Using Spanning Graphs for Efficient Image Registration

Mert R. Sabuncu, *Member, IEEE*, and Peter Ramadge, *Fellow, IEEE*

***Abstract*—We provide a detailed analysis of the use of minimal spanning graphs as an alignment method for registering multimodal images. This yields an efficient graph theoretic algorithm that, for the first time, jointly estimates both an alignment measure and a viable descent direction with respect to a parameterized class of spatial transformations. We also show how prior information about the interimage modality relationship from prealigned image pairs can be incorporated into the graph-based algorithm. A comparison of the graph theoretic alignment measure is provided with more traditional measures based on plug-in entropy estimators. This highlights previously unrecognized similarities between these two registration methods. Our analysis gives additional insight into the tradeoffs the graph-based algorithm is making and how these will manifest themselves in the registration algorithm's performance.**

***Index Terms*—Entropy, estimation, image registration.**

I. Introduction and Prior Work

IMAGE registration is the process of bringing images into spatial alignment. This is often a critical precurser to fusing information and identifying dependence. If the images are obtained through different sensing modalities, the problem is called multimodal. There are two forms of uncertainty involved in this form of registration: uncertainty in the modal relationship and uncertainty in the spatial alignment. Registration attempts to reduce uncertainty in the spatial alignment and in the process also reduces the uncertainty in the modal relationship. The wide range of registration applications has resulted in many proposed solutions; see, for example, [10] and [18].

Generally speaking, a registration method comprises of three coupled components: an alignment measure that quantifies the quality of spatial alignment, a group of spatial transformations that defines the possible alignments, and an optimization scheme that searches the group for the transformation that optimizes the alignment measure.

Our work focuses on an alignment measure based on certain minimal graphs. This approach is in turn motivated by the wide use of entropy and related quantities as multimodal image alignment measures [7], [21]. Mutual information of pixel intensity values has been the most popular entropic measure [26], [17]. Variations of this measure, such as normalized mutual information [24], mutual information of gradient-based features [3], and mutual information combined with a gradient term [20], have also been studied. More recently, Rényi entropy has also been applied to image registration [1], [11], [19], [23], [27], [28]. For a detailed survey of information-theoretic image registration approaches, see [21].

Motivated by the use of entropic alignment measures, in [12], Hero *et al.* propose to use minimal graphs to obtain a direct estimate of Rényi entropy and define an alignment measure. However, it is not clear how to efficiently search over the transformation space for the optimum of such graph-based measures, since they are nondifferentiable.

One of the main contributions of this paper is an analysis of minimal graph alignment measures and the joint determination of both the alignment measure and a descent direction with respect to alignment parameters. In particular, we show how to efficiently use this descent direction for fast optimization.

To benchmark our graph theoretic registration method, we provide a comparison with image registration methods based on plug-in estimators of entropy. In a "plug-in" entropy estimator [2], the probability distribution underlying the data is estimated [8] (e.g., using Parzen windowing or histograms) and used in the entropy formula to obtain an entropy estimate. Provided a suitable density estimator is employed, it is relatively straightforward using this method to also obtain the gradient of the entropy estimate with respect to a vector of registration parameters.

We obtain closed-form expressions for a descent direction of the minimal entropic graph estimator and compare this with those for the gradient of plug-in methods. This approach leads to some interesting insights on how the different estimators weight the data and some predictions of likely performance.

An additional problem in multimodal registration is how to incorporate prior knowledge about the modal relationship. This prior knowledge may take the form of aligned image pairs from earlier in a sequence of registration problems or may come from a set of training examples. Leventon *et al.* [15] proposed estimating the underlying joint prior intensity distribution of registered image pairs using training data and then employing a maximum likelihood (ML) approach to define the alignment measure for new image pairs. Subsequently, Chung *et al.* [5] proposed to measure the quality of registration using the Kullback–Leibler (KL) divergence between the joint intensity distribution of prealigned data (training pdf) and of the new images (test pdf). Registration is then accomplished by minimizing this KL divergence, which was shown to be superior to [15]'s ML

Manuscript received June 15, 2007; revised January 15, 2008. This research was supported by in part by Siemens Corporate Research, Princeton, NJ. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Peyman Milanfar.

M. R. Sabuncu is with the CSAIL, Massachusetts Institute of Technology, Cambridge, MA 02139 USA (e-mail: msabuncu@csail.mit.edu).

P. Ramadge is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: ramadge@princeton.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2008.918951

approach. The KL divergence approach was recently extended to nonrigid image registration in [9]. Alternatively, [16] proposes to employ the Jensen–Shannon (JS) divergence to quantify the discrepancy between the training and test pdfs. These studies indicate experimentally that incorporating prior information can produce a registration function with a wider basin of attraction, making the algorithm more robust to initialization and reducing the time required to achieve registration.

Our second main contribution is to incorporate prior information within the graph theoretic image registration framework. Our method employs the Jensen–Rényi (JR) divergence [27] to define an alignment measure based on the discrepancy between the training and test pdfs. This is similar to the KL and JS divergence approaches of [5], [9], and [16]. We then estimate the JR divergence-based alignment measure using an entropic graph. Interestingly, this measure is a natural extension of the original entropic graph measure that only uses the test images.

In summary, our main contributions are to explore and expand the use of minimal entropic graphs as an alignment measure, to develop efficient descent-based optimization methods for these minimum entropic graphs, to compare this method both analytically and experimentally with popular plug-in estimators, and to develop methods for incorporating prior information into the entropic graph-based registration framework.

II. Background: Entropy and Image Registration

The basic idea behind the use of entropic measures for multimodal image alignment is that when two images, $I_1, I_2 : \mathbb{R}^d \mapsto \mathbb{R}$ (where d is typically 2 or 3), are spatially aligned, there is a strong statistical dependence between corresponding pixel values. This is captured, for example, in the joint entropy (or a related measure, e.g., mutual information) of the *paired* pixel process (I_1, I_2). Thus, image registration can be posed as the following optimization problem:

$$\Phi^* = \arg\min_{\Phi} \hat{H}(I_1(\mathbf{x}), I_2(\Phi(\mathbf{x}))) \tag{1}$$

where $\Phi : \mathbb{R}^d \mapsto \mathbb{R}^d$ is a spatial transformation and $\hat{H}$ is an estimate of an entropic measure, e.g., Shannon's entropy. Note that this formulation requires the joint entropy of the paired process to be estimated from the samples of $(I_1(\mathbf{x}), I_2(\Phi(\mathbf{x})))$ given by the current alignment of the images determined by $\Phi(\cdot)$.

In this paper, we explore different estimators of the α-Rényi entropy [defined in the following section in (3)] as alignment measures. This is similar to Shannon's entropy-based approaches, such as mutual information. In these algorithms nonparametric estimation is usually the weapon of choice. The so-called "plug-in" entropy estimator [2] is the most popular technique and can be straightforwardly applied to estimate various information theoretic quantities. It is a two-step process based on estimating the pdf of the observed samples (using e.g., Parzen windowing or histograms) and then plugging this estimate into the expression for the entropic measure [21]. The second step requires the evaluation of an expectation, which is usually achieved with an approximation.

Alternatively, an estimate of an entopic measure can be obtained by using so-called entropic graphs [12]. These estimates are based on computing a minimal graph on a set of samples. A monotonic function of the total edge length of the minimal graph then provides a direct estimate of the underlying Rényi entropy.

In the following subsections, we provide some additional background on these two forms of entropy estimators.

A. Plug-In Estimators

Let $S \in \mathbb{R}^2$ be a random variable with the pdf $p(\mathbf{s})$ and let $\mathcal{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_N\}$ be a set of independent samples of S. Our goal is to estimate its entropy given the set of samples $\mathcal{S}$.

We first use a Parzen-window estimator [8] to estimate the density of S from its samples

$$\hat{p}(\mathbf{s}; \mathcal{S}) = \frac{1}{N} \sum_{i=1}^{N} K(\mathbf{s} - \mathbf{s}_i) \tag{2}$$

where $K: \mathbb{R}^2 \mapsto \mathbb{R}$ is a continuous density. This estimator corresponds to using a "blurred" histogram as an estimate of the underlying pdf. Note that additional conditions on K determine the convergence rate of this estimator and in practice most kernel functions are selected to be symmetric, i.e., $K(\mathbf{s}) = K(-\mathbf{s})$.

Using the above density estimate, we then evaluate the expression for the desired entropic quantity, which requires the evaluation of an expectation. Two approaches have been used to compute this expectation. The first uses a sample mean and was employed by Viola *et al.* in [26]. For example, the α-Rényi entropy of S is defined as

$$H_\alpha(p) = H_\alpha(S) = \frac{1}{1-\alpha} \log \mathbb{E}_p(p^{\alpha-1}(s)) \tag{3}$$

where $\alpha > 0$ and $\mathbb{E}_p$ denotes expectation. This formula can be expressed in terms of the α-information potential

$$V_\alpha(p) = \mathbb{E}_p(p^{\alpha-1}(s)) \tag{4}$$

as $H_\alpha(S) = (1/1-\alpha)\log V_\alpha(p)$. An estimate of the α-information potential can be obtained from i.i.d. samples of S using a Parzen-window estimator of the density $p(\cdot)$ and a sample mean approximation of the expectation. This yields

$$\begin{aligned} \hat{V}_M(\mathcal{S}, \alpha) &= \frac{1}{N} \sum_{i=1}^{N} \hat{p}^{\alpha-1}(\mathbf{s}_i; \mathcal{S}) \\ &= \frac{1}{N^\alpha} \sum_{i=1}^{N} \left(\sum_{j=1}^{N} K(\mathbf{s}_i - \mathbf{s}_j) \right)^{\alpha-1}. \end{aligned} \tag{5}$$

We have chosen this particular example because it will be useful in our later analysis.

The second approach to evaluating the expectation, a histogram-based method, approximates the infinite integral in the expectation using a finite sum [25]. If we ignore the nonlinearity introduced by "binning" and impose some conditions on K, the histogram-based estimate can be thought of as an approximation of the sample mean estimate. For example, in the case of estimating the α information potential, let $\hat{V}_H(\cdot)$ denote the histogram-based estimate and $\hat{V}_M(\cdot)$ denote the sample mean estimate. Then $\hat{V}_H(\mathcal{S}, \alpha) \approx \hat{V}_M(q(\mathcal{S}), \alpha)$, where $q(\cdot)$ is a quantizer.

Suppose the class of spatial transformations $\{\Phi_{\mathbf{t}}\}$ to be used for registration is parameterized by $\mathbf{t} = (t_1, \ldots, t_T) \in \mathbb{R}^T$ and let

$$\mathcal{S}_{\mathbf{t}} \triangleq \{(I_1(\mathbf{x}), I_2(\Phi_{\mathbf{t}}(\mathbf{x}))) : \mathbf{x} \in \Omega\} \tag{6}$$

where Ω is a finite subset of $\mathbb{R}^d$. Based on (1), image registration can be formulated as

$$\mathbf{t}^* = \arg\min_{\mathbf{t}} \hat{V}_M(\mathcal{S}_{\mathbf{t}}, \alpha) \tag{7}$$

for some $\alpha \in (0, 1)$. This optimization can be solved using a gradient-descent type strategy, where the derivative of the alignment measure with respect to the transformation parameters $\mathbf{t}$ is computed.

For a vector $\mathbf{v}$, let v_i denote the ith component of $\mathbf{v}$ and $\nabla_{\mathbf{v}}$ denote the gradient w.r.t. $\mathbf{v}$. Using the chain rule, the gradient of a similarity measure can be written in the following form:

$$\nabla_{\mathbf{t}} \hat{V}(\mathcal{S}, \alpha) = \sum_{\mathbf{s}_j \in \mathcal{S}} [\nabla_{\mathbf{s}_j} \hat{V}(\mathcal{S}, \alpha)]' [\nabla_{\mathbf{t}} \mathbf{s}_j] \tag{8}$$

where $'$ denotes transposition. The first term in the summation is a 2-D gradient vector of the alignment measure with respect to sample values. The second term is the $2 \times T$-dimensional Jacobian matrix of the sample value with respect to the transformation parameters. Its value depends on the images, the interpolation method and the geometric transformation, but not on the alignment measure. Hence, the first term is of particular interest when comparing different alignment measures.

An advantage of the sample mean plug-in estimator is that it is readily differentiated. The gradient of (5) with respect to $\mathbf{s}_j$ can be written as

$$\nabla_{\mathbf{s}_j} \hat{V}_M(\mathcal{S}, \alpha) = (\alpha - 1) \sum_{k \neq j} n_M(\mathcal{S}, \alpha, j, k) \times \mathbf{f}_M(\mathbf{s}_j, \mathbf{s}_k) \tag{9}$$

where

$$n_M(\mathcal{S}, \alpha, j, k) = N^{-2}(\hat{p}(\mathbf{s}_j)^{\alpha-2} + \hat{p}(\mathbf{s}_k)^{\alpha-2}) \tag{10}$$

and

$$\mathbf{f}_M(\mathbf{s}_j, \mathbf{s}_k) = \nabla K(\mathbf{s}_j - \mathbf{s}_k). \tag{11}$$

B. Entropic Graph Estimators

Let $G = (E, \mathcal{S})$ be a graph with the finite vertex set $\mathcal{S} \subset \mathbb{R}^2$ and edge set E. Each edge $e = (\mathbf{s}_1, \mathbf{s}_2) \in E$ has Euclidean length $\|e\| = \|\mathbf{s}_1 - \mathbf{s}_2\|$. For $\gamma \in \mathbb{R}$, the γ-weight of G is $W_\gamma(G) \triangleq \sum_{e \in E} \|e\|^\gamma$.

Let $\mathcal{G}_C(\mathcal{S})$ denote the family of graphs conforming to a specified topological constraint C and having the common vertex set $\mathcal{S}$. $\mathcal{G}_C(\mathcal{S})$ might, for example, be all spanning trees, all k-neighbor graphs, all TSP graphs, etc. We assume that C is fixed and will not explicitly indicate it henceforth.

For a fixed family of graphs $\mathcal{G}(\mathcal{S})$, define the minimal γ-weight of $\mathcal{G}(\mathcal{S})$ to be

$$W_\gamma^*(\mathcal{S}) = \min_{G \in \mathcal{G}(\mathcal{S})} W_\gamma(G) \tag{12}$$

and let $G^*(\mathcal{S}) = \arg\min_{G \in \mathcal{G}(\mathcal{S})} W_\gamma(G)$ denote a graph in $\mathcal{G}(\mathcal{S})$ of minimal γ-weight. Note that $G^*(\mathcal{S})$ may not be unique.

These constructions are of interest since they yield an estimate of α-Rényi entropy. Let S be a random variable taking values in $[0,1]^2$ with Lebesgue density $p(\mathbf{s})$ and let $\mathcal{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_N\}$ be a set of independent samples of S. Let $\gamma = 2(1-\alpha)$ and set

$$\tilde{H}_\alpha(S) = \frac{1}{1-\alpha} \log \left(\frac{W_{2(1-\alpha)}^*(\mathcal{S})}{N^\alpha} \right). \tag{13}$$

Hero *et al.* [13] show that, for all $\alpha \in (0,1)$, $\tilde{H}_\alpha(S) + \log \beta / (1-\alpha)$ is a strongly consistent estimator of $H_\alpha(S)$, where β is a constant that depends on the topological constraint C, and the parameter α, but *not* on p_S. These results are based on the general framework developed in [22], which provides convergence results for some Euclidean length functionals of specific graphs.

A corresponding graph-theoretic estimate of the α information potential is

$$\hat{V}_G(\mathcal{S}, \alpha) = \beta \frac{W_{2(1-\alpha)}^*(\mathcal{S})}{N^\alpha}. \tag{14}$$

Now we come to an important point: the entropic graph estimator is not differentiable (see Lemma 2 in appendix for details). This was thought to be a major disadvantage of using these estimators for image registration. We can illustrate the nondifferentiability using spanning trees as the topological constraint. Consider the vertex set $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3\}$ with edges and parameterized lengths: $\|e_{12}\| = 2 - t, \|e_{23}\| = t + 2$ and $\|e_{13}\| = 1$. It is easy to show that at $t = 0^-$, the MST consists of e_{23} and e_{13}, whereas at $t = 0^+, e_{12}$ and e_{13} belong to the MST. Thus, $dW(0^-)/dt = 1$ and $dW(0^+)/dt = -1$. Since the left and right derivatives are not equal, the derivative of the MST weight does not exist at $t = 0$. This is the starting point of our investigation.

III. Descent Direction for the Entropic Graph Estimator

In this section, we show that the gradient of the γ-weight of any of the minimal graphs $W_\gamma(G^*)$ is a descent direction for the γ-weight W_γ^* given by (12). This result can be used to efficiently optimize the entropic graph estimator.

Let us consider a family of entropic graphs on these points (e.g., MSTs) and find the minimal weight graph $G^*(E_{\mathbf{t}}^*, \mathcal{S}_{\mathbf{t}})$ in this class. Similar to (7) and based on (1) and (14), we can employ the corresponding weight $W_{\gamma,\mathbf{t}}^* \triangleq W_\gamma^*(\mathcal{S}_{\mathbf{t}})$ as an alignment measure and formulate registration as

$$\mathbf{t}^* = \arg\min_{\mathbf{t}} W_{\gamma,\mathbf{t}}^* \tag{15}$$

for some γ, where γ and α are related as $\gamma = 2(1-\alpha)$.

We assume (for simplicity of notation) that the cardinality of the set $S_{\mathbf{t}}$ does not depend on $\mathbf{t}$. Hence, as we change $\mathbf{t}$, the points in $S_{\mathbf{t}}$ move around in $[0,1]^2$ and, in general, the topology of a minimal graph changes as does its weight. The fact that the topology of a minimal graph changes is precisely what leads to the problem of nondifferentiability.

Suppose that we have $\mathcal{S}_{\mathbf{t}_0}$, a minimal graph $G^*(E^*_{\mathbf{t}_0}, \mathcal{S}_{\mathbf{t}_0})$ and its weight $W^*_{\gamma,\mathbf{t}_0}$ at alignment $\mathbf{t}_0$. Now let us fix the graph G^* and consider the gradient of the weight of G^* with respect to $\mathbf{t}$ at $\mathbf{t}_0$. This yields the following result.

Theorem 1: Let $\mathbf{u} \in \mathbb{R}^T$ be a unit vector that satisfies

$$\sum_{e \in E^*_{\mathbf{t}_0}} [\nabla_{\mathbf{t}} \|e(\mathbf{t}_0)\|^\gamma]' \mathbf{u} < 0. \qquad (16)$$

Then there exists $\epsilon > 0$ such that $W^*_\gamma(\mathbf{t}_0 + h\mathbf{u}) \leq W^*_\gamma(\mathbf{t}_0)$ for all $0 \leq h \leq \epsilon$.

Proof: If (16) exists and is negative, by vector calculus there exists $\epsilon > 0$ such that

$$\sum_{e \in E^*_{\mathbf{t}_0}} \|e(\mathbf{t}_0 + h\mathbf{u})\|^\gamma \leq \sum_{e \in E^*_{\mathbf{t}_0}} \|e(\mathbf{t}_0)\|^\gamma = W^*_\gamma(\mathbf{t}_0) \qquad (17)$$

for all $0 \leq h \leq \epsilon$. By definition, we have

$$W^*_\gamma(\mathbf{t}_0 + h\mathbf{u}) \leq \sum_{e \in E^*_{\mathbf{t}_0}} \|e(\mathbf{t}_0 + h\mathbf{u})\|^\gamma. \qquad (18)$$

Hence, combining (17) and (18), we get $W^*_\gamma(\mathbf{t}_0 + h\mathbf{u}) \leq W^*_\gamma(\mathbf{t}_0)$. ■

Choose a minimal graph $G^*(E^*_{\mathbf{t}_0}, \mathcal{S}_{\mathbf{t}_0})$. Define

$$\begin{aligned}\mathbf{d}_\gamma(G^*(E^*_{\mathbf{t}_0}, \mathcal{S}_{\mathbf{t}_0})) &= -\nabla_{\mathbf{t}} W_\gamma(G^*(E^*_{\mathbf{t}_0}, \mathcal{S}_{\mathbf{t}_0})) \\ &= -\sum_{e \in E^*_{\mathbf{t}_0}} \nabla_{\mathbf{t}} \|e(\mathbf{t}_0)\|^\gamma \end{aligned} \qquad (19)$$

the steepest descent direction for the chosen $W_\gamma(G^*)$. It is easy to see that, when nonzero and finite, $\mathbf{d}_\gamma / \|\mathbf{d}_\gamma\|$ satisfies the condition in (16) and, therefore, is a descent direction for W^*_γ. Note that, if zero length edges exist, i.e., some sample values coincide, and $\gamma < 1$, then (19) does not exist and (16) is never satisfied. In practice, the direction we choose for this problematic case is

$$\bar{\mathbf{d}}_\gamma(G^*(E^*_{\mathbf{t}_0}, \mathcal{S}_{\mathbf{t}_0})) \triangleq -\sum_{e \in E^*_{\mathbf{t}_0}, \|e\| \neq 0} \nabla_{\mathbf{t}} \|e(\mathbf{t}_0)\|^\gamma \qquad (20)$$

which is the steepest descent direction for the graph that excludes the zero-length edges, i.e., the minimal graph on the unique samples. Note that $\bar{\mathbf{d}}_\gamma = \mathbf{d}_\gamma$, when $\mathbf{d}_\gamma$ exists and is finite.

More complex schemes for finding a descent direction are also possible, e.g., selecting several minimal graphs G^* and averaging the corresponding descent directions. However, we focus our analysis on the descent direction obtained from one of the minimal entropic graphs G^*. Correspondingly, for a fixed $G^*(\mathcal{S})$, we define the pseudo-gradient, $\mathbf{g}_j(G^*(\mathcal{S}))$, of the entropic graph estimate of the α-information potential, $\hat{V}_G(\mathcal{S})$ (14) w.r.t. $\mathbf{s}_j$ as

$$\mathbf{g}_j(G^*(\mathcal{S})) \triangleq (\alpha - 1) \sum_{\mathbf{s}_k \in \mathcal{S}} n_G(\mathcal{S}, \alpha, j, k) \mathbf{f}_G(\mathbf{s}_j, \mathbf{s}_k) \qquad (21)$$

where

$$n_G(\mathcal{S}, \alpha, j, k) = \frac{\beta}{N^\alpha} A(G^*(\mathcal{S}))(j, k) \qquad (22)$$

is the network weight and

$$\mathbf{f}_G(\mathbf{s}_j, \mathbf{s}_k) = \begin{cases} 2\|\mathbf{s}_j - \mathbf{s}_k\|^{-2\alpha}(\mathbf{s}_j - \mathbf{s}_k), & \text{if } \|\mathbf{s}_j - \mathbf{s}_k\| > 0 \\ 0, & \text{else} \end{cases} \qquad (23)$$

is the sample pair attraction. $A(G)$ is the adjacency matrix of the graph G, which contains the topology information. The (i, j)th entry $A(G)(i, j)$ is the number of edges connecting vertices i and j. For example, if samples $\mathbf{s}_i$ and $\mathbf{s}_j$ are connected in the minimum spanning tree (MST) of $\mathcal{S} = \{\mathbf{s}_i\}$, then $A(G)(i, j) = 1$, since in an MST there is only one edge between connected vertices. Note, we have put both the gradient of the sample mean plug-in estimator (9)–(11) and descent direction (pseudo-gradient) of the entropic graph estimator (21)–(23) into a common comparative form involving a pairwise attraction and a corresponding network weight.

IV. Comparison: Entropic Graph Estimator Versus Plug-In Estimator

In this section, we compare the "gradient" expressions for the two entropy estimators. This analysis provides some useful insights on the performance of registration algorithms formulated similar to (7) and (15) which use entropy estimators and their "gradients." The optimization problem is typically solved with an iterative descent scheme. The transformation parameters are updated as $\mathbf{t}_{m+1} = \mathbf{t}_m + \lambda_m \sum_j (\sum_k n_{jk} \mathbf{f}_{jk}) \nabla_{\mathbf{t}} \mathbf{s}^m_j$, where λ_m is a step size, $\mathbf{f}_{jk}$ is the sample pair attraction, n_{jk} is the network weight, $\nabla_{\mathbf{t}} \mathbf{s}^m_j$ is the gradient of the jth sample w.r.t the transformation parameters, and $\mathbf{t}_m$ is the value of $\mathbf{t}$ at the mth iteration. n_{jk} and $\mathbf{f}_{jk}$ are summarized for the two entropy estimators in Table I. Their product represents the influence of this sample pair interaction on the total gradient. In the remainder, we use $\mathcal{S}_{\mathbf{t}}$ as defined in (6), i.e., a set of intensity sample pairs from the two images at an alignment determined by the transformation parameters $\mathbf{t} \in \mathbb{R}^T$. Also, to keep the discussion clear, we make the following practical assumptions.

- The kernel used for the plug-in estimator is a 2-D separable Gaussian, $G_\sigma(x, y) = g_\sigma(x) g_\sigma(y)$, where $g_\sigma(\cdot)$ is a zero mean Gaussian with variance σ^2.
- The family of spanning tree graphs is used to compute a minimal entropic graph. Note that an MST has a 0–1 adjacency matrix, i.e., $A(G^*)(i, j) = 0, 1$ for all i, j.

A. Plug-In Estimator

First, let us consider the sample mean plug-in estimator. The computation time of the estimator and its gradient is $\mathcal{O}(N^2)$, where N is the total number of samples.[1] Fig. 1 shows the attraction field magnitude $|\mathbf{f}_{ji}|$ acting on a sample $\mathbf{s}_j$ as a function of distance $\|\mathbf{s}_j - \mathbf{s}_i\|$. With the plug-in estimator, the attraction field does not depend on α, but the network weight does. Also, the attractive force between two samples is zero when they coincide, achieves a maximum value at a close distance σ and becomes negligible when they are far apart.

[1] Some practical entropy-based registration algorithms employ histogram-based fast approximations of the plug-in estimate. Assuming the number of histograms is $\mathcal{O}(N^m)$, this entropy estimate has a computational complexity of $\mathcal{O}(N^{m+1})$. Note $m < 1$ and typically around 1/3.

TABLE I
COMPARISON OF THE INFLUENCE OF SAMPLE PAIR ($\mathbf{s}_j$ AND $\mathbf{s}_k$) INTERACTIONS ON THE UPDATE EQUATION. NOTE $c = 1/2\sigma^2$ AND $\mathbf{u}_{jk}$ IS THE UNIT VECTOR POINTING FROM $\mathbf{s}_j$ TO $\mathbf{s}_k$

	Plug-in	Entropic Graph
$\mathbf{f}_{jk}$	$2e^{-c\|\mathbf{s}_j-\mathbf{s}_k\|^2}\|\mathbf{s}_j-\mathbf{s}_k\|\mathbf{u}_{jk}$	$2\|\mathbf{s}_j-\mathbf{s}_k\|^{1-2\alpha}\mathbf{u}_{jk}$
n_{jk}	$N^{-2}[\hat{p}(s_j)^{\alpha-2}+\hat{p}(s_k)^{\alpha-2}]$	$A(G^*)(j,k) = 1$ or 0

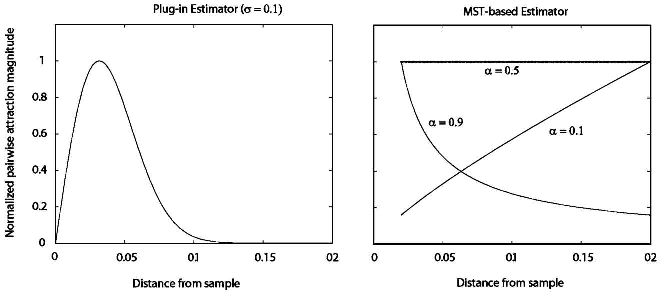

Fig. 1. Attraction field magnitude profiles as a function of distance from sample. The profiles have been normalized so that their maximum value is 1. Notice the very different profiles for different alpha values in the entropic graph estimator. For alpha greater than 0.5, it diverges to plus infinity as one approaches the origin.

To analyze the network effect consider a cluster of points, where a cluster can be thought of as a set of points within a relatively small diameter. Let $\mathbf{s}_c$ and N_c denote the mean value and number of samples within the cluster, respectively. The total net force[2] generated by this cluster and acting on a sample $\mathbf{s}_j$ is approximately

$$N_c N^{-2}[\hat{p}(\mathbf{s}_c)^{\alpha-2}+\hat{p}(\mathbf{s}_j)^{\alpha-2}]e^{-c\|\mathbf{s}_j-\mathbf{s}_c\|^2}\|\mathbf{s}_j-\mathbf{s}_c\|\mathbf{u}_{jc}$$

where $\mathbf{u}_{jc}$ is the unit vector pointing from the sample $\mathbf{s}_j$ to the cluster center $\mathbf{s}_c$. Assuming all $\mathbf{s} \in \mathcal{S}$ are independent samples of a sufficiently smooth density $p(\,\cdot\,)$, by the law of large numbers $N_c \propto Np(\mathbf{s}_c)$ and the total net force is approximately proportional to

$$N^{-1}n_M^c(p(\mathbf{s}_c);p(\mathbf{s}_k))e^{-c\|\mathbf{s}_j-\mathbf{s}_c\|^2}\|\mathbf{s}_j-\mathbf{s}_c\|\mathbf{u}_{jc} \qquad (24)$$

where $n_M^c(p(\mathbf{s}_c);p(\mathbf{s}_k)) = p(\mathbf{s}_c)^{\alpha-1}+p(\mathbf{s}_c)p(\mathbf{s}_j)^{\alpha-2}$ is the total network weight between a cluster and a point. Note that n_M^c is a monotonically increasing function of $p(\mathbf{s}_c)$ when $p(\mathbf{s}_c) > p(\mathbf{s}_j)$, and a monotonically decreasing function of $p(\mathbf{s}_j)$. Thus, we observe that low probability samples are attracted to high probability, i.e., more crowded, clusters with a force increasing with the number of samples in the cluster.

B. Entropic Graphs

The computation time of the MST estimator is $\mathcal{O}(N\log N)$. One advantage of this estimator is that once the MST is computed, the computation of the gradient for any α value is $\mathcal{O}(N)$ and negligible in practice. As intersample distance $\|\mathbf{s}_j-\mathbf{s}_k\|$ approaches zero, the sample pair attraction $\mathbf{f}_G(\mathbf{s}_j,\mathbf{s}_k)$ diverges to $+\infty$ for $\alpha > 0.5$, but converges to 0 for $\alpha \leq 0.5$. Fig. 1 shows the attraction field magnitude profiles for the entropic graph estimator with three different α values. When $\alpha > 0.5$, the attraction field achieves arbitrarily large magnitudes around the origin and monotonically decreases at a slower pace than the plug-in estimator as one moves away from the origin. When $\alpha < 0.5$, however, it is zero at the origin and monotonically increases as one moves away. The network effect, on the other hand, is either 1, if the two samples are connected in the minimal graph; or 0, otherwise. Thus, only a small subset of the sample pair interactions actually influence the gradient.

C. Samples, Gradients, and Image Registration

When digital images are uniformly sampled, coarse structures typically have a large representation, whereas fine detail structures are weakly represented. Thus, with a pair of images, sample clusters typically correspond to partially overlapping coarse image structures. Outliers, i.e., isolated samples that don't belong to a cluster, are usually due to a misaligned region, a point that has no correspondence, or noise. The goal of a registration algorithm can be viewed as "to pull in" outliers toward reliable clusters. Lacking any other useful information, it is natural to trust clusters rather than outliers when driving the registration algorithm.

At bad image alignment, we expect samples from fine detail structures to have arbitrarily scattered values. In an entropic graph estimator, by weighting shorter edges more heavily (with $\alpha > 0.5$), clusters of points drive the algorithm. However, for a given sample, the entropic graph estimator relies on a small subset of its neighbors, ignoring other samples. This is potentially too aggressive. On the other hand, in the plug-in estimator, all sample pair attractions are taken into account, and for a given sample the attractions to different clusters are weighted averaged (24), where the weights are proportional to the number of samples within the cluster and the inverse of the distance to that cluster. This observation leads to the following interpretation: the number of samples within a cluster is used as a measure of confidence about these samples being from a correctly aligned region and samples are "pulled into" local high probability regions. *Based on this interpretation, we expect the plug-in estimator to be more robust against bad initialization.*

On the other hand, the lower computational complexity of the entropic graph estimator makes this approach attractive for applications where speed is of concern. These predictions are empirically tested in Section VII.

V. IMPLEMENTATION OF AN MST ALIGNMENT MEASURE

In our implementation, we employ spanning trees as the entropic graph family $\mathcal{G}$. The alignment measure is the minimum spanning tree (MST) weight function $W_\gamma^{MST}(\mathbf{r}) \triangleq W_\gamma^{MST}(\mathcal{S}_\mathbf{r})$. We employ Kruskal's algorithm preceded by a Delaunay triangulation to compute the MST. The computational complexity of this implementation is $\mathcal{O}(N\log N)$, where N is the number of samples. Extension of these ideas to other entropic graphs, e.g., TSP, Steiner tree, nearest neighbor graphs, etc., is also possible.

In the entropic graph estimator, only with $\alpha \geq 0.5$ is the attractive field's magnitude decreasing as one moves away from the origin (see Fig. 1). Thus, consistent with our decision to trust clusters, we choose $\alpha \geq 0.5$ in our implementation. However, for $\alpha \geq 0.5$, very close samples undesirably dominate the computation of the function gradient (21). Hence, we apply a hard

[2] Net force equals attraction force times network weight.

threshold on $\mathbf{f}_G$ (23) and assign a zero value when $\|\mathbf{s}_j - \mathbf{s}_k\|$ is smaller than some small tolerance value. This threshold roughly corresponds to the width of the Gaussian kernel used in the plug-in estimator.

Experimental evidence suggests that α values closer to 1 yield better registration accuracy, whereas smaller α values, i.e., closer to 0.5, yield a wider capture range. In our implementation, we start the algorithm with $\alpha \approx 0.6$ and gradually increase to 0.9. To minimize the chance of getting trapped in local optima, we employ a multiresolution pyramid scheme, where the algorithm starts at a coarse resolution and works its way up to the finest resolution. At each level, the initial alignment is obtained from the result of the previous level. In addition, we use quantization within each level to aggregate information. Image intensity values are quantized, initially using a small number of quantization levels. Note that this approach leads to a significant amount of coinciding samples. To handle this issue, we compute the MST over the unique sample set and replace the sample gradient with the average gradient of coinciding samples. The number of quantization levels is then gradually increased. An advantage of this multiscale quantization approach is the speed-up of the MST computation. Our experiments suggest that the scheme also increases the capture range. Similar to [26], we employ a stochastic gradient descent optimization strategy, where at each iteration a random selection of less than 1% of the pixels is used to approximate the gradient. A software package that contains this implementation is freely available at: http://people.csail.mit.edu/msabuncu/sw/mst/.

VI. Incorporating Prior Knowledge

In this section, we present a method to incorporate prior knowledge about the modality relationship from prealigned image pairs into the MST-based registration algorithm. The goal is to improve the performance of the registration algorithm. These prealigned images can be provided by an expert as a training set or one can use previously registered image pairs within an image sequence. The modality relationship is assumed to be invariant across all image pairs and, hence, information about the modality relationship gained from prior alignments is useful in the registration of new images.

Our approach is parallel to [5], [9], and [16], where the alignment measure is defined using an information-theoretic divergence to quantify the discrepancy between the intensity distribution of the images to be aligned (test) and the prealigned training image pair(s). In [16], the authors propose the employment of the Jensen-Shannon divergence (JSD). Compared to the popular KL divergence, JSD has the advantage of being symmetric and well-defined for zero-probability regions. In this paper, we investigate a generalization of JSD, namely the Jensen-Rényi divergence (JRD) as an alignment measure.

JRD was used for mono-modal image registration in [11] and [28], while [1] extended its usage to a multimodal application. In these approaches, JRD is computed between the (1- D, conditional) probabilities of (scalar) pixel intensity values in the second image conditioned on the intensity values of the first. This can be viewed as a generalization of standard mutual information registration approaches and [28] shows that it is maximized at correct alignment. In [1], the authors propose a normalization strategy that is successfully applied to the particularly difficult MR-PET registration problem.

We, on the other hand, employ JRD to measure the discrepancy between the (2-D) test joint intensity distribution and prior distribution from the prealigned images. In this setting, our goal is to minimize (not maximize) JRD.

JRD is a distance measure between multiple probability distributions. For two distributions p_X and p_Y and a fixed $\alpha \in (0,1), \pi \in [0,1]$, it is defined as

$$J_{\alpha,\pi}(p_X, p_Y) = H_\alpha(\pi p_X + (1-\pi)p_Y) \\ -[\pi H_\alpha(p_X) + (1-\pi)H_\alpha(p_Y)].$$

Since H_α is concave, $J_{\alpha,\pi}(p_X, p_Y) > 0$ when $p_X \neq p_Y$ and $J_{\alpha,\pi}(p_X, p_Y) = 0$ when $p_X = p_Y$ (a.e.).

Let $I_1^\dagger(\mathbf{x})$ and $I_2^\dagger(\mathbf{x})$ denote two aligned training images from different modalities. For a given test transformation $\Phi(\mathbf{x})$, assume that each pixel intensity value in the image pairs $\mathcal{S}^\dagger = \{(I_1^\dagger, I_2^\dagger)\}$ and $\mathcal{S}^\Phi = \{(I_1, I_2 \circ \Phi)\}$ are sets of i.i.d samples from $p^\dagger$ and p^Φ, respectively. Then the distance between these distributions is a useful way of determining the quality of the current alignment. In particular

$$J_{\alpha,\pi}(p^\dagger, p^\Phi) = H_\alpha(\pi p^\dagger + (1-\pi)p^\Phi) \\ -\pi H_\alpha(p^\dagger) - (1-\pi)H_\alpha(p^\Phi) \quad (25)$$

can be employed as a supervised alignment measure that incorporates prior training data.

In practice, however, relying heavily on the prior distribution to determine the quality of alignment makes the algorithm's performance sensitive to noise. Also, note the negative marginal entropy term, $H_\alpha(p^\Phi)$, in (25). This term suggests that in some cases decreasing $J_{\alpha,\pi}(p^\Phi, p^\dagger)$ may increase the marginal entropy. Recall that in previous sections we motivated $H_\alpha(p^\Phi)$ as a *blind* alignment measure, i.e., to evaluate the quality of alignment based only on the test images and the algorithm was to minimize $H_\alpha(p^\Phi)$.

Based on these observations, we investigate the following hybrid measure that combines the blind and supervised alignment measures:

$$Q_\alpha(I_1, I_2 \circ \Phi) = J_{\alpha,\pi}(p^\dagger, p^\Phi) + \nu H_\alpha(p^\Phi) \quad (26)$$

where $\pi, \nu \in [0,1]$ are free parameters. Choosing $\pi = |\mathcal{S}^\Phi|/(|\mathcal{S}^\dagger| + |\mathcal{S}^\Phi|)$ presents a practical advantage:[3] the entropy of the mixture distribution [i.e., the first term on the right-hand side of (25)] becomes easily estimable. Moreover, $\nu = 1 - \pi$ cancels out the $H_\alpha(p^\dagger)$ term in (25). Since $H_\alpha(p^\dagger)$ does not depend on the current alignment, it can be removed from the objective function. With the chosen weights, the resulting expression then simplifies to

$$R_\alpha(I_1, I_2 \circ \Phi) = H_\alpha((1-\pi)p^\dagger + \pi p^\Phi) \quad (27)$$

which can be estimated using the *pooled* sample set $\mathcal{S} = \mathcal{S}^\dagger \cup \mathcal{S}^\Phi$. We can assume that the samples in $\mathcal{S}$ are drawn from a mixture distribution equal to $(1-\pi)p^\dagger + \pi p^\Phi$, where $\pi = |\mathcal{S}^\Phi|/(|\mathcal{S}^\dagger| + |\mathcal{S}^\Phi|)$. Using the entropic spanning graph estimator

[3] $|\cdot|$ denotes set cardinality.

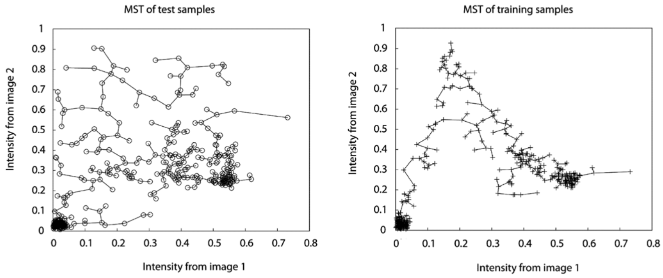

Fig. 2. MSTs of the test sample set and the training sample set (from prealigned images).

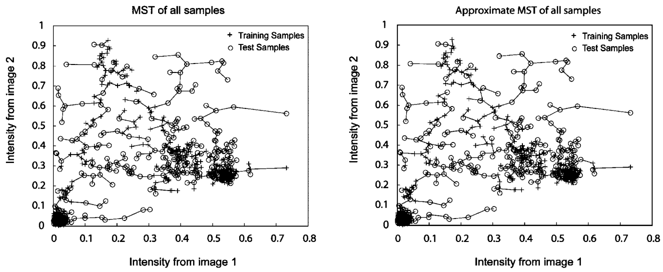

Fig. 3. MST and approximate MST (with $k = 1$, see Section VI-A for a description) of the pooled sample set (from $\mathrm{observed} + \mathrm{prealigned}$ images).

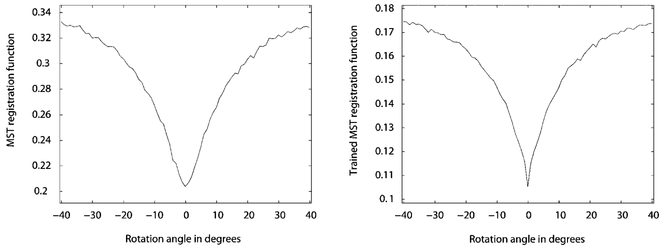

Fig. 4. MST measure versus rotational angle for two cases: (right) with training samples from a prealigned image pair and (left) no training samples; 3-D images from the Brainweb [6] data-sets were used. The training image pair was obtained with a different noise realization.

on the pooled sample set, $W^*_{2(1-\alpha)}(\mathcal{S})$, as defined in (12), yields a consistent estimate of (27) and can be used as an alignment measure (e.g., see Figs. 2 and 3, where the prealigned images are a simulated t1-t2 MR image pair [6]; for the test case, the second image was artificially rotated by 10°). Since the training samples are "stationary," i.e., are independent of the alignment, they behave like anchors, pulling in the observed samples. This "anchoring effect" improves the capture range of the algorithm. Fig. 4 shows the effect on the alignment measure profile. Interestingly, Zöllei and Wells recently presented an approach to incorporate prior knowledge into an information-theoretic framework by pooling the training data with the test data [29]. This result was derived using a probabilistic model and Dirichlet prior. The final result is strikingly similar to the entropy of the pooled sample set in (27).

A. Computational Issues

The additional computational load of introducing a large set of training samples is important. The following theorem indicates that an MST of the training samples, computed offline, can be used to decrease the computational overhead.

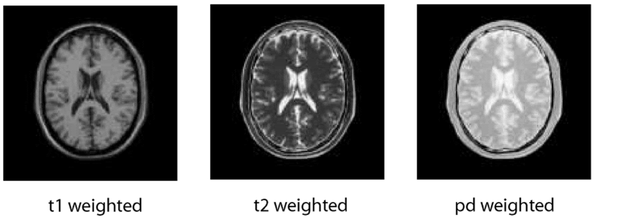

Fig. 5. Axial slices of the Brainweb volumes.

Let $T^\dagger$ and T^Φ be the edges in the MSTs of $\mathcal{S}^\dagger$ and $\mathcal{S}^\Phi$, respectively. Let F^L be the minimum spanning forest (MSF)[4] of the edges that connect samples from $\mathcal{S}^\dagger$ to $\mathcal{S}^\Phi$.

Theorem 2: The edges in a MST of $\mathcal{S}^\Phi \cup \mathcal{S}^\dagger$ are a subset of $T^\dagger \cup T^\Phi \cup F^L$.

Proof: If an edge is not in $T^\dagger \cup T^\Phi \cup F^L$, it is the longest edge in a cycle and, thus, by Kruskal's algorithm, cannot be in the final MST. ■

In our implementation, for large training sets, we replace F^L by the set of edges, E_{NN}, that connect each sample in $\mathcal{S}^\Phi$ to its k-nearest neighbors in $\mathcal{S}^\dagger$. This approach yields a fast approximate MST algorithm that uses edges in $T^\dagger \cup T^\Phi \cup E_{NN}$. In our experience, the output tree is a good approximation of the complete MST (see Fig. 3). This approach reduces the computational complexity from $\mathcal{O}(N^\dagger \log N^\dagger)$ to $\mathcal{O}(\log N^\dagger)$, where $N^\dagger = |\mathcal{S}^\dagger|$ is the number of training samples.

VII. Empirical Results

A. Three-Dimensional Simulations

To perform an objective comparison of our algorithms and other well-known registration methods, we employed a simulated 3-D MR data set [6]. This is a realistic simulation of three different MR modality images of a healthy subject's brain: t1 weighted, t2 weighted, and proton density (pd) weighted. Fig. 5 shows representative axial slices. All three images are of $181 \times 217 \times 181$ (1 mm^3 voxel size) resolution, contain about 5% noise and 20% intensity nonuniformity. Here, we provide results for five different 3-D rigid registration algorithms.[5]

- **NMI**: Histogram-based normalized mutual information [24] with an implementation of the Nelder–Mead simplex optimization method [14]. We used [4]'s partial volume histogramming approach that produced the smoothest alignment measure profiles and most accurate results.
- **JRD**: The normalized alignment measure based on the JR divergence proposed in [1] with the same optimization and histogramming methods as **NMI**. Unsurprisingly, this method yielded very similar results to the JRD-based technique of [28], [11], which are not included.
- **RPI**: A sample mean-based estimate of the joint Rényi entropy (7). A stochastic gradient descent strategy was employed.
- **MST**: An MST-based estimate of the joint Rényi entropy (15). A stochastic gradient descent strategy was employed.

[4] The MSF of a graph G is a union of the MSTs of the connected components.

[5] Rigid-body transformations were parameterized with six parameters: Three translations and three rotations.

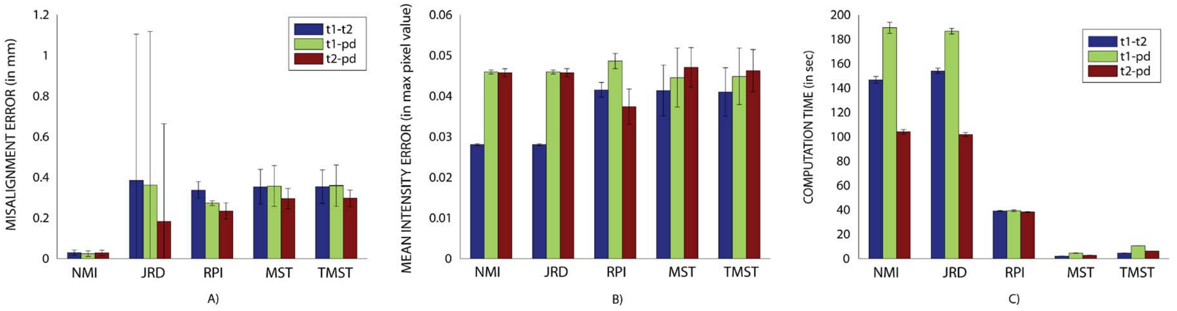

Fig. 6. Average performance of the five algorithms in a good initial alignment scenario. A) Registration accuracy as measured by the average distance between voxel locations of the transformed grid output by the algorithms and of the grid transformed by the ground truth (in millimeters). B) The mean square intensity difference between the moving image transformed by the algorithms' final results and ground truth. C) Run times of single-threaded MEX/Matlab implementations on a dual-core PC at 2.66 GHz and with 4-GB RAM. Error bars show standard deviations.

- **TMST**: A trained MST-based alignment measure, which is an estimate of the entropy of a mixture of the test samples and training samples (27). The training samples came from a different noise realization and subvoxel misalignment.

Now, we present registration results for two different scenarios: good and bad initialization, and three modality pairings: t1-t2, t1-pd, and t2-pd. The random experiments were repeated 100 times for each modality pairing.

We simulated a "ground truth" transformation by independently drawing from a uniform distribution on $[-10, 10]$ for all six transformation parameters (in mm for translations, and degrees for rotations). This transformation was then used to warp the second image which yielded an average initial misalignment of up to 28 mm per voxel. Fig. 6 summarizes the performance of all five algorithms for this "good initialization" scenario. We see that all five algorithms on average achieve subvoxel accuracy (less than 1 mm) and the quality of alignment is comparable (especially, as measured by the mean square intensity error between the output moving image and ground truth). In computational speed, however, it is obvious that the stochastic gradient descent approach [26] yields substantially faster algorithms (**RPI**, **MST**, and **TMST**). There are two reasons for this: first, the gradient information is used when searching the transformation space and second, at each iteration the algorithm only uses a fraction (less than 1%) of the total voxels which speeds up the computation of the alignment measure and its gradient. Moreover, as predicted, the MST-based algorithm achieves the fastest registration (with less than 5 s). In the "bad initialization" scenario, the ground truth transformation was generated by independently drawing the six transformation parameters (in mm for translations, and degrees for rotations) from a uniform distribution on $[-30, -10] \cup [10, 30]$. This yielded a maximum average initial misalignment of 85 mm per voxel. Our experiments showed that, when an algorithm converged to a correct result, the performance (in speed and accuracy) was comparable to the results presented for the "good initialization" scenario. However, as expected, not every algorithm converged always to a correct result. Fig. 7 shows the convergence rates for each algorithm. The convergence rate is defined as the fraction of instances where an algorithm achieved a subvoxel (less than 1 mm) alignment accuracy. Notice that **NMI** and **RPI** converged almost always to the correct result. **NMI** was initially proposed as an "overlap-invariant" alignment measure robust to bad initial alignments [24]. Also, in Section IV-C, we had theoretically predicted **RPI** to perform well with bad initialization since it employed a plug-in based entropy estimator. Another interesting observation that supports our predictions is that, with training (i.e., by pooling samples from a prealigned image pair), the convergence rate of the **MST** algorithm can be improved substantially (for t2-pd, we observe an increase from 25% to 85%).

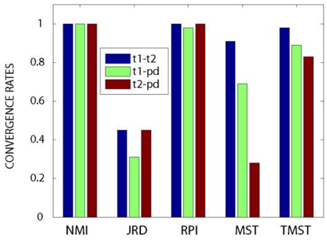

Fig. 7. Convergence rates: Initialized with a misalignment of up to 85 mm, the fraction of instances each algorithm output, an alignment of subvoxel accuracy.

B. Three-Dimensional PET-MR Registration

Here, we present a result from a real world application: 3-D intrapatient MR-PET[6] rigid registration using the proposed **MST** algorithm. Fig. 8 shows the volumes before and after registration. Visual inspection suggests that the algorithm achieves voxel-resolution accuracy in aligning these two modalities. Fig. 9 shows the MSTs computed on pixel intensity values before and after rigid-body registration. The final result was obtained using a single-threaded MEX/Matlab implementation of the proposed MST algorithm. The run-time was approximately 2.0 s on a Dual-Core PC at 2.66 GHz and with 4-GB RAM.

VIII. Conclusion

In this paper, we investigated and expanded the employment of Rényi entropy for multimodal image registration. We developed a common framework where two popular entropy estimators, namely the plug-in estimator and entropic graph method, were analytically comparable. This comparison provided valuable insight on how these techniques weight data which lead to predictions of likely performance when applied

[6] Both data-sets were re-sampled onto a 128 × 128 × 128 grid prior to registration.

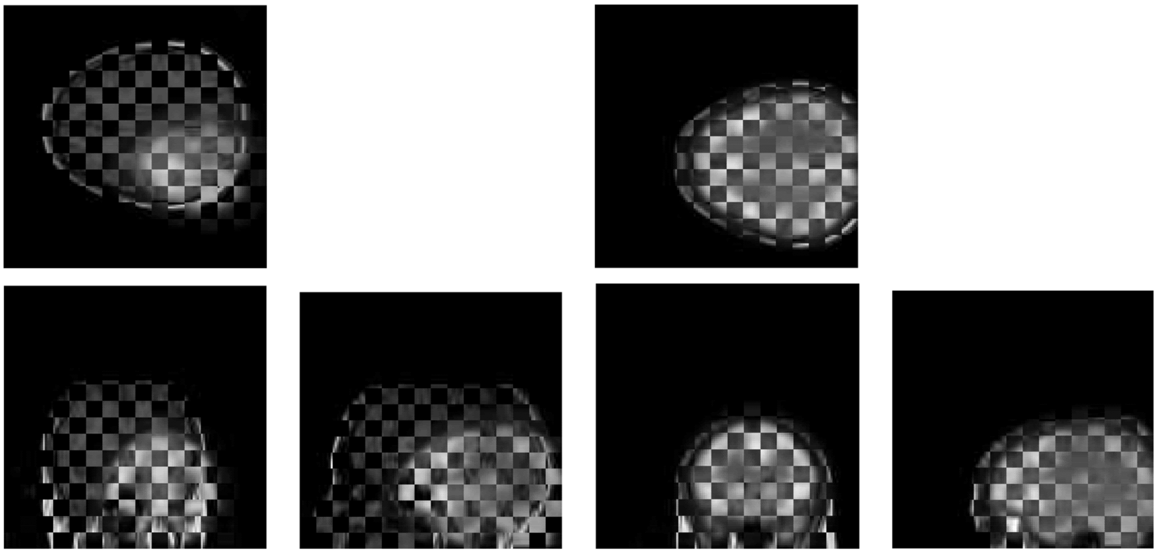

Fig. 8. Checkerboard representations of the *patient 17* MR and PET data sets (left) before and (right) after EMST-based rigid-body registration: transverse, sagittal, and coronal views.

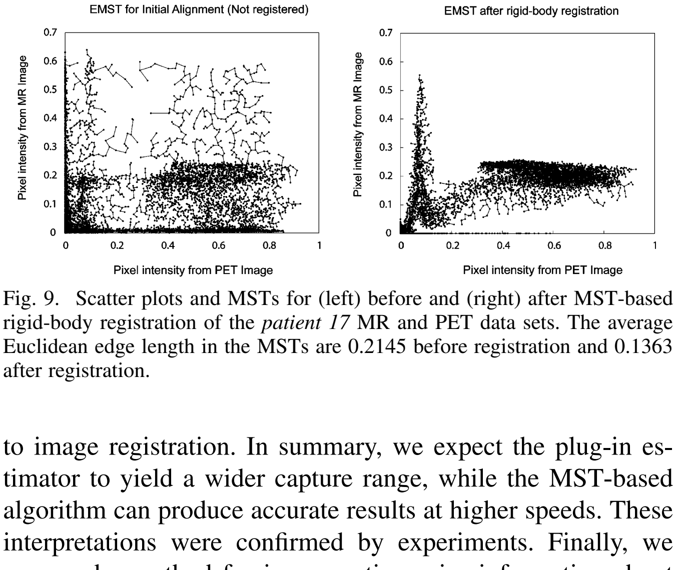

Fig. 9. Scatter plots and MSTs for (left) before and (right) after MST-based rigid-body registration of the *patient 17* MR and PET data sets. The average Euclidean edge length in the MSTs are 0.2145 before registration and 0.1363 after registration.

to image registration. In summary, we expect the plug-in estimator to yield a wider capture range, while the MST-based algorithm can produce accurate results at higher speeds. These interpretations were confirmed by experiments. Finally, we proposed a method for incorporating prior information about the modality relationship from prealigned image pairs into the entropic graph-based registration framework. Empirical results suggest that the employment of this knowledge improves the capture range of the alignment measure and makes it more robust against bad initial alignment.

One important point to note is that the presented MST-based registration framework is extendible to more than two images. This will potentially allow for efficient group-wise image registration algorithms, where the group alignment measure will be a function of the MST computed over the samples in a space with a dimensionality equal to the number of images, not just two. This extension, we believe, is a promising direction for future research.

Appendix

A. Differentiability of the Entropic Graph Estimate

Let $\mathcal{S}^0 = \{\mathbf{s}_1^0,\ldots,\mathbf{s}_N^0\}$ be a set of N samples in $[0,1]^2$ and $\mathbf{u}$ be a unit vector in $\mathbb{R}^2$. Define $\mathcal{G}^*(\mathcal{S}^0) \triangleq \{G^*(\mathcal{S}^0)\}$, the set of all minimal graphs on $\mathcal{S}^0$. The following lemma states that after a slight perturbation of the value of a sample in $\mathcal{S}^0$, some of the current minimal graphs remain as minimal graphs and no other graph can become a minimal graph.

Lemma 1: For any $k \in \{0,\ldots,N\}$, there exists an $\epsilon > 0$ such that $\mathcal{G}^*(\{\mathbf{s}_1^0,\ldots,\mathbf{s}_k^0 + h\mathbf{u},\ldots,\mathbf{s}_N^0\}) \subset \mathcal{G}^*(\mathcal{S}^0)$, for all $|h| \le \epsilon$.

Proof: Let $\delta \triangleq \min_{G\in\mathcal{G}\setminus\mathcal{G}^*}(W_\gamma(G(\mathcal{S}^0)) - W_\gamma^*(\mathcal{S}^0))$. Note $\gamma > 0$. If $|h| \le (\|e\|^\gamma + \delta/2N)^{1/\gamma} - \|e\|$ for all $\|e\|$ in G, then using the triangle inequality on each edge, it is easy to show that the change in $W_\gamma(G)$ is upper bounded by $\delta/2$. Recall that $\|e\| < \sqrt{2}$, since all $\mathbf{s} \in [0,1]^2$. Set $\epsilon = \max((\delta/2N)^{1/\gamma}, (2^{\gamma/2} + \delta/2N)^{1/\gamma} - \sqrt{2})$. Then for $|h| < \epsilon$ and all $G_1, G_2 \in \mathcal{G}(\mathcal{S}^0)$, the change in $W_\gamma(G_1) - W_\gamma(G_2)$ will be upper bounded by δ. Thus, if $G \notin \mathcal{G}^*(\mathcal{S}^0), G$ will not achieve a $W_\gamma(G)$ smaller than $W_\gamma^*(\mathcal{S}^0)$.

Now, let us look at the partial derivative of a power weighted edge length, $\|e_{ij}\| \triangleq \|\mathbf{s}_i - \mathbf{s}_j\|$

$$\frac{\partial(\|e_{ij}\|^\gamma)}{\partial s_{ic}} = \begin{cases} \gamma\|e_{ij}\|^{\gamma-2}(s_{ic} - s_{jc}), & \text{if } \mathbf{s}_i \neq \mathbf{s}_j, \\ 0, & \text{if } \mathbf{s}_i = \mathbf{s}_j \text{ and } \gamma \ge 1 \\ \pm\infty, & \text{if } \mathbf{s}_i = \mathbf{s}_j \text{ and } \gamma < 1 \end{cases}$$

for $i, j = 1 \cdots N$ and $c = 1, 2$. Note that the derivative does not exist if the samples are coinciding and $\gamma < 1$. Elsewhere, it is well defined.

Lemma 2: For a $\mathbf{s}_k \in \mathcal{S}, \nabla_{\mathbf{s}_k} W_\gamma^*(\mathcal{S})$ exists if and only if $\nabla_{\mathbf{s}_k} W_\gamma(G^*(\mathcal{S}))$ exists and is equal for all $G^*(\mathcal{S})$.

Proof: Using the formal definition of the right derivative

$$\begin{aligned} \partial W_\gamma^*(\mathcal{S})/\partial s_{kc}\big|_{s_{kc}=s_{kc}^{0+}} &= \lim_{h\to 0^+} \\ &\frac{W_\gamma^*\left(\{\mathbf{s}_1^0,\ldots,\mathbf{s}_k^0 + hu_{dc},\ldots,\mathbf{s}_N^0\}\right) - W_\gamma^*(\mathcal{S}^0)}{h} \\ &= \min_{G\in\mathcal{G}^*(\mathcal{S}^0)} \partial W_\gamma(G)/\partial s_{kc}\big|_{s_{kc}=s_{kc}^{0+}}. \end{aligned} \quad (28)$$

Similarly, the left derivative is equal to the maximum of the left derivatives among all the $G^*(\mathcal{S}^0)$s. Now, consider the two cases.

1) $\mathbf{s}_k$ has a unique value $\mathbf{s}_k^0$. Then, $\nabla_{\mathbf{s}_k} W_\gamma(G^*(\mathcal{S}^0))$ exists for all $G^*(\mathcal{S}^0)$. Here, $\partial W_\gamma^*(\mathcal{S})/\partial s_{kc}$ exists if and only if the maximum and minimum derivatives are equal for all $c \in \{1,2\}$.
2) $\mathbf{s}_k$ is not unique, i.e., there are other samples with the same value. Then it is easy to see that all minimal spanning graphs $G^*(\mathcal{S}^0)$s contain at least one zero length edge with $\mathbf{s}_k$ as an endpoint. If $0 < \gamma < 1$, then the right and left derivatives of this edge length are $+\infty$ and $-\infty$, respectively. Thus, $\nabla_{\mathbf{s}_k} W_\gamma^*(\mathcal{S})$ does not exist. If $\gamma > 1$, the edge length derivatives exist and the argument from 1 holds.

REFERENCES

[1] A. Bardera, M. Feixas, and I. Boada, "Normalized similarity measures for medical image registration," in *Proc. SIPE Medical Imaging: Image Processing*, J. Fitzpatrick and M. Sonka, Eds., 2004, vol. 5370, pp. 108–118.

[2] J. Beirlant, E. Dudewicz, L. Gyorfi, and E. van der Meulen, "Nonparametric entropy estimation: An overview," *Int. J. Math. Statist. Sci.*, vol. 6, pp. 17–39, 1997.

[3] T. Butz and J. Thiran, "Affine registration with feature space mutual information," in *Proc. MICCAI*, W. Niessen and M. Viergever, Eds., 2001, pp. 549–556.

[4] H. Chen and P. Varshney, "Mutual information-based CT-MR brain image registration using generalized partial volume joint histogram estimation," *IEEE Trans. Med. Imag.*, vol. 22, no. 9, pp. 1111–1119, Sep. 2003.

[5] A. Chung, W. Wells, A. Norbash, and W. Grimson, "Multi-modal image registration by minimising Kullback-Leibler distance," in *Proc. MICCAI*, T. Dohi and R. Kikinis, Eds., 2002, pp. 525–532.

[6] C. Cocosco, V. Koolokian, R. Kwan, and A. Evans, "Brainweb: Online interface to a 3D MRI simulated brain database," in *Proc. NeuroImage 3rd Int. Conf. Functional Mapping of the Human Brain*, Copenhagen, Denmark, 1997, vol. 5, p. S425, (4, part2/4).

[7] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[8] R. Duda, P. Hart, and D. Stork, *Pattern Classification*. New York: Wiley-Interscience, 2001.

[9] C. Guetter, C. Xu, F. Sauer, and J. Hornegger, "Learning based nonrigid multi-modal image registratyion using Kullback–Leibler divergence," in *Proc. MICCAI*, J. Duncan and G. Gerig, Eds., 2005, pp. 255–262.

[10] J. Hajnal, D. Hill, and D. Hawkes, Eds., *Medical Image Registration*. Boca Raton, FL: CRC, 2001.

[11] A. Hamza and H. Krim, "Image registration and segmentation by maximizing the Jensen–Rényi divergence," *Proc. EMMCVPR*, pp. 247–263, 2003.

[12] A. Hero, B. Ma, O. Michel, and J. Gorman, "Applications of entropic spanning graphs," *IEEE Signal Process. Mag.*, vol. 19, no. 5, pp. 85–95, May 2002.

[13] A. Hero and O. Michel, "Asymptotic theory of greedy approximations to minimal k-point random graphs," *IEEE Trans. Inf. Theory*, vol. 45, no. 6, pp. 1921–1938, Jun. 1999.

[14] J. Lagarias, J. A. Reeds, M. H. Wright, and P. E. Wright, "Convergence properties of the Nelder–Mead simplex method in low dimensions," *SIAM J. Optim.*, vol. 9, no. 1, pp. 112–147, 1998.

[15] M. Leventon and W. Grimson, "Multi-modal volume registration using joint intensity distribution," in *Proc. MICCAI*, W. Welles, E. Colchester, and S. Delp, Eds., 1998, pp. 1057–1066.

[16] R. Liao, C. Guetter, C. Xu, Y. Sun, A. Khamane, and F. Sauer, "Learning-based 2D/3D rigid registration using Jensen–Shannon divergence for image-guided surgery," in *Proc. MIAR*, G. Yang, Ed., 2006, pp. 228–235.

[17] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Seutens, "Multimodality image registration by maximization of mutual information," *IEEE Trans. Med. Imag.*, vol. 16, no. 2, pp. 187–198, Feb. 1997.

[18] J. Maintz and M. Viergever, "A survey of medical image registration," *Med. Image Anal.*, vol. 2, no. 1, pp. 1–36, 1998.

[19] H. Neemuchwala, A. Hero, and P. Carson, "Image matching using alpha-entropy measures and entropic graphs," *Signal Process.*, vol. 85, no. 2, 2002.

[20] J. Pluim, J. Maintz, and M. Viergever, "Image registration by maximization of combined mutual information and gradient information," *IEEE Trans. Med. Imag.*, vol. 19, no. 8, pp. 809–814, Aug. 2000.

[21] J. Pluim, J. Maintz, and M. Viergever, "Mutual information based registration of medical images: A survey," *IEEE Trans. Med. Imag.*, vol. 22, no. 8, pp. 986–1004, Aug. 2003.

[22] C. Redmond and J. E. Yukich, "Asymptotics for Euclidean functionals with power weighted edges," *Stochastic Process. Appl.*, vol. 6, pp. 289–304, 1996.

[23] M. Sabuncu and P. Ramadge, "Spatial information in entropy-based image registration," in *Biomedical Image Registration*. New York: Springer-Verlag, 2003.

[24] C. Studholme, D. L. G. Hill, and D. Hawkes, "An overlap invariant entropy measure of 3D medical image alignment," *Pattern Recognit.*, vol. 32, no. 1, pp. 71–86, 1999.

[25] P. Thvenaz and M. Unser, "Optimization of mutual information for multiresolution image registration," *IEEE Trans. Image Process.*, vol. 9, no. 12, pp. 1083–1100, Dec. 2000.

[26] P. Viola and W. Wells, "Alignment by maximization of mutual information," *Int. J. Comput. Vis.*, vol. 24, no. 2, pp. 137–154, 1997.

[27] Y. He, A. Hamza, and H. Krim, "Information divergence measure for ISAR image registration," in *Proc. SPIE*, Oct. 2001, vol. 4379, pp. 199–208.

[28] Y. He, A. Hamza, and H. Krim, "A generalized divergence measure for robust image registration," *IEEE Trans. Signal Process.*, vol. 51, no. 5, pp. 1211–1220, May 2003.

[29] L. Zöllei and W. Wells, "Multi-modal image registration using Dirichlet-encoded prior information," in *Proc. WBIR*, J. Pluim, B. Likar, and F. Gerritsen, Eds., 2006, pp. 34–42.

Mert R. Sabuncu (M'99) received the B.Sc. degree from the Electrical and Electronics Engineering Department, Middle East Technical University, Ankara, Turkey, in 2001, and the Ph.D. degree from the Department of Electrical Engineering, Princeton University, Princeton, NJ, in 2006.

Between 2003–2006, he was an intern and temporary research staff at Siemens Corporate Research, Princeton. In August 2006, he joined the medical vision group at MIT's Computer Science and Artificial Intelligence Lab as a Postdoctoral Researcher. His current research interests are in image processing, medical image analysis, pattern recognition, and fMRI.

Peter Ramadge (S'79–M'82–SM'92–F'96) received the B.Sc., B.E., and M.E. degrees from the University of Newcastle, Australia, and the Ph.D. degree from the Department of Electrical Engineering at the University of Toronto, Toronto, ON, Canada.

He joined the faculty of Princeton University, Princeton, NJ, in September 1984, where he is currently the Professor and Chair of the Department of Electrical Engineering. He has been a Visiting Professor at the Massachusetts Institute of Technology, Cambridge, and a Visiting Research Scientist at IBM's Tokyo Research Laboratory, Tokyo, Japan. His current research interests are in high-dimensional signal processing, fMRI imaging, medical imaging, and video/image processing.

Dr. Ramadge is a member of SIAM. He has received several honors and awards including a paper selected for inclusion in the IEEE book *Control Theory: Twenty Five Seminal Papers* (1932–1981); an Outstanding Paper Award from the Control Systems Society of the IEEE; the Convocation Medal for Professional Excellence from the University of Newcastle, Australia; an Engineering Council Teaching Award from the School of Engineering and Applied Science, Princeton University; an IBM Faculty Development Award; and the University Medal from Newcastle University, Australia.