Multi-Modal Volume Registration by Maximization of Mutual Information

William M. Wells III [12] Paul Viola [13] Ron Kikinis [2]
sw@ai.mit.edu viola@salk.edu kikinis@bwh.harvard.edu

Keywords: Multi-Modal Image Registration, Mutual Information, Information Theory
Categories: Registration for Medical Applications

Abstract

A new information-theoretic approach is presented for finding the registration of volumetric medical images of differing modalities. Registration is achieved by adjustment of the relative pose until the mutual information between images is maximized. In our derivation of the registration procedure, few assumptions are made about the nature of the imaging process. As a result the algorithms are quite general and can foreseeably be used with a wide variety of imaging devices.

This approach works directly with raw images; no preprocessing or feature detection is required. As opposed to feature-based techniques, all of the information in the scan is used to evaluate the registration. This technique is however more flexible and robust than other intensity based techniques like correlation. Additionally, it has an efficient implementation that is based on stochastic approximation.

Experiments are presented that demonstrate the approach registering magnetic resonance (MR) images with computed tomography (CT) images, and with positron-emission tomography (PET) images.

1 Introduction

Multi-modal medical image registration is an important capability for surgical applications. For example, in neurosurgery it is currently useful to identify tumors with magnetic resonance images (MRI), yet the established stereotaxy technology uses computed tomography (CT) images. Being able to register these two modalities allows one to transfer the coordinates of tumors from the MR images into the CT stereotaxy. It is similarly useful to transfer functional information from SPECT or PET into MR or CT for anatomical reference, and for stereotactic exploitation.

Consider the problem of registering two different MR images of the same individual. When perfectly aligned these signals should be quite similar. One simple measure of the quality of a hypothetical registration is the sum of squared differences between voxel values. This measure can be motivated with a probabilistic argument. If the noise inherent in an MR image were Gaussian, independent and identically distributed, then the sum of squared differences is directly proportional to the likelihood that the two images are correctly registered. Unfortunately, squared difference is not an effective measure for the registration of different modalities. Even when perfectly registered, MR and CT images taken from the same individual are quite different. In fact MR and CT are useful in conjunction precisely because they are different.

This is not to say the MR and CT images are completely unrelated. They are after all both informative measures of the properties of human tissue. Using a large corpus of data, or some physical theory, it might be possible to construct a function $F(\cdot)$ that predicts CT from the corresponding MR value, at least approximately. Using F we could evaluate registrations by computing $F(\mathrm{MR})$ and comparing it via sum of squared differences (or some similar measure) to the CT image. If the CT and MR images were not correctly registered, then F would not be good at predicting one from the other.

Given that both MR and CT are informative of the same underlying phenomena, there will be mutual information between the MR image and the CT image. We propose to finesse the problem of finding and computing F by dealing with this mutual information directly. Such a technique would attempt to find the registration by maximizing the information that one volumetric image provides about the other. We will present an algorithm that does just this. It requires no a priori model of the relationship between the modalities, it only assumes that one volume provides the most information about the other one when they are correctly registered.

2 Description of Method

2.1 Registration by Maximization of Mutual Information

In the following derivation we will refer to the two volumes of image data that are to be registered as the *reference volume* and the *test volume*. A voxel of the reference volume is denoted $u(x)$, where x are the coordinates of the voxel. A voxel of the test volume is denoted similarly as $v(y)$. Given that T is a transformation from the coordinate frame of the reference volume to the test volume, $v(T(x))$ is the test voxel associated with reference voxel $u(x)$.

We seek an estimate of the transformation that registers the reference volume u and test volume v by maximizing their mutual information,

$$\hat{T} = \arg\max_T I(u(x),\ v(T(x))) \ . \tag{1}$$

Here we treat x as a random variable over coordinate locations in the reference volume. In the registration

[1] Massachusetts Institute of Technology, Artificial Intelligence Laboratory
[2] Harvard Medical School and Brigham and Women's Hospital, Department of Radiology
[3] The Salk Institute, Computational Neurobiology Laboratory

algorithm described below, we will draw samples from x in order to approximate I and its derivative.

Mutual information is defined in terms of entropy in the following way [1] :

$$I(u(x),\ v(T(x))) \equiv H(u(x)) + H(v(T(x))) - H(u(x), v(T(x))) \quad (2)$$

. $H(\cdot)$ is the entropy of a random variable, and is defined as $H(x) \equiv -\int p(x)\ln p(x)dx$, while the joint entropy of two random variables x and y is $H(x,y) \equiv -\int p(x,y)\ln p(x,y)dxdy$. Entropy can be interpreted as a measure of uncertainty, variability, or complexity.

The mutual information defined in Equation 2 has three components. The first term on the right is the entropy in the reference volume, and is not a function of T. The second term is the entropy of the part of the test volume into which the reference volume projects. It encourages transformations that project u into complex parts of v. The third term, the (negative) joint entropy of u and v, contributes when u and v are functionally related. It encourages transformations where u explains v well. Together the last two terms identify transformations that find complexity and explain it well. This is the essence of mutual information.

2.2 Estimating Entropies and their Derivatives

The entropies described above are defined in terms of integrals over the probability densities associated with the random variables $u(x)$ and $v(T(x))$. When registering medical image data we will not have direct access to these densities. In this section we describe a differentiable estimate of the entropy of a random variable that is calculated from samples.

The entropy of a random variable z may be expressed as an expectation of the negative logarithm of the probability density: $H(z) = E_z(-ln\ p(z))$.

Our first step in estimating the entropies from samples is to approximate the underlying probability density $p(z)$ by a superposition of Gaussian densities centered on the elements of a sample A drawn from z: $p(z) \approx \frac{1}{N_A}\sum_{z_j \in A} G_\psi(z - z_j)$, where $G_\psi(x) \equiv (2\pi)^{\frac{-n}{2}}|\psi|^{\frac{-1}{2}}\exp(-\frac{1}{2}x^T\psi^{-1}x)$, and N_A is the number of trials in the sample A. This method of density estimation is widely known as the *Parzen Window* method. It is described in the textbook by Duda and Hart[2].

Next we approximate statistical expectation with the sample average over another sample B drawn from z: $E_z(f(z)) \approx \frac{1}{N_B}\sum_{z_i \in B} f(z_i)$.

We may now write an approximation for the entropy of a random variable z as follows,

$$H(z) \approx \frac{-1}{N_B}\sum_{z_i \in B}\ln\frac{1}{N_A}\sum_{z_j \in A} G_\psi(z_i - z_j) \ . \quad (3)$$

In order to find maxima of mutual information, we calculate the derivative of entropy with respect to the transformation T. After some manipulation, this may be written compactly as follows,

$$\frac{d}{dT}H(z(T)) \approx \quad (4)$$

$$\frac{1}{N_B}\sum_{z_i \in B}\sum_{z_j \in A} W_z(z_i, z_j)(z_i - z_j)^T\psi^{-1}\frac{d}{dT}(z_i - z_j) \ ,$$

using the following definition:

$$W_z(z_i, z_j) \equiv \frac{G_\psi(z_i - z_j)}{\sum_{z_k \in A} G_\psi(z_i - z_k)} \ .$$

The weighting factor $W_z(z_i, z_j)$ takes on values between zero and one. It will approach one if z_i is significantly closer to z_j than it is to any other element of A. It will be near zero if some other element of A is significantly closer to z_i. Distance is interpreted with respect to the squared Mahalanobis distance (see [2]) $D_\psi(z) \equiv z^T\psi^{-1}z$. Thus, $W_z(z_i, z_j)$ is an indicator of the degree of match between its arguments, in a "soft" sense. It is equivalent to using the "softmax" function of neural networks [3] on the negative of the Mahalanobis distance to indicate correspondence between z_i and elements of A.

The summand in Equation 4 may also be written as: $W_z(z_i, z_j)\ \frac{d}{dT}\frac{1}{2}D_\psi(z_i - z_j)$. In this form it is apparent that to reduce entropy, the transformation T should be adjusted such that there is a reduction in the average squared distance between those values which W indicates are nearby, i.e., clusters should be tightened.

2.3 Stochastic Maximization of Mutual Information

The entropy approximation described in Equation 3 may now be used to evaluate the mutual information of the reference volume and test volume (Equation 2). In order to seek a maximum of the mutual information, we will calculate an approximation to its derivative,

$$\frac{d}{dT}I(T) = \frac{d}{dT}H(v(T(x))) - \frac{d}{dT}H(u(x), v(T(x))) \ .$$

Using Equation 4, and assuming that the covariance matrices of the component densities used in the approximation scheme for the joint density are block diagonal: $\psi_{uv}^{-1} = \mathrm{DIAG}(\psi_{uu}^{-1}, \psi_{vv}^{-1})$, we can obtain an estimate for the derivative of the mutual information as follows:

$$\widehat{\frac{dI}{dT}} = \frac{1}{N_B}\sum_{x_i \in B}\sum_{x_j \in A}(v_i - v_j)^T$$
$$[W_v(v_i, v_j)\psi_v^{-1} - W_{uv}(w_i, w_j)\psi_{vv}^{-1}]\frac{d}{dT}(v_i - v_j) \ .$$

The weighting factors are defined as

$$W_v(v_i, v_j) \equiv \frac{G_{\psi_v}(v_i - v_j)}{\sum_{x_k \in A} G_{\psi_v}(v_i - v_k)} \ , \text{and}$$

$$W_{uv}(w_i, w_j) \equiv \frac{G_{\psi_{uv}}(w_i - w_j)}{\sum_{x_k \in A} G_{\psi_{uv}}(w_i - w_k)} \ ,$$

using the following notation (and similarly for indices j and k),

$$u_i \equiv u(x_i)\ ,\ v_i \equiv v(T(x_i))\ ,\ \text{and}\ w_i \equiv [u_i, v_i]^T \ .$$

If we are to increase the mutual information, then the first term in the brackets may be interpreted as acting to increase the squared distance between pairs of samples that are nearby in test volume intensity, while the second term acts to decrease the squared distance between pairs of samples whose intensities are nearby in *both* volumes. It is important to emphasize that these distances are in the space of intensities, rather than coordinate locations.

The term $\frac{d}{dT}(v_i - v_j)$ will generally involve gradients of the test volume intensities, and the derivative of transformed coordinates with respect to the transformation. In the simple case that T is a linear operator, the following outer product expression holds: $\frac{d}{dT}v(T(x_i)) = \nabla v(T(x_i))x_i^T$.

2.3.1 Stochastic Maximization Algorithm

We seek a local maximum of mutual information by using a stochastic analog of gradient descent. Steps are repeatedly taken that are proportional to the approximation of the derivative of the mutual information with respect to the transformation:

Repeat:

$$
\begin{aligned}
A &\leftarrow \{\text{sample of size } N_A \text{ drawn from } x\} \\
B &\leftarrow \{\text{sample of size } N_B \text{ drawn from } x\} \\
T &\leftarrow T + \lambda \widehat{\frac{dI}{dT}}
\end{aligned}
$$

The parameter λ is called the *learning rate*. The above procedure is repeated a fixed number of times or until convergence is detected.

A good estimate of the derivative of the mutual information could be obtained by exhaustively sampling the data. This approach has serious drawbacks because the algorithm's cost is quadratic in the sample size. For smaller sample sizes, less effort is expended, but additional noise is introduced into the derivative estimates.

Stochastic approximation is a scheme that uses noisy derivative estimate instead of the true derivative for optimizing a function (see [4], [5], and [6]). Convergence can be proven for particular linear systems, provided that the derivative estimates are unbiased, and the learning rate is annealed (decreased over time). In practice, we have found that successful registration may be obtained using relatively small sample sizes, for example $N_A = N_B = 50$. We have proven that the technique will always converge to a transformation estimate that is close to locally optimal [7].

It has been observed that the noise introduced by the sampling can effectively penetrate small local minima. Such local minima are often characteristic of continuous registration schemes, and we have found that local minima can be overcome in this manner in these applications as well. We believe that stochastic estimates for the gradient usefully combine efficiency with effective escape from local minima.

2.4 Estimating the Covariance

The covariance matrices, ψ, of the Gaussians used in the Parzen density estimate can be estimated as described in [8]. In the experiments described here, these parameters are not difficult to adjust, nor is the method especially sensitive to their values.

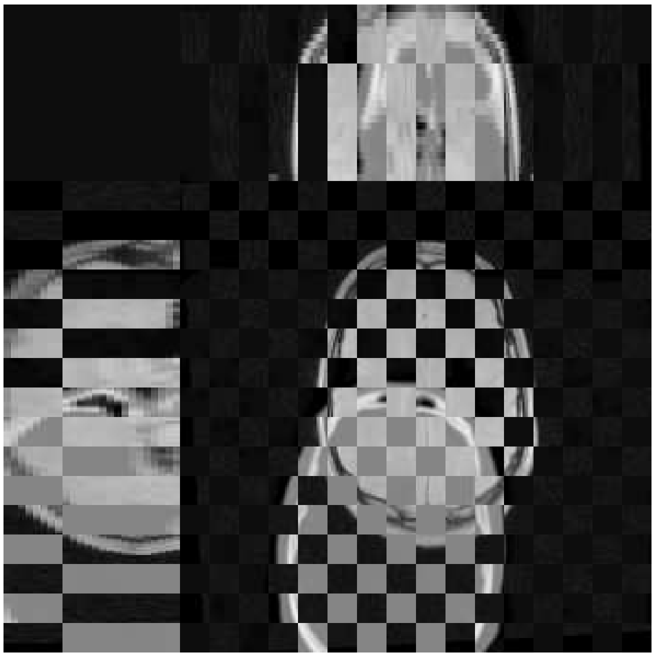

Figure 1: An initial condition for MR-CT registration by maximization of mutual information displayed as a checkerboard composite of the three central slices.

3 Experiments

3.1 MRI – CT Registration [4]

In this section we describe a series of experiments where the method was used to register magnetic resonance (MR) images and computed tomography (CT) images from the same person. Figures 1, 2 and 3 illustrate the data, initial configuration and final configuration for representative MR-CT Registration.

The MRI data consisted of 24 proton-density cross sections of 256x256 pixels each. The pixel dimensions were 1.25 mm squared and the slice spacing was 4 mm. The CT data was 29 slices of 512x512, the pixel dimensions were .653595 mm square, while the slice spacing was 4 mm. The MR data served as the reference volume, while the CT data served as the test volume.

The registration was performed in a coarse-fine fashion on a hierarchy of data volumes that had been generated by successive smoothing and reduction. This strategy was used to increase the capture range of the method: at the lower resolutions, there was less tendency to become trapped in local minima, but the accuracy was reduced.

Smoothing (in each dimension applicable) was performed by convolving with the binomial kernel {1,4,6,4,1}, and subsequent reduction was accomplished by deleting alternating samples. This scheme generates an approximation to a "Gaussian Pyramid" representation of the data [9].

Rigid transformations were used, they were represented by displacement vectors and quaternions. At each

[4] The images were provided as part of the project, "Evaluation of Retrospective Image Registration", National Institutes of Health, Project Number 1 R01 NS33926-01, Principal Investigator, J. Michael Fitzpatrick, Vanderbilt University, Nashville, TN.

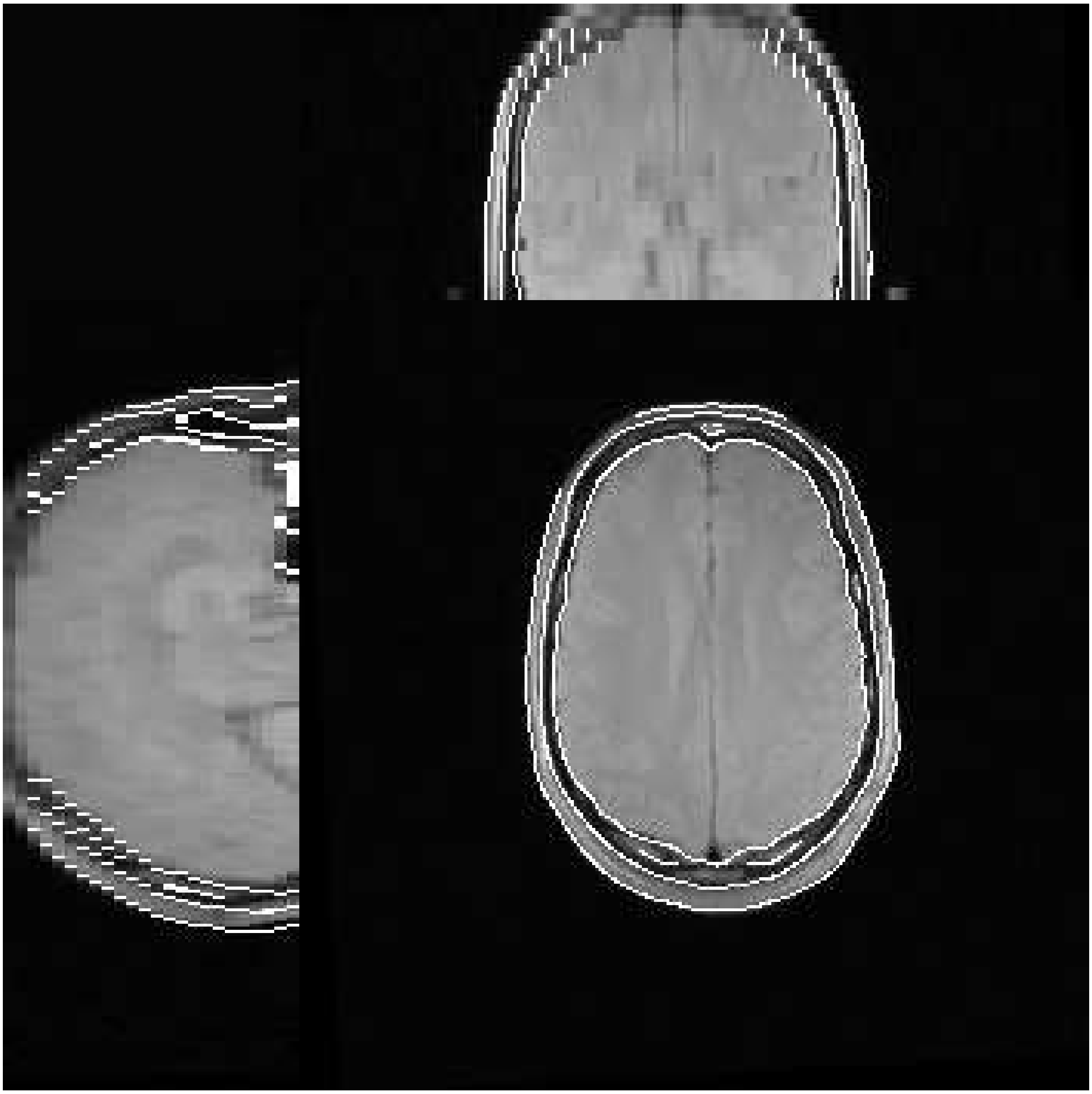

Figure 3: A final configuration for MR-CT registration by maximization of mutual information. The three central slices of the MRI data are shown with the edges from the registered CT data overlaid.

Level	XY Reduction		Z Reduction		Iterations	σ_{uu}	σ_{vv}	σ_v	λ_T	λ_R
	MR	CT	MR	CT						
1	8:1	16:1	1:1	1:1	10000	2.0	2.0	4.0	1	.0001
2	4:1	8:1	1:1	1:1	5000	2.0	2.0	4.0	.2	.00005
3	2:1	4:1	1:1	1:1	5000	2.0	2.0	4.0	.1	.00002
4	1:1	2:1	1:1	1:1	5000	2.0	2.0	4.0	.05	.00001
5	1:1	2:1	1:1	1:1	5000	2.0	2.0	4.0	.02	.000005

Table 1: MR – CT Registration Parameter Table

$\triangle$T	$\triangle\theta$	INITIAL				FINAL				TRIALS	SUCCESS
X Y Z		σ_X	σ_Y	σ_Z	$\overline{\mid\triangle\theta\mid}$	σ_X	σ_Y	σ_Z	$\overline{\mid\triangle\theta\mid}$		
$\pm$mm	$^\circ$	mm			$^\circ$	mm			$^\circ$		%
25	20	14.14	14.27	14.81	10.72	1.00	1.70	1.09	2.70	111	90
100	20	57.43	56.36	51.60	8.92	1.06	1.97	1.16	2.96	87	41
25	45	17.00	16.8	17.64	22.42	1.05	1.34	.98	2.42	70	68
10	10	5.63	5.90	5.89	5.11	1.44	2.05	1.12	3.18	20	100

Table 2: MR – CT Registration Results Table

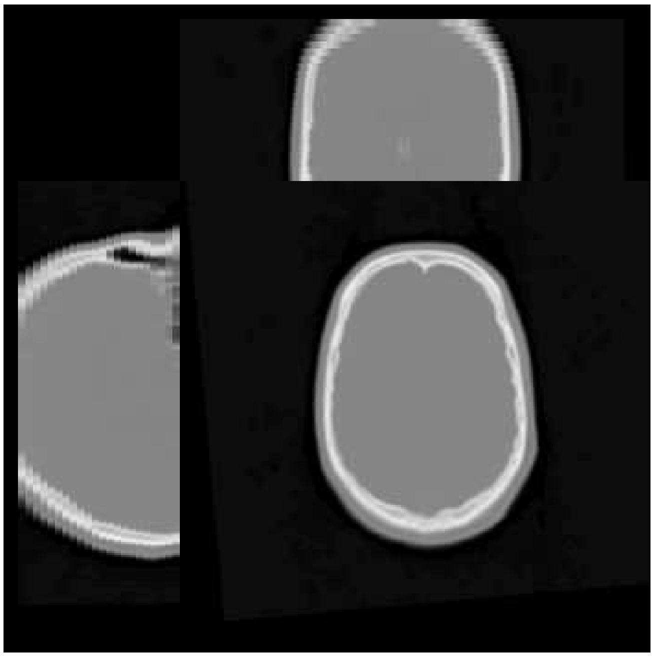

Figure 2: The three central slices of the CT data used in the MR-CT experiments.

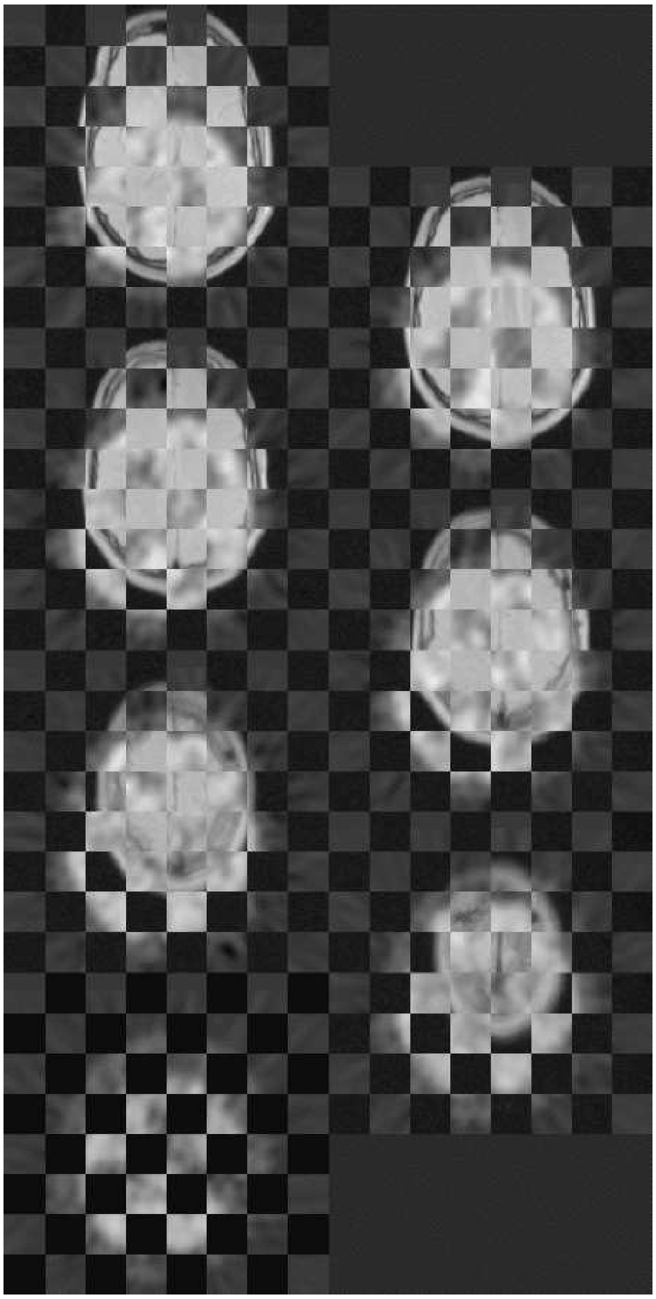

Figure 4: An initial condition for MR-PET registration as a checkerboard composite of The three central slices.

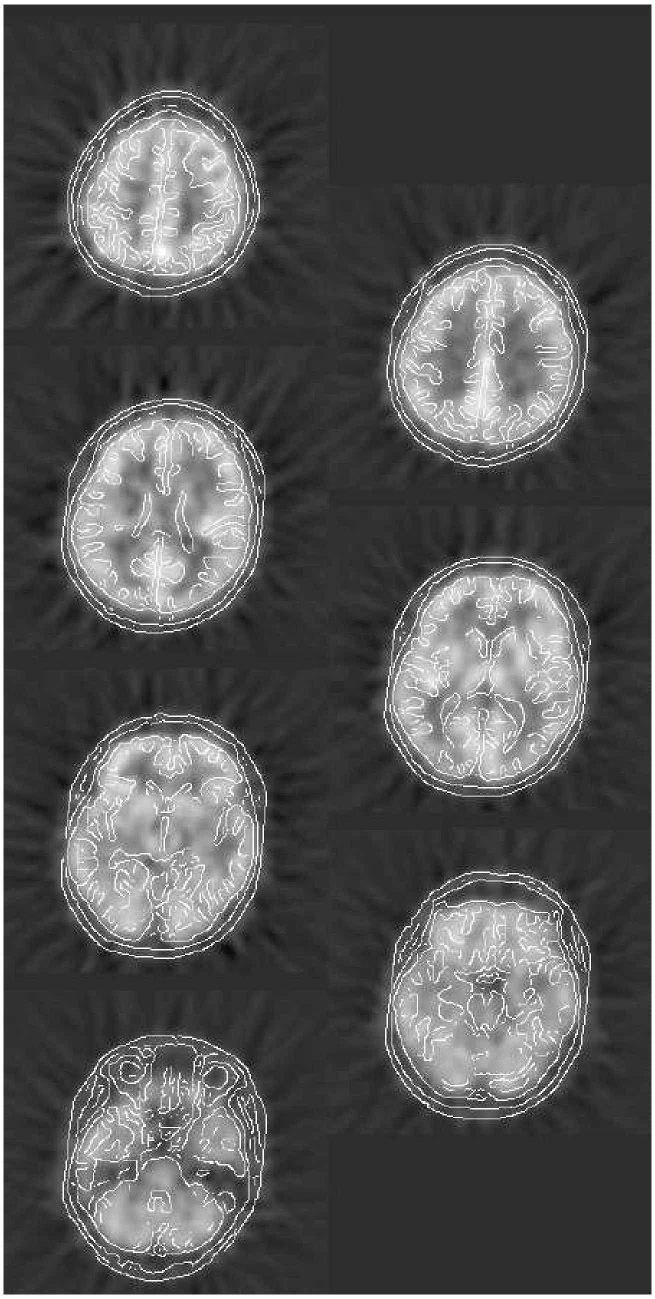

Figure 5: A final configuration for MR-PET registration by maximization of mutual information. The original PET slices are shown with the edges from the registered MRI overlaid

iteration an incremental change in position and orientation was computed. The incremental rotation was represented by a small-angle approximation of a rotation quaternion that is linear in three parameters.

The reference volume data voxels were sampled uniformly, and tri-linear interpolation was used to sample the test volume at non-integral coordinates. The test volume gradient was approximated (without interpolation) by the first differences of the data surrounding the location. If the transformation of a reference volume coordinate projected outside of the test volume, the value zero was used for the test volume intensity.

The parameter settings used in the registration experiments are listed in Table 1. The two intensities are both scalars, so we have listed standard deviations for the Parzen kernels rather than covariances. Different learning rates were used for rotations and translations, they are λ_R and λ_T respectively.

Table 2 summarizes a series of randomized experiments that were performed to gain an indication of the reliability, accuracy, and repeatability of the registration. From a "good" position and orientation (*pose*), a random uniformly distributed offset is added to each translational axis ($\triangle$T) after the reference volume has been rotated about a randomly selected axis by a random uniformly selected angle $\triangle\theta$. The distributions of the final and initial poses can be evaluated by comparing the standard deviations of the location of the center, computed separately in X, Y and Z. Furthermore, the average rotation angle from an "average" rotation is computed ($\overline{| \triangle \theta |}$). Finally, the number of poses that failed to converge near the correct solution is reported. The final statistics were only evaluated over the "good" poses.

Running time for each full registration was approximately six minutes on a Digital Equipment Corporation Alpha 3000/600.

3.2 MRI – PET Registration [5]

A preliminary experiment was performed to investigate the utility of the method for the registration of MR images with positron-emission tomography (PET) images. The PET data consisted of seven slices of 256x256 pixels each, the interslice spacing was 12 mm, while the pixel size was 1 mm square.

The MRI data consisted of 120 slices of 256x256 pixels each, the voxels measured 1.3 mm cubed.

The experiments closely followed the procedures described above. The MR image served as the test volume volume while the PET images were the reference volume. The parameters used are summarized in Table 3.

Repeated trials were not performed here, however a representative run is illustrated in Figures 4 and 5 which illustrate the data and final configuration for an MR-PET Registration. These results are at least visually satisfying, the activity imaged in the PET data follows the brain anatomy apparent in the MRI.

It was observed in these experiments that if the initial position of the PET activity was "above" the MRI brain anatomy, then there was a tendency for the optimization to become trapped in a local minimum where

[5] Images are courtesey of Dr. Jael Travere of Cyceron Center (CEA, Caen, FRANCE).

Level	XY Reduction		Z Reduction		Iterations	σ_{uu}	σ_{vv}	σ_v	λ_T	λ_R
	MR	PET	MR	PET						
1	8:1	8:1	8:1	1:1	10000	2.0	2.0	4.0	.1	.00001
2	8:1	4:1	8:1	1:1	10000	2.0	2.0	4.0	.05	.000005
3	4:1	2:1	4:1	1:1	5000	2.0	2.0	4.0	.02	.000002
4	2:1	1:1	2:1	1:1	5000	2.0	2.0	4.0	.01	.000001

Table 3: MR – PET Registration Parameter Table

the PET activity was "stuck to the scalp tissue" in the MRI. One reason this problem arises is because the MRI data is anatomical, while the PET data is functional. A variety of methods could be used to overcome this difficulty – one approach would be to first isolate the brain in the MRI, semi-automatic methods for doing this are available [10].

4 Discussion and Related Work

We have presented a metric for evaluating the registration of multi-modal image data that uses intensity information directly. The metric has been rigorously derived from information theory. While intensity based, it is more robust than traditional correlation – since it is insensitive to negating the test volume data, as well as a variety of non-linear transformations [8], which would defeat conventional intensity-based correlation.

The use of mutual information as a registration method and the stochastic search technique we use appeared in [8], The experiments there were primarily registration of video images to 3D object models. A "toy" medical image problem was described: that of 2D registration of the two components of a dual-echo MRI slice.

The registration of medical images by optimization in transformation space has been investigated by many researchers. The use of correlation for the registration of MRI and CT has been investigated by Van den Elsen [11].

Pelizzari et al. have used surface-based methods to register PET and MRI imagery [12]. Jiang et al. [13] have applied a robust variant of chamfer matching to register surfaces from multi-modal medical images. Malandain et al. [14] have described a physically-based method for registration of medical images, including PET to MR, that is based on potentials of attraction. Grimson et al. have used surface-based methods to register MRI to laser measurements of the skin, as well as to register MRI to MRI [15]. While such approaches are often useful, the need for reliable segmentation can be a drawback for surface-based registration methods. In addition, the skin surface may be the least geometrically accurate part of MRI data, due to susceptibility artifacts.

Registration by extremizing properties of the joint signal has been used by Hill and Hawkes et al. [16] to register MRI, CT, and other medical image modalities. They use third order moments of the joint histogram to characterize the clustering of the joint data. We believe that mutual information is perhaps a more direct measure of the salient property of the joint data at registration, and demonstrate an efficient means of estimating and extremizing it.

It has been argued by Hill and Hawkes et al. that joint entropy is not an effective measure of the registration between medical images. They support this claim with experimental evidence which shows that the minimum of joint entropy, as they compute it, does not reliably coincide with the correct registration. This result may actually be due to their measure of entropy. As with each of the approaches they present, their measure of entropy is computed from a histogram of the joint distribution (frequently called the feature space). They compute entropy as $H = -\sum p(x) log p(x)$ where the summation is over histogram bins. From this formulation it is clear that H is only an estimator of entropy to the extent that the histogram is an estimate of the joint distribution.

The effectiveness of the histogram density estimate is significantly dependent on bin size. Furthermore, histograms enforce a bias toward piecewise constant density estimates. Hill and Hawkes et al. make no attempt to evaluate the quality of their histogram density estimates and little attempt to adjust bin size. Thus, their criticism is of a potentially inaccurate histogram-based entropy estimate. In contrast, the Parzen density estimator provides many advantages over the histogram density estimator: (1) it is smoothly differentiable; (2) it is better suited to smoothly varying densities; (3) while the minimum of the histogram entropy estimate is very dependent on the bin size, our entropy estimates are not terribly dependent on the smoothing parameters[6]; and (4) the Parzen density estimate allows us to prove that the stochastic gradient descent registration procedure will converge to solutions close to minimum entropy solutions even when the number of samples used is quite small [7].

Recently, Collignon et al. [17] described the use of joint entropy as a criterion for registration of CT and MRI data. They graphically demonstrated a good minimum by probing the criterion, but no search techniques were described. They also describe the use of Parzen density estimation for computing entropy, and their graphs illustrate a reduction in ripple artifacts when Parzen windowing is used. Later work employing mutual information and Powell's optimization method appears in [18].

We believe that mutual information provides some advantage over joint entropy by providing larger capture range – this behavior was apparent in the experiments we performed. It arises because of the additional influence of the term that rewards for complexity (entropy) in the portion of the test volume into which the reference

[6] Experiments presented in [7] demonstrate that minima of joint entropy are stable as the smoothing parameters change across several orders of magnitude

volume is transformed.

Woods [19] has suggested a measure of registration between MR and PET based on the assumption that when registered the range of PET values associated with a particular value of MR should be minimized. The overall measure is a sum of the standard deviations of the PET values associated with each value of MR. When viewed in a theoretical light, Woods' measure of registration is closely related to the conditional entropy of the test volume given the reference volume. We have shown that a very similar approach is a measure of conditional entropy when the test volume is conditionally Gaussian [7]. Woods' measure is most effective when the test volume is in fact conditionally Gaussian: for each value in the reference volume there is a uni-modal distribution of test volume values. Woods' technique can break down when there is a bi-modal or multi-modal distribution of test volume values. This is a common occurrence when matching CT and MR: indistinguishable tissue in CT can map to significantly different tissues in MR. In addition, differing levels of imaged activation may normally occur in brain compartments. In contrast, our mutual information measure can easily handle data that is conditionally multi-modal. Another source of concern regarding Wood's measure is sensitivity to noise and outliers. Like other quadratic measures, an otherwise good match can be swamped out by a few outliers. Our mutual information measure is robust in the face of outliers, since it does not involve higher order moments of the distribution.

Additional technical details on the relationship between mutual information and other measures of registration may be found in [7].

Entropy is playing an ever increasing role within the field of neural networks. There has been work using entropy and information in vision problems. None of these technique uses a non-parametric scheme for density/entropy estimation as we do. In most cases the distributions are assumed to be either binomial or Gaussian. Entropy and mutual information plays a role in the work of Linsker [20], Becker and Hinton [21] and Bell and Sejnowski [22].

Acknowledgments

We were partially inspired by the work of Hill and Hawkes on registration of medical images.

The experiments reported here occured while author Wells was visiting INRIA [7] Sophia-Antipolis. He thanks Nicholas Ayache for the opportunity to interact with project Epidaure, and the following people for fruitful discussions and contributions the the success of our experiments: Grégoire Malandain, and Xavier Pennec.

References

[1] Athanasios Papoulis. *Probability, Random Variables, and Stochastic Processes.* McGraw-Hill, Inc., third edition, 1991.

[2] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis.* John Wiley and Sons, 1973.

[3] John S. Bridle. Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. In David S. Touretzky, editor, *Advances in Neural Information Processing 2*, pages 211–217. Morgan Kaufman, 1989.

[4] B. Widrow and M.E. Hoff. Adaptive switching circuits. In *1960 IRE WESCON Convention Record*, volume 4, pages 96–104. IRE, New York, 1960.

[5] Lennart Ljung and Torsten Söderström. *Theory and Practice of Recursive Identification.* MIT Press, 1983.

[6] Simon Haykin. *Neural Networks: A comprehensive foundation.* Macmillan College Publishing, 1994.

[7] Paul A. Viola. *Alignment by Maximization of Mutual Information.* PhD thesis, Massachusetts Institute of Technology, 1995.

[8] P.A. Viola and W.M. Wells III. Alignment by maximization of mutual information. In *Proceedings of the 5th International Conference on Computer Vision*, 1995.

[9] P.J. Burt and E. H. Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, 31(4):532–540, 1983.

[10] H.E. Cline, W.E. Lorensen, R. Kikinis, and F. Jolesz. Three-Dimensional Segmentation of MR Images of the Head Using Probability and Connectivity. *JCAT*, 14(6):1037–1045, 1990.

[11] PA Van den Elsen. *Multimodality Matching of Brain Images.* PhD thesis, Utrecht University, 1993.

[12] CA Pelizzari, GTY Chen, DR Spelbring, RR Weichselbaum, and CT Chen. Accurate three dimensional registration of ct, pet and/or mr images of the brain. *J. Comp. Assis. Tomogr.*, 13:20 – 26, 1989.

[13] H. Jiang, R.A. Robb, and K.S. Holton. A New Approach to 3-D Registration of Multimodality Medical Images by Surface Matching. In *Visualization in Biomedical Computing*, pages 196–213. SPIE, 1992.

[14] G. Malandain, S. Fernandez-Vidal, and J.M. Rocchisani. Physically based rigid registration of 3-d free-form objects: Application to medical imaging. Technical Report 2453, Institut National de Recherche en Informatique et en Automatique, Sophia-Antipolis, 1005.

[15] W.E.L. Grimson, T. Lozano-Pérez, W. Wells, et al. An Automatic Registration Method for Frameless Stereotaxy, Image Guided Surgery, and Enhanced Realigy Visualization. In *Proceedings of the Computer Society Conference on Computer Vision and Pattern Recognition*, Seattle, WA., June 1994. IEEE.

[16] Derek LG Hill, Colin Studholme, and David J. Hawkes. Voxel Similarity Measures for Automated Image Registration. In *Proceedings of the Third Conference on Visualization in Biomedical Computing*, pages 205 – 216. SPIE, 1994.

[17] A. Collignon, D. Vandermuelen, P. Suetens, and G. Marchal. 3D Multi-Modality Medical Image Registration Using Feature Space Clustering. In N. Ayache, editor, *Computer Vision, Virtual Reality and Robotics in Medicine*, pages 195 – 204. Springer Verlag, 1995.

[18] A. Collignon et al. Automated multi-modality image registration based on information theory. In *Proceedings of the Information Processing in Medical Imaging Conference*, pages 263–274. Kluwer Academic Publishers, 1995.

[19] RP Woods, JC Mazziotta, and SR Cherry. Mri-pet registration with automated algorithm. *J. Comp. Assis. Tomogr.*, 17:536 – 546, 1993.

[20] R. Linsker. From basic network principles to neural architecture. *Proceedings of the National Academy of Sciences, USA*, 83:7508–7512, 8390–8394, 8779–8783, 1986.

[21] Suzanna Becker and Geoffrey E. Hinton. Learning to make coherent predictions in domains with discontinuities. In John E. Moody, Steven J. Hanson, and Richard P. Lippmann, editors, *Advances in Neural Information Processing Systems*, volume 4, Denver 1991, 1992. Morgan Kaufmann, San Mateo.

[22] Anthony J. Bell. An information-maximisation approach to blind separation. In *Advances in Neural Information Processing Systems*, volume 7, Denver 1994, 1995. Morgan Kaufmann, San Francisco. To appear.

[7] Institut National de Recherche en Informatique et en Automatique