Medical Image Analysis (1999) volume 4, number ?, pp 1–17

Rigid Point Feature Registration Using Mutual Information

Anand Rangarajan[1*], Haili Chui[2] and James S. Duncan[1]

[1]Departments of Diagnostic Radiology and Electrical Engineering, Yale University
[2]Department of Electrical Engineering, Yale University

Abstract

We have developed a new mutual information-based registration method for matching unlabeled point features. In contrast to earlier mutual information-based registration methods which estimate the mutual information using image intensity information, our approach uses the point feature location information. A novel aspect of our approach is the emergence of correspondence (between the two sets of features) as a natural by-product of joint density estimation. We have applied this algorithm to the problem of geometric alignment of primate autoradiographs. We also present preliminary results on 3D robust matching of sulci derived from anatomical MR. Finally, we present an experimental comparison between the mutual information approach and other recent approaches which explicitly parameterize feature correspondence.

Keywords: point feature registration, rigid alignment, mutual information, similarity transformation, spatial mapping, correspondence, joint probability, softassign

Received ?; revised ?; accepted ?

1. INTRODUCTION

In developing a taxonomy of registration methods, it is common to make a basic distinction between feature-based (Besl and McKay, 1992) and voxel-based matching methods (Van den Elsen et al., 1993). Within the class of feature-based matching, the typical problems encountered are: choice and extraction of the features from the underlying image, choice of the distance measure and parameterization of the spatial mapping between features extracted from the images, design of the algorithm that matches homologous features, minimizes the distance measure and returns a good spatial mapping. Within the class of voxel-based matching, the typical problems encountered are: choice and parameterization of the intensity mapping, choice of the distance measure and parameterization of the spatial mapping, design of the algorithm that matches intensities and returns a good spatial mapping.

Recently (Wells III et al., 1996; Maes et al., 1997), there has been tremendous interest in using mutual information-based similarity measures in voxel based registration. Two important reasons behind this new enthusiasm are the comparative ease of matching intensity images across modalities (for example MR and PET) and the ability to discard intensity patterns in either modality that are not relevant to registration (robustness property). Mutual information-based registration begins with the estimation of the joint probability of the intensities of corresponding voxels in the two images. Typically, the joint probability is estimated via histogramming (Maes et al., 1997), Parzen windows (Wells III et al., 1996) or other non-parametric statistical methods. Since the joint probability is a function of the spatial mapping between the two images, mutual information methods re-estimate the joint probability at each step before updating/improving the spatial mapping. Needless to say, common intensity patterns provide all the information necessary for estimating the joint probability. In this sense, voxel-based mutual information methods can be seen as a powerful extension of standard intensity correlation methods.

The use of information theoretic measures such as mutual information has obviously benefited voxel-based registration. It is natural to ask whether mutual information can play a similar role in feature-based matching as well. We believe it can and the present paper demonstrates that mutual information

*Corresponding Author
(e-mail: anand@noodle.med.yale.edu)

can be used to parameterize and solve the correspondence problem in feature-based registration. That mutual information can be successfully employed in feature-based matching is a surprising fact. While the existence of highly correlated intensity patterns—the source of the mutual information in the first place—could be assumed in voxel-based matching, in the case of matching unlabeled point features, it cannot. Note that when we have access to a feature image (for example a gradient magnitude intensity image), we can still maximize the mutual information using intensity information. On the other hand, when we merely have two sets of unlabeled features, we have to rely on the existence of common structural (and not intensity) patterns. The joint probability now has to be defined w.r.t. the features (under the current spatial mapping) and not the voxel intensities. It turns out that the joint probability between the two sets of features plays a crucial role in determining the feature *correspondences* or homologies and the *outliers* or non-homologies. It also turns out that the joint probability between the feature sets bears a remarkable similarity to the fuzzy correspondence matrix used in our earlier work on feature matching (Rangarajan, 1997; Rangarajan et al., 1997a).

Once the joint probability between the two sets of features is computed, the mutual information, being a function of the joint probability, can be easily computed as well. However, we do not have recourse to the familiar methods of histogramming or Parzen windows. Instead we rely on the *maximum entropy* (Jaynes, 1982) principle to estimate the joint probability. Maximum entropy methods are well known for their ability to produce probability distributions and/or densities that are least biased. In our case, once we specify the distance measure between the features, the maximum entropy principle can be invoked to produce a joint probability estimate. Just as in the earlier voxel-based matching case, the joint probability is a function of the distance measure (between the feature sets) which in turn depends on the spatial mapping.

Given the joint probability, we can estimate the mutual information between the feature sets and maximize it. It turns out that the maximization is somewhat cumbersome. Based on our considerable experience with fuzzy correspondence-based feature matching, we use an alternative strategy: use mutual information as a prior in the estimation of the joint probability itself. The final energy function used is a function of both the spatial mapping and the joint probability, exactly analogous to previous methods in feature matching which are functions of the correspondence and the spatial mapping. This permits us to derive an extremely simple *discrete-time* algorithm which eschews step-sizes and canned nonlinear function optimization methods. The resulting algorithm is very similar to the Expectation-Maximization (EM) algorithm used in estimating mixture densities. Essentially it is an alternating algorithm that alternates between estimation of the joint probabilities and the spatial mapping. We have derived the algorithm for the case of 2-D feature matching. It should be fairly straightforward to extend our approach to include other distance measures and to 3-D. We predict that such extensions will continue to bear a striking resemblance to our work employing affine (Gold et al., 1998) and piecewise affine transformations (Pappu et al., 1996) in 2D and dual quaternion parameterized spatial mappings (Gold et al., 1998) in 3D. With this connection established between mutual information and point feature matching, it should be possible to set up a combined (voxel- and feature-based) registration method that is now unified under the rubric of mutual information.

2. MUTUAL INFORMATION POINT MATCHING

The mutual information between two unlabeled point-sets is a function of the chosen spatial mapping (for example, rigid, similarity, affine). Below, we keep the formulation at a general level. Later, we will specialize to the case of similarity transformations in 2D.

2.1. The Spatial Mapping

Denote the point-sets by $X_i,\ i = 1,2,\dots,N_1$ and $Y_j,\ j = 1,2,\dots,N_2$ respectively. The point sets are assumed to be in $\mathcal{R}^2$ or $\mathcal{R}^3$. N_1 and N_2 are the numbers of points in the sets X and Y respectively. Assume for the moment that $N_1 = N_2 = N$ and that the correspondence of X_i to Y_i is known. Then, a suitable choice for the distance measure between X and Y is

$$D(T) = \sum_{i=1}^{N} \|X_i - TY_i\|^2. \tag{1}$$

In (1), T is the spatial mapping (rigid, similarity, affine) linking the two point-sets. Note that (1) assumes that the point-to-point correspondences are known and that there are no outliers in either point-set. We now turn to the situation where the correspondences are unknown and where outliers are present—points in X and Y that do not have homologies in the other set.

2.2. Correspondence, Joint Probability and Maximum Entropy

In our previous work (Rangarajan, 1997), we parameterized the point-to-point feature correspondences via *permutations*. Outliers or non-homologous features were represented using binary outlier variables. As mentioned in the Introduction, in the present work, the joint probability between feature index i in X and feature index j in Y characterizes correspondence. However, there is no need for a separate binary outlier array to represent the non-homologies—a point feature in X is approximately an outlier if the joint probabilities between that feature and all features in Y are very low. The same holds for outlier features in Y.

The joint probability between feature i in X and feature j in Y is written as $\Pr(I = i,\ i \in \{1, \ldots, N_1\}, J = j,\ j \in \{1, \ldots, N_2\})$ and is denoted by P_{ij}. Note that the random variables in the above joint probability are *not* locations. Instead, we have random variables defined on *indices* (I and J). The joint probability P_{ij} is the association probability between indices. For instance, if we picked point 5 from X and point 7 from Y, P_{57} would be a measure of association or correspondence between those two point features. The marginal probability $Q_i = \sum_{j=1}^{N_2} P_{ij}$ ($R_j = \sum_{i=1}^{N_1} P_{ij}$) determines the degree of association between a point i (j) in X (Y) and the entire set Y (X). If the marginal probability of any point (in either set) is low, it can be regarded as an outlier.

We generalize the earlier Euclidean distance measure between *corresponding* points. Define a distance D_{ij} between i in X and j in Y:

$$D_{ij}(T) \stackrel{\text{def}}{=} \|X_i - TY_j\|^2.$$

We have defined a cross-product of distances between each point feature i in X and j in Y. The cross-product distance plays an important role in the estimation of the joint probabilities P_{ij}. Given the distance measure D, we may define the *expected* value of the overall point matching distance:

$$E(P,T) = \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} P_{ij} D_{ij}(T) \tag{2}$$

A well known method of estimating probability distributions is the method of *maximum entropy* (Jaynes, 1982). Recall that the joint probability P_{ij} between feature i in X and feature j in Y is $\Pr(I = i,\ i \in \{1, \ldots, N_1\}, J = j,\ j \in \{1, \ldots, N_2\})$. The above joint probability characterizes the likelihood of obtaining a pair of features (i, j), i from X and j from Y. Intuitively, the likelihood should be large if i and j are homologies and small if they are not. Also, note that (under a good spatial mapping), the distance measure D_{ij} should be small for homologies and large for non-homologies. The maximum entropy method attempts to maximize the number of correspondence possibilities while being constrained by the aforementioned expected point matching energy.

The entropy is

$$-\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} P_{ij}\log P_{ij} \tag{3}$$

The method of maximum entropy works by maximizing the entropy subject to the problem constraints which are

$$P_{ij} \geq 0 \text{ and } \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} P_{ij} = 1.$$

(The constraints on P reflect the fact that it is a joint probability.) As before, the expected value of the energy is

$$E(P,T) = \sum_{i=1}^{N_1}\sum_{j=1}^{N_2} P_{ij} D_{ij}(T)$$

Setting the expected value of the energy $E(P,T) = d$ is tantamount to setting the noise level in the system. Since $D_{ij}(T) = \|X_i - TY_j\|^2$, the expected value of D sums over the product space of the features. When alignment occurs, the expected value of D is just the residual noise *provided the joint probability matrix P_{ij} returns an adequate correspondence*. Note that the expected value of the energy can be broken down into two terms—the alignment error on corresponding features and a cross-product distance between all pairs of non-homologous features. When the features are misaligned, the expected value of D will be higher due to the presence of alignment error (as well as residual noise from the non-homologous features).

The second constraint on P is the familiar $\sum_{ij} P_{ij} = 1$. Using the method of Lagrange multipliers, we get

$$\min_P \max_{\lambda,\alpha} E_{\text{ME}}(P,\lambda,\alpha) = \alpha\left(\sum_{ij} P_{ij}D_{ij}(T) - d\right) + \lambda\left(\sum_{ij} P_{ij} - 1\right) + \sum_{ij} P_{ij}\log P_{ij}. \tag{4}$$

In (4), α is a Lagrange parameter satisfying the constraint $E(P,T) = d$. Unfortunately, in real alignment situations, the alignment noise d (and the outliers) are unknown making it difficult to estimate α. The solution for the joint probability is:

$$P_{ij}(T) = \exp(-\alpha D_{ij}(T) - \lambda). \tag{5}$$

With this solution, note that the joint probability is also a function of the spatial mapping parameters T. Solving for λ using the constraint $\sum_{ij} P_{ij} = 1$, we get

$$P_{ij}(T) = \frac{\exp(-\alpha D_{ij}(T))}{\sum_{ij}\exp(-\alpha D_{ij}(T))}$$

It is possible to solve for α using the constraint $\sum_{ij} P_{ij} D_{ij}(T) = d$ (when d is known). Since D is actually a function of the spatial mapping T, solving for α by setting the noise level to a fixed value d is valid only when the features are aligned. When we solve for α, we get

$$\sum_{ij} D_{ij}(T)\frac{\exp(-\alpha D_{ij}(T))}{\sum_{ij}\exp(-\alpha D_{ij}(T))} = d \tag{6}$$

This is a transcendental equation for α. A solution for α can be obtained using a numerical search technique. However, we stress that this is useful only in toy situations where the amount of point jitter and the number of outliers are known (in order to estimate d). It is more reasonable to treat α as an independent parameter that determines the expected amount of residual noise after alignment. We have more to say about this issue in Section 4.

2.3. Mutual Information as a Prior

The mutual information between the point-sets is a function of the joint probability P:

$$MI(P) = \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} P_{ij} \log \frac{P_{ij}}{\sum_{k=1}^{N_1} P_{kj} \sum_{l=1}^{N_2} P_{il}} \quad (7)$$

Equation (7) is the Kullback-Leibler (Cover and Thomas, 1991) distance measure between the joint probability P and the marginal probabilities $\sum_i P_{ij}$ and $\sum_j P_{ij}$.

From the previous section, we know that the joint probability P can be estimated using the maximum entropy principle. From (5), we see that the joint probability P is a function of the cross-product distance measure D (which in turn is a function of the spatial mapping T). The normal approach using mutual information maximization would be to use the joint probability estimate in (5) and maximize the mutual information in (7). Since P is ultimately a function of the spatial mapping T, the mutual information will also be a function of the spatial mapping and can be maximized using an appropriate nonlinear optimization algorithm. While this can be accomplished, the actual optimization is somewhat cumbersome (especially in 3D). We have an alternative strategy that results in an extremely simple algorithm while still maximizing the mutual information.

Before embarking on the design of our mutual information maximization algorithm, we reiterate that the mutual information measure in (7) is an adequate point matching distance measure. In Section 4, we detail some of the properties of this new distance measure.

The basic idea in this paper is to consider the mutual information as a prior that modulates the maximum entropy approach. Note that in (4), the entropy is defined w.r.t. the joint probability P_{ij}. The same entropy is also present as the joint entropy in the above mutual information distance measure [see (7)]. Now, instead of obtaining the joint probability P via a maximum entropy approach and subsequently maximizing the mutual information, we let the estimate of the joint probability be directly influenced by the mutual information:

$$E_{\mathrm{MI}}(P,T,\lambda,\alpha) = \alpha \left(\sum_{ij} P_{ij} D_{ij}(T) - d \right) + \lambda \left(\sum_{ij} P_{ij} - 1 \right)$$
$$+ \sum_{ij} P_{ij} \log P_{ij} - \kappa MI(P). \quad (8)$$

In (8), $\kappa > 0$ is a new parameter which acts as a weight on the mutual information *vis-a-vis* the entropy and the distance measure.If $\kappa = 1$, the separate entropy term and the joint entropy in the mutual information perfectly match one another. Our approach to minimizing the energy function in (8) is detailed next.

3. DERIVING THE ALIGNMENT ALGORITHM

As it stands, it is not clear how to minimize the energy function in (8). The entropy and the mutual information share a joint entropy term in common. A straightforward minimization (by differentiating the energy function and setting the result to zero) will be very dependent on the value of the parameter κ. Indeed, when $\kappa = 1$, the joint entropy term vanishes. Eschewing a straightforward minimization, we instead exploit the convexity property of the mutual information by first carrying out an algebraic transformation.

An algebraic transformation (Mjolsness and Garrett, 1990)—essentially a Legendre transformation—simplifies energy functions by introducing auxiliary variables. Consider the algebraic transformations shown below:

$$\begin{aligned} -\frac{x^2}{2} &= \min_\sigma \left(-\sigma x + \frac{\sigma^2}{2} \right) \\ -x \log x &= \min_\sigma (-x \log \sigma + \sigma - x) \\ \log x &= \min_\sigma \left(\frac{x}{\sigma} + \log \sigma - 1 \right) \end{aligned} \quad (9)$$

In the algebraic transformations shown above, three concave functions were transformed with the help of an auxiliary variable σ. In all cases the energy functions on the right are now linear functions of x. The associated minimizations w.r.t. σ are trivial to perform, yielding $\sigma = x$ in all three instances. The algebraic transformation is easily extended to vectors.

Since the joint entropy is convex, its negative is concave. We exploit this property in the following algebraic transformations:

$$\begin{aligned} -\kappa \sum_{ij} P_{ij} \log P_{ij} &= \min_\sigma \kappa \sum_{ij} (-P_{ij} \log \sigma_{ij} + \sigma_{ij} - P_{ij}) \\ \kappa \sum_{ij} P_{ij} \log \sum_k P_{kj} &= \min_\tau \kappa \sum_{ij} P_{ij} \left(\frac{\sum_k P_{kj}}{\tau_j} + \log \tau_j - 1 \right) \\ \kappa \sum_{ij} P_{ij} \log \sum_l P_{il} &= \min_\rho \kappa \sum_{ij} P_{ij} \left(\frac{\sum_l P_{il}}{\rho_i} + \log \rho_i - 1 \right) \end{aligned} \quad (10)$$

With these algebraic transformations in place, we may write down the final energy function:

$$F(P,T,\sigma,\tau,\rho) = \alpha\left(\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} P_{ij}\|X_i - TY_j\|^2 - d\right)$$
$$+\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} P_{ij}\log P_{ij} - \kappa\sum_{i=1}^{N_1}\sum_{j=1}^{N_2}(P_{ij}\log\sigma_{ij} - \sigma_{ij} + P_{ij})$$
$$+\kappa\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} P_{ij}\left(\frac{\sum_k P_{kj}}{\tau_j} + \log\tau_j - 1\right)$$
$$+\kappa\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} P_{ij}\left(\frac{\sum_l P_{il}}{\rho_i} + \log\rho_i - 1\right)$$
$$+\lambda\left(\sum_{i=1}^{N_1}\sum_{j=1}^{N_2} P_{ij} - 1\right) \quad (11)$$

In (11), α and λ are two Lagrange parameters enforcing the constraints on the expected value of the point matching distance measure and the probability sum respectively. Once again, we reiterate that we will not attempt to estimate α given a desired residual noise level d—instead a maximum value of α will be chosen and gradually approached by the algorithm. We explain this in greater detail in Section 4. Consequently, the energy function has two free parameters, α and κ corresponding to the expected level of residual mismatch (noise) in alignment and the weight of the mutual information prior respectively.

We may now set all variables in the above energy function either by direct minimization (P,T,σ,τ,ρ), or direct maximization (λ):

$$\begin{aligned}
\frac{\partial F}{\partial P_{ij}} &= 0 \Rightarrow P_{ij} = \exp(-\alpha\|X_i - TY_j\|^2 - \lambda + \kappa\log\sigma_{ij} \\
&\quad +\kappa\log\rho_i + \kappa\log\tau_j) \\
\frac{\partial F}{\partial \sigma_{ij}} &= 0 \Rightarrow \sigma_{ij} = P_{ij} \\
\frac{\partial F}{\partial \rho_i} &= 0 \Rightarrow \rho_i = \sum_j P_{ij} \\
\frac{\partial F}{\partial \tau_j} &= 0 \Rightarrow \tau_j = \sum_i P_{ij} \\
\frac{\partial F}{\partial \lambda} &= 0 \Rightarrow \sum_{ij} P_{ij} = 1 \qquad (12)
\end{aligned}$$

The spatial mapping parameters T can be solved by a standard least-squares (Golub and Van Loan, 1989) descent algorithm on the first term in (11). Note that T does not appear anywhere other than in the first term which is a quadratic function of T. Due to the algebraic transformation, no general purpose nonlinear function optimization methods are necessary. Application of the algebraic transformation results in an Expectation-Maximization (EM) (Dempster et al., 1977) like alternating sequence of updates (Rangarajan et al., 1996).

The above update equations can be drastically simplified. Since σ,ρ,τ are dummy variables, they can be replaced by their 'P variable' counterparts. Care must be taken to separate the updates of variables that depend on each other. The updates must take place in different time steps in an alternating manner. For instance, σ,ρ,τ all depend only on P. Consequently, after P is updated, these dummy variables may be updated. Since P depends on σ,ρ,τ, the update equation for P must use the "older" values of σ,ρ,τ. This results in a vast simplification.

At each time step n, the joint probability update equation simplifies to

$$P_{ij}^{(n)} = \frac{\left(\frac{P_{ij}^{(n-1)}}{\sum_i P_{ij}^{(n-1)}\sum_j P_{ij}^{(n-1)}}\right)^{\kappa}\exp(-\alpha\|X_i - TY_j\|^2)}{\sum_{ij}\left(\frac{P_{ij}^{(n-1)}}{\sum_i P_{ij}^{(n-1)}\sum_j P_{ij}^{(n-1)}}\right)^{\kappa}\exp(-\alpha\|X_i - TY_j\|^2)} \quad (13)$$

Note that the prior weighting parameter κ appears as an exponentiating factor of the "old" values of P. We will have more to say about κ in the next section (Section 4). The overall algorithm is described in the pseudo-code below.

Mutual Information Point Feature Matching
Initialization: T to zero, P_{ij} to $\frac{1+\varepsilon_{ij}}{N_1N_2}$, α to α_{min}
Begin A: Do A until $\alpha = \alpha_{max}$.

Begin B: Do B until change in T is very small.

- Distance measure computation
  $D_{ij} \leftarrow \|X_i - TY_j\|^2$.
- Joint probability update

$$P_{ij}^{(n)} = \frac{\left(\frac{P_{ij}^{(n-1)}}{\sum_i P_{ij}^{(n-1)}\sum_j P_{ij}^{(n-1)}}\right)^{\kappa}\exp(-\alpha D_{ij}(T))}{\sum_{ij}\left(\frac{P_{ij}^{(n-1)}}{\sum_i P_{ij}^{(n-1)}\sum_j P_{ij}^{(n-1)}}\right)^{\kappa}\exp(-\alpha D_{ij}(T))}$$

Update T using a standard least-squares algorithm

End B

Increase α according to an "annealing" schedule
End A

4. RESULTS

In this section, we examine the properties of the new mutual information-based distance measure in (7) for point matching. After studying some of the noise resistant properties of the new distance, we conduct comparisons with two competing algorithms: the recently developed softassign matching algorithm (Rangarajan et al., 1997b) and the iterated closest point (ICP) algorithm (Besl and McKay, 1992; Feldmar and Ayache, 1996). We have chosen the problem of autoradiograph alignment to test our mutual information-based feature matching since we have considerable experience in the use of manual and automated algorithms in this area. Unless otherwise noted, in all experiments, the free parameter κ in the mutual information matching algorithm was set to one.

In Figure 1 we have shown the mutual information distance measure in (7) for several values of α. 100 points were randomly generated in a unit square. 20 points were deleted and 20 points added. Additive white Gaussian noise (AWGN) was then added. The figure clearly shows that the mutual information distance measure is quite smooth over a large range of α (a range from 0.1-30). Note that the minimum value of the (negative) of the mutual information occurs a bit further away from zero (the true value) and then progressively improves as α is increased. To study the effect of changing α on the mutual information, we attempted to solve the transcendental constraint equation in (6) in a synthetic example where the noise is under our control. 100 points were generated in a unit square. As before 10 points were deleted and 10 added. AWGN was added first with $\sigma = 0.01$. The constraint equation behavior is shown on the left in Figure 2. The value of α obtained by solving (6) is approximately 100. Next, we changed the noise standard deviation to $\sigma = 0.04$. The plot is significantly different as seen on the right in Figure 2. A value of $\alpha \approx 10$ is now obtained. Since we almost never know the amount of noise in a real registration problem, these values of α should be taken with a grain of salt. In most experiments, we have used a value of $\alpha = 10$ as the maximum value and approached it via an annealing schedule.

Next, we compared the new mutual information distance measure with "equivalent" softassign (Rangarajan et al., 1997b), Hausdorff distance (Huttenlocher et al., 1993) and ICP (Besl and McKay, 1992; Feldmar and Ayache, 1996) distance measures. Both ICP and softassign do not have a single distance measure on the spatial mapping. However, in both cases, it is possible to solve for the correspondence (and outliers) as a function of the spatial mapping T. In ICP, the correspondence is a straightforward nearest neighbor distance measure with a parameter K controlling the effect of outliers—the larger the value of K, the less the number of outliers rejected. In softassign, the correspondence is a fuzzy "match" matrix which is obtained by exponentiating the Euclidean distance and then performing row and column normalizations (Rangarajan et al., 1997b). The degree of fuzziness depends on a deterministic annealing parameter which we roughly equate with α. The Hausdorff distance between two point-sets X and $Y(T)$ is defined as $H(X,Y(T)) = \max(h(X,Y(T)),h(Y(T),X))$ where $h(X,Y(T)) = \max_{i\in\{1,\ldots,N_1\}} \min_{j\in\{1,\ldots,N_2\}} D_{ij}(T)$. We conducted both low-noise and high-noise experiments shown in Figures 3 and 4. The results show that ICP becomes increasingly brittle as a distance measure when the noise is increased (and outliers are present). It is easier for ICP to fall into a local minimum. (Our synthetic comparison experiments using a primate autoradiograph slice bear this out as well.) The Hausdorff distance seems somewhat brittle presumably due to the use of the max and min operators. The Hausdorff distance also does not vary smoothly with the spatial mapping parameters. These results are anecdotal and more work is needed to quantify the performances of these different distance measures.

Primate autoradiograph slice 522 is shown in Figure 5. This slice will be considered the principal slice with other slices being aligned to it. First we run a Canny edge detector (Canny, 1986) on each slice. The edge detector is a single scale (Gaussian filter with width σ) edge detector incorporating hysteresis and non-maximum suppression. The edge detector output for slice 522 is shown in Figure 6. A point-set is obtained from the edge image of slice 522 by thresholding. Since many thousands of points result, the points are first clustered in order to reduce the point count. The degree of clustering was chosen to yield approximately 100 points in the point-set, shown in Figure 6. The other slices (372, 422, 472, 572, 621 and 672) are shown in Figure 10. The difference in slice number (for example between slices 372 and 522) are indicative of the actual distance in the primate brain. In our data set, slices 372 and 672 are furthest from slice 522. (Each slice is approximately 20μm thick which gives us a "stack" span of approximately 6mm.) The edge images and point-sets (with the same level of clustering) are shown in Figures 11 and 12 respectively.

We performed 1800 experiments using slice 522 as the "X" point-set. Basically, we compared mutual information with ICP and softassign for the following cases: zero, 10%, 20% and 30% missing and extra points for noise standard deviations ranging from 0.0 to 0.08 in steps of 0.01. 50 experiments were run at each noise standard deviation. Examples of the noisy point-sets resulting from this procedure are shown in Figure 7. T was restricted to a similarity transformation in 2D (rotation, translation and scale).

The transformation parameters $\{t_x,t_y,\theta,s\}$ were bounded in the following way: $-0.5 < t_x,t_y < 0.5$, $-45^\circ < \theta < 45^\circ$, $0.5 < s < 2$. Each parameter was chosen independently and uniformly from the possible ranges. We used the error measure $e_\tau = 3|\frac{\tau^{\text{actual}} - \tau^{\text{estimated}}}{\tau^{\text{range}}}|$ where e_τ is the error measure for parameter τ (rotation, translation or scale) and τ^{range} is the aforementioned range of permissible values of τ. Dividing by τ^{range} is preferable to dividing by τ^{actual} since the latter inappropriately weights small τ^{actual} values. The reported error is the average error over all three parameters.

We executed the mutual information alignment algorithm almost exactly as described in the previous section. ICP and softassign were executed as reported in our previous work (Rangarajan et al., 1997b). The ICP robustness parameter K was set to 5. The results are shown in Figure 8. The mutual information registration algorithm performance is in between softassign and ICP. However, the mutual information algorithm was much faster than softassign. So while the performance is Softassign $>$ mutual information $>$ ICP, the speed is ICP $>$ mutual information $>$ softassign. All experiments were executed in Matlab^{TM} on an Intel PII 400 running linux. Indeed all experiments we have run indicate that mutual information-based alignment is midway between softassign and ICP in terms of speed and accuracy. There are intriguing theoretical and experimental affinities between mutual information and softassign. For example, we have shown the joint probability matrix P_{ij} as an image in Figure 9 along with the softassign match matrix at two different temperatures. Note that the mutual information "correspondences" are always fuzzy, whereas softassign converges to a discrete correspondence matrix.

Next we compared mutual information with softassign on a real autoradiograph alignment task. In all experiments (except where we explicitly tested the change in performance with parameter variation), the parameter settings were $\kappa = 1$ and $\alpha_{max} = 10$. α was increased according to a doubling schedule. The value of $\kappa = 1$ was chosen such that the mutual information term exactly matched the entropy term. Later in this section, we present the results of varying κ while keeping α fixed. The alignment results are shown in Figure 13. The maximum rotation angle between any two slices in the "stack" was about 15 degrees. The scale factor between the first and the last slice in the stack was about 1.4. As can be seen from the figure, the alignment improves as we proceed to the middle of the stack and again worsens as we proceed to the end. This is due to the center slice being the reference image for alignment. Now, we compare the results with the earlier results obtained using the softassign algorithm shown in Figure 14. While the results of the softassign algorithm are visually slightly better, the mutual information alignment algorithm was much faster, by a factor of 5 or more.

We now present an experimental confirmation of the close affinity between the joint probability matrix and correspondence. To demonstrate this, we took only about 25% of the points in slice 572 and matched it against the reference slice 522. Since one point-set is much sparser than the other, we can expect to see an overlap in the correspondence structure of the joint probability. A contour plot of the joint probability matrix is shown in Figure 15. The sparsity structure of the joint-probability matrix is strongly evocative of a correspondence matrix. The various small specks of contour islands seen in the joint probability plot indicate that the joint probability matrix P plays a similar role here to the fuzzy correspondence matrix in the softassign algorithm.

Next, we present a simple study on the variation of κ shown in Figure 16. The three plots (from left to right) correspond to $\kappa = 1, 1.1$ and 0.9 respectively. If $\kappa > 1$, the joint probability is very peaked and narrow. Whereas, if $\kappa < 1$, the joint probability is wide and shallow (for the same point). The case of $\kappa = 1$ is clearly in between. However, the other cases are important since they suggest that the emergence of the correspondence structure of the joint probability can be speeded up or slowed down by varying κ. Since $\kappa > 1$ corresponds to a larger mutual information weight in the prior, it is not surprising that the correspondence structure of P is sharper than the case of $\kappa = 1$. The opposite is true when $\kappa < 1$. In Figure 17, we depict the growth of the mutual information $\sum_{ij} P_{ij} \log \frac{P_{ij}}{\sum_i P_{ij} \sum_j P_{ij}}$ for the three cases of κ discussed above. The rate of growth depends on the value of κ as expected. What is more surprising perhaps is the monotonic increase in the mutual information. A theoretical analysis of the algorithm's convergence properties may shed some light on the growth of the mutual information.

Finally, we present very preliminary results for 3D point matching. The application domain is the automated matching of sulcal point-sets derived from different patients' anatomical MR. Sulcal extraction is done using an interactive tracing tool on an SGI graphics platform (Rambo et al., 1998). A ray-casting technique allows drawing in 3D space by projecting 2D coordinates of the tracing onto the exposed cortical surface. A screenshot of the tool is shown on the left in Figure 18. The inter-hemispheric fissure and 10 other major sulci (superior frontal, central, post-central, Sylvian and superior temporal on both hemispheres) were extracted as point features. A sulcal point-set extracted from one subject is shown on the right in Figure 18.

Visualization of the 3D point matching process is quite difficult. To this end, we first present a simpler example of 3D affine mutual information point matching by matching a sulcal point-set with a spatially warped version of itself. The initial condition and the final match are shown in Figures 19 and 20 respectively. Note the large spatial mapping applied. We executed the mutual information point matching algorithm in essentially the same manner as previously described. The parameter α was initialized to 0.05 and an annealing rate schedule of 0.99 was used. We also varied[a] the parameter κ

[a]We acknowledge an anonymous reviewer who suggested that we vary κ in this manner in order to improve the outlier rejection property of this approach.

from 1.0 to 1.1 using the same annealing schedule of 0.99. This had the effect of accelerating the growth of the mutual information as the algorithm progressed. The results in Figure 20 show a good match. Finally, we executed the same algorithm on sulcal point-sets derived from two different patients. The initial condition and the final result are shown in Figures 21 and 22 respectively. The correspondences can be more clearly seen in this result. In this experiment, we noticed that annealing on κ had the beneficial effect of strong outlier rejection. We stress that these results are very preliminary. They merely indicate that extending the mutual information point matching algorithm to 3D (using an affine mapping) is fairly straightforward.

5. REVIEW, DISCUSSION AND CONCLUSION

The energy function (11) resulting from our approach is closely related to several other energy functions (for feature-based matching) that have appeared in the literature. Below, we briefly review some of these other methods and relate them to the mutual information-based approach developed here.

The mutual information-based approach has its closest parallels with the softassign fuzzy correspondence approaches in (Lu and Mjolsness, 1994; Rangarajan, 1997; Rangarajan et al., 1997a), the statistical feature matching approaches in (Hinton et al., 1992; Wells, 1997; Grimson et al., 1994), the eigenvector-based approach in (Scott and Longuet-Higgins, 1991; Shapiro and Brady, 1992) and the iterated closest point algorithm (ICP) (Besl and McKay, 1992; Feldmar and Ayache, 1996). All of these approaches use (or come very close to using) a pairing matrix to represent correspondence between the features in one point-set and the features in the other. It should be reiterated that none of these methods use a feature image (for example a gradient magnitude image map). Rather, all of these methods parse the original image into a list of feature locations and then try and determine the spatial mapping that best matches the two lists of features.

The methods of (Lu and Mjolsness, 1994; Grimson et al., 1994) are equivalent to using the following energy function for feature matching:

$$E(M,T) = \alpha \sum_{ij} M_{ij} D_{ij}(T) + \sum_{ij} M_{ij} \log M_{ij} \tag{14}$$

In (14), T denotes an arbitrary set of spatial mapping parameters (rigid, similarity, affine) as before. In (Lu and Mjolsness, 1994), this energy function is derived by evaluating the partition function for the energy function $\alpha \sum_{ij} M_{ij} D_{ij}(T)$ over all integer valued matrices M satisfying the constraint $\sum_{ij} M_{ij} = N$ where N is the size of the point-set. In (Grimson et al., 1994), the energy function is directly written out as an exponential form of T, but is identical to the energy function above (after an algebraic transformation). It should be clear that the above energy function (regardless of its origin) can be reinterpreted as resulting from the maximum entropy principle (as derived in this paper).

The energy functions in (Hinton et al., 1992; Wells, 1997) are quite similar if not identical to an approach based on estimating mixture densities. A mixture model can be specified for feature matching by regarding one of the feature sets as the origin of the other. One then writes down a mixture likelihood $\Pr(X|Y,T) = \prod_i \sum_j \pi_j \Pr(X_i|Y_j,T)$. Here π is the occupation probability of X in Y. While this approach is asymmetric in the sense that it always regards one feature set as a model and the other as data (and hence is very suitable for object recognition), it can be quite effective when one feature set is very sparse compared to the other. Then we can imagine that the sparse locations in the first feature set act as seed locations from which the other feature set emerges (after a spatial mapping has been applied to the first set). What is remarkable is the close similarity between this approach and ours. After an Expectation-Maximization (EM) transformation to a complete data space, the final energy function for mixtures is

$$E(M,T) = -\sum_{ij} M_{ij} \log(\pi_j \Pr(X_i|Y_j,T)) + \sum_{ij} M_{ij} \log M_{ij} \tag{15}$$

with the constraints $\sum_j M_{ij} = 1$, $\sum_j \pi_j = 1$. Here the complete data M is the classification matrix. It assigns every point feature in X to one feature in Y and not vice versa. The EM algorithm returns a fuzzy classification, not unlike the joint probability matrix used here. The term $-\log(\pi_j \Pr(X_i|Y_j,T))$ plays the role of a distance measure.

The softassign correspondence approach in (Rangarajan, 1997; Rangarajan et al., 1997a) explicitly parameterizes correspondence as a binary match matrix M. However, just as in the other approaches, the correspondence matrix is softened or made fuzzy during the optimization procedure. Unlike the other approaches, the softassign approach guarantees a one-to-one match between features in one image and features in the other. The energy function used is

$$E(M,T) = \sum_{ij} M_{ij} D_{ij}(T) - \gamma \sum_{ij} M_{ij} + \frac{1}{\beta} \sum_{ij} M_{ij} \log M_{ij} \tag{16}$$

with the constraints $\sum_i M_{ij} \le 1$ and $\sum_j M_{ij} \le 1$. The parameter γ is a robustness parameter. Due to the linear nature of the objective (in M) and the linear constraints, the algorithm is guaranteed to return a binary-valued matrix M in the limit as the deterministic annealing parameter β approaches ∞. We have already conducted visual comparisons between the mutual information approach and the softassign approach.

While the latter (at present) seems more accurate on alignment problems it is much slower. In addition, the softassign algorithm has a free parameter γ which is difficult to set. In comparison, the mutual information algorithm has almost no free parameters since only an annealing schedule has to be prescribed for α.

Finally, the ICP algorithm (Besl and McKay, 1992; Feldmar and Ayache, 1996) can also be included in this theoretical comparison. In ICP, the correspondence is explicitly computed using a nearest neighbor measure. This represents a limiting case of using a pairing matrix to assign correspondence. ICP is usually much faster than the other methods. Based on our experience (as shown in Section 4), the mutual information alignment algorithm stands midway in terms of speed and accuracy between ICP and the softassign alignment algorithms.

In summary, we have derived a new mutual information maximization algorithm for feature-based registration. While the experimental results were reported for 2D alignment using a similarity transformation, the actual algorithm is quite general and is not *a priori* restricted to a particular choice of spatial mapping. Mutual information maximization methods have proven to be quite useful in intermodality voxel-based registration. We speculate that mutual information maximization methods may also be useful in medical image registration domains where well defined features exist and can be easily extracted. This work also allows for the interesting possibility of combining intensity- and feature-based registration using an overall mutual information maximization criterion.

ACKNOWLEDGEMENTS

This project is partially supported by a grant from the Whitaker Foundation. A.R. would like to thank Fred Bookstein for an interesting discussion.

REFERENCES

Besl, P. J. and McKay, N. D. (1992). A method for registration of 3-D shapes. *IEEE Trans. Patt. Anal. Mach. Intell.*, 14(2):239–256.

Canny, J. (1986). A Computational Approach to Edge Detection. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 8(6):679–698.

Cover, T. and Thomas, J. (1991). *Elements of Information Theory*. John Wiley and Sons, New York, NY.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. Ser. B*, 39:1–38.

Feldmar, J. and Ayache, N. (1996). Rigid, affine and locally affine registration of free-form surfaces. *Intl. J. Computer Vision*, 18(2):99–119.

Gold, S., Rangarajan, A., Lu, C. P., Pappu, S., and Mjolsness, E. (1998). New algorithms for 2-D and 3-D point matching: pose estimation and correspondence. *Pattern Recognition*, 31(8):1019–1031.

Golub, G. and Van Loan, C. (1989). *Matrix Computations*. Johns Hopkins University Press, 2nd edition.

Grimson, E., Lozano-Perez, T., Wells III, W., Ettinger, G., White, S. J., and Kikinis, R. (1994). An automatic registration method for frameless stereotaxy, image guided surgery and enhanced reality visualization. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 430–436. IEEE Press.

Hinton, G., Williams, C., and Revow, M. (1992). Adaptive elastic models for hand-printed character recognition. In Moody, J., Hanson, S., and Lippmann, R., editors, *Advances in Neural Information Processing Systems 4*, pages 512–519. Morgan Kaufmann, San Mateo, CA.

Huttenlocher, D. P., Klanderman, G. A., and Rucklidge, W. J. (1993). Comparing images using the Hausdorff distance. *IEEE Trans. Patt. Anal. Mach. Intell.*, 15(9):850–863.

Jaynes, E. T. (1982). On the rationale of maximum-entropy methods. *Proc. IEEE*, 70:939–952.

Lu, C. P. and Mjolsness, E. (1994). Two-dimensional object localization by coarse-to-fine correlation matching. In Cowan, J., Tesauro, G., and Alspector, J., editors, *Advances in Neural Information Processing Systems 6*, pages 985–992. Morgan Kaufmann, San Francisco, CA.

Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., and Suetens, P. (1997). Multi-modality image registration by maximization of mutual information. *IEEE Trans. Med. Imag.*, 16(2):187–198.

Mjolsness, E. and Garrett, C. (1990). Algebraic transformations of objective functions. *Neural Networks*, 3:651–669.

Pappu, S., Gold, S., and Rangarajan, A. (1996). A framework for non-rigid matching and correspondence. In Touretzky, D. S., Mozer, M. C., and Hasselmo, M. E., editors, *Advances in Neural Information Processing Systems 8*, pages 795–801. MIT Press, Cambridge, MA.

Rambo, J., Zeng, X., Schultz, R., Win, L., Staib, L., and Duncan, J. (1998). Platform for visualization and measurement of gray matter volume and surface area within discrete cortical regions from MR images. *NeuroImage*, 7(4):795.

Rangarajan, A. (1997). Self annealing: Unifying deterministic annealing and relaxation labeling. In *Energy Minimization Methods in Computer Vision and Pattern Recognition (EMM-CVPR '97)*, pages 229–244. Springer.

Rangarajan, A., Chui, H., and Bookstein, F. (1997a). The softassign Procrustes matching algorithm. In *Information Processing in Medical Imaging (IPMI '97)*, pages 29–42. Springer.

Rangarajan, A., Chui, H., Mjolsness, E., Pappu, S., Davachi, L., Goldman-Rakic, P., and Duncan, J. (1997b). A robust point matching algorithm for autoradiograph alignment. *Medical Image Analysis*, 4(1):379–398.

Rangarajan, A., Gold, S., and Mjolsness, E. (1996). A novel optimizing network architecture with applications. *Neural Computation*, 8(5):1041–1060.

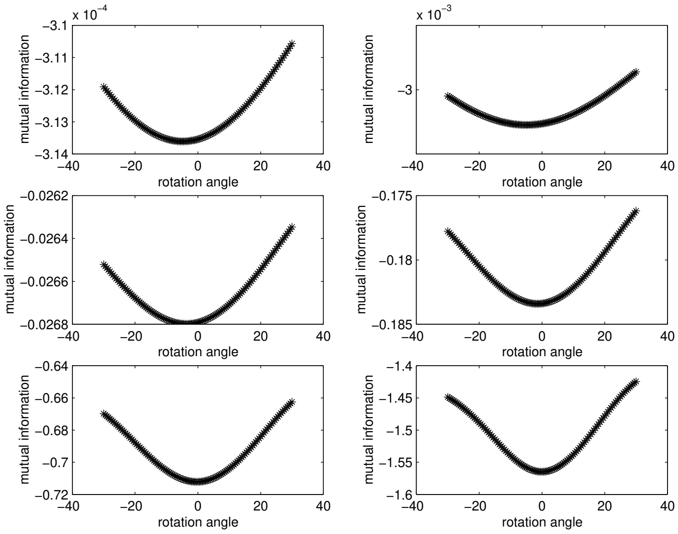

Figure 1. Mutual Information distance for different values of α. 100 points, 20 missing and extra, $\sigma = 0.02$. Top Left: $\alpha = 0.1$. Top Right: $\alpha = 0.3$. Middle Left $\alpha = 1$. Middle Right: $\alpha = 3$. Bottom Left: $\alpha = 10$, Bottom Right: $\alpha = 32$.Note the gradual progress toward the correct alignment angle $\theta = 0$.

Scott, G. and Longuet-Higgins, C. (1991). An algorithm for associating the features of two images. *Proc. Royal Society of London*, B244:21–26.

Shapiro, L. and Brady, J. (1992). Feature-based correspondence: an eigenvector approach. *Image and Vision Computing*, 10:283–288.

Van den Elsen, P., Pol, E., and Viergever, M. (1993). Medical image matching—A review with classification. *IEEE Engineering in Medicine and Biology*, pages 26–38.

Wells, W. (1997). Statistical approaches to feature-based object recognition. *Intl. J. Computer Vision*, 21(1/2):63–98.

Wells III, W., Viola, P., Atsumi, H., Nakajima, S., and Kikinis, R. (1996). Multi-modal volume registration by maximization of mutual information. *Medical Image Analysis*, 1(1):35–52.

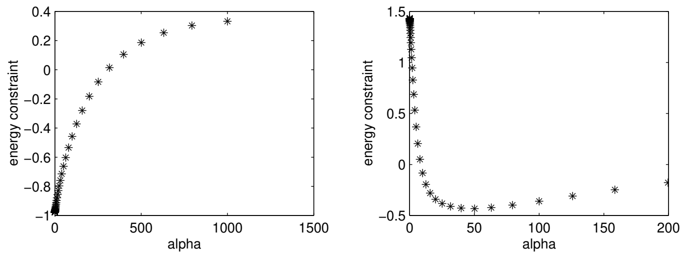

Figure 2. Mutual Information constraint equation. 100 points, 10 missing, 10 extra. Left: $\sigma = 0.01$. Right: $\sigma = 0.04$. Note the different behavior of the constraint equation for the two noise variance choices. The significant point is when the constraint equation crosses zero.

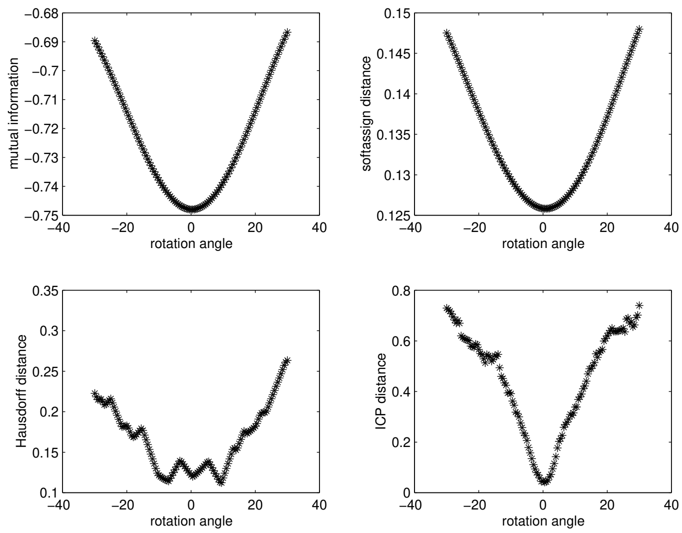

Figure 3. Comparison of different distance measures. 100 points, 20 missing, 20 extra, $\sigma = 0.01$. Top Left: Mutual Information with $\alpha = 10$. Top Right: Softassign with $\alpha = 10$. Bottom Left: Hausdorff distance. Bottom Right: ICP with $K = 10$.

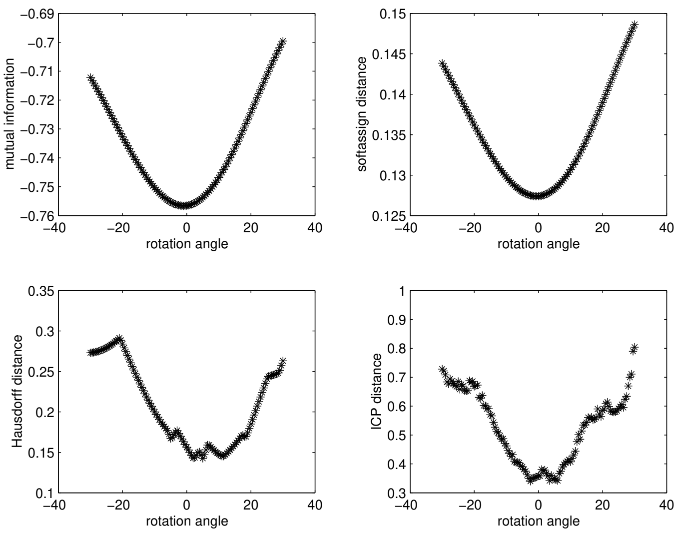

Figure 4. Comparison of different distance measures. 100 points, 20 missing, 20 extra, $\sigma = 0.05$. Top Left: Mutual Information with $\alpha = 10$. Top Right: Softassign with $\alpha = 10$. Bottom Left: Hausdorff distance. Bottom Right: ICP with $K = 10$. Note the increased number of local minima in the ICP plot relative to the previous figure.

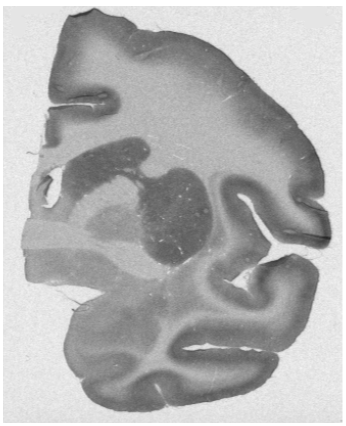

Figure 5. Primate autoradiograph slice number 522

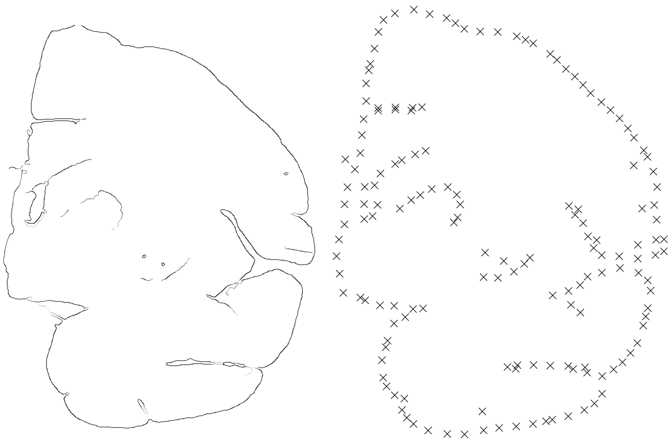

Figure 6. Left: Canny edges corresponding to primate autoradiograph slice 522. Right: Clustered point set corresponding to slice 522

Figure 7. Synthetic noise added to slice 522. Top Left: $\sigma = 0.02$. Top Right: $\sigma = 0.04$. Bottom Left: $\sigma = 0.06$. Bottom Right: $\sigma = 0.08$. The apparent change in scale is due to the increased number of points deviating from the slice outer boundary.

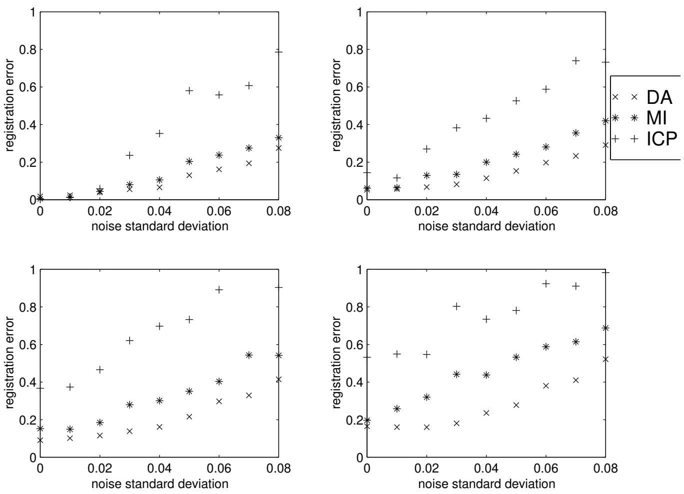

Figure 8. Comparison of different algorithms for different numbers of missing and extra points. Slice 522 was used as the basic point-set. Top Left: No missing and extra points. Top Right: 10% missing and extra points. Bottom Left: 20% missing and extra points. Bottom Right: 30% missing and extra points.

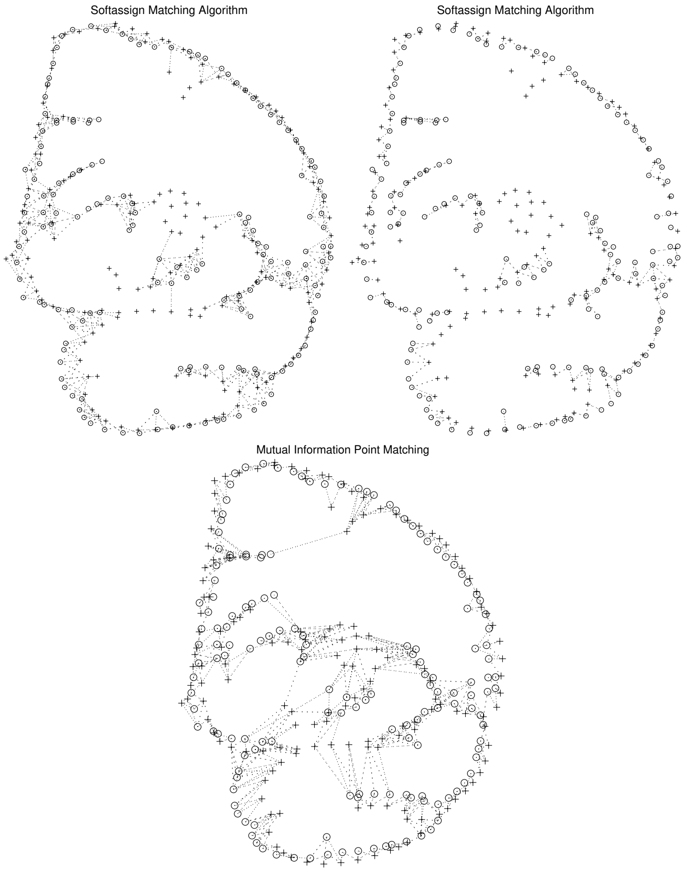

Figure 9. Comparison of correspondence structure between deterministic annealing and mutual information. Top: Two correspondence plots from deterministic annealing at different temperatures. Bottom: Joint probability plot from mutual information at end of algorithm.

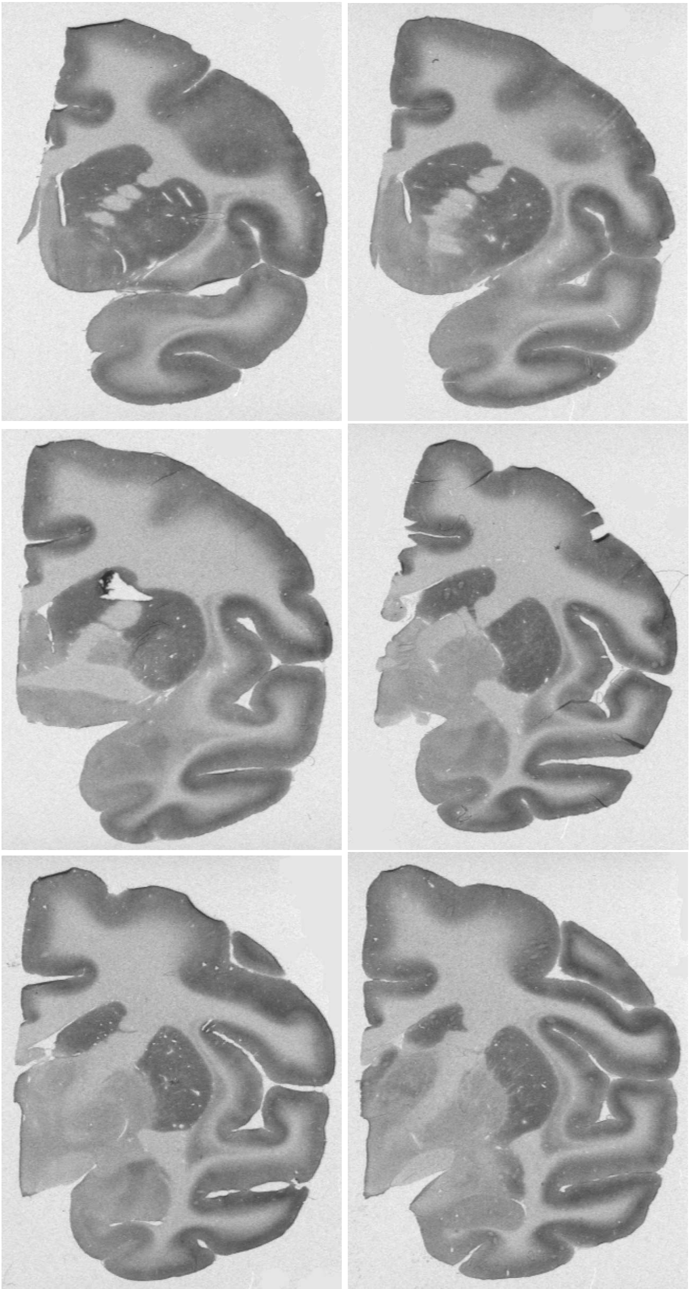

Figure 10. Primate autoradiograph slices 372, 422, 472, 572, 621 and 672 from top left to bottom right

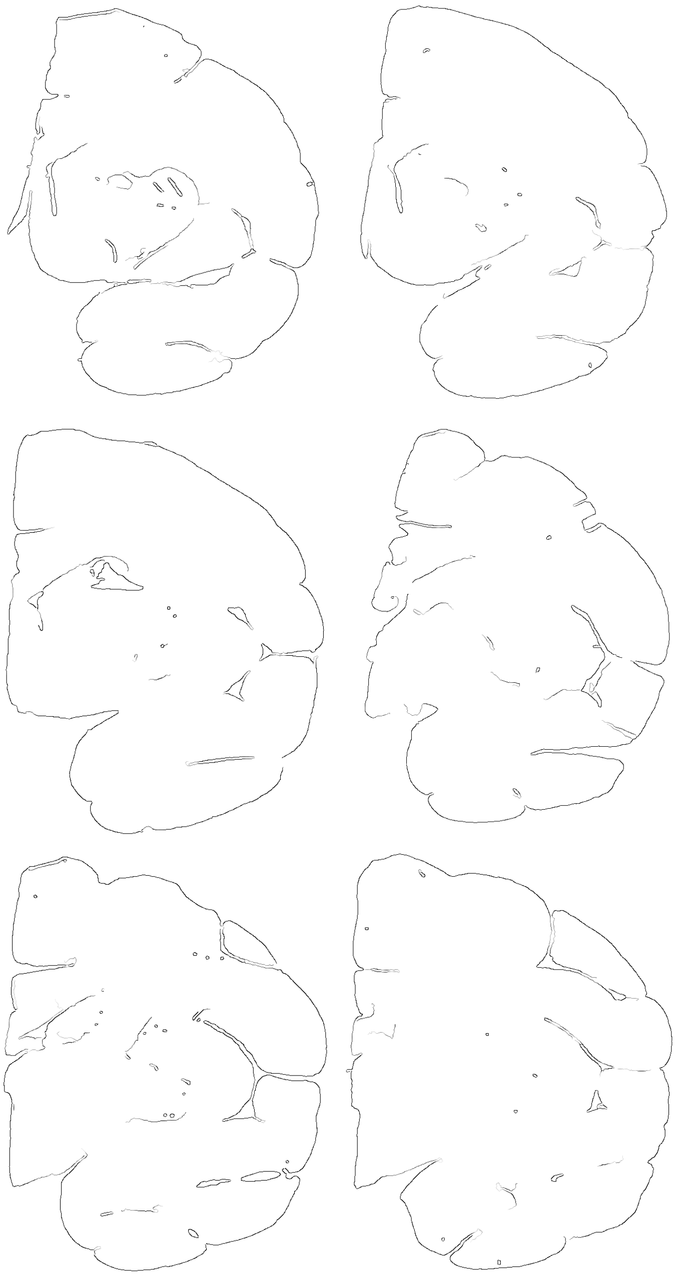

Figure 11. Canny edge images corresponding to slices 372, 422, 472, 572, 621 and 672 from top left to bottom right

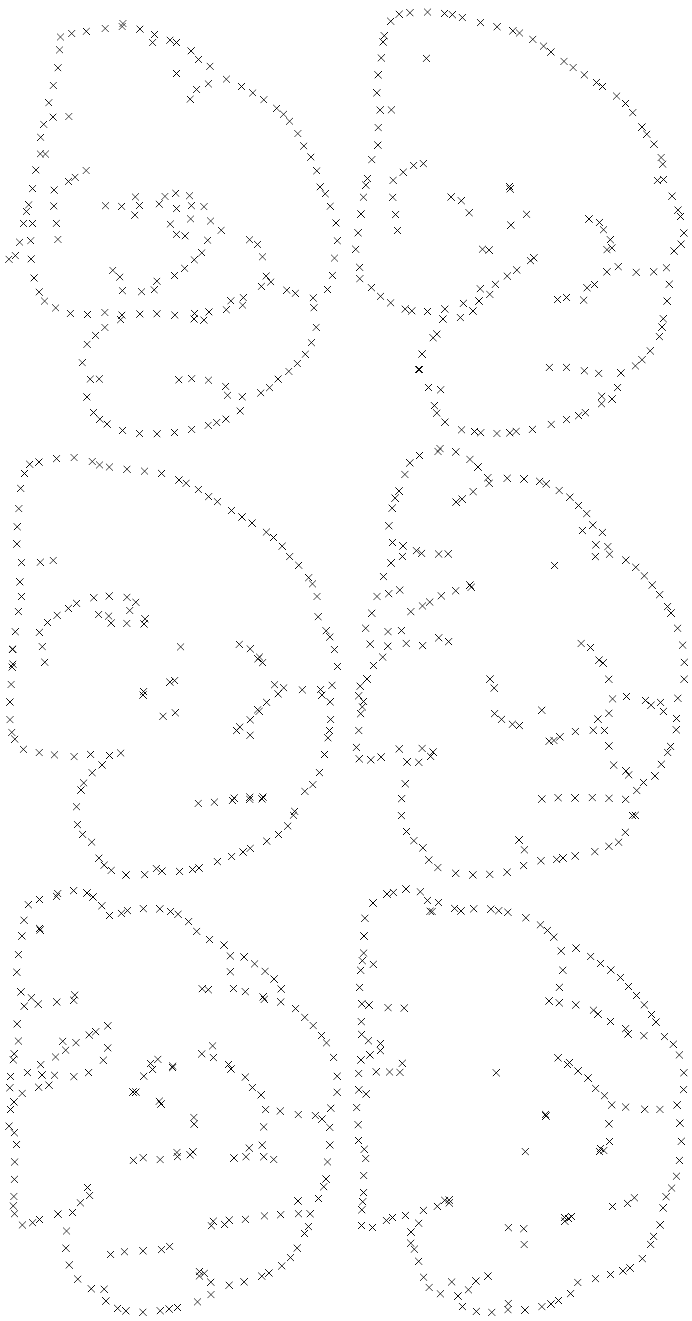

Figure 12. Point-sets corresponding to slices 372, 422, 472, 572, 621 and 672 from top left to bottom right

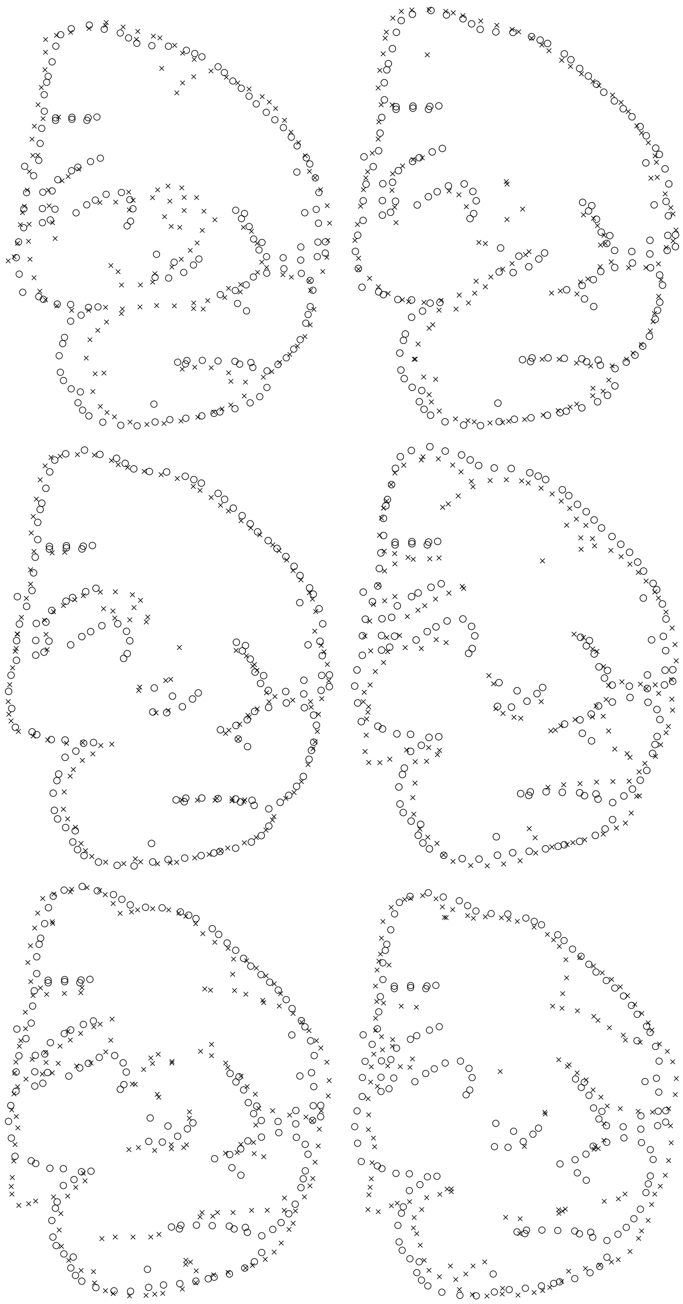

Figure 13. Mutual Information-based alignment for slices 372, 422, 472, 572, 621 and 672 aligned with slice 522 from top left to bottom right

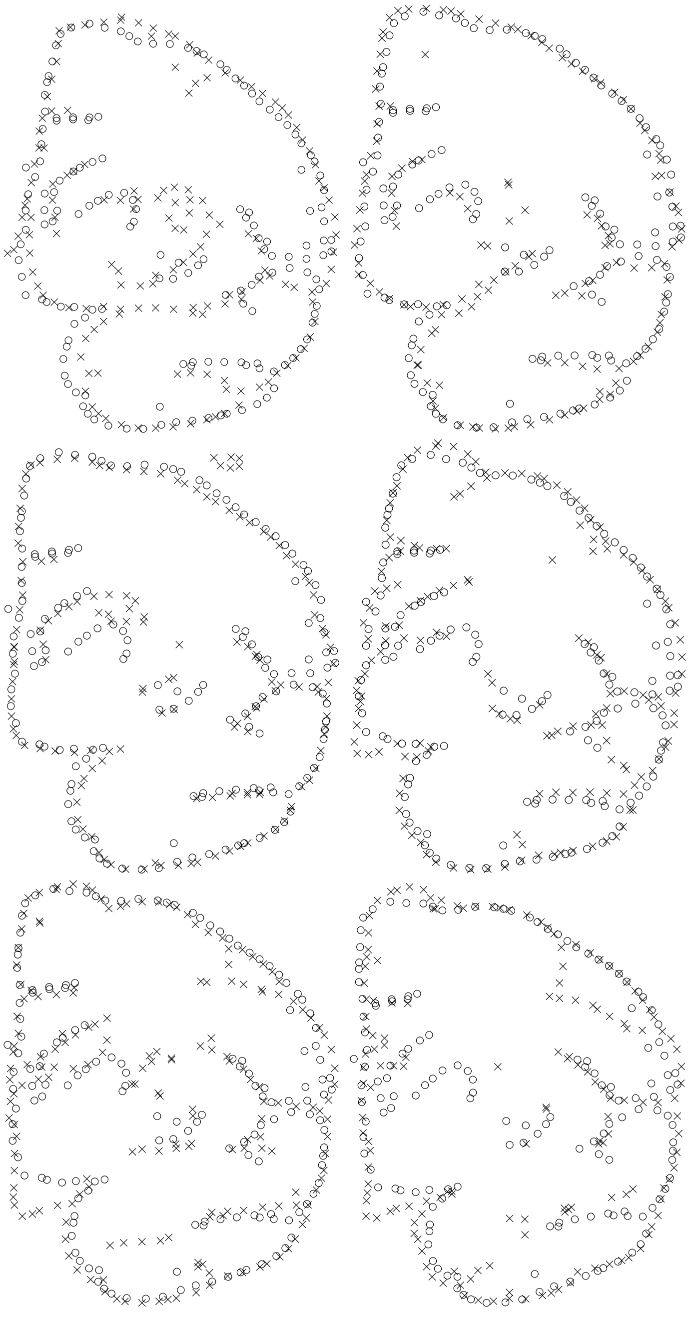

Figure 14. Softassign-based alignment for slices 372, 422, 472, 572, 621 and 672 aligned with slice 522 from top left to bottom right

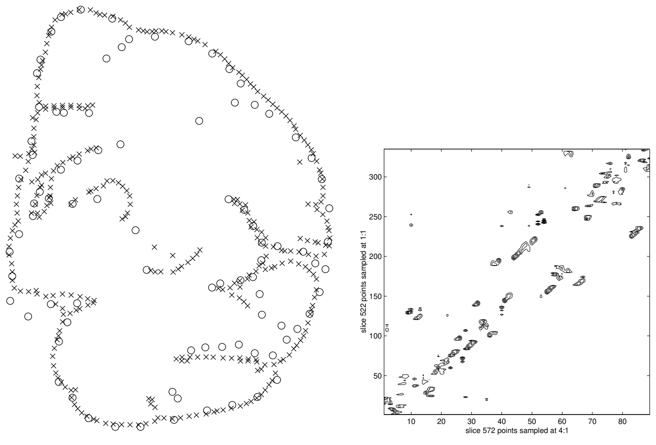

Figure 15. Mutual Information-based alignment for slice 572 against slice 522 with a 1:4 sampling of points. Left: Overlay. Right: Contour plot of the joint probability matrix $\{P_{ij}\}$.

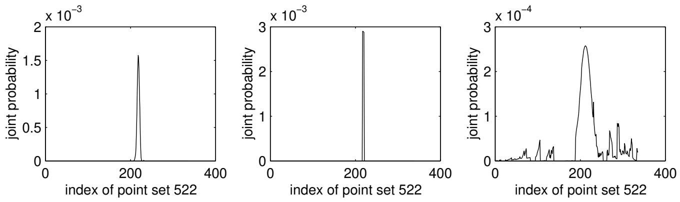

Figure 16. Profiles of the joint probability matrix $\{P_{ij}\}$ for all points in slice 522 against point 50 in slice 572 with a 1:4 sampling of points. Left: $\kappa = 1.0$. Middle: $\kappa = 1.1$. Right: $\kappa = 0.9$

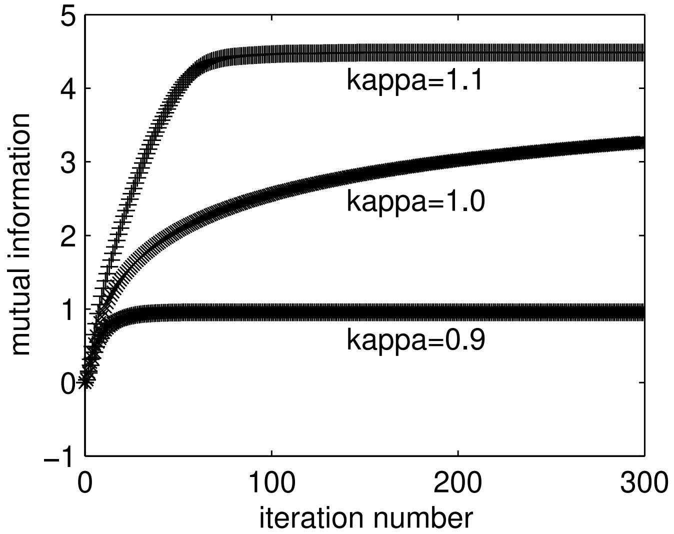

Figure 17. Growth of the mutual information $\sum_{ij} P_{ij} \log \frac{P_{ij}}{\sum_i P_{ij} \sum_j P_{ij}}$ versus iteration number for the alignment of slice 572 against slice 522 with a 1:4 sampling of points

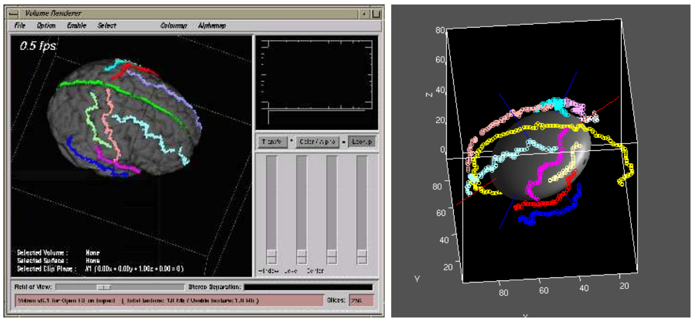

Figure 18. Left: A screenshot of the sulcal tracing tool with some traced sulci on the 3D MR brain volume. Right: Sulci extracted and displayed as point-sets. Each sulcus is shown in a different color.

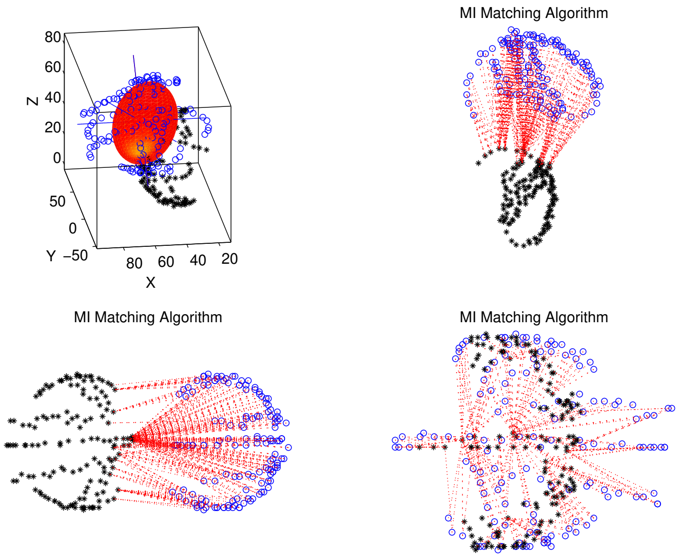

Figure 19. Initial condition for matching two sulcal point-sets. The second sulcal point-set is derived by applying an arbitrary affine spatial mapping to the first set.

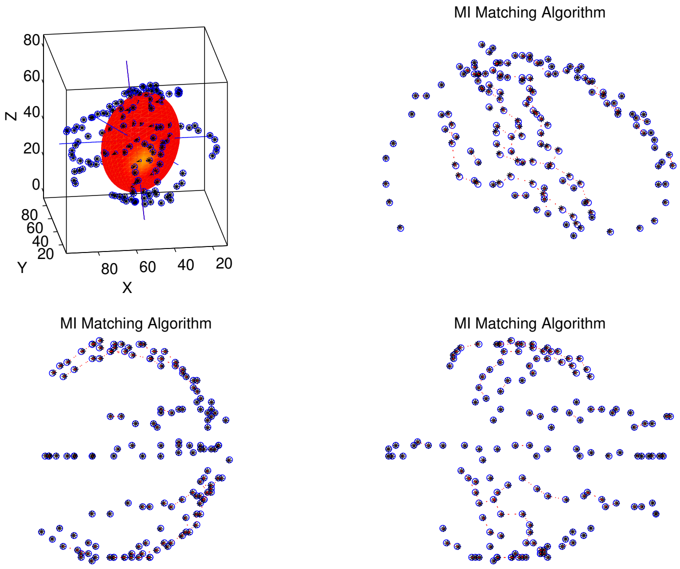

Figure 20. The result of matching the sulcal point-sets.

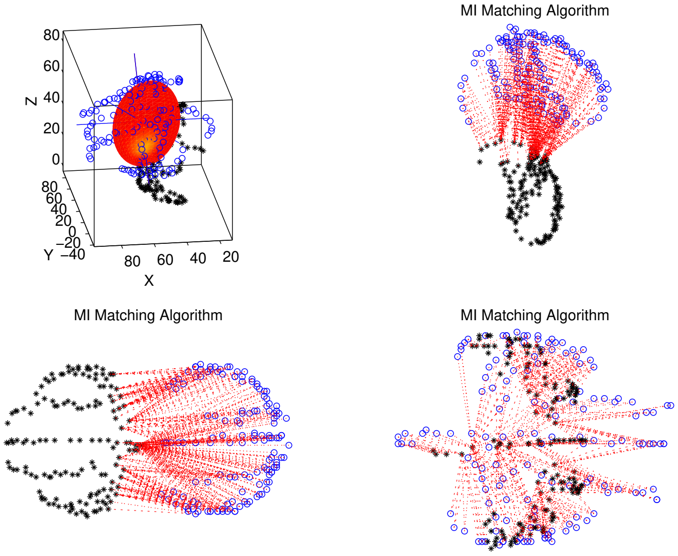

Figure 21. Initial condition for matching two sulcal point-sets derived from two different patients.

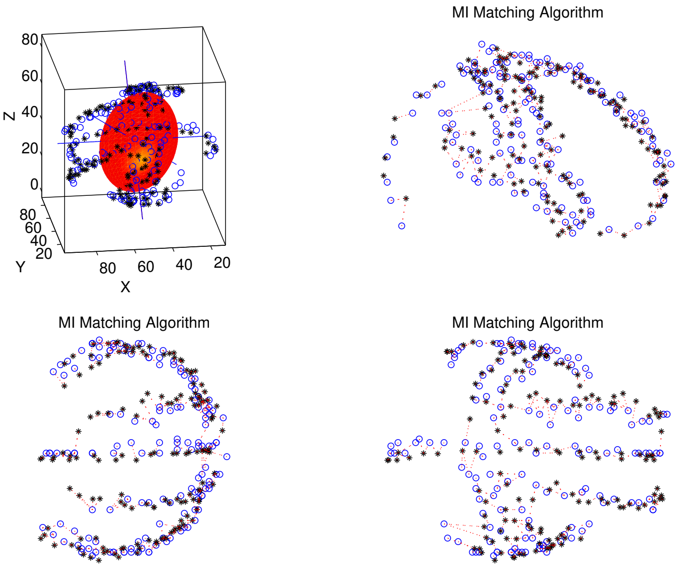

Figure 22. The result of matching the sulcal point-sets.