$ in the unit square in $\mathcal{R}^{\in}$ (left) and the corresponding k-Nearest Neighbor graph $(k = 4)$ (right).

A related k-NN graph is the graph where edges connecting two points are counted only once. Such a graph eliminates one of the edges from each point pair that are mutual k-nearest neighbors. A kNN graph can be built by pruning such that every unique edge contributes only once to the total length. The resultant graph has the an identical appearance to the initial unpruned k-NN graph, when plotted on the page. However, the cumulative length of the edges in the graphs differ, and so does their β factor (See Figure 11). We call this special pruned k-NN graph, the "Single-Count k-NN graph".

IV. Entropic Graph Estimate of Henze-Penrose Affinity

Friedman and Rafsky [27] presented a multivariate generalization of the Wald-Wolfowitz [85] runs statistic for the two sample problem. The Wald-Wolfowitz test statistic is used to decide between the following hypothesis based on a pair of samples $X, O \in \mathbb{R}^d$ with densities f_x and f_o respectively:

$$H_0: f_x = f_o \tag{27}$$

$$H_1: f_x \neq f_o,$$

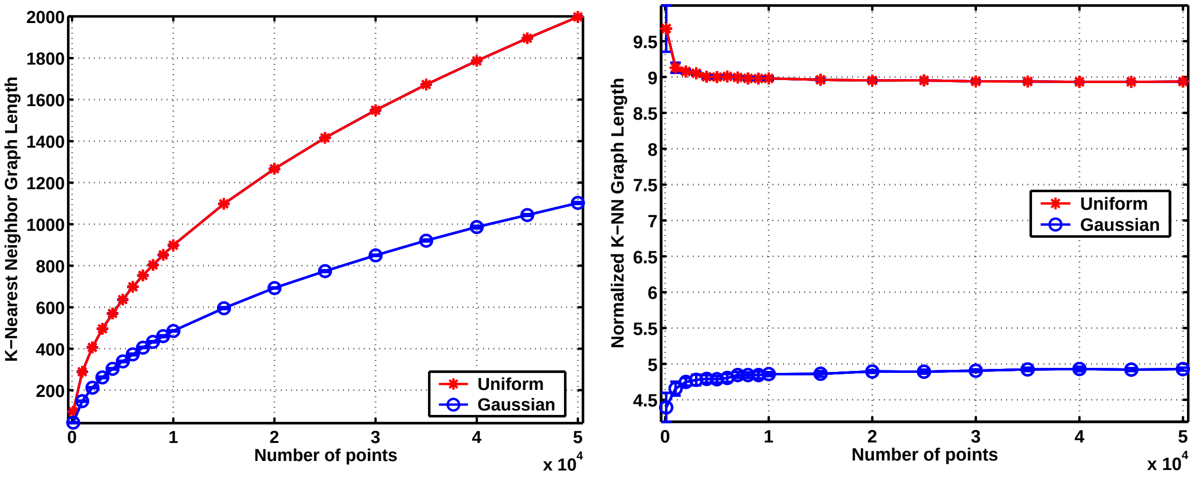


Fig. 10. Mean Length functions L_n of kNN graph implemented with $\gamma = 1$ (left) and $L_n/\sqrt{n}$ (right) as a function of n for uniform and Gaussian distributed points.

The test statistic is applied to an i.i.d. random sample $\{X_i\}_{i=1}^m, \{O_i\}_{i=1}^n$ from f_x and f_o. In the univariate Wald Wolfowitz test ($p = 1$), the $n+m$ 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.

In Friedman and Rafsky's paper [27], the MST was used to obtain a multivariate generalization of the Wald-Wolfowitz test. This procedure is called the Friedman-Rafsky (FR) test and is similar to the MST for estimating the the α-Jensen difference. It is constructed 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 is accepted for smaller values of the FR test statistic. As shown in [30], the FR test statistic N converges to the Henze-Penrose affinity (16) between the distributions f_x and f_o. The limit can be converted to the HP divergence by replacing N by the multivariate run length statistic $R_\ell^{FR} = n + m - 1 - N$.

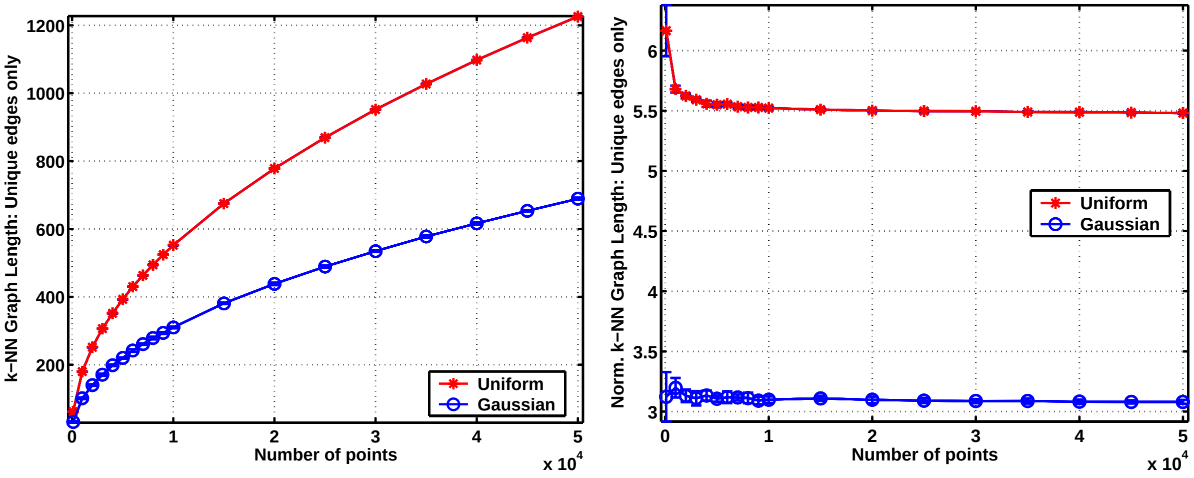
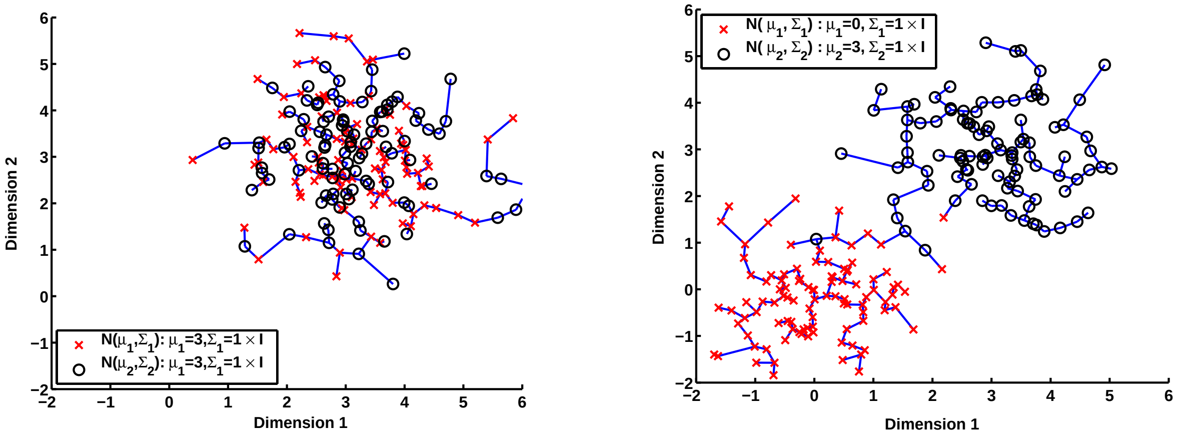


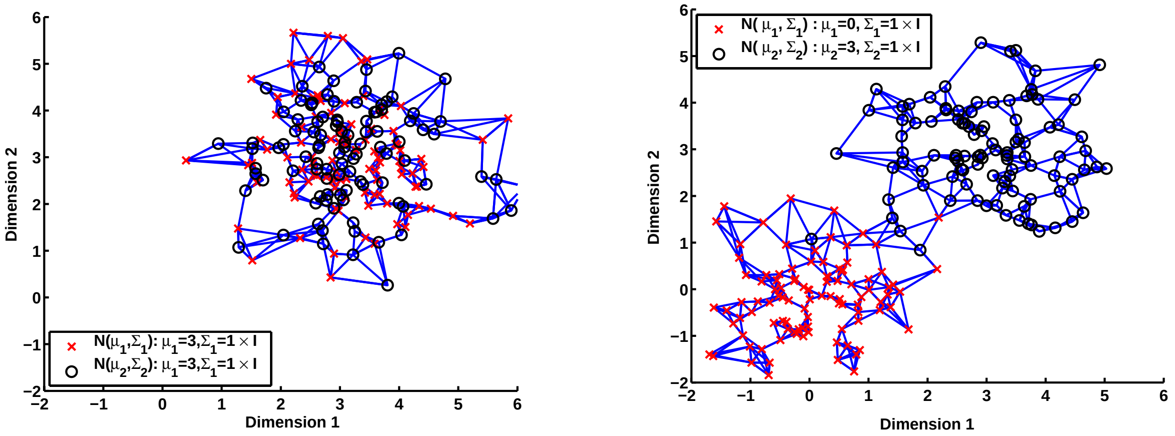
Fig. 11. Mean Length functions L_n of Singe-Count kNN graph implemented with $\gamma = 1$ (left) and $L_n/\sqrt{n}$ (right) as a function of n for uniform and normal distributed points.



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

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

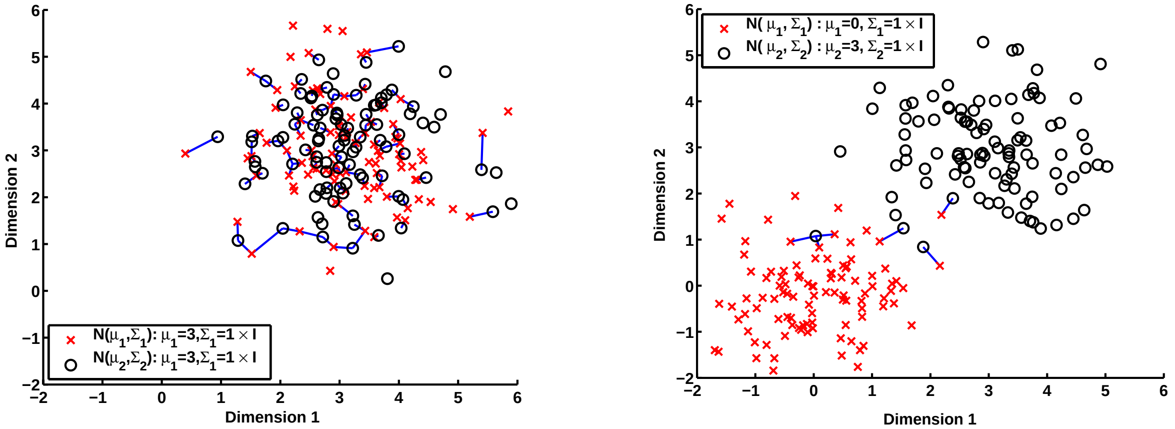
Fig. 12. Illustration of MST for Gaussian case. Two bivariate normal distributions $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_1, \Sigma_1)$ are used. The 'x' labeled points are samples from $f_1(x) = \mathcal{N}(\mu_1, \Sigma_1)$, whereas the 'o' labeled points are samples from $f_2(o) = \mathcal{N}(\mu_2, \Sigma_2)$. (left) $\mu_1 = \mu_2$ and $\Sigma_1 = \Sigma_2$ and (right) $\mu_1 = \mu_2 - 3$ while $\Sigma_1 = \Sigma_2$.



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

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

Fig. 13. Illustration of kNN 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 samples from $f_1(x) = \mathcal{N}(\mu_1, \Sigma_1)$, whereas the 'o' labeled points are samples from $f_2(o) = \mathcal{N}(\mu_2, \Sigma_2)$. (left) $\mu_1 = \mu_2$ and $\Sigma_1 = \Sigma_2$ and (right) $\mu_1 = \mu_2 - 3$ while $\Sigma_1 = \Sigma_2$.



(a) Henze-Penrose $\mu_1 = \mu_2$ and $\Sigma_1 = \Sigma_2$

(b) Henze-Penrose $\mu_2 = \mu_1 + 3$ and $\Sigma_1 = \Sigma_2$

Fig. 14. Illustration of Henze-Penrose affinity 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 samples from $f_1(x) = \mathcal{N}(\mu_1, \Sigma_1)$, whereas the 'o' labeled points are samples from $f_2(o) = \mathcal{N}(\mu_2, \Sigma_2)$. (left) $\mu_1 = \mu_2$ and $\Sigma_1 = \Sigma_2$ and (right) $\mu_1 = \mu_2 - 3$ while $\Sigma_1 = \Sigma_2$.

For illustration of these graph constructions we consider two bivariate normal distributions with density functions f_1 and f_2 parametrized by their mean and covariance $(\mu_1, \Sigma_1), (\mu_2, \Sigma_2)$. Graphs of the α-Jensen divergence calculated using MST (Figure 12), kNNG (Figure 13), and the Henze-Penrose affinity (Figure 14) are shown for the case where $\mu_1 = \mu_2, \Sigma_1 = \Sigma_2$. The 'x' labeled points are samples from $f_1(x) = \mathcal{N}(\mu_1, \Sigma_1)$, whereas the 'o' labeled points are samples from $f_2(o) = \mathcal{N}(\mu_2, \Sigma_2)$. μ_1 is then decreased so that $\mu_1 = \mu_2 - 3$.

V. Entropic Graph Estimators of α-GA mean and α-MI

Assume for simplicity that the target and reference feature sets $O = \{o_i\}_i$ and $X = \{x_i\}_i$ have the same cardinality $m = n$. Here i denotes the i^{th} pixel location in target and reference images. An entropic graph approximation to α-GA mean divergence (15) between target and reference is:

$$\widehat{\alpha D_{GA}} = \frac{1}{\alpha - 1} \log \frac{2^{\gamma/d}}{2n} \sum_{i=1}^{2n} \min\left\{ \left(\frac{e_i(o)}{e_i(x)}\right)^{\gamma/2}, \left(\frac{e_i(x)}{e_i(o)}\right)^{\gamma/2} \right\}, \tag{28}$$

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

Likewise, an entropic graph approximation to the α-MI (7) between the target and the reference is:

$$\widehat{\alpha MI} = \frac{1}{\alpha - 1} \log \frac{1}{n^\alpha} \sum_{i=1}^{n} \left(\frac{e_i(o \times x)}{\sqrt{e_i(o) e_i(x)}} \right)^{2\gamma}, \tag{29}$$

where $e_i(o \times x)$ is the distance from the point $z_i = [o_i, x_i] \in \mathbb{R}^{2d}$ to its nearest neighbor in $\{Z_j\}_{j \neq i}$ and $e_i(o)$ $(e_i(x))$ is the distance from the point $o_i \in \mathbb{R}^d$, $(x_i \in \mathbb{R}^d)$ to its nearest neighbor in $\{O_j\}_{j \neq i}(\{X_j\}_{j \neq i})$.

The estimators (28) and (29) are derived from making a nearest neighbor approximation to the volume of the Voronoi cells constituting the kNN density estimator after plug-in to formulas (15) and (7), respectively. The details are given in the appendix. The theoretical convergence properties of these estimators are at present unknown.

Natural generalizations of (28) and (29) to multiple (> 2) images exist. The computational complexity of the α-MI estimator (29) grows only linearly in the number of images to be registered while that of the α-GA estimator (28) grows as linear log linear. Therefore, there is a significant complexity advantage to implementing α-MI via (29) for simultaneous registration of a large number of images.

VI. Feature-based Matching

While scalar single pixel intensity level is the most popular feature for MI registration, it is not the only possible feature. As pointed out by Leventon and Grimson [48], single pixel MI does not take into account joint spatial behavior of the coincidences and this can cause poor registration, especially in multi-modality situations. Alternative scalar valued features [11] and vector valued features [61], [73] have been investigated for mutual information based image registration. We will focus on local basis projection feature vectors which generalize pixel intensity levels.

Basis projection features are extracted from an image by projecting local sub-images onto a basis of linearly independent sub-images of the same size. Such an approach is widely adopted in image matching applications, in particular with DCT or more general 2D wavelet bases [82], [21], [74], [50], [22]. Others have extracted a basis set adapted to image database using principal components (PCA) or independent components analysis (ICA) [49], [41].

A. ICA Basis Projection Features

The ICA basis is especially well suited for our purposes since it aims to obtain vector features which have statistically independent elements that can facilitate estimation of α-MI and other entropic measures. Specifically, in ICA an optimal basis is found which decomposes the image X_i into a small number of approximately statistically independent components (sub-images) $\{S_j\}$:

$$X_i = \sum_{j=1}^{p} a_{ij} S_j. \tag{30}$$

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 (30) obtained by projecting the image onto the basis.

In Figure 15 we illustrate the ICA basis selected for the MRI image database. ICA was implemented using Hyvarinen and Oja's [41] `FastICA` code (available from [40]) which uses a fixed-point algorithm to perform maximum likelihood estimation of the basis elements in the ICA data model (30). Figure 15 shows a set of 64 16×16 basis vectors which were estimated from over 100,000 16×16 training sub-images randomly selected from 5 consecutive image slices each from two MRI volumes scan of the brain, one of the scans was T1 weighted whereas the other is T2 weighted. Given this ICA basis and a pair of to-be-registered $M \times N$ images, coefficient vectors are extracted by projecting each 16×16 neighborhood in the images onto the basis set. For the 64 dimensional ICA basis shown in Figure 15 this yields a set of MN vectors in a 64 dimensional vector space which will be used to define features.

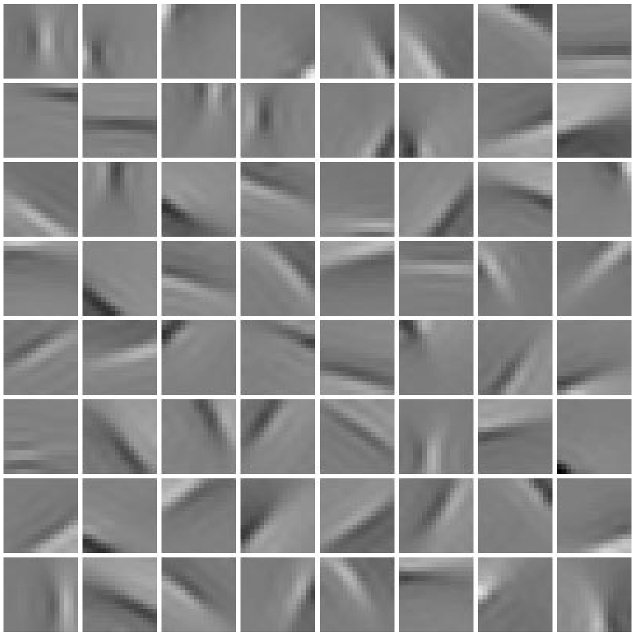


Fig. 15. 16×16 *ICA basis set obtained from training on randomly selected* 16×16 *blocks in 10 T1 and T2 time weighted MRI images. Features extracted from an image are the 64-dimensional vectors obtained by projecting* 16×16 *sub-images of the image on the ICA basis.*

B. Multiresolution Wavelet basis features

Coarse-to-fine hierarchical wavelet basis functions describe a linear synthesis model for the image. The coarser basis functions have larger support than the finer basis; together they incorporate global and local spatial frequency information in the image. The multiresolution properties of the wavelet basis offer an alternative to the ICA basis, which is restricted to a single window size. Wavelet basis are commonly used for image registration [87], [78], [43] and we briefly review them here.

A multiresolution analysis of the space of Lebesgue measurable functions, $\mathcal{L}^2(\mathbb{R})$, is a set of closed, nested subspaces V_j, $j \in \mathbb{Z}$. A wavelet expansion uses translations and dilations of one fixed function, the wavelet $\psi \in L^2(R)$. ψ is a wavelet if the collection of functions $\{\psi(x-l)|l \in Z\}$ is a Riesz basis of V_0 and its orthogonal complement W_0. the The continuous wavelet transform of a function $f(x) \in \mathcal{L}^2(R)$ is given by:

$$\mathcal{W}f(a,b) =< f, \psi_{a,b} >; \psi_{a,b} = \frac{1}{\sqrt{|a|}}\psi(\frac{x-b}{a}), \tag{31}$$

where $a, b \in \mathbb{R}, a \neq 0$.

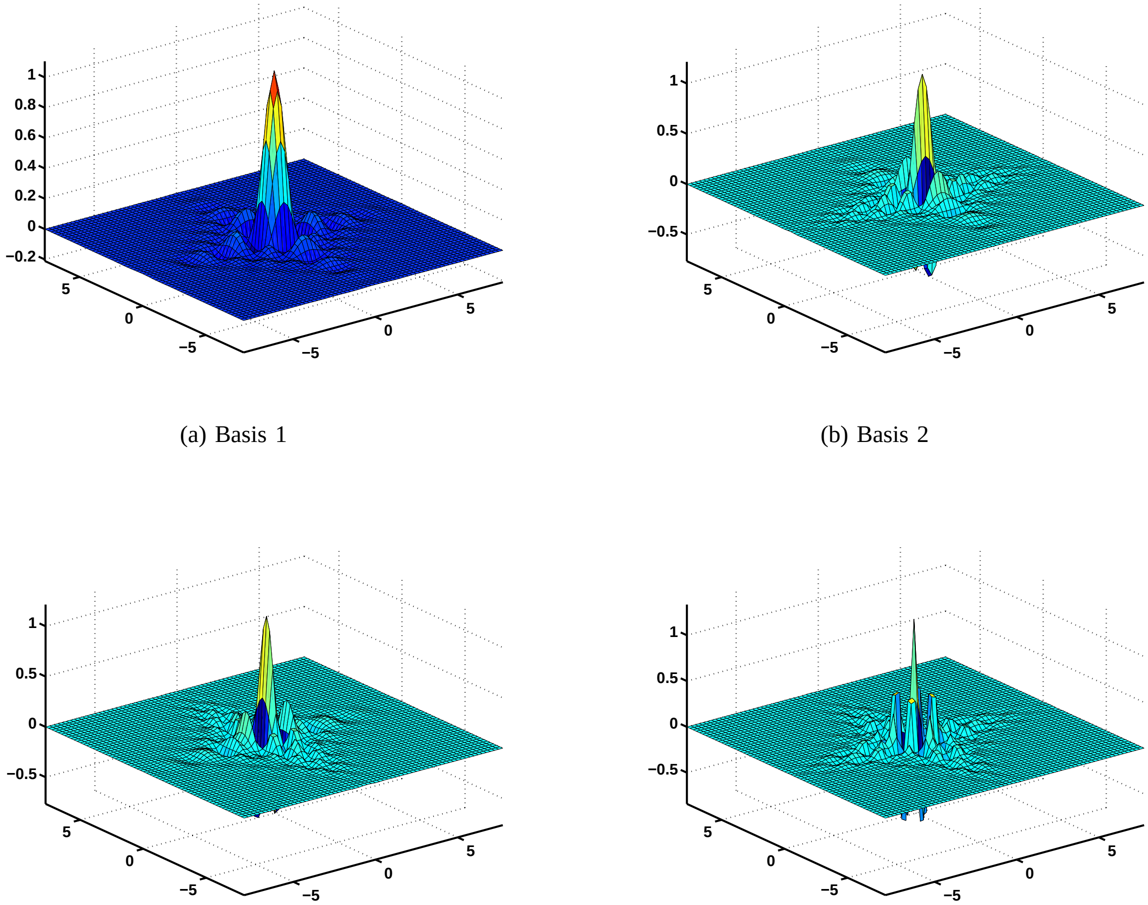
For discrete wavelets, the dilation and translation parameters, b and a, are restricted to a discrete set, $a = 2^j, b = k$ where j and k are integers. The dyadic discrete wavelet transform is then given as:

$$\mathcal{W}f(j,k) =< f, \psi_{j,k} > \psi_{j,k} = 2^{-j/2}\psi(2^{-j}x - k) \tag{32}$$

where $j, k \in \mathbb{Z}$. Thus the wavelet coefficient of f at scale j and translation k is the inner product of f with the appropriate basis vector at scale j and translation k. The 2D discrete wavelet analysis is obtained by a tensor product of two multiresolution analysis of $\mathcal{L}^2(\mathbb{R})$. At each scale, j, we have one scaling function subspace and three wavelet subspaces. The discrete wavelet transform of an image is the projection of the image onto the scaling function V_0 subspaces and the wavelet subspaces W_0. The corresponding coefficients are called the approximate and detail coefficients, implying the low and high pass characteristics of the basis filters. The process of projecting the image onto the successively coarser spaces continues to achieve the approximation desired. The difference information sensitive to vertical, horizontal and diagonal edges are treated as the three dimensions of each feature vector. Several members of the discrete Meyer basis used in this work are plotted below in Figure (16)

VII. Computational Considerations

A popular sentiment about graph methods, such as the MST and the kNN graph, is that they could be computationally taxing. However, since the early days, graph theory algorithms have evolved and several variants with low time-memory complexity have been found. Henze-Penrose and the α-GA mean divergence metrics are based directly on the MST and kNNG and first require the solution of these combinatorial optimization problems. This section is devoted to providing insight into the formulation of these algorithms and the assumptions that lead to faster, lower complexity variants of these algorithms.



(a) Basis 1

(b) Basis 2

(c) Basis 3

(d) Basis 4

Fig. 16. 2D Discrete Meyer Wavelet basis from coarse (a) to fine (d).

A. Reducing time-memory complexity of the MST

The MST problem has been studied since the early part of this century. Due to its widespread applicability in other computer science, optimization theory and pattern recognition related problems there have been and continue to be sporadic improvements in the time-memory complexity of the MST problem. Two principal algorithms exist for computing the MST, the Prim algorithm [67] and the Kruskal algorithm [45]. For sparse graphs the Kruskal algorithm is the fastest general purpose MST computation algorithm. Kruskal's algorithm maintains a list of edges sorted by their weights and grows the tree one edge at a time. Cycles are avoided within the tree by discarding edges that connect two sub-trees already joined through a prior established path. The time complexity of the Kruskal algorithm is of order $O(E \log E)$ and the the memory requirement is $O(E)$, where E is the initial number of edges in the graph. Recent algorithms have been proposed that offer advantages over these older algorithms at the expense of increased complexity. A review can be found in [5]

An initial approach may be to construct the MST by including all the possible edges within the feature set. This results in N^2 edges for N points; a time requirement of $O(N^2)$ and a memory requirement of $O(N^2 \log N)$. The number of points in the graph is the total number of d-dimensional features participating in the registration from the two images. If each image has $M \times N$ features (for eg. pixels), the total number of points in the graph is $2 \times M \times N \approx$ 150,000 for images of size 256 $\times$ 256 pixels. The time and memory requirements of the MST is beyond the capabilities of even the fastest available desktop processors.

The earliest solution can be attributed to Bentley and Friedman [10]. Using a method to quickly find nearest neighbors in high dimensions they proposed building a minimum spanning tree using the assumption that local neighbors are more likely to be included in the MST than distant neighbors. Several improvements have been made on this technique, and have been proposed in [14] and [56]. For our experiments we have been motivated by the adapted the original Bentley method, as explained below. This method achieves significant acceleration by sparsification of the initial graph before tree construction.

We have implemented a method for sparsification that allows MST to be constructed for several hundred thousand points in a few minutes of desktop computing time. This implementation uses a disc windowing method for constructing the edge list. Specifically, we center disc's at each point under consideration and pick only those neighbors whose distance from the point is less than the radius of the disc (See Figure 17 for illustration). A list intersection approach similar to [62] is adopted to prune unnecessary edges within the disc. Through a combination of list intersection and disc radius criterion we reduce the number of edges that must be sorted and simultaneously ensure that the MST thus built is valid. We have empirically found that for uniform distributions, a constant disc radius is best. For non uniform distributions, the disc radius is better selected as the distance to the k^{th}-nearest neighbor (kNN). Figure 18 shows the bias of modified MST algorithm as a function of the radius parameter and the number of nearest neighbors for a uniform density on the plane.

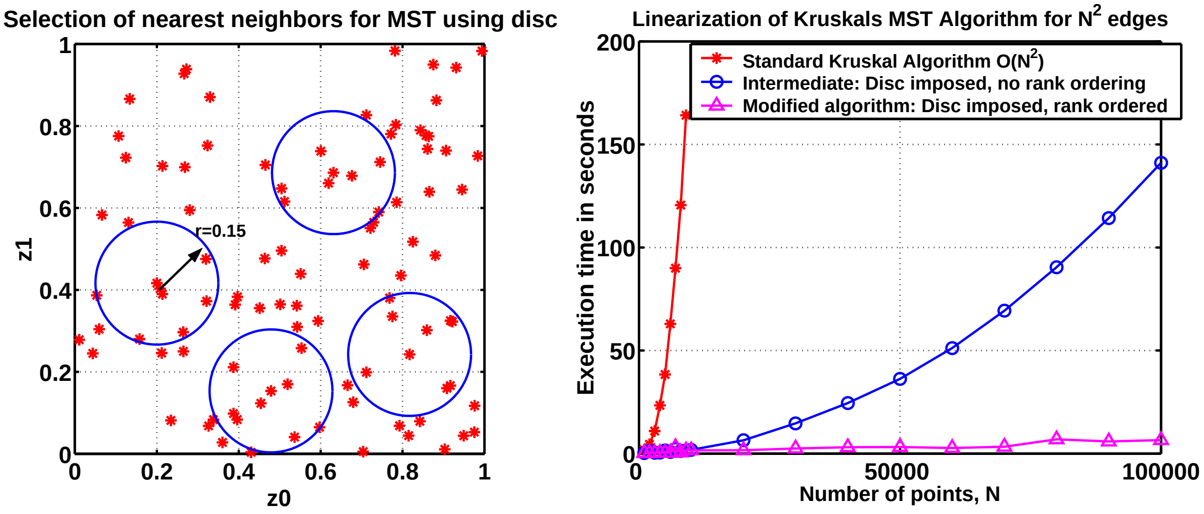


Fig. 17. Disc-based acceleration of Kruskal's MST algorithm from $n^2 \log n$ to $n \log n$ (left) and comparison of computation time for Kruskal's standard MST algorithm with respect to our accelerated algorithm (right).

It is straightforward to prove that, if the radius is suitably specified, our MST construction yields a valid minimum spanning tree. Recall that the Kruskal algorithm ensures construction of the exact MST [45]. Consider a point p_i in the graph.

(1) If point p_i is included in the tree, then the path of its connection to the tree has the lowest weight amongst all possible non-cyclic connections. To prove this is trivial. The disc criterion includes lower weight edge before considering an edge with a higher weight. Hence, if a path is found by imposing the disc, that path is the smallest possible non-cyclic path. The non-cyclicity of the path is ensured in the Kruskal algorithm through a standard Union-Find data set.

(2) If a point p_i is not in the tree, it is because all the edges between p_i and its neighbors considered using the disc criterion of edge inclusion have total edge weight greater than disc radius or have led to a cyclic path. Expanding the disc radius would then provide the path which is lowest in weight and non-cyclic.

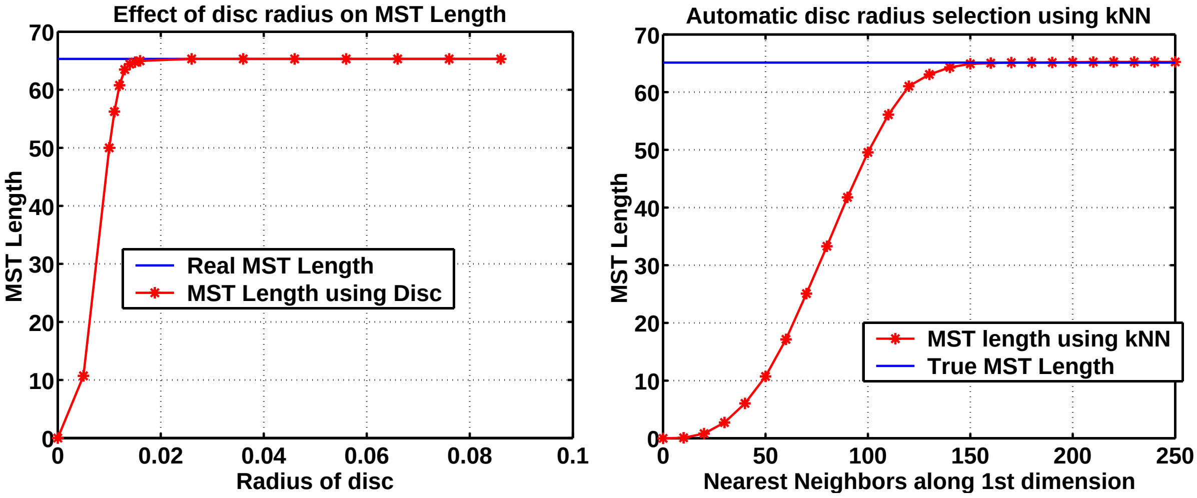


Fig. 18. Bias of the $n \log n$ MST algorithm as a function of radius parameter (left) and as a function of the number of nearest neighbors (right) for uniform points in the unit square.

B. Reducing time-memory complexity of the kNN Graph

Time memory considerations in the nearest neighbor graph have prompted researchers to come up with various exact and approximate graph algorithms. With its wide-spread usage, it is not surprising that several fast methods exist for nearest neighbor graph constructions. Most of them are expandable to construct k-NN graphs. One of the first fast algorithms for constructing NNG was proposed by Bentley [9], [8]. A comprehensive survey of the latest methods for nearest neighbor searches in vector spaces is presented in [12]. A simple and intuitive method for nearest neighbor search in high dimensions is presented in [62].

Though compelling, the methods presented above focus on retrieving the exact nearest neighbors. One could hypothesize that for applications where the accuracy of the nearest neighbors is not critical, we could achieve significant speed-up by accepting a small bias in the nearest neighbors retrieved. This is the principal argument presented in [1]. We conducted our own experiments on the approximate NN method using the code provided in [55] (Figure 19). We conducted benchmarks on uniformly points distributed in 8 dimensional space. If the error incurred in picking the incorrect k-th nearest neighbor $\leq \epsilon$, the cumulative error in the length of the kNNG is plotted in Figure 19. Compared to an exact kNN search using k-d trees a significant reduction ($> 85\%$) in time can be obtained, through approximate NN methods, incurring a 15% cumulative graph length error.

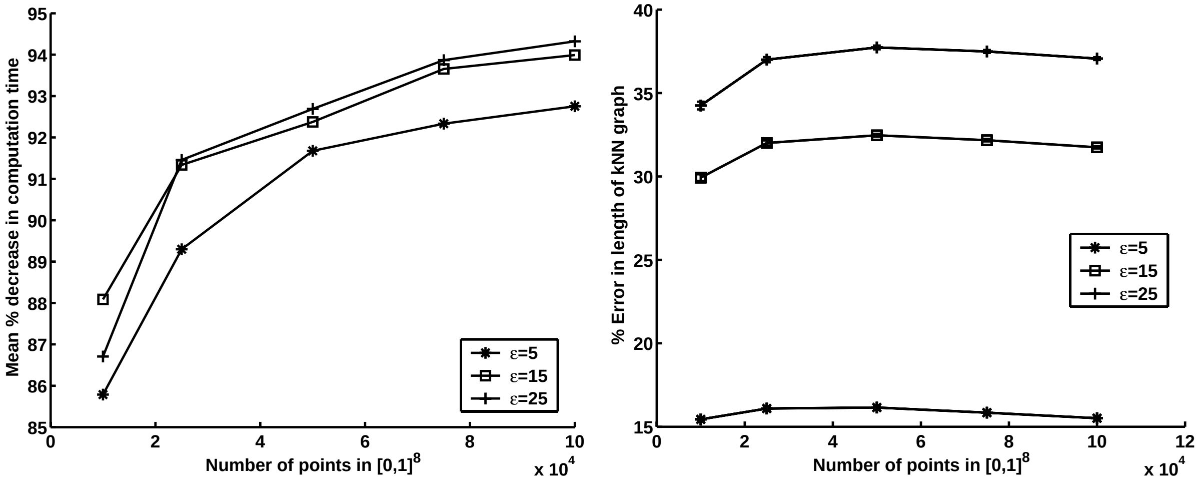


Fig. 19. Approximate k-NNG: (left) Decrease in computation time to build approximate kNNG for different ϵ, expressed as a percentage of time spent computing the exact kNNG over a uniformly distributed points in $[0, 1]^8$. An 85% reduction in computation time can be obtained by incurring a 15% error in cumulative graph length. (right) Corresponding error incurred in cumulative graph length.

VIII. Applications: Multisensor satellite image fusion

In this section, we shall illustrate entropic graph based image registration for a remote sensing example. Images of sites on the earth are gathered by a variety of geostationary satellites. Numerous sensors gather information in distinct frequency bands in the electromagnetic spectrum. These images help predict daily weather patterns, environmental parameters influencing crop cycles such as soil composition, water and mineral levels deeper in the Earth's crust, and may also serve as surveillance sensors meant to monitor activity over hostile regions. A satellite may carry more than one sensor and may acquire images throughout a period of time. Changing weather conditions may interfere with the signal. Images captured in a multisensor satellite imaging environment show linear deformations due to the position of the sensors relative to the object. This transformation is often linear in nature and may manifest itself as relative translational, rotational or scaling between images. This provides a good setting to observe different divergence measures as a function of the relative deformation between images. We simulated linear rotational deformation in order to reliably test the image registration algorithms presented above.

Figure 20 shows two images of downtown Atlanta, captured with visible and thermal sensors, as a part of the ‘Urban Heat Island’ project [68] that studies the creation of high heat spots in metropolitan areas across the USA. Pairs of visible light and thermal satellite images were also obtained from NASA’s Visible Earth website [57]. The variability in imagery arises due to the different specialized satellites used for imaging. These include weather satellites wherein the imagery shows heavy occlusion due to clouds and other atmospheric disturbances. Other satellites focus on urban areas with roads, bridges and high rise buildings. Still other images show entire countries or continents, oceans and large geographic landmarks such as volcanoes and active geologic features. Lastly, images contain different landscapes such as deserts, mountains and valleys with dense foliage.

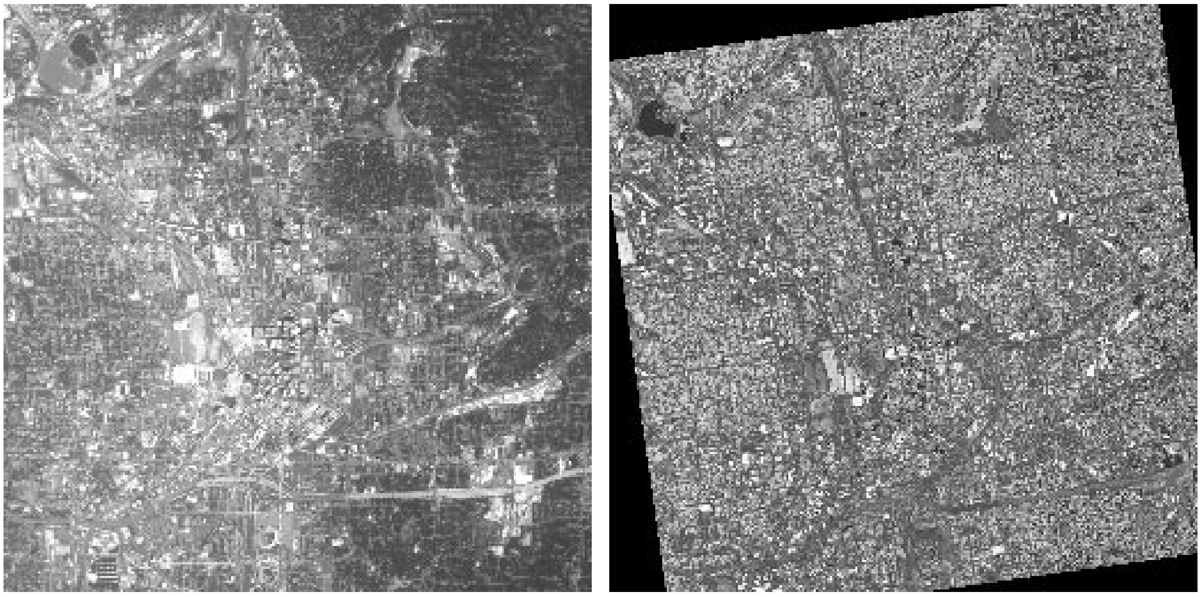


Fig. 20. Images of downtown Atlanta obtained from Urban Heat Island project [68]. (a) Thermal image (b) Visible-light image under artificial rotational transformation

A. Deformation and feature definition

Images are rotated through $0°$ to $32°$, with a step size adjusted to allow a finer sampling of the objective function near $0°$. The images are projected onto a Meyer wavelet basis, and the coefficients are used as features for registration. A feature sample from an image I in the database is represented as tuple consisting of the coefficient vector, and a two dimensional vector identifying the spatial coordinates of the origin of the image region it represents. For example $\{\underline{W_{(i,j)}}, x_{(i,j)}, y_{(i,j)}\}$ represents the a tuple from position $\{i, j\}$ in the image. Now, $\underline{W_{(i,j)}} \equiv \{w_{(i,j)}^{Low-Low}, w_{(i,j)}^{Low-High}, w_{(i,j)}^{High-Low}, w_{(i,j)}^{High-High}\}$, where the super-script identifies the frequency band in the wavelet spectrum. Features from both the images $\{Z_1, Z_2\}$ are pooled together to form a joint sample pool $\{Z_1 \bigcup Z_2\}$. The MST and k-NN graph are individually constructed on this sample pool.

Figure 21 shows the rotational mean-squared registration error for the images in our database, in the presence of additive noise. Best performance under the presence of noise can be seen through the use of the α-MI estimated using wavelet features and kNN graph. Comparable performances are seen through the use of Henze-Penrose and αGeometric-Arithmetic mean divergences, both estimated using wavelet features. Interestingly, the single pixel Shannon MI has the poorest performance which may be attributed to its use of poorly discriminating scalar intensity features. Notice that the α-GA, Henze-Penrose affinity, and α-MI(Wavelet-kNN estimate), all implemented with wavelet features, have significantly lower MSE compared to the other methods.

Further insight into the performance of these wavelet-based divergence measures may be gained by considering the mean objective function over 750 independent trials. Figure 22.a shows the α-MI, HP affinity and the α-GA affinity and Fig. 22.b shows the α-Jensen difference divergence calculated using the kNN graph and the MST. The sensitivity and robustness of the dissimilarity measures can be evaluated by observing the divergence function near zero rotational deformation (Figure 22).

IX. Applications: Local feature matching

The ability to discriminate differences between images with sensitivity to local differences is pivotal to any image matching algorithm. Previous work in these techniques has been limited to simple pixel based mutual information (MI) and pixel correlation techniques. In [65], local measures of MI outperform

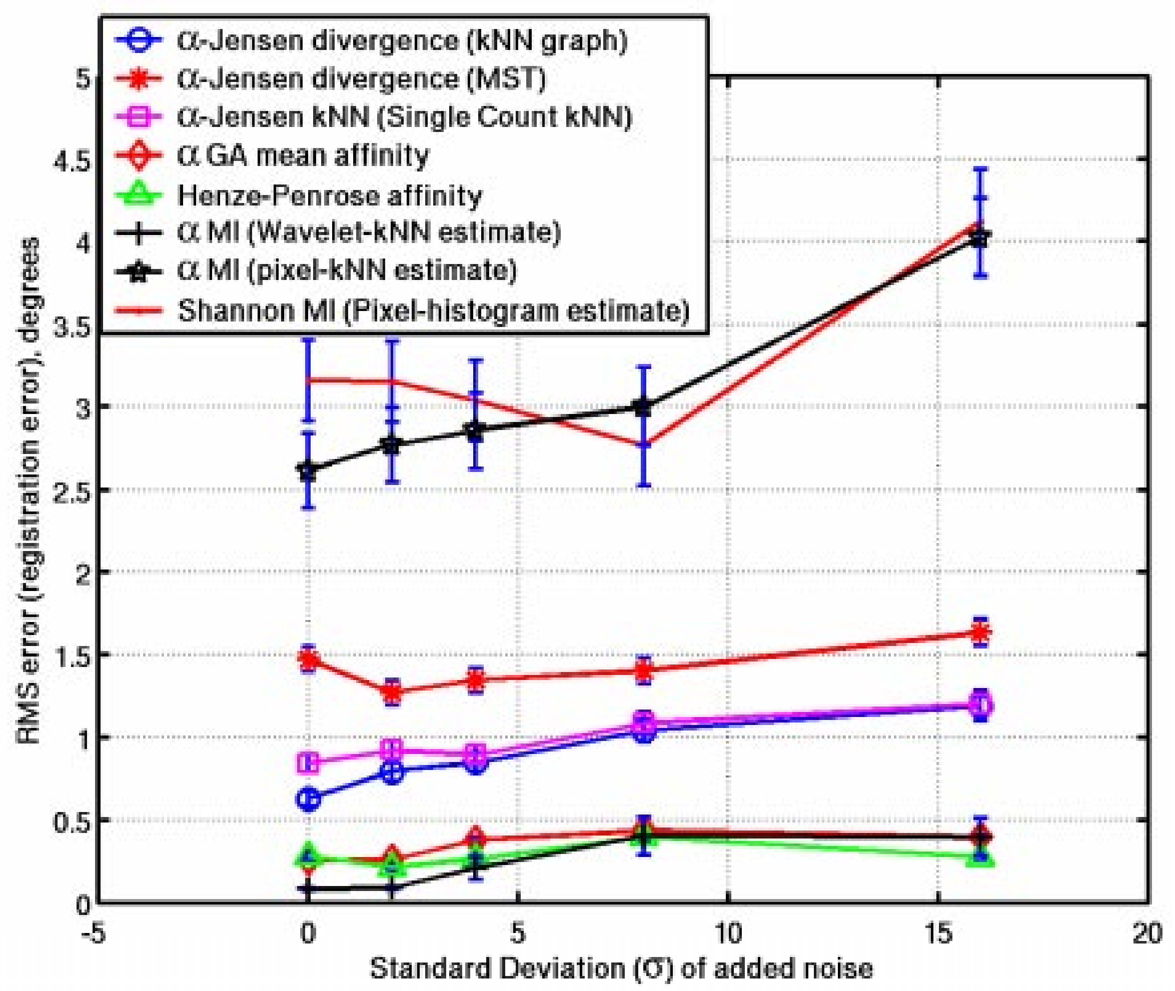
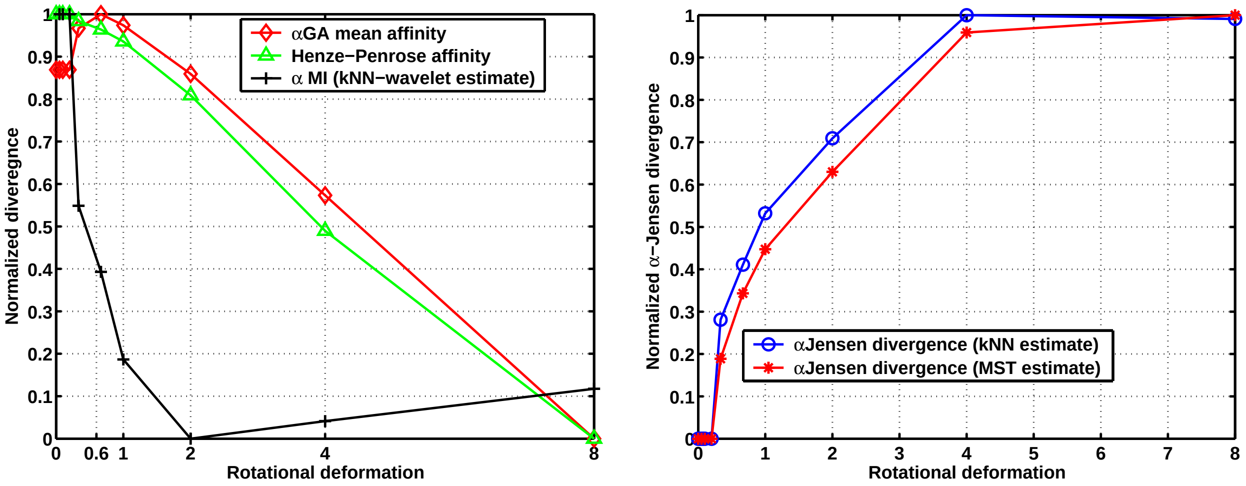


Fig. 21. Rotational root mean squared error obtained from rotational registration of multisensor satellite imagery using six different image similarity/dissimilarity criteria. Standard error bars are as indicated. These plots were obtained from Monte Carlo trials consisting of adding i.i.d. Gaussian distributed noise to the images prior to registration.

global MI in the context of adaptive grid refinement for automatic control point placement. However, the sensitivity of local MI deteriorates rapidly as the size of the image window decreases below 40×40 pixels in 2D.

The main constraints on these algorithms, when localizing differences, are (1) limited feature resolution of single pixel intensity based features, and (2) histogram estimators $h(X, Y)$ of joint probability density $f(X, Y)$ are noisy when computed with a small number of pixel features and are thus poor estimators of



(a) Average α-GA affinity, HP affinity and α-MI (kNN-wavelet estimate. Rotation angle estimated by maximizing noisy versions of these objective functions.

(b) Average α-Jensen divergence (kNN and MST estimate on wavelet features). Rotation angle estimated by minimizing noisy versions of these objective functions.

Fig. 22. Average affinity and divergence, over all images, in the vicinity of zero rotation error: (left) α-Jensen (kNN) and α-Jensen (MST), (right) α-GA mean affinity, HP affinity and α-MI estimated using wavelet features and kNN graph.

$f(X,Y)$ used by the algorithm to derive joint entropy $H(X,Y)$. Reliable identification of subtle local differences within images, is key to improving registration sensitivity and accuracy [59]. Stable unbiased estimates of local entropy are required to identify sites of local mismatch between images. These estimates play a vital role in successfully implementing local transformations.

A. *Deformation localization*

Iterative registration algorithms apply transformations to a sequence of images while minimizing some objective function. We demonstrate the sensitivity of our technique by tracking deformations that correspond to small perturbations of the image. These perturbations are recorded by the change in the mismatch metric.

Global deformations reflect a change in imaging geometry and are modeled as global transformations on the images. However, global similarity metrics are ineffective in capturing local deformations in medical images that occur due to physiological or pathological changes in the specimen. Typical examples are: change in brain tumor size, appearance of micro-calcifications in breast, non-linear displacement of soft tissue due to disease and modality induced inhomogeneities such as in MRI and nonlinear breast compression in XRay mammograms. Most registration algorithms will not be reliable when the size of the mismatch site is insufficiently small, typically $(m \times n) \leq 40 \times 40$ [65]. With a combination of ICA and α-entropy we match sites having as few as 8×8 pixels. Due to the limited number samples in the feature space, the faster convergence properties of the MST are better suited to this problem. Although we do not estimate other divergence measures, α-Jensen calculated using the MST provides a benchmark for their performance.

In Figure 23, multimodal synthesized scan of T1 and T2 weighted brain MRI each of size 256×256 pixels [16] are seen. The original target images shall be deformed locally (see below) to generate a deformed target image.

1) Locally deforming original image using B-Splines: B-spline deformations are cubic mapping functions that have local control and injective properties [13]. The 2D uniform tensor B-spline function F, is defined with a 4×4 control lattice ϕ in $\mathbb{R}^2$ as:

$$F(u,v) = \sum_{i=0}^{3}\sum_{j=0}^{3} B_i(u)B_j(v)\phi_{ij}, \tag{33}$$

where $0 \leq u,v \leq 1$, ϕ_{ij} is the spatial coordinates of the lattice and B_i are the standard B-Spline basis functions. The uniform B-Spline basis functions used here are quite common in computer graphics literature and may be found in [13] are defined as:

$$\begin{aligned} B_0(u) &= \frac{(1-u)^3}{6}, \\ B_1(u) &= \frac{3u^3 - 6u^2 + 4}{6}, \\ B_2(u) &= \frac{-3u^3 + 3u^2 + 3u + 1}{6}, \\ B_3(u) &= \frac{u^3}{6}. \end{aligned} \tag{34}$$

Given that the original images have 256×256 pixels, we impose a grid(Φ) of 10×10 control points on I_{tar}. Since the aim is to deform I_{tar} locally, not globally, we select a sub-grid (ϕ) of 4×4 control points in the center of I_{tar}. We then diagonally displace, by $\ell = 10$ mm, only one of the control points in ϕ, to generate deformed grid ϕ_{def}. I_{tar} is then reconstructed according to ϕ_{def}. The induced deformation is measured as $||\phi_{def} - \phi||$. Figure 23 shows the resultant warped image and difference image, $I_{tar} - T(I_{tar})$. For smaller deformations, Φ is a finer grid of 20×20 points, from which ϕ is picked. A control point in ϕ is then displaced diagonally by $\ell = 1, 2, \ldots 10$ to generate ϕ_{def}. When $\ell \leq 3$, noticeable deformation spans only 8×8 pixels.

2) Feature discrimination algorithm: We generate a d-dimensional feature set $\{Z_i\}_{i=1}^{m \times n}$, $m \times n \geq d$ by sequentially projecting sub-image block (window) $\{\Gamma_j\}_{j=1}^{M \times N}$ of size $m \times n$ onto a d-dimensional basis function set $\{S_k\}$ extracted from the MRI image, as discussed in Section VI-A. Raster scanning through I_{ref} we select sub-image blocks $\{\Gamma_i^{ref}\}_{i=1}^{M \times N}$. For this simulation exercise, we pick only the sub-image block Γ^{tar} from $T(I_{tar})$ corresponding to the particular pixel location $k = (128, 128)$. $\Gamma_{128,128}^{tar}$ corresponds to the area in I_{tar} where the B-Spline deformation has been applied.

The size of the ICA basis features is 8×8, i.e. the feature dimension is, $d = 64$. The MST is constructed over the joint feature set $\{Z_i^{ref}, Z_j^{tar}\}$. When suitably normalized with $1/n^\alpha, \alpha = 0.5$, the length of the MST becomes an estimate of $H_\alpha(Z_i^{ref}, Z_j^{tar})$. We score all the sub-image blocks $\{\Gamma_i^{ref}\}_{i=1}^{M \times N}$ with respect to the sub-image block $\Gamma_{128,128}^{tar}$. Let O_ℓ be the resultant $M \times N$ matrix of scores, at deformation ℓ. The objective function surface O_ℓ is a similarity map between $\{\Gamma^{ref}\}_{i=1}^{M \times N}$ and Γ^{tar}. When two sites are compared, the resulting joint probability distribution depends on the degree of mismatch. The best match is detected by searching for the region in I_{ref} that corresponds to Γ^{tar} as determined by the MST length. As opposed to the one-to-all block matching approach adopted here, one could also perform a block-by-block matching, where each block Γ_i^{ref} is compared with its corresponding block Γ_i^{tar}.

B. Local Feature matching Results

Figure 23 shows O_{10} for $m \times n = 8 \times 8$, 16×16 and 32×32. Similar maps can be generated for $\ell = \ell_1, \ell_2, \ldots \ell_p$. The gradient $\nabla(O) = O_{\ell_1} - O_{\ell_2}$ reflects the change in H_α, the objective function, when I_{tar} experiences an incremental change in deformation, from $\ell = \ell_1 \rightarrow \ell_2$. This gradient, at various sub-image block size is seen in Figure 23, where $\ell_1 = 0$ and $\ell_2 = 10$. For demonstration purposes in Figure 23, we imposed a large deformation to I_{tar}. Smaller deformations generated using a control grid spanning only 40×40 pixels are used to generate Figure 24. It shows the ratio of the gradient of the objective function:

$$R = \frac{\frac{1}{(m\times n)}\sum_{i=1}^{m\times n}|\nabla(O(i))|}{\frac{1}{(M\times N-m\times n)}\sum_{i=1}^{M\times N-m\times n}|\nabla(O(i))|}, \tag{35}$$

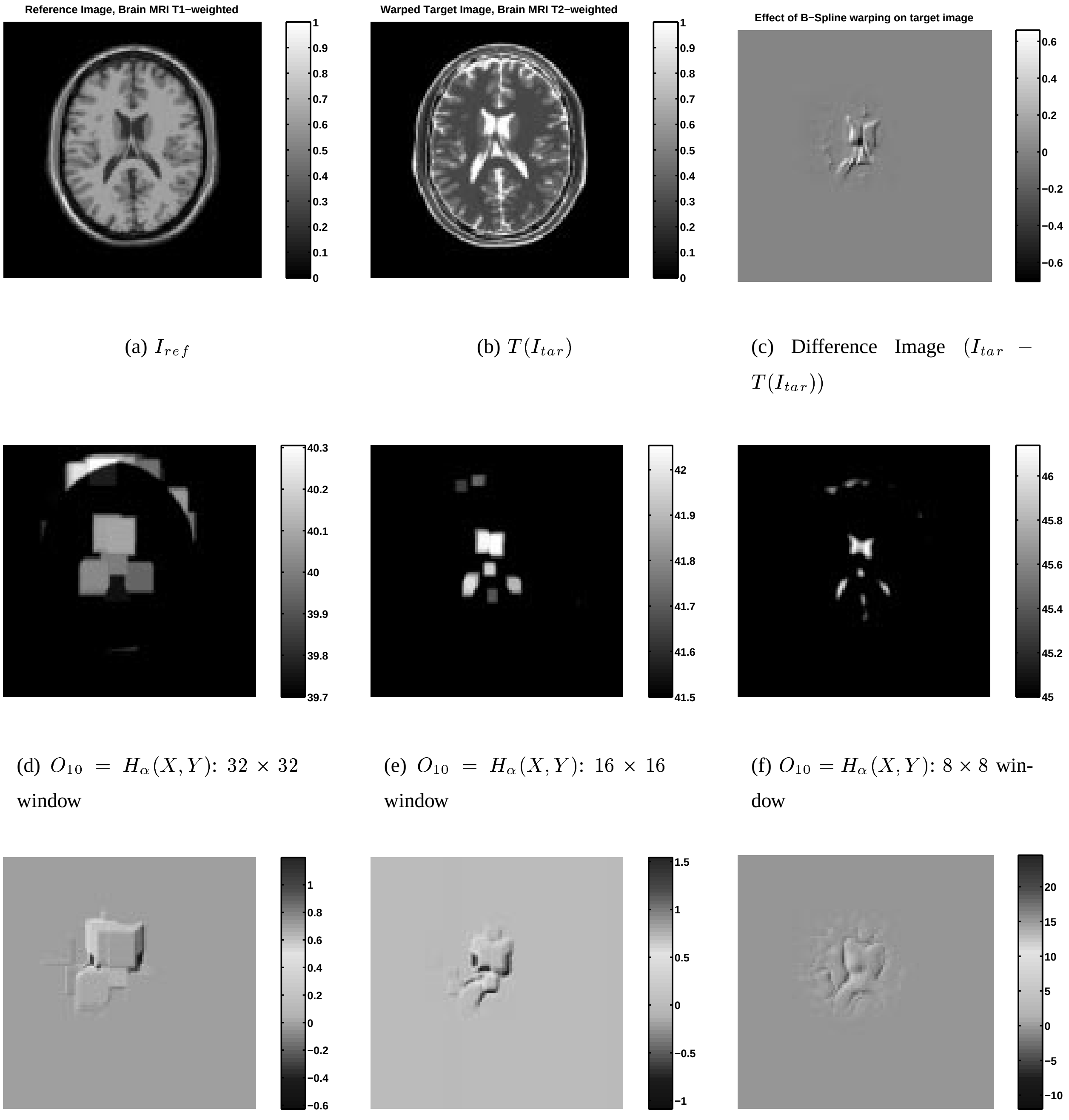
over the deformation site v/s background in the presence of additive Gaussian noise.

Figure 25 shows the similarity map O_ℓ when constructed using a histogram estimate of joint entropy calculated over sub-image size $m \times n$ (8). At lower sub-image sizes, the estimate displays bias and several local minima even under noise free conditions. It is thus unsuitable for detection of local deformation of I_{tar}.

The framework presented here could be extended to (1) enhance registration performance by sensitizing it to local mismatch, (2) automatically track features of interest, such as tumors in brain or micro-calcifications in breast across temporal image sequences, (3) reliably match or register small images or image regions so as to improve disease diagnosis by locating and identifying small pathological changes in medical image volumes and (4) automate control point placement to initiate registration.

X. Conclusion

In this paper we presented several techniques to extend the multisensor image fusion problem to high dimensional feature spaces. Rényi's α-entropy is estimated directly in high dimensions through the use of entropic graph methods. These include the use of Euclidean functionals such as the α-Jensen, the HP divergence and GA mean divergence. Graph theory methods such as the MST and the kNN graph are central to our approach due to their quasi additive properties in estimating Euclidean functionals. These methods provide a robust and viable alternative to traditional pixel intensity histograms used for estimating MI. Higher dimensional features used for this work are the Wavelet basis and ICA, where features are 64 dimensional. Our methods are validated through a demonstration of registration of multisensor satellite and medical imagery.



(a) I_{ref}

(b) $T(I_{tar})$

(c) Difference Image $(I_{tar} - T(I_{tar}))$

(d) $O_{10} = H_\alpha(X, Y)$: 32×32 window

(e) $O_{10} = H_\alpha(X, Y)$: 16×16 window

(f) $O_{10} = H_\alpha(X, Y)$: 8×8 window

(g) Local $\nabla(H_\alpha) = O_{10} - O_0$: 32×32 window

(h) Local $\nabla(H_\alpha) = O_{10} - O_0$: 16×16 window

(i) Local $\nabla(H_\alpha) = O_{10} - O_0$: 8×8 window

Fig. 23. B-Spline deformation on MRI images of the brain. (a) Reference image, (b) Warped target (c) True Deformation, (d) $O_{10} = H_\alpha$ as seen with a 32×32 window, (e) 16×16 window and (f) 8×8 window. (g) $\nabla(O) = \nabla(H_\alpha) = O_{10} - O_0$ as seen with a 32×32, (h) 16×16 and (i) 8×8 window.

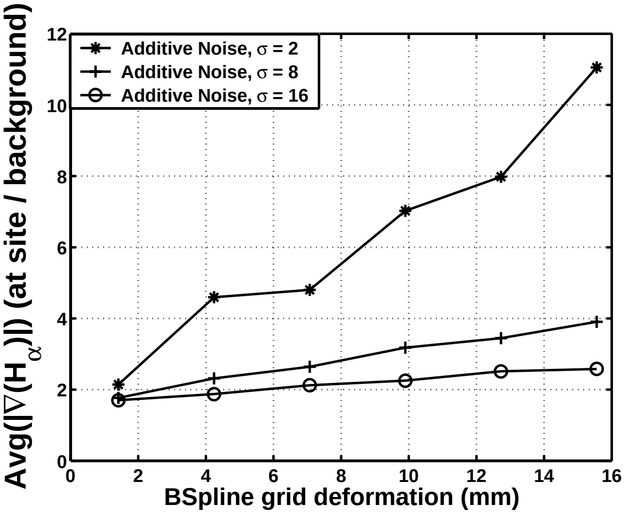


Fig. 24. Ratio of $\nabla(H_\alpha) = \nabla O$ calculated over deformation site v/s background image for smaller deformation spanning $m \times n \geq 8 \times 8$.

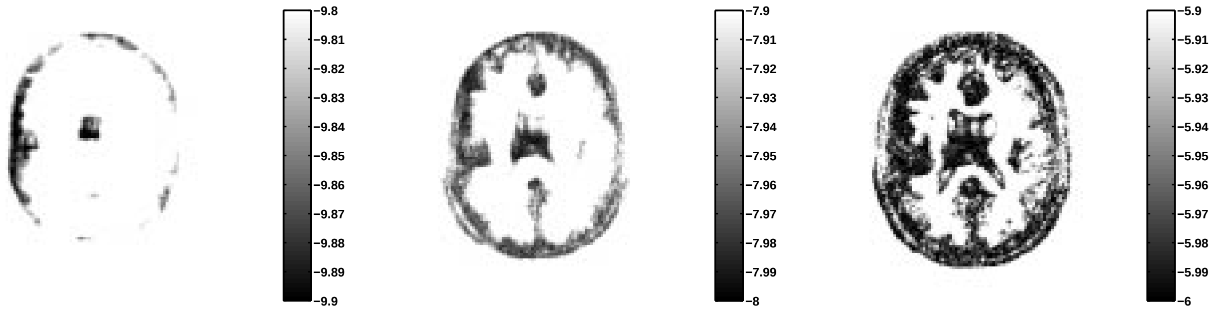
XI. Appendix

Here we give a derivation of the entropic graph estimators of α-GA (28) and α-MI (29) estimators. The derivation is based on the heuristic equivalences (44) and (45) and the convergence properties are, at present, unknown.

First consider estimating $\alpha D_{GA}(f,g) = (\alpha-1)^{-1}\log I_{GA}(f,g)$, where $I_{GA}(f,g)$ is the integral in (15), by $\widehat{\alpha D_{GA}} = (\alpha-1)^{-1}\log\widehat{I_{GA}}$ where:

$$\widehat{I_{GA}} = \frac{1}{m+n}\sum_{i=1}^{m+n}\left(\frac{\hat{f}^p(z_i)\hat{g}^q(z_i)}{\hat{h}(z_i)}\right)^{1-\alpha}. \tag{36}$$

Here $\hat{h}(z)$ is an estimate of the common pdf $pf(z)+qg(z)$ of the i.i.d. pooled unordered sample $\{Z_i\}_{i=1}^{m+n} = \{\{O_i\}_{i=1}^m\{,X_i\}_{i=1}^n\}$, $p = m/(m+n), q = 1-p$, and $\hat{f}$, $\hat{g}$ are estimates of the common densities f, g of the i.i.d. samples $\{O_i\}_{i=1}^m$ and $\{X_i\}_{i=1}^n$, respectively. We assume that the support set of f, g, h is contained in a bounded region $\mathcal{S}$ of $\mathbb{R}^d$ and that $m = \rho n$ for some fixed $\rho > 0$ ($p = \rho/(1+\rho)$). If $\hat{f}$, $\hat{g}$, $\hat{h}$ are consistent, i.e., they converge (a.s.) as $n \to \infty$ to f, g, h then by the strong law of large



(a) Shannon MI $O_{10} = MI(X,Y)$: 32×32

(b) Shannon MI $O_{10} = MI(X,Y)$: 16×16

(c) Shannon MI $O_{10} = MI(X,Y)$: 8×8

Fig. 25. Performance of Shannon MI, computed using pixel intensity histograms, on deformed MRI images: (a) 32×32 window, (b) 16×16 window and (c) 8×8 window.

numbers $\widehat{I_{GA}}$ converges (a.s) to

$$E[\widehat{I}_{GA}] = E\left[\left(\frac{f^p(z_i)g^q(z_i)}{h(z_i)}\right)^{1-\alpha}\right] \tag{37}$$

$$= \int_{\mathcal{S}} \left(\frac{f^p(z)g^q(z)}{h(z)}\right)^{1-\alpha} h(z)dz, \tag{38}$$

Taking the log of expression (38) and dividing by $\alpha - 1$, we obtain $\alpha D_{GA}(f,g)$ in (15) so that $\widehat{\alpha D_{GA}}$ is asymptotically unbiased and its variance goes to zero.

Next divide the samples $\{Z_i\}_{i=1}^{m+n}$ into two disjoint sets of samples Z_{train} and test samples Z_{test}, where we have preserved the relative number $\rho = m/n$ of labels in each of these sets. Using the training sample construct the Voronoi partition density estimators

$$\begin{aligned}\hat{h}(z) &= \frac{\mu(\Pi_z(z))}{\lambda(\Pi_z(z))}\\ \hat{f}(z) &= \frac{\mu(\Pi_o(z))}{\lambda(\Pi_o(z))}\\ \hat{g}(z) &= \frac{\mu(\Pi_x(z))}{\lambda(\Pi_x(z))},\end{aligned} \tag{39}$$

where $\Pi_Z(z)$, $\Pi_O(z)$, $\Pi_X(z)$ are the cells of the Voronoi partition of $\mathcal{S} \in \mathbb{R}^d$ containing the point $z \in \mathbb{R}^d$ and constructed from training samples $Z_{train} \equiv \{O_{train}, X_{train}\}$, O_{train} and X_{train} respectively using

K-means or other algorithm. Here μ and λ are the (normalized) counting measure and Lebesgue measure respectively, i.e $\mu(\Pi)$ is the number of points in the set Π divided by the total number of points and $\lambda(\Pi)$ is the volume of the set Π. Let $\{K_z, K_o, K_x\}$ be the number of cells in the partitions $\{\Pi_z, \Pi_o, \Pi_x\}$ respectively and let n_{train} be the number of training samples. The Voronoi partition density estimators are asymptotically consistent as $k, n_{train} \to \infty$ and $k/n_{train} \to 0$, for $k \in \{K_z, K_o, K_x\}$ [63].

Therefore, under these conditions and defining $\tilde{Z}_i = Z_{test}(i)$,

$$\widehat{\alpha D_{GA}} = \frac{1}{\alpha - 1} \log \left(\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \left(\frac{\hat{f}^p(\tilde{z}_i) \hat{g}^q(\tilde{z}_i)}{\hat{h}(\tilde{z}_i)} \right)^{1-\alpha} \right), \tag{40}$$

is an asymptotically consistent estimator.

We next consider a similar plug-in estimator of α-MI. Assume that the concatenated pair of feature vectors $Z_i = [O_i, X_i] \in \mathbb{R}^{2d}$ are collected, $i = 1, \ldots, n$. The plug-in estimator $\widehat{\alpha MI} = (\alpha - 1) \log \widehat{I_{MI}}$ is constructed, where

$$\widehat{I_{MI}} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\hat{f}_o(o_i) \hat{f}_x(x_i)}{\hat{f}_{ox}(o_i, x_i)} \right)^{1-\alpha}, \tag{41}$$

and $\hat{f}_{ox}$ is an estimate of the joint density of $[O_i, X_i] \in \mathbb{R}^{2d}$, $\hat{f}_o$ and $\hat{f}_x$ are estimates of the marginal densities of O_i and X_i, respectively. Again, if $\hat{f}_o$, $\hat{f}_x$ and $\hat{f}_{ox}$ are consistent then it is easily shown that $\widehat{I_{MI}}$ converges to the integral in the expression (7) for α-MI:

$$\int_{\mathcal{S} \times \mathcal{S}} \int \left(\frac{\hat{f}_o(u) \hat{f}_x(v)}{\hat{f}_{ox}(u, v)} \right)^{1-\alpha} f_{ox}(u, v) du dv, \tag{42}$$

where $\mathcal{S}$ is a bounded set containing the support of densities f_o and f_x. Similarly, separating $\{[O_i, X_i]\}_{i=1}^n$ into training and test samples, we obtain an asymptotically consistent estimator:

$$\widehat{\alpha MI} = \frac{1}{\alpha - 1} \log \left(\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \left(\frac{\hat{f}_o(\tilde{o}_i) \hat{f}_x(\tilde{x}_i)}{\hat{f}_{ox}(\tilde{o}_i, \tilde{x}_i)} \right)^{1-\alpha} \right). \tag{43}$$

The entropic graph αMI estimator (29) is obtained by specializing to the case $n_{train} = 0, n_{test} = n$, in which case $\mu(\Pi_O(z)) = \mu(\Pi_X(z)) = \mu(\Pi_{O \times X}(z)) = 1/n$, and using the Voronoi cell volume approximations

$$\begin{aligned} \lambda(\Pi_X(z_i)) &\asymp e_i^d(x) \\ \lambda(\Pi_Y(z_i)) &\asymp e_i^d(o) \\ \lambda(\Pi_{O\times X}(z_i)) &\asymp e_i^{2d}(o\times x) \end{aligned} \tag{44}$$

where $\asymp$ denotes "proportional to." The quantity $e_i(o\times x)$ is the distance of the closest point to $Z_i = [O_i, X_i] \in \mathbb{R}^{2d}$ in the cartesian product space $O\times X$ and $e_i(o)$, $e_i(x)$ are the componentwise NN distances $\min_{j\neq i}\|O_i - O_j\|$ and $\min_{j\neq i}\|X_i - X_j\|$.

The entropic graph αGA divergence estimator (28) can be similarly obtained. Again specializing to the case $n_{train} = 0, n_{test} = n+m$ we have: $\mu(\Pi_Z(z)) = (m+n)^{-1}$, $\mu(\Pi_O(z)) = 1/m$ and $\mu(\Pi_X(z)) = 1/n$. Making the following Voronoi cell volume approximations

$$\begin{aligned} \lambda(\Pi_X(z_i)) &\asymp e_i^d(x) \\ \lambda(\Pi_Y(z_i)) &\asymp e_i^d(o) \\ \lambda(\Pi_Z(z_i)) &\asymp e_i^d(z) = \min\{e_i^d(o), e_i^d(x)\} \end{aligned} \tag{45}$$

where $e_i(o)$, $e_i(x)$ denote the distances of the point $Z_i \in \mathbb{R}^d$ to the nearest "O" and "X" labeled points, respectively, in the pooled sample $\{Z_i\}_{i\neq j}$ Substitution of these relations into (40) gives

$$\widehat{\alpha D_{GA}} = \frac{1}{\alpha-1}\log\left(\frac{c}{m+n}\sum_{i=1}^{m+n}\min\left\{\left(\frac{e_i(o)}{e_i(x)}\right)^{\gamma q}, \left(\frac{e_i(x)}{e_i(o)}\right)^{\gamma p}\right\}\right), \tag{46}$$

where $c = ((m+n)/(m^p n^q))^{1-\alpha} = ((1+\rho)/\rho^p)^{\gamma/d}$ which reduces to the expression (28) upon specializing to the case $m = n$ ($\rho = 1$).

XII. Acknowledgments

We would like to acknowledge Sakina Zabuawala, EECS graduate student at University of Michigan, for benchmarking the approximate NN algorithm.

References

[1] S. Arya, D. M. Mount, N. S. Netanyahu, R. Silverman, and A. Y. Wu, "An optimal algorithm for approximate nearest neighbor searching fixed dimensions," *Journal of the ACM*, vol. 45, no. 6, pp. 891–923, 1998.

[2] J. Ashley, R. Barber, M. Flickner, D. Lee, W. Niblack, and D. Petkovic, “Automatic and semiautomatic methods for image annotation and retrieval in qbic,” in *Proc. SPIE Storage and Retrieval for Image and Video Databases III*, pp. 24–35, 1995.

[3] R. Baraniuk, P. Flandrin, A. J. E. M. Jensen, and O. Michel, “Measuring time frequency information content using the Rényi entropies,” *IEEE Trans. on Inform. Theory*, vol. IT-47, no. 4, , April 2001.

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

[5] C. F. Bazlamaçci and K. Hindi, “Minimum-weight spanning tree algorithms: A survey and empirical study,” *Computers and Operations Research*, vol. 28, pp. 767–785, 2001.

[6] J. Beardwood, J. H. Halton, and J. M. Hammersley, “The shortest path through many points,” *Proc. Cambridge Philosophical Society*, vol. 55, pp. 299–327, 1959.

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

[8] J. L. Bentley, “Multidimensional binary search trees in database applications,” *IEEE Trans. Software Engineering*, vol. SE-5, no. 4, pp. 333–340, 1979.

[9] J. L. Bentley, “Multidimensional binary search trees used for associative searching,” *Communications of the ACM*, vol. 18, no. 9, pp. 509–517, Sept. 1975.

[10] J. L. Bentley and J. H. Friedman, “Fast algorithms for constructing minimal spanning trees in coordinate spaces,” *IEEE Trans. on Computers*, vol. C-27, no. 2, pp. 97–105, 1978.

[11] T. Butz and J. Thiran, “Affine registration with feature space mututal information,” in *Lecture Notes in Computer Science 2208: MICCAI 2001*, pp. 549–556, Springer-Verlag berlin Heidelberg 2001, 2001.

[12] E. Chávez, G. Navarro, R. Baeza-Yates, and J. L. Marroquín, “Searching in metric spaces,” *ACM Computing Surveys*, vol. 33, no. 3, pp. 273–321, Sept. 2001.

[13] Y. Choi and S. Lee, “Injectivity conditions of 2D and 3D uniform cubic B-Spline functions,” *Graphical Models*, vol. 62, pp. 411–427, 2000.

[14] K. L. Clarkson, “An algorithm for geometric minimum spanning trees requiring nearly linear expected time,” *Algorithmica*, vol. 4, pp. 461–469, 1989.

[15] S. R. Cloude and E. Pottier, “An entropy based classification scheme for land applications of polarimetric SAR,” *IEEE Trans. on Geoscience and Remote Sensing*, vol. 75, pp. pp. 68–78, 1997.

[16] C. A. Cocosco, V. Kollokian, R. K. S. Kwan, and A. C. Evans, “Brainweb: Online interface to a 3D MRI simulated brain database,” *NeuroImage*, vol. 5, no. 4, , 1997.

[17] N. Cristiani and J. Shaw-Taylor, *Suport Vector Machines and other kernel-based learning methods*, Cambridge U. Press, 2000.

[18] I. Csiszár, “Information-type measures of divergence of probability distributions and indirect observations,” *Studia Sci. Math. Hung.*, vol. 2, pp. 299–318, 1967.

[19] J. S. de Bonet and P. Viola, “Structure driven image database retrieval,” in *Advances in neural information processing*, volume 10, 1997.

[20] A. Dembo and O. Zeitouni, *Large deviations techniques and applications*, Springer-Verlag, NY, 1998.

[21] M. N. Do and M. Vetterli, “Texture similarity measurement using Kullback-Liebler distance on wavelet subbands,” in *IEEE Int. Conf. on Image Processing*, pp. 367–370, Vancouver, BC, 2000.

[22] D. Dunn, W. Higgins, and J. Wakeley, “Texture segmentation using 2d gabor elementary functions,” *IEEE Trans. Pattern Anal. Mach. Intelligence*, vol. 16, no. 2, pp. 130–149, 1994.

[23] Equinox Corporation. *Human Identification at Distance Project*.

[24] Y. Erdi, K. Rosenzweig, A. Erdi, H. Macapinlac, Y. Hu, L. Braban, J. Humm, O. Squire, C. Chui, S. Larson, and E. Yorke, “Radiotherapy treatment planning for patients with non-small cell lung cancer using pet,” *Radiotherapy and Oncology*, vol. 62, no. 1, pp. 51–60, 2002.

[25] V. Erdogmus, J. Prncipe, and L. Vielva, “Blind deconvolution with minimum rényi’s entropy,” in *EUSIPCO*, Toulouse, France, 2002.

[26] B. Frieden and A. T. Bajkova, “Reconstruction of complex signals using minimum rényi information,” in *Proc. of Meeting of Intl. Soc. for Optical Engin. (SPIE)*, volume 2298, 1994.

[27] J. H. Friedman and L. C. Rafsky, “Multivariate generalizations of the Wald-Wolfowitz and Smirnov two-sample tests,” *Annals of Statistics*, vol. 7, no. 4, pp. 697–717, 1979.

[28] S. Gilles, “Description and experimentation of image matching using mutual information,” Technical report, Oxford University, 1996. `www-rocq.inria.fr/˜gilles/IMMMI/mutual_info.ps.gz`.

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

[30] N. Henze and M. Penrose, “On the multivariate runs test,” *Annals of Statistics*, vol. 27, pp. 290–298, 1999.

[31] A. O. Hero, J. Costa, and B. Ma, “Asymptotic relations between minimal graphs and alpha entropy,” Technical Report 334, Comm. and Sig. Proc. Lab. (CSPL), Dept. EECS, University of Michigan, Ann Arbor, Mar, 2003. `www.eecs.umich.edu/˜hero/det_est.html`.

[32] A. O. Hero, B. Ma, and O. Michel, “Imaging applications of stochastic minimal graphs,” in *IEEE Int. Conf. on Image Processing*, Thessaloniki, Greece, October 2001.

[33] A. O. Hero, B. Ma, O. Michel, and J. D. Gorman, “Alpha-divergence for classification, indexing and retrieval,” Technical Report 328, Comm. and Sig. Proc. Lab. (CSPL), Dept. EECS, University of Michigan, Ann Arbor, July, 2001. `www.eecs.umich.edu/˜hero/det_est.html`.

[34] A. 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`.

[35] A. 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`.

[36] A. 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.

[37] D. Hill, P. Batchelor, M. Holden, and D. Hawkes, “Medical image registration,” *Phys. Med. Biol.*, vol. 26, pp. R1–R45, 2001.

[38] R. Hoffman and A. K. Jain, “A test of randomness based on the minimal spanning tree,” *Pattern Recognition Letters*, vol. 1, pp. 175–180, 1983.

[39] J. Huang, S. Kumar, M. Mitra, and W. Zhu, "Spatial color indexing and applications," in *Proc. of IEEE Int'l Conf. Computer Vision ICCV'98*, pp. 602–608, Bombay, India, 1998.

[40] A. Hyvärinen. *Fast ICA Code*. `www.cis.hut.fi/projects/ica/fastica/`.

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

[42] M. Jenkinson, P. Bannister, M. Brady, and S. Smith, "Improved methods for the registration and motion correction of brain images," Technical report, Oxford University, 2002.

[43] K. Johnson, A. Cole-Rhodes, I. Zavorin, and J. L. Moigne, "Multi-resolution image registration of remotely sensed imagery using mutual information," in *Proc. of SPIE OE/Aerospace Sensing, Wavelet Applications VIII*, Orlando, FL, 2001.

[44] T. Kieu and P. Viola, "Boosting image retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2000.

[45] J. B. Kruskal, "On the shortest spanning subtree of a graph and the traveling salesman problem," *Proc. Amer. Math. Soc.*, vol. 7, pp. 48–50, 1956.

[46] S. Kullback and R. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, pp. 79–86, 1951.

[47] 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.

[48] M. E. Leventon and W. E. L. Grimson, "Multi-modal volume registration using joint intensity distributions," Technical report, MIT AI Laboratory, 1998. `www.ai.mit.edu/projects/vision-surgery`.

[49] M. Lewicki and B. Olshausen, "Probabilistic framework for the adaptation and comparison of image codes," *J. Opt. Soc. Am.*, vol. 16, no. 7, pp. 1587–1601, 1999.

[50] W. Ma and B. Manjunath, "Netra: A toolbox for navigating large image databased," in *Proc. of IEEE Int. Conf. on Image Processing*, volume 1, pp. 568–571, 1997.

[51] F. Maes, D. Vandermeulen, and P. Suetens, "Medical image registration using mutual information," *Proceedings of the IEEE*, vol. 91, no. 10, pp. 1699–1722, 2003.

[52] J. B. Maintz and M. Viergever, "A survey of medical image registration," *Medical Image Analysis*, vol. 2, no. 1, pp. 1–36, 1998.

[53] 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.

[54] O. Michel, R. Baraniuk, and P. Flandrin, "Time-frequency based distance and divergence measures," in *IEEE International time-frequency and Time-Scale Analysis Symposium*, pp. 64–67, Oct 1994.

[55] D. Mount and S. Arya. *Approximate Nearest Neighbor Code*. `http://www.cs.umd.edu/˜mount/ANN`.

[56] G. Narasimhan, J. Zhu, and M. Zachariasen, "Experiments with computing geometric minimum spanning trees," in *Proc. of Second Workshop on Algorithm Engineering and Experiments*, pp. 183–196, 2000.

[57] NASA Visible Earth internet site.

[58] H. Neemuchwala, A. Hero, and P. Carson, "Image registration using entropic graph matching criteria," in *Proc. of Asilomar SS&C Conference*, Monterey, CA, November 2002.

[59] H. Neemuchwala, A. Hero, P. Carson, and C. Meyer, “Local feature matching using entropic graphs,” in *Proc. of the IEEE International Symposium on Biomedical Imaging*, Washington, DC, April, 2004.

[60] 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.

[61] H. Neemuchwala, A. Hero, and P. Carson, “Feature coincidence trees for registration of ultrasound breast images,” in *IEEE Int. Conf. on Image Processing*, Thesaloniki, Greece, October 2001.

[62] S. A. Nene and S. K. Nayar, “A simple algorithm for nearest neighbor search in high dimensions,” *IEEE Trans. on Pattern Anal. and Machine Intell.*, vol. 19, , 1997.

[63] A. Nobel and R. Olshen, “Termination and continuity of greedy growing for tree-structured vector quantizers,” *IEEE Trans. on Inform. Theory*, vol. IT-42, no. 1, pp. 191–205, 1996.

[64] B. Olshausen, *Sparse codes and spikes*, MIT Press, 2001.

[65] H. Park and C. Meyer, “Grid refinement in adaptive non-rigid registration,” *Lecture Notes in Computer Science, MICCAI 2003 (to appear)*, 2003.

[66] C. Penney, J. Weese, J. Little, D. Hill, and D. Hawkes, “A comparison of similarity measures for used in 2-D-3-D medical image registration,” *IEEE Trans. on Medical Imaging*, vol. 17, no. 4, pp. 586–595, 1998.

[67] R. C. Prim, “Shortest connection networks and some generalizations,” *Bell Syst. Tech. Journ.*, vol. 36, pp. 1389–1401, 1957.

[68] Project Atlanta.

[69] 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*, volume III, Oct. 1998.

[70] 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.

[71] A. Rényi, “On measures of entropy and information,” in *Proc. 4th Berkeley Symp. Math. Stat. and Prob.*, volume 1, pp. 547–561, 1961.

[72] T. Rohlfing, J. West, J. Beier, T. Liebig, C. Tachner, and U. Thornale, “Registration of functional and anatomical mri: accuracy, assessment and applications in navigated neurosurgery,” *Computer Aided Surgery*, vol. 5, no. 6, pp. 414–425, 2000.

[73] D. Rueckert, M. Clarkson, D. Hill, and D. Hawkes, “Non-rigid registration using higher order mutual information,” in *Proc. SPIE*, volume 3979, pp. 438–447, 2000.

[74] A. Srivastava, A. B. Lee, E. P. Simoncelli, and S. C. Zhu, “On advances in statistical modeling of natural images,” *Journal of Mathematical Imaging and Vision*, vol. 18, no. 1, , Jan. 2003.

[75] J. M. Steele, *Probability theory and combinatorial optimization*, volume 69 of *CBMF-NSF regional conferences in applied mathematics*, Society for Industrial and Applied Mathematics (SIAM), 1997.

[76] R. Stoica, J. Zerubia, and J. M. Francos, “Image retrieval and indexing: A hierarchical approach in computing the distance between textured images,” in *IEEE Int. Conf. on Image Processing*, Chicago, Oct. 1998.

[77] R. Stoica, J. Zerubia, and J. M. Francos, “The two-dimensional wold decomposition for segmentation and indexing in image libraries,” in *Proc. IEEE Int. Conf. Acoust., Speech, and Sig. Proc.*, Seattle, May 1998.

[78] H. Stone, J. L. Moigne, and M. McGuire, “The translation sensitivity of wavelet-based registration,” *IEEE Trans. Pattern Anal. Mach. Intelligence*, vol. 21, no. 10, pp. 1074–1081, 1999.

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

[80] A. Toga, *Brain Warping*, Academic Press, ISBN: 0126925356, 1999.

[81] N. Vasconcelos and A. Lippman, “A Bayesian framework for content-based indexing and retrieval,” in *IEEE Data Compression Conference*, Snowbird, Utah, 1998. `nuno.www.media.mit.edu/people/nuno/`.

[82] N. Vasconcelos and A. Lippman, “Bayesian representations and learning mechanisms for content based image retrieval,” in *SPIE Storage and Retrieval for Media Databases 2000*, San Jose, CA, 2000. `nuno.www.media.mit.edu/people/nuno/`.

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

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

[85] W. Wald and J. Wolfowitz, “On a test whether two samples are from the same population,” *Ann. Math. Statist.*, vol. 11, pp. 147–162, 1940.

[86] W. J. Williams, M. L. Brown, and A. O. Hero, “Uncertainty, information, and time-frequency distributions,” in *Proc. of Meeting of Intl. Soc. for Optical Engin. (SPIE)*, volume 1566, pp. 144–156, 1991.

[87] Y. Wu, T. Kanade, C. Li, and J. Cohn, “Image registration using wavelet-based motion model,” *International Journal of Computer Vision*, vol. 38, no. 2, pp. 129–152, 2000.

[88] J. E. Yukich, *Probability theory of classical Euclidean optimization*, volume 1675 of *Lecture Notes in Mathematics*, Springer-Verlag, Berlin, 1998.