Why Does Mutual-Information Work for Image Registration? A Deterministic Explanation

Hemant D. Tagare, *Senior Member, IEEE* and Murali Rao

Abstract—This paper proposes a deterministic explanation for mutual-information-based image registration (MI registration). The explanation is that MI registration works because it aligns certain image partitions. This notion of aligning partitions is new, and is shown to be related to Schur- and quasi-convexity. The partition-alignment theory of this paper goes beyond explaining mutual- information. It suggests other objective functions for registering images. Some of these newer objective functions are not entropy-based. Simulations with noisy images show that the newer objective functions work well for registration, lending support to the theory. The theory proposed in this paper opens a number of directions for further research in image registration. These directions are also discussed.

Index Terms—Image Rregistration, medical image registration, mutual information, convexity

1 Introduction

Since its introduction in 1995 [1], [2], mutual-information-based registration (MI registration) has been extraordinarily successful. But why does it work so well? A clear *deterministic* (non-probabilistic) mathematical explanation has remained elusive. We hope to shed some light on the situation by proposing that MI registration works because it aligns certain partitions of the images. In fact, a simple theory of image registration can be developed solely based on the idea of partition-alignment. Formulating this partition-alignment theory, and showing how MI registration is an instance of it, are the two main goals of this paper.

MI registration algorithms calculate mutual-information between images either by histogramming their joint intensities, or by using a kernel estimator for the joint density. The histogramming method is what our theory explains. We hope to extend the partition-alignment theory to kernel estimators in the future.

The partition-alignment approach goes beyond just explaining MI. The approach shows that a very broad class of objective functions can be used to register images in the same spirit as MI. Many objective functions in this class are unusual; they are not associated with any known mathematical entropy. In Section 6, we experimentally evaluate the performance of a few of these non-entropy objective functions and show that they work well.

This paper approaches partition-alignment and MI registration from a purely deterministic point of view. But that does not preclude a probabilistic interpretation. A reader who is knowledgeable about the modern (Kolmogorov) definition of a random variable will immediately see the connection. We take the domain of any image to be a probability space and the image itself to be a random variable defined on this space.

In spite of such an interpretation, a deterministic presentation of the partition-alignment theory seems to be the simplest, and we stick with it.

Similarly, areas and volumes (Lebesgue-measures) of partitions are important to this theory, but a measure-theoretic presentation is not likely to appeal to many readers. We avoid measure theory, too. The reader should bear in mind that all sets, partitions, and functions used below are assumed to be measurable. A reader unfamiliar with measure theory can proceed without trepidation. Intuitive ideas of areas and volumes are sufficient to grasp the theory fully.

1.1 Literature Review

Image registration using mutual-information was first proposed in [1], [2]. The literature on mutual-information-based registration is vast, and we only review the literature that is relevant either to extending the notion of mutual-information-based registration or to explaining it. Discussions of classical mutual-information-based registration can be found in textbooks [4], [5]. Even though mutual-information was originally proposed as a new criterion for image registration, subsequent research has suggested that for certain image models, mutual-information-based registration may be derived from the maximum-likelihood principle [6].

Mutual-information registration methods require an estimate of image entropy. A survey of various entropy estimation methods is available in [7]. A critique of these methods from the point of view of image registration is available in [8].

Many extensions of mutual-information-based registration methods are motivated by a desire to overcome the limitations of histogram-based and kernel-based entropy estimates. One limitation—caused by the fact that the MI methods approximate image entropy with the entropy of a discrete random variable, and this discrete approximation

- *H. D. Tagare is with the Department of Diagnostic Radiology, the Department of Biomedical Engineering, and the Department of Electrical Engineering, Yale University, New Haven, CT 06520. E-mail: hemant.tagare@yale.edu.*
- *M. Rao is with the Department of Mathematics, University of Florida, Gainesville, FL 32611. E-mail: mrao@ufl.edu.*

Manuscript received 4 Nov. 2013; revised 25 July 2014; accepted 3 Aug. 2014. Date of publication 2 Oct. 2014; date of current version 8 May 2015.
Recommended for acceptance by T. Cootes.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2014.2361512

need not converge to the continuous case—is addressed by the cumulative residual entropy [9]. The cumulative residual entropy has been successfully used in medical image registration [10].

Entropic-graph methods are proposed in [11], [12], [13], [14], [15] for estimating information divergence using graph-theoretic structures such as nearest-neighbor graphs or minimal spanning trees. They have the advantage that histogramming or kernel density estimation is not required, making registration with high-dimensional feature vectors tractable.

A popular variant of mutual-information is the normalized mutual-information [16]. It is useful when the transformation maps a significant portion of one of the images outside the domain of the other image. Another variant is the conditional mutual-information [17] which is useful for general diffeomorphic image registration.

F-information measures are suggested in [18] as a generalization of mutual-information for image registration. The simulation study of f-information measures in [18] concludes that many of the f-information measures give registrations that are as accurate as mutual-information registrations. In fact, some appear to be more accurate. These results are similar to ours in the sense that mutual-information is found to be one member of a large family of objective functions, all of which appear to be useful for image registration.

The Jensen-Renyi divergence is proposed in [19] for image registration. Simulations in [19] show that when used to register multi-modal brain images, the Jensen-Renyi divergence produces a sharper peak in the registration objective function than mutual-information.

Not directly related to registration using image intensities, is the work on registering labeled point sets using various entropy measures. A recent algorithm for this is [20].

Prior knowledge about non-linear relations between intensities in the two images can also be exploited for registration, giving computationally simple, correlation-like algorithms. In [21], for example, piecewise-linear relations between image intensities are exploited to match CT and MR brain and spine images.

The approach that comes closest in spirit to our approach is the one in [8]. The authors of [8] clearly identify the relation between joint and marginal image probability densities and the areas bounded by image level sets. Using a clever change of variables, this relation is further exploited to create an accurate density estimator for differentiable images. Plugging this new estimator into MI registration is shown to improve registration accuracy.

Image partitions are central to our formulation. Image partitions are also discussed in [22] which proposes a partition-based region overlap criterion for image registration. In [22], image partitions are classified into different categories, and registration is achieved by "maximizing the number of pixels in regions which are a priori registrable (application-dependent control structures)". However, the objective function of [22] is different from mutual information, and the theory of that paper does not explain why mutual-information works.

All of the above-mentioned extensions of mutual-information, except [22], explicitly adopt a probabilistic point of view. To our knowledge there is no available literature that proposes a deterministic explanation of why mutual-information-based methods work for image registration.

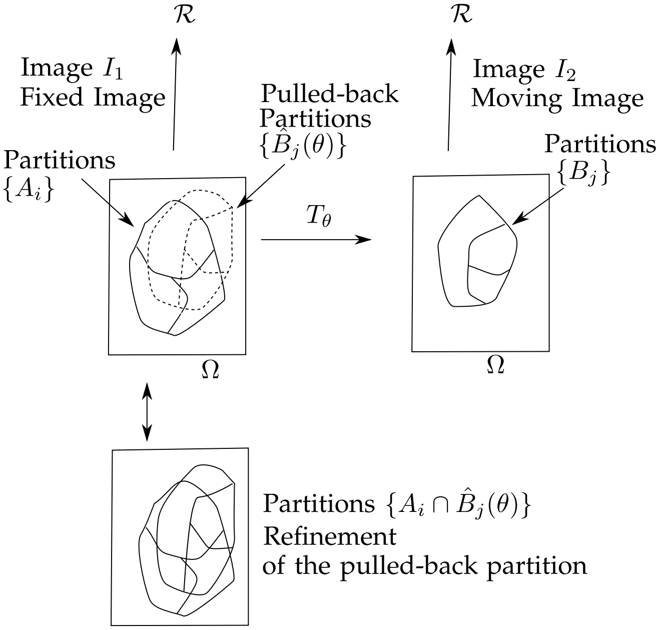


Fig. 1. The fixed and moving images and their partitions.

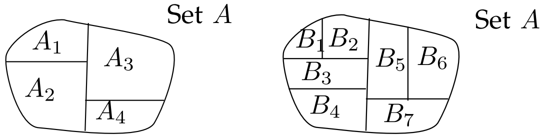
2 Image Registration via Partition Alignment

The main idea of our theory is as follows; Fig. 1 illustrates it:

1) Histogramming an image into n bins is equivalent to partitioning the domain of each image into n subsets. Each subset is the set containing pixels which contribute to exactly one bin of the histogram. Different subsets contribute to different bins.

   In Fig. 1, two images are shown as functions from their domains, Ω, to the real line. The partitions of the images are illustrated as regions $\{A_i\}$ and $\{B_j\}$ in the respective domains.
2) A parameterized transformation from the domain of one image to the domain of the other pulls back the partition of the second image (the "moving" image) onto the first image (the "fixed" image). The pulled-back partitions are illustrated as regions $\{\hat{B}_j(\theta)\}$ in the domain of the fixed image in Fig. 1, where θ is the parameter of the transformation.
3) The intersection of the pulled-back partition with the partition of the fixed image refines the pulled-back partition. The severity of this refinement—that is, loosely speaking, the amount by which the pulled-back partition is further split into smaller chunks—depends on the transformation.
4) The transformation that causes the "least severe refinement" of the pulled-back partition registers the two images.

Step 4 above requires an index of (i.e. a real-valued number which measures) the severity of a partition, and as we shall see, mutual-information works for image registration because it is such an index.



(a) A 4-partition of set A

(b) A 7-partition of the same set A. This is a refinement of the partition shown in (a).

Fig. 2. Partitions and their refinements.

There is one key point about this explanation: the number of partitions in both images are not required to be the same, nor are the partitions required to align exactly. The only requirement is that the refinement of the pulled-back partition by the fixed-image partition be as mild as possible (as less severe as possible). Speaking very loosely, this happens when some, but not necessarily all, of the boundaries of the pulled-back partition lie close to, or coincide with, some of the boundaries of the fixed partition. This "slack" in alignment is why MI works for registering images even when the images come from different modalities.

The above theory is explained in detail in the rest of the paper: Partitions are discussed in Section 3. The intuition that leads to an index of severity of a partition is also in Section 3. Section 4 explains how the index of severity is used to construct registration objective functions. Section 5 shows that MI registration is an instance of such an objective function. Section 6 contains numerical simulations of alternative objective functions. Section 7 contains a discussion of several implications of the theory and concludes the paper.

3 Partitions and Their Severity

3.1 Preliminaries

We begin with preliminary definitions borrowed from standard set theory. Let A be a subset of $\mathcal{R}^n$. Then, the *size of* A (the measure of A) means "the length of A" when the dimension $n = 1$, "the area of A" when the dimension $n = 2$, "the volume of A" when the dimension $n = 3$, etc. The size of A is denoted $\mu(A)$. If B is a subset of A, then the *relative size* of B is $\mu^*(B) = \mu(B)/\mu(A)$. The relative size is always between 0 and 1.

An *n-partition* of A is a family $\{A_1, A_2, \ldots, A_n\}$ of n disjoint subsets of A having the property that their union is A (i.e., $\cup_i A_i = A$, while $A_i \cap A_j = \emptyset$ for $i \neq j$). Fig. 2a shows a four-partition of a set A by subsets $A_1, \ldots, A_4$.

Suppose $\{A_1, A_2, \ldots, A_n\}$ and $\{B_1, \ldots, B_m\}$ are n- and m-partitions of the same set A with $m > n$. Then, the second partition is a refinement of the first partition if every A_i can be written as a union of one or more B_j's. Thus, the seven-partition in Fig. 2b is a refinement of the four partition in Fig. 2a.

An *image* is a function from the unit cube $\Omega \subset \mathcal{R}^n$ to $\mathcal{R}$. Our interest is in registering two images by a parametrized transformation T_θ from Ω to Ω, where θ is the parameter of the transformation.

3.2 Severity of a Partition

Now we turn to developing ideas about the "severity" of a partition.

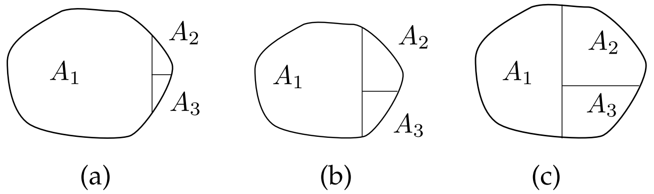


Fig. 3. Partitions with increasing severity from left to right.

Fig. 3 shows a number of three-partitions of a set A. Intuitively speaking, the partitioning of A becomes progressively more "severe" from "a" to "c" in the figure. That is, A appears to be shattered more severely by the partition as we go from "a" to "c". How shall we make this intuition more precise? Notice that when the partition is not severe, such as in Fig. 3a, the largest subset in the partition, which is A_1, is quite similar to the original set A. As the partitioning gets more severe, as in Fig. 3c, the largest subset deviates more from the original set A. This suggests that one way of capturing severity of a partition is to simply consider the relative size of the largest subset. The greater this relative size, the smaller the severity of partition.

We need not stop at comparing just the largest subset, we can go on to compare smaller subsets too. Given two n-partitions of A, we can rank the subsets of each partition in decreasing order of relative size and compare the top k subsets for $k = 1, \ldots, n$. If for every k, the sum of the relative sizes of the largest k subsets of one of the partitions is always greater than or equal to the sum of the relative sizes of the largest k subset of the other partition, then it is reasonable to declare the first partition to be less severe than or equally severe to the second. This is true of the partitions in Figs. 3a, 3b, 3c, where the partition in Fig. 3a is less severe in this sense than the partition in Fig. 3b which in turn is less severe than the partition in Fig. 3c.

Suppose that the vector of relative sizes of an n-partition $\{A_i\}$ is $\alpha = (\mu^*(A_1), \mu^*(A_2), \ldots, \mu^*(A_n))$. Let $\alpha_{[i]}$ be the ith largest component of α, i.e. $\alpha_{[1]} \geq \alpha_{[2]} \geq \cdots \geq \alpha_{[n]}$. Similarly, suppose $\{B_i\}$ is another n-partition of the same set and β is the vector of its relative sizes. Then, according to the above discussion, the partition $\{A_i\}$ is less severe than or equally severe to the partition $\{B_i\}$ if

$$\begin{aligned} \alpha_{[1]} &\geq \beta_{[1]}, \\ \alpha_{[1]} + \alpha_{[2]} &\geq \beta_{[1]} + \beta_{[2]}, \\ \alpha_{[1]} + \cdots + \alpha_{[n-1]} &\geq \beta_{[1]} + \cdots + \beta_{[n-1]}. \end{aligned} \tag{1}$$

And since the relative sizes always add up to 1

$$\alpha_{[1]} + \cdots + \alpha_{[n]} = \beta_{[1]} + \cdots + \beta_{[n]}. \tag{2}$$

Equations (1-2) are precisely the classical definition of majorization, showing that the idea of a severity of a partition is related to the theory of majorization. To proceed, we review the theory of majorization, suspending the discussion of partitions until Section 3.4.

3.3 Majorization

Let $x = (x_1, \ldots, x_n) \in \mathcal{R}^n$, and as above, let $x_{[i]}$ be the ith largest coordinate of x. Then,

Definition. *If $x, y \in \mathcal{R}^n$, then x* majorizes *y, denoted $x \succeq y$, if*

$$\begin{aligned} x_{[1]} &\geq y_{[1]}, \\ x_{[1]} + x_{[2]} &\geq y_{[1]} + y_{[2]}, \\ \cdots &\geq \cdots \\ x_{[1]} + \cdots + x_{[n-1]} &\geq y_{[1]} + \cdots + y_{[n-1]} \\ x_{[1]} + \cdots + x_{[n]} &= y_{[1]} + \cdots + y_{[n]}. \end{aligned} \tag{3}$$

If $x \succeq y$, but $(x_{[1]}, x_{[2]}, \ldots, x_{[n]}) \neq (y_{[1]}, y_{[2]}, \ldots, y_{[n]})$, then at least one of the inequalities in (3) is strict, and x strictly majorizes *y, which is denoted $x \succ y$.*

A comprehensive account of the theory of majorization is given by Marshall [23]. A more concise account of the theory is available in [24].

Majorization has several interesting properties. The following are especially relevant to our discussion:

M1. Majorization is a partial ordering of the elements of $\mathcal{R}^n$. Two arbitrary elements x and $y \in \mathcal{R}^n$ need not satisfy the relations (3), and thus neither $x \succeq y$ nor $y \succeq x$ may hold. When that happens, we will say that the elements x and y are *incomparable*. When x and y are comparable, either $x \succeq y$ or $y \succeq x$.

M2. Since majorization is a relation between elements of $\mathcal{R}^n$, it is also a relation between elements of any subset $A \subset \mathcal{R}^n$.

M3. A key property is the existence of functions which are compatible with majorization. Suppose that ϕ is a real-valued function on $A \subset \mathcal{R}^n$, then ϕ is *compatible with majorization in A* if $x \succeq y$ $(x, y \in A)$ implies $\phi(x) \geq \phi(y)$. Such functions are called *Schur-convex functions*. In addition, ϕ is *strictly Schur-convex* if $x \succ y$ implies $\phi(x) > \phi(y)$.

There are infinitely many Schur-convex functions. Many important properties as well as specific families of Schur-convex functions are given in [23]. The following propositions identify families of Schur-convex functions that are useful to us. These propositions are directly taken from [23], or are minor modifications thereof:

S1. Any symmetric (strictly-) convex function defined on a set A is (strictly-) Schur-convex on A or any subset of A. The converse is not true. There are many Schur-convex functions that are not convex.

S2. If symmetric real-valued functions $\zeta_1, \ldots, \zeta_k$ have A as domain and are (strictly-) convex, and h is a real-valued (strictly-) increasing function on $\mathcal{R}^k$, then $\psi(x) = h(\zeta_1(x), \ldots, \zeta_k(x))$ is (strictly-) Schur-convex on A.

3.4 The Index of Severity

We can rephrase the previous discussion of severity of partitions from Section 3.2 in terms of majorization as follows: One n-partition is (strictly-) less severe than another n-partition if the relative sizes of the first partition (strictly-) majorize the relative sizes of the second partition.

The idea of Schur-convex functions suggests that we could construct functions which are compatible with this notion of severity. To do this, note that the relative sizes of the subsets of any n-partition belong to the set $S^n = \{x \in \mathcal{R}^n \mid \sum_i x_i = 1, x_i \geq 0\}$. This set is an $n-1$ dimensional simplex in $\mathcal{R}^n$ with n vertices. The vertices are located at all permutations of $(1, 0, \ldots, 0)$. We want a function $\phi : S^n \rightarrow \mathcal{R}$ to serve as an index of severity, so that the severity of any n-partition $\{A_1, \ldots, A_n\}$ is measured as $\phi(\mu^*(A_1), \ldots, \mu^*(A_n))$. It is reasonable to ask that ϕ be strictly Schur-convex on S^n. In fact, we can list all of the properties we would like ϕ to have starting from the simplest:

1) *Symmetry*. The labeling of the subsets of an n-partition as $\{A_1, \ldots, A_n\}$ is arbitrary. Any re-labeling of the same partition does not alter the severity of a partition. Since relabeling a partition only permutes the arguments of ϕ, a basic requirement is that ϕ should be invariant to permutations of its arguments, i.e. ϕ should be a symmetric function.
2) *Smoothness*. ϕ should be sufficiently smooth. We will take this to mean that ϕ should be at least continuous.
3) *Strict Schur-convexity*. As discussed above, ϕ should be strictly Schur-convex on S^n. This implies that ϕ is an "inverted" index of severity in the sense that larger values of $\phi(\mu^*(A_1), \ldots, \mu^*(A_n))$ indicate a less severe partitioning.
4) *Quasi-convexity*. Requiring ϕ to be strictly Schur-convex only constrains ϕ to behave well with respect to n-partitions that are comparable with majorization. We can additionally specify how ϕ should behave with respect to n-partitions that are incomparable with majorization. Suppose $u, w \in S^n$ are relative sizes of any two n-partitions, and that v is a point in S^n in the line segment from u to w. Then v has relative sizes that are a weighted average (a convex combination) of the relative sizes of u and w, and it is reasonable to ask that the severity of v be in between the severity of u and w. In particular, we can ask that the severity of v be greater than the smallest severity of u and w. Keeping in mind that ϕ is an "inverted" index, this can be written as $\phi(v) \leq \max(\phi(u), \phi(w))$, for $v = \alpha u + (1-\alpha)w$, all $0 \leq \alpha \leq 1$ and all $u, w \in S^n$. This implies that ϕ is a quasi-convex function on S^n.

Quasi-convexity is discussed in detail in [26], which gives the definition:

Definition. *A function $\phi : A \rightarrow \mathcal{R}$ is* quasi-convex *if its sub-level sets*

$$S_\alpha = \{x \in A \,|\, \phi(x) \leq \alpha\}$$

are convex.

The condition in requirement 4 above appears to be different from this definition but, as discussed in [26], the two are completely equivalent.

The following two facts about quasi-convexity are needed a little later:

Q1. Every convex function is quasi-convex. But the converse is not true.

Q2. If ζ is a convex function defined on a set A and h is a monotone increasing function, then $h \circ \zeta$ is a quasi-convex function on A.

Following this discussion, we may ask that an index of severity have all of the above-mentioned properties 1-4. However, the properties of symmetry and strict Schur-convexity are not independent of each other. To understand the relation between them, we need the following definition: A set $A \subset \mathcal{R}^n$ is *symmetric* if for $x \in A$, any y that is obtained

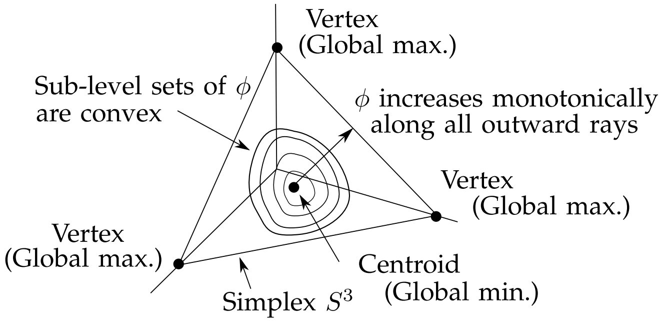


Fig. 4. The Structure of ϕ.

by permuting components of x also belongs to A. It is straightforward to see that S^n is a symmetric set ($x \in S^n$ if and only if $x_i \geq 0$ for all i, and $\sum_i x_i = 1$, and both of these relations are independent of permutations of the components of x). A property of (strictly-) Schur-convex functions is that when such functions are defined on a symmetric set, then they are necessarily symmetric [23]. Since S^n is symmetric, it is sufficient to require that ϕ be strictly Schur-convex; symmetry is a consequence of it.

Thus, we only need require ϕ to be smooth, strictly Schur-convex, and quasi-convex. The existence of such functions is easily established. There are infinitely many such functions.

This long discussion leads us to define

Definition. *A function $\phi : S^n \rightarrow \mathcal{R}$ is an* index of severity *or a* measure of severity *of an n-partition, if ϕ is smooth, strictly Schur-convex, and quasi-convex on S^n.*

What do such ϕ functions look like? Their structure is explicitly given by Theorem 1.

Theorem 1. *Any index of severity ϕ has the following properties:*

1) *ϕ has a single global minimum at the centroid of S^n, which is the point $c = (1/n, \ldots, 1/n)^T$.*
2) *ϕ has n global maxima, one at each vertex of S^n.*
3) *ϕ has no other local minima or maxima in S^n.*
4) *ϕ strictly increases along every half-line segment (ray) from the centroid of S^n.*
5) *Sub-level sets of ϕ are symmetric and convex.*

A proof of this theorem is given below in Section 3.6. Fig. 4 provides an illustration of Theorem 1 for $n = 3$.

Theorem 1 shows that the indices of severity are rather nice functions—they are almost convex, but not quite.

3.5 Examples

Below are some examples of functions that are valid indices of severity. We use these functions in the experiments reported in Section 6. The functions in equations (4)-(8) below are symmetric and strictly convex in the unit cube $[0,1]^n$ in $\mathcal{R}^n$, and hence on S^n. Therefore, they are valid indices of severity. The function in equation (9) below is the kth-root of the function in equation (8) (for $p = 2$) for $1 > k > 0$. For these values of k, the kth-root is strictly monotonically increasing, but non-convex. By property S2 and Q2 above, function in equation (9) is strictly Schur-convex and quasi-convex and hence a valid index of severity. But the function in equation (9) is not convex.

1) Negative Entropy:

$$\phi(u_1, \ldots, u_n) = \sum_i u_i \log u_i. \tag{4}$$

2) Negative Geometric Mean:

$$\phi(u_1, \ldots, u_n) = -\sqrt[n]{\prod (u_i + \epsilon)} \text{ for } \epsilon > 0. \tag{5}$$

For $\epsilon = 0$, this function is convex, but not strictly convex, on S^n. Hence ϵ is required to be strictly greater than 0.

3) Log-Sum-Exponential:

$$\phi(u_1, \ldots, u_n) = \log\left(\sum_i e^{u_i}\right). \tag{6}$$

4) $\mathcal{L}_p$ norm:

$$\phi(u_1, \ldots, u_n) = \sqrt[1/p]{|u_1|^p + \cdots + |u_n|^p} \text{ for } p > 1. \tag{7}$$

For $p = 1$ this function is a constant function on S^n, and is thus convex but not strictly convex. Requiring $p > 1$ makes the function strictly convex.

5) $\mathcal{L}_p$ deviation:

$$\phi(u_1, \ldots, u_n) = \sqrt[1/p]{\left|u_1 - \frac{1}{n}\right|^p + \cdots + \left|u_n - \frac{1}{n}\right|^p} \quad \text{for } p > 1. \tag{8}$$

6) kth-root $\mathcal{L}_2$ deviation:

$$\phi(u_1, \ldots, u_n) = \sqrt[k/2]{\left|u_1 - \frac{1}{n}\right|^2 + \cdots + \left|u_n - \frac{1}{n}\right|^2} \quad \text{for } 0 < k < 1. \tag{9}$$

3.6 Proof of Theorem 1

We now prove theorem 1. The proof is somewhat technical. To avoid losing continuity with the main argument of this paper, the reader may wish to skip this section at the first reading.

To prove Therorem 1, we begin with some preliminary results: The point $c = (1/n, 1/n, \ldots, 1/n)^T$ is the centroid of S^n. Given any point $x \neq c$ in S^n, a standard result in majorization is that $x \succ c$ [23]. We need a simple extension of this result. For $0 \leq \alpha \leq 1$, the set of points $(1-\alpha)x + \alpha c$ defines the line segment in S^n from x to c, and

Proposition 1. *For any* $0 < \alpha < 1$,

$$x \succ (1-\alpha)x + \alpha c \succ c.$$

Proof. We first note that for any $0 < \beta \leq 1$, $(\beta x)_{[i]} = \beta x_{[i]}$. That is, the ith ordered coordinate of the vector βx is β times the ith ordered coordinate of x. This is simply the result of β being a non-negative scalar.

Suppose $0 < \alpha < 1$, then $((1-\alpha)x)_{[i]} = (1-\alpha)x_{[i]}$, and since αc has the same value for each coordinate, $((1-\alpha)x + \alpha c)_{[i]} = ((1-\alpha)x)_{[i]} + \alpha c_{[i]} = (1-\alpha)x_{[i]} + \alpha c_{[i]}$.

Next, since $x \succ c$,

$$\sum_{i=1}^{k} x_{[i]} \geq \sum_{i=1}^{k} c_{[i]} \text{ for } k = 1, \dots, n-1 \tag{10}$$

$$\sum_{i=1}^{n} x_{[i]} = \sum_{i=1}^{n} c_{[i]} = 1, \tag{11}$$

with at least one of the inequalities in (10) being strict.

Because S^n is convex, the point $x_\alpha = (1-\alpha)x + \alpha c$ is in S^n for $0 < \alpha < 1$. Hence $\sum_{i=1}^{n} x_{\alpha[i]} = 1$. Multiplying the inequality of (10) by $(1-\alpha)$ on both sides and moving the term containing α from the right hand side to the left hand side gives

$$(1-\alpha)\sum_{i=1}^{k} x_{[i]} + \alpha \sum_{i=1}^{k} c_{[i]} \geq \sum_{i=1}^{k} c_{[i]} \text{ for } k = 1, \dots, n-1. \tag{12}$$

Because $(1-\alpha)x_{[i]} + \alpha c_{[i]} = ((1-\alpha)x + \alpha c)_{[i]} = x_{\alpha[i]}$, the inequality (12) can be written as

$$\sum_{i=1}^{k} x_{\alpha[i]} \geq \sum_{i=1}^{k} c_{[i]} \text{ for } k = 1, \dots, n-1. \tag{13}$$

Further observing that the steps (12-13) preserve strict inequalities, we have $x_\alpha \succ c$.

Similar to the above argument, multiplying the inequality of (10) by $(1-\alpha)$ on both sides, moving the term containing α from the left hand side to the right hand side, gives $x \succ x_\alpha$, establishing the proposition. □

An elementary consequence of Proposition 1 is

Lemma 1. *Set $x_K = x$ and $x_0 = c$. Further let $x_j, j = 1, \dots, K-1$ be distinct points on the line segment from x_0 to x_K such that x_j is contained in the open segment (x_{j-1}, x_{j+1}). Then, $x_K \succ x_{K-1} \succ \cdots \succ x_1 \succ x_0$.*

Proof. x_{K-1} is contained in the segment (x_{K-2}, x_K), and hence is contained in the segment (x_0, x_K). Since $x_K \succ x_0$ and x_{K-1} is distinct from x_K and x_0, applying Proposition 1 gives, $x_K \succ x_{K-1} \succ x_0$. Next, considering the point x_{K-2} and the segment (x_0, x_{K-1}), and repeating the same argument gives $x_{K-1} \succ x_{K-2} \succ x_0$, from which follows $x_K \succ x_{K-1} \succ x_{K-2} \succ x_0$. Iterating this argument establishes the result. □

Using these results we can establish Theorem 1.

Proof of Theorem 1:

1) Every non-centroid point x strictly majorizes the centroid c, and since ϕ is strictly Schur-convex $\phi(x) > \phi(c)$, which shows that there is a single global minimum at c.

2) Because all vertices of S^n are permutations of $(1, 0, \dots, 0)$, the symmetry of ϕ implies that ϕ takes the same value at all vertices. Further, since $(1, 0, \dots, 0)$ strictly majorizes all non-vertex points of S^n, the function ϕ has a global maximum at $(1, 0, \dots, 0)$ and hence a global maximum at every vertex. There are no other global maxima.

3,4) Every half-line segment (ray) from the centroid c intersects the boundary of S^n at some point, say x. Since x is a non-centroid point, x strictly majorizes the centroid. Further if $x_1, \dots, x_n$ are any sequence of distinct points in the interior of the line segment from c to x, then Lemma 1 shows that $x \succ x_n \succ x_{n-1} \succ \cdots \succ x_1 \succ c$. Since ϕ is strictly Schur-convex $\phi(x) > \phi(x_n) > \phi(x_{n-1}) > \cdots > \phi(x_1) > \phi(c)$, showing that ϕ is monotonically strictly increasing along the ray, establishing 4).

To establish 3), we first show that a non-centroid point of S^n cannot be a local minimum of ϕ. Because S^n is convex, every non-centroid point x of S^n can be connected to the centroid by a line segment. This line segment has a non-empty interior, and by the argument given above establishing part 4 of the theorem, ϕ takes values in the interior of the segment that are strictly less than $\phi(x)$ showing that any non-centroid point x is not a local minimum of ϕ.

Next we show that non-vertex points of S cannot be local maxima of ϕ. Let x be a non-vertex point of S^n. Then x has at least two coordinates that are not zero. Suppose that $x_{i_1}, \dots, x_{i_k} (k \geq 2)$ are all coordinates of x that are greater than zero. Since they must sum to 1 and $k \geq 2$, these coordinates are all strictly less than 1. Consider the simplex

$$\tilde{S} = \left\{ (x_{i_1}, \dots, x_{i_k}) \,|\, x_{i_m} \geq 0 \text{ for } m = 1, \dots, k, \sum_{m=1}^{m} x_{i_m} = 1 \right\}.$$

$\tilde{S}$ is a subset of S^n (but may be contained completely in the boundary of S^n), and because the definition of $\tilde{S}$ mirrors the definition of S^n, all of the properties established for S so far hold for $\tilde{S}$. Further, because the coordinates of x in $\tilde{S}$ are strictly greater than zero and strictly less than 1, x is in the interior of $\tilde{S}$. Consider two cases: First, suppose x is the centroid of $\tilde{S}$. Then by the reasoning of part 1 of this proof, x is the single global minimum of ϕ restricted to $\tilde{S}$. That is, every neighborhood of x in $\tilde{S}$, and hence in S^n, has points where ϕ takes values strictly greater than $\phi(x)$, showing that x is not a local maximum of ϕ. If x is not the centroid of $\tilde{S}$, then consider the ray from the centroid of $\tilde{S}$ passing through x. Because x is an interior point of $\tilde{S}$, this ray intersects the boundary of $\tilde{S}$ at a point $y \neq x$ and ϕ is strictly increasing along this ray. The line segment (x, y) has a non-empty interior and ϕ takes values strictly greater than $\phi(x)$ in the interior of the line segment showing that x is not a local maximum of ϕ. This establishes that non-vertex points of S cannot be local maxima of ϕ.

5) This is a direct consequence of quasi-convexity and symmetry of ϕ.

4 The Severity of a Partition and Image Registration

Suppose we have an m-partition $\{A_i\}$ and an n-partition $\{B_j\}$ of the unit cube $\Omega \subset \mathcal{R}^k$ (illustrated for $k = 2$ in Fig. 1) and a transformation $T_\theta : \Omega \to \Omega$ parametrized by θ. Each set B_j can be pulled back onto the first copy of Ω using the transformation T_θ as the set $\hat{B}_j(\theta) = T_\theta^{-1}(B_j)$, where we explicitly denote the dependence on θ. The sets $\hat{B}_j(\theta)$ are a partition of Ω and their intersections $A_i \cap \hat{B}_j(\theta)$ are a refinement of this partition. For a fixed j, let $\mu^*_{ij}(\theta)$ be the relative size of $A_i \cap \hat{B}_j(\theta)$ (relative to $\hat{B}_j(\theta)$). Then, $\phi(\mu^*_{1j}(\theta), \mu^*_{2j}(\theta), \ldots, \mu^*_{mj}(\theta))$ measures the severity of the partitioning of $\hat{B}_j(\theta)$ by $\{A_i \cap \hat{B}_j(\theta)\}$, with higher values indicating a less severe partition. Thus the weighted sum of $\phi(\mu^*_{1j}(\theta), \mu^*_{2j}(\theta), \ldots, \mu^*_{mj}(\theta))$ over all sets $\hat{B}_j(\theta)$ with weight equal to the size of $\hat{B}_j(\theta)$ is the average severity of the refined partition $\{A_i \cap \hat{B}_j(\theta)\}$. This weighted sum is:

$$J(\theta) = \sum_{j=1}^{n} \mu_j(\theta) \phi\big(\mu^*_{1j}(\theta), \mu^*_{2j}(\theta), \ldots, \mu^*_{mj}(\theta)\big), \tag{14}$$

where $\mu_j(\theta)$ is the size of $\hat{B}_j(\theta)$. The value of θ that maximizes this function gives the transformation that best aligns the partitions $\{\hat{B}_j(\theta)\}$ to the partitions $\{A_i\}$ in an average sense, thereby registering the partitions.

Going from aligning partitions to registering images is straightforward. Given two images $I_1 : \Omega \to \mathcal{R}$ and $I_2 : \Omega \to \mathcal{R}$ use a segmentation algorithm to m- and n-partition the domains of the images into subsets $\{A_i\}$ and $\{B_j\}$ respectively. Maximizing the objective function of equation (14) with respect to θ using these partitions registers the two images.

The mutual-information objective function does exactly this, as shown below.

5 Image Registration with Mutual Information

To start, consider the histogramming process. Let $I : \Omega \to \mathcal{R}$ be an image and $B : \mathcal{R} \to \{1, \ldots, n\}$ be the binning function that maps an image intensity to the index of the appropriate histogram bin. Then $\alpha_i = (B \circ I)^{-1}(i)$ is the set of pixels of image I that contribute to the histogram in bin i. The value of the histogram in this bin is the size of the set α_i, which is $\mu_i = \mu(\alpha_i)$. Finally, the discrete-valued random variable corresponding to binning is $X = B \circ I$, and $p(X = i) = \mu_i$.

Applying this to the image I_1 with a histogram of m bins and to the transformed image $I_2 \circ T_\theta$ with a histogram of n bins gives the following:

1) The set of pixels of image I_1 that contribute to the histogram in bin i is $\sigma_i = (B \circ I_1)^{-1}(i)$. The value of the histogram in this bin is the size of the set σ_i, which is $\mu_i = \mu(\sigma_i)$. The discrete-valued random variable is $X = B \circ I_1$, and $p(X = i) = \mu_i$.
2) The set of pixels of image $I_2 \circ T_\theta$ that contribute to the histogram in bin j is $\tau(\theta)_j = (B \circ I_2 \circ T_\theta)^{-1}(j)$. The value of the histogram in this bin is the size of the set $\tau(\theta)_j$, which is $\mu(\theta)_j = \mu(\tau(\theta)_j)$. The discrete-valued random variable is $Y(\theta) = B \circ I_2 \circ T_\theta$, and $p(Y(\theta) = j) = \mu(\theta)_j$. We are explicitly denoting the dependence on θ for all quantities associated with the pulled-back image.

Finally, letting the size of the intersection of $\sigma_i \cap \tau(\theta)_j$ be $\mu(\theta)_{i,j} = \mu(\sigma_i \cap \tau(\theta)_j)$ gives the joint probability

$$p(X = i, Y(\theta) = j) = \mu(\theta)_{i,j}.$$

Therefore, the mutual-information image registration objective function $MI(\theta)$ is

$$\begin{aligned} MI(\theta) &= H(X) + H(Y(\theta)) - H(X, Y(\theta)) \\ &= -\sum_{i=1}^{m} p(X = i) \log p(X = i) \\ &\quad - \sum_{j=1}^{n} p(Y(\theta) = j) \log p(Y(\theta) = j) \\ &\quad + \sum_{i,j=1,1}^{m,n} p(X = i, Y(\theta) = j) \log p(X = i, Y(\theta) = j). \end{aligned} \tag{15}$$

It is now straightforward to establish the following result:

Theorem 2. $MI(\theta) = \text{const.} + J(\theta)$, *where const. is a term that is independent of θ and $J(\theta)$ is the function in equation (14) with negative entropy as the index of severity.*

Proof. Using a standard result from information theory

$$MI(\theta) = H(X) - H(X \,|\, Y(\theta)), \tag{16}$$

The first term on the right hand side of the above equation depends only on X and is independent of θ. The second term on the right hand side of the above equation is the negative conditional entropy of X given $Y(\theta)$

$$\begin{aligned} &- H(X \,|\, Y(\theta)) \\ &= \sum_{j=1}^{n} p(Y(\theta) = j) \\ &\quad \times \left\{ \sum_{i=1}^{m} p(X = i \,|\, Y(\theta) = j) \log p(X = i \,|\, Y(\theta) = j) \right\} \\ &= \sum_{j=1}^{n} \mu_j(\theta) \phi\left(\frac{\mu_{1,j}(\theta)}{\mu_j(\theta)}, \ldots, \frac{\mu_{m,j}(\theta)}{\mu_j(\theta)} \right) \\ &= \sum_{j=1}^{n} \mu_j(\theta) \phi\big(\mu^*_{1,j}(\theta), \ldots, \mu^*_{m,j}(\theta)\big) \\ &= J(\theta), \end{aligned}$$

where ϕ is the negative entropy function of equation (4). This establishes the theorem. □

Theorem 2 is our deterministic explanation of why mutual information works for image registration. To summarize: the two images are partitioned by histogramming and the mutual information objective function finds the transformation that minimizes the average severity of the refinement of the pulled-back partition of the moving image. The negative entropy function serves as an index of severity.

6 Experimental Results

According to the partition-alignment idea, other indices of severity besides negative entropy should also be able to register images. In this section, we experimentally investigate

TABLE 1
The ϕ Functions, Their Parameter Values, and Short-Form Names

Short form name for Figures and Tables	Defining Eq. for ϕ	Parameter Value
Mutual Info	(4)	None
Neg. Geom. Mean	(5)	$\epsilon = 0.1$
Log. Exp.	(6)	None
$\mathcal{L}_p$ Norm ($p = 2$)	(7)	$p = 2$
$\mathcal{L}_p$ Norm ($p = 3$)	(7)	$p = 3$
$\mathcal{L}_p$ Dev. ($p = 1.5$)	(8)	$p = 1.5$
Root $\mathcal{L}_2$ Dev.	(9)	$k = 0.5$
Root $\mathcal{L}_2$ Dev.	(9)	$k = 0.75$

the performance of the objective function J of equation (14) for image registration with ϕ set to functions of equation (4-9) of Section 3.5. To be concise, we identify the objective function simply by the ϕ function used in its formulation. Thus, we refer to the image registration with the $\mathcal{L}_2$ norm as the ϕ function simply as "registration with $\mathcal{L}_2$". When we use negative entropy as ϕ, we call the registration "mutual-information registration" since the equivalence of the two is established by Theorem 2.

Some of the ϕ functions of Section 3.5 require parameters. The values of the parameters used in the experiments are given in Table 1. Note that the $\mathcal{L}_p$-norm is used with two different values of p. In the figures and tables that follow, we refer to the ϕ functions by short-form names; these names too are given in Table 1.

In the experiments, the fixed and the moving images were partitioned by histogramming with equal number of bins. Multiple experiments were conducted with the number of bins systematically varied from 10–15. The conclusions from all of these experiments were essentially identical, so we only report the results of experiments with 15 bins. Higher number of bins may also be used, but 15 bins are sufficient for accurate registration in our experiments.

The objective function $J(\theta)$ of equation (14) is not differentiable with respect to θ because the $\mu^*_{i,j}(\theta)$'s are not differentiable with respect to θ. In order to make the objective function differentiable, we smoothed the binning function by altering its definition in the following way. Recall from Section 5 that the binning function B maps real numbers to the discrete set $\{1, \ldots, N\}$. Instead, we created smooth binning functions $\tilde{B}_i, i = 1, \ldots, N$ that mapped the image intensities to $[0, 1]$ as shown in Fig. 5. The functions $\tilde{B}_i$ are a partition of unity. The various measures used in the objective function are defined in terms of these smooth binning functions as follows:

$$\mu_j(\theta) = \int \tilde{B}_j \circ I_2(\theta) \circ T(\theta)(u)\, du, \tag{17}$$

$$\mu_{i,j}(\theta) = \int \left\{\tilde{B}_j \circ I_2(\theta) \circ T(\theta)(u)\right\} \times \left\{\tilde{B}_i \circ I_1(u)\right\} du, \tag{18}$$

$$\mu^*_{i,j}(\theta) = \frac{\mu_{i,j}(\theta)}{\mu_j(\theta)}. \tag{19}$$

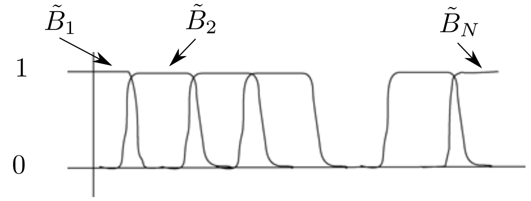


Fig. 5. The partition of unity functions $\tilde{B}_i$.

With these definitions, the objective function in equation (14) is continuously differentiable.

The transformation $T(\theta)$ is taken as the 3d affine transformation $u \to Au + t$, where $u \in \mathcal{R}^3$, A is a 3×3 invertible matrix, and $t \in \mathcal{R}^3$ is a translation. The parameters θ of the transform are the three translation components and the nine elements of the matrix A.

Instead of maximizing the objective function $J(\theta)$ with θ, we equivalently minimize $-J(\theta)$ with respect to θ. The minimization is carried out by a multi-scale algorithm where the volumes are down-sampled by factors of 4, 2, and 1. Minimization proceeds from the coarsest scale to the finest by using the minimum in the coarser-scale as initialization for the next finer scale. No downsampling is used at the last (the finest) scale. The minimum at this scale is taken as the registration. The minimization at every scale is carried out by the Broyden-Fletcher-Goldfarb-Shanno (BFGS) method which is a popular quasi-Newton minimization method. BFGS is used in two alternating phases. The first phase minimizes with respect to the translation t, and the second phase minimizes with respect to the matrix A. BFGS worked quite reliably with all of the objective functions in all experiments.

Three-dimensional T1 and PD (proton density) brain images obtained from the MNI phantom [25] were used as data (with default parameters TR/TE = 18ms/10 ms). Because they are three-dimensional, we refer to them as the T1 and PD volumes rather than T1 and PD images. Both volumes were obtained at an isotropic 1 mm resolution without additive noise and without intensity non-uniformity. A scatter plot of the joint intensities of the T1 and PD (Fig. 6) reveals a non-linear and multi-valued relation. Some T1-values (e.g. T1 = 400, marked by the vertical dashed line in Fig. 6) correspond to more than a single PD-value. Similarly some PD-values (e.g. PD = 5200, marked by the horizontal dashed line in Fig. 6) correspond to multiple T1-values. This lack of a simple functional relation

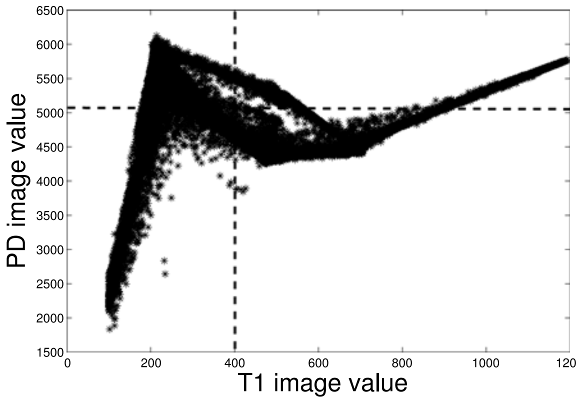
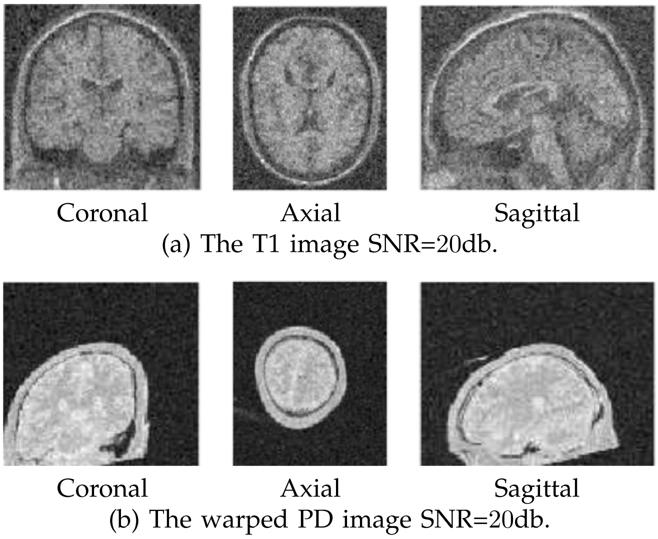


Fig. 6. Scatter plot of joint T1-PD image values.



(a) The T1 image SNR=20db.

(b) The warped PD image SNR=20db.

Fig. 7. The T1 and warped PD-images at SNR $= 20$db.

makes the T1- and PD-volumes good candidates for registration with mutual-information.

The PD volume was warped by a known affine transformation, then zero mean Gaussian noise was added to it and the T1 volume. The amount (i.e. the variance) of the noise was set to give signal-to-noise ratios (SNRs) of 20 db. The warped, noisy PD volume was registered back to the noisy T1 volume minimizing the objective function $-J(\theta)$ with all of the ϕ functions of Table 1. This was repeated 20 times with different realizations of noise.

The MNI phantom gives images which are already registered. As mentioned above, we first warped the PD image by $Au+t$ and then registered the warped image back to the T1 by A^*u+t^*. If the registration is exact, then the composition of these two warps is identity, giving $A^*A=I$ and $t^*=-At$. In our experiments, the translations were almost exactly recovered, hence we evaluated the accuracy of registration by checking how close A^*A was to I by calculating the Frobenius norm of A^*A-I.

6.1 Simulations

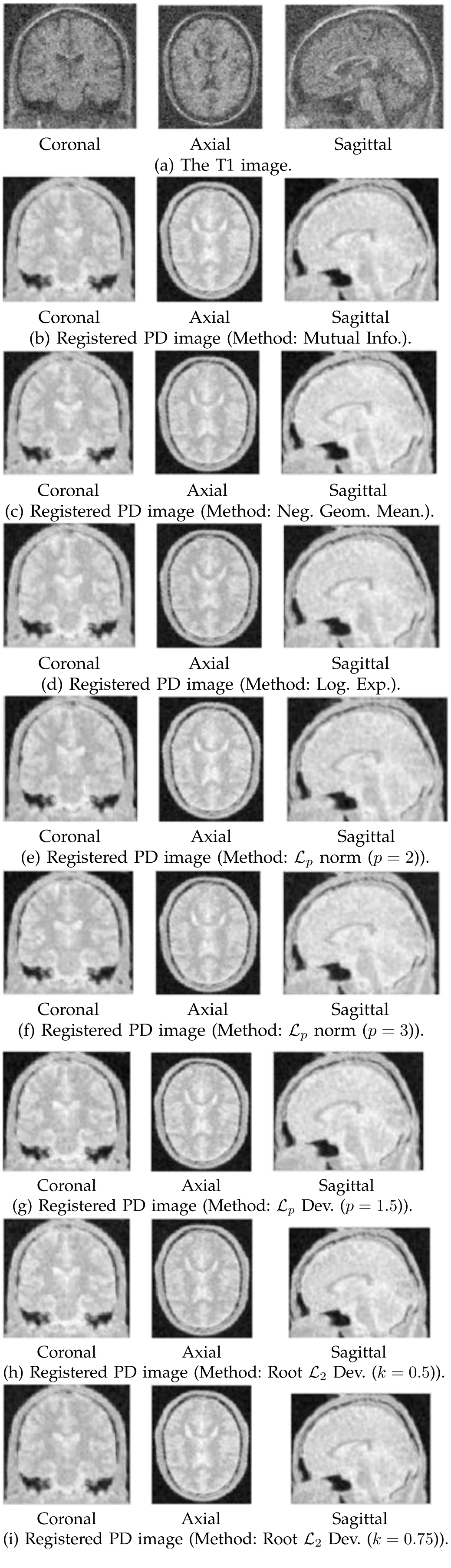
Fig. 7 shows the axial, coronal, and sagittal slices for the T1 volume and the warped PD volume for SNR $= 20$ db. The amount of noise is visually quite significant. Fig. 8a repeats the T1 slices for reference and fig. 8b-i show slices of the noisy PD volume after registration to the T1 volume using the ϕ functions of Table 1. Visually at least, the registrations in Fig. 8b-i appear to be indistinguishable from each other.

Fig. 9 shows box plots of the Frobenius norms of A^*A-I for different ϕ's. Each box plot shows the median, the 25 percent and the 75 percent quantiles, and the entire range of the values of the Frobenius norm.

Several observations can be made from Fig. 9:

1) The medians of the Forbenius norms of A^*A-I for all ϕ functions are quite similar. The median values are given in the second column of Table 2.

   Since A^*A is a 3×3 matrix, the value in the second column Table 2 divided by 9 is roughly the error in each element of A^*A. All of these errors are less than 0.01.



(a) The T1 image.

(b) Registered PD image (Method: Mutual Info.).

(c) Registered PD image (Method: Neg. Geom. Mean.).

(d) Registered PD image (Method: Log. Exp.).

(e) Registered PD image (Method: $\mathcal{L}_p$ norm ($p=2$)).

(f) Registered PD image (Method: $\mathcal{L}_p$ norm ($p=3$)).

(g) Registered PD image (Method: $\mathcal{L}_p$ Dev. ($p=1.5$)).

(h) Registered PD image (Method: Root $\mathcal{L}_2$ Dev. ($k=0.5$)).

(i) Registered PD image (Method: Root $\mathcal{L}_2$ Dev. ($k=0.75$)).

Fig. 8. The registered PD-images for different ϕ functions.

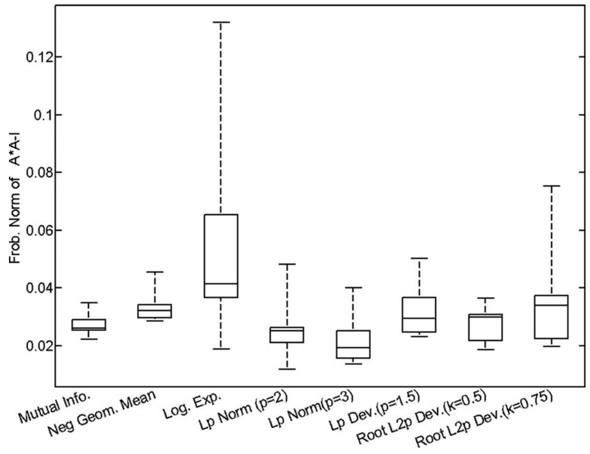


Fig. 9. The Frobenius norm of $A^*A - I$.

TABLE 2
The Median and the Median Absolute Deviation (MAD) of the Frobenius Norms of $A^*A - I$ for Different ϕ

Forbenius Norms of $A^*A - I$		
ϕ	Median	MAD
Mutual Info.	0.026	0.0012
Neg. Geom. Mean	0.032	0.0026
Log. Exp.	0.040	0.0125
$\mathcal{L}_p$ Norm ($p = 2$)	0.025	0.0025
$\mathcal{L}_p$ Norm ($p = 3$)	0.019	0.0032
$\mathcal{L}_p$ Dev. ($p = 1.5$)	0.027	0.0043
Root $\mathcal{L}_2$ Dev. ($k = 0.5$)	0.030	0.0044
Root $\mathcal{L}_2$ Dev. ($k = 0.75$)	0.032	0.0073

2) The smallest median value occurs for $\mathcal{L}_p$ norm for $p = 3$. Moreover, the median values for the $\mathcal{L}_p$ norm for $p = 2, 3$ are less than the median value for mutual information.
3) The median values for root $\mathcal{L}_2$ deviations ($k = 0.5$ and $k = 0.75$) are comparable to the median value of mutual information showing that convexity of ϕ is not necessary for image registration.
4) The medians of the absolute deviation from the median (column 3 of Table 2) are all significantly smaller than the medians, showing that all of the objective functions are reasonably robust to noise realizations.

The above results—especially the facts that most ϕ functions have a performance that is similar to mutual-information, and that the convexity of ϕ does not seem to be essential—support the theory of this paper. These results suggest that it is reasonable to interpret mutual-information registration as an instance of an algorithm that minimizes the average severity of partitions.

7 Implications and Conclusion

We now discuss some implications of our theory:

1) The theory shows the fundamental importance of partitioning to registration. However, the means by which partitioning is done is not constrained by the theory. This suggests that multi-modal registration may be achieved with many segmentation techniques, thus making it possible to use level set techniques or Markov Random Field techniques for multi-modal registration. This could vastly enlarge the class of images that can be registered. For example, it may be possible to register ultrasound images which can be partitioned more sensibly with techniques other than intensity-based histogramming.
2) It is a common experience that the number of bins have to be carefully chosen when using MI registration. A poor choice of bins often leads to failure. Our theory explains why this might happen: As a concrete example, consider registering a T1-weighted brain image to a fractional anisotropy (FA) brain image. A desirable result of this registration is to map the high FA regions to white matter tracts in the T1 image. Now suppose that both images are noisy, and we happen to choose a small number of bins for histogramming. With a small number of bins, the intensity that separates gray from white matter is likely to fall in the middle on a bin rather than at the boundary of a bin. If this happens, the T1 image partition will not conform to the boundary of gray and white matter, and the high FA regions will not map exactly to the white matter tracts, since the white matter tracts are not delineated by any of the T1 partitions, causing a registration “failure”.

   As the number of bins increases, the gray-white separating intensity will get closer to a histogram bin boundary (just because there are more bins) fixing this anomaly. If the number of bins increases even further, many spurious T1-partitions will appear as noise islands. Aligning these partitions will not map high FA regions exactly to white matter either, causing another registration “failure”. This explains the need for choosing the right number of bins.
3) The theory also opens the question of the “best” ϕ function to use in registration. The ideas of Section 3.2 suggest that ϕ should be strictly Schur-convex and quasi-convex in order to be an index of severity. There are infinitely many such functions. Although they all are valid indices of severity, they are unlikely to perform equally well for a given class of images. It is likely that finding the best ϕ will require some sort of a model for the class of images to be registered, and a precise notion of what metric to measure the “best” with. This idea is very appealing from an applications point of view because in applications it is common to repeatedly register the same class of images, e.g., head MRI images with head CT images.

We hope to pursue these implications in the future.

In conclusion, this paper provides a deterministic explanation of why mutual information works for image registration. Our theory suggests that MI registration works because it minimizes the average severity of refinement of the partitions of the “moving” image. The notion of severity of a partition is new, and as suggested in this paper, it is precisely measured using majorization, Schur-, and quasi-convexity. This theory opens the possibility for further advances in multi-modal registration of complex images.

ACKNOWLEDGMENTS

The authors acknowledge useful discussions with and comments from Yunho Kim. Lisa Berlinger, Nicha Dvornek, and Chenxi Huang helped with the manuscript.

REFERENCES

[1] A. Collignon, F. Maes, D. Delaere, D. Vandermeulen, P. Suetens, and G. Marchal, "Automated multi-modality image registration based on information theory," in *Information Processing in Medical Imaging*, Y. Bizais, C. Barillot, and R. Di Paola, Eds. Norwell, MA, USA: Kluwer Academic, 1995, pp. 263–274.

[2] P. Viola and W. M. I. I. I Wells, "Alignment by maximization of mutual information," in *Proc. 5th Int. Conf. Comput. Vis.*, 1995, pp. 16–23.

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

[4] J. Hajnal and D. Hill, *Medical Image Registration*. Boca Raton, FL, USA: CRC Press, 2001.

[5] J. Modersitzki, *FAIR: Flexible Algorithms for Image Registration*. Philadelphia, PA, USA: SIAM, 2009.

[6] A. Roche, G. Malandain, and N. Ayache, "Unifying maximum likelihood approaches in medical image registration," *Int. J. Imag. Syst. Tech.*, vol. 11, pp. 71–80, 2000.

[7] J. Bierlant, E. Dudewicz, L. Gyorfi, and E. C. van der Meulen, "Nonparametric entropy estimation: An overview," *Int. J. Math. Stat. Sci.*, vol. 6, no. 1, pp. 17–39, 1997.

[8] A. Rajwade, A. Banerjee, and A. Rangarajan, "Probability density estimation using isocontours and isosurfaces: Applications to information theoretic image registration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 3, pp. 475–491, Mar. 2009.

[9] M. Rao, Y. Chen, B. C. Vemuri, and F. Wang, "Cumulative residual entropy: A new measure of information," *IEEE Trans. Inf. Theory*, vol. 50, no. 6, pp. 1220–1228, Jun. 2004.

[10] F. Wang and B. C. Vemuri, "Non-rigid multi-modal image registration using cross-cumulative residual entropy," *Int. J. Comput. Vis.*, vol. 74, pp. 201–215, 2007.

[11] B. Ma, A. Hero, J. Gorman, and O. Michel, "Image registration with minimum spanning tree algorithm," in *Proc. IEEE Int. Conf. Image Process.*, 2000, vol. 1, pp. 481–484.

[12] A. Costa and A. Hero, "Entropic graphs for manifold learning," in *Proc. IEEE Asilomar Conf. Signals, Syst. Comput.*, 2003, vol. 1, pp. 316–320.

[13] H. Neemuchwala and A. Hero, "Entropic graphs for registration," in *Multisensor Image Fusion and Its Applications*, R. S. Blum and Z. Liu, Eds. New York, NY, USA: Marcel Dekker, 2005.

[14] H. Neemuchwala, A. Hero, S. Zabuawala, and P. Carson, "Image registration methods in high-dimensional space," *Int. J. Imag. Syst. Tech.*, vol. 16, no. 5, pp. 130–145, Mar. 2007.

[15] M. Sabuncu and P. Ramadge, "Gradient-based optimization of an EMST image registration function," in *Proc. IEEE Int. Conf. Acoust. Speech. Signal Process.*, 2005, vol. 2, pp. 253–256.

[16] C. Studholme, D. L. G. Hill, and D. J. Hawkes, "An overlap invariant entropy measure of 3D medical image alignment," *Pattern Recognit.*, vol. 32, pp. 71–86, 1999.

[17] D. Loeckx, P. Slagmolen, F. Maes, D. Vandermeulen, and P. Seutens, "Nonrigid image registration using conditional mutual information," *IEEE Trans. Med. Imag.*, vol. 29, no. 1, pp. 19–29, Jan. 2010.

[18] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "f-information measures in medical image registration," *IEEE Trans. Med. Imag.*, vol. 25, no. 12, pp. 1508–1516, Dec. 2004.

[19] Y. He, A. B. Hamza, and H. Krim, "A generalized divergence measure for robust image registration," *IEEE Trans. Signal Process.*, vol. 51, no. 5, pp. 1211–1220, May 2005.

[20] N. J. Tustison, S. P. Awate, G. Song, T. S. Cook, and J. C. Gee, "Point set registration using Havrda-Charvar-Tsallis entropy measures," *IEEE Trans. Med. Imag.*, vol. 30, no. 2, pp. 451–460, Feb. 2011.

[21] P. A. van den Elsen, E.-J. D. Pol, T. S. Sumanaweera, P. F. Hemler, S. Napel, and J. R. Adler, "Grey value correlation techniques used for automatic matching of CT and MR brain and spine images," *Proc. SPIE*, vol. 2359, pp. 227–237, Sep. 1994.

[22] P. Gerlot-Chiron and Y. Bizais, "Registration of multimodality medical images using a region overlap criterion," *Graph. Models Image Process.*, vol. 54, no. 5, pp. 396–406, 1992.

[23] A. W. Marshall, I. Olkin, and B. C. Arnold, *Inequalities: Theory of Majorization and Its Applications*, (Springer Series Statistics). New York, NY, USA: 2010.

[24] M. J. Steele, *The Cauchy-Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities*. Washington, DC, USA: Math. Assoc. Amer., 2004.

[25] [Online]. Available: http://www.bic.mni.mcgill.ca/brainweb/

[26] S. Boyd and L. Vendeberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

Hemant D. Tagare received the PhD degree from Rice University. He is a professor in the Department of Diagnostic Radiology, the Department of Biomedical Engineering, and the Department of Electrical Engineering at Yale University. His research interests are in bio-medical signal and image analysis. He is a senior member of the IEEE.

Murali Rao is a professor in the Department Mathematics at the University of Florida, Gainesville. His research interests are in probability and information theory and its applications.