Medical Image Analysis (1996) volume 1, number 1, pp 35–51

Multi-modal volume registration by maximization of mutual information

William M. Wells III[1,2*], Paul Viola[2,3], Hideki Atsumi[4], Shin Nakajima[5] and Ron Kikinis[5]

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

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 position and orientation until the mutual information between the 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 image data; no pre-processing or segmentation is required. 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. Surgical applications of the registration method are described.

Keywords: information theory, multi-modality volume registration, mutual information

Received October 5, 1995; revised November 22, 1995; accepted February 2, 1996

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 positron-emission tomography (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 negatively proportional to the likelihood that the two images are correctly registered. Unfortunately, squared difference and the closely related operation of correlation are not effective measures 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

*Corresponding author
(e-mail: sw@ai.mit.edu)

value, at least approximately. Using F we could evaluate registrations by computing F(MR) and comparing it via sum of squared differences (or correlation) with 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. While theoretically it might be possible to find F and use it in this fashion, in practice prediction of CT from MR is a difficult and under-determined problem.

Given that both MR and CT are informative of the same underlying anatomy, 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.

The paper is organized as follows. The method of registration by maximization of mutual information is described in section 2. The formulation is defined in terms of entropies of the image data, and an approach for estimating these entropies is described, along with a stochastic search algorithm. Experimental results involving MRI–CT and MRI–PET registration are reported in section 3. Section 4 describes the use of our alignment technology to assist in neurosurgical applications. Section 5 includes an analysis of an idealized multi-modal registration problem. In this section we also discuss issues of robustness with respect to occlusion. The paper concludes with a section describing related work and a summary.

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 the x are the coordinates of the voxel. A voxel of the test volume is denoted similarly as $v(x)$. Given that T is a transformation from the coordinate frame of the reference volume to the test volume, $v(T(x))$ is the test volume voxel associated with the reference volume voxel $u(x)$. Note that in order to simplify some of the subsequent equations we will use T to denote both the transformation and its parameterization.

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

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

Here we treat x as a random variable over coordinate locations in the reference volume. In the registration 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 (see Papoulis, 1991, for example):

$$I(u(x),\ v(T(x))) \equiv h(u(x)) + h(v(T(x))) - h(u(x), v(T(x))). \tag{2}$$

$h(\cdot)$ is the entropy of a random variable, and is defined as $h(x) \equiv -\int p(x) \ln p(x) \mathrm{d}x$, while the joint entropy of two random variables x and y is $h(x, y) \equiv -\int p(x, y) \ln p(x, y) \mathrm{d}x\, \mathrm{d}y$. 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. This term is discussed in relation to an idealized example in section 5. The negative joint entropy 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 a sample.

Our first step in estimating entropy from a sample is to approximate the underlying probability density $p(z)$ by a superposition of functions centered on the elements of a sample A drawn from z:

$$p(z) \approx P^*(z) \equiv \frac{1}{N_{\mathrm{A}}} \sum_{z_j \in A} R(z - z_j) \tag{3}$$

where N_{A} is the number of trials in the sample A and R is a *window* function which integrates to 1. $P^*(z)$ is widely known as the *Parzen window* density estimate. It is described in Duda and Hart (1973).

In our subsequent analysis we will assume that the window function is a Gaussian density function. This will simplify some of our subsequent analysis, but it is *not* necessary. Any differentiable function could be used. Another good choice is

the Cauchy density. The Gaussian density function is

$$G_\psi(z) \equiv (2\pi)^{\frac{-n}{2}} |\psi|^{\frac{-1}{2}} \exp(-\tfrac{1}{2} z^{\mathrm{T}} \psi^{-1} z)\,,$$

where ψ is the (co-)variance of the Gaussian. The Parzen density estimate and the Parzen window functions can be defined over either scalar or vector data. When z is a vector ψ is the covariance matrix of a multi-dimensional Gaussian density.

Unfortunately, evaluating the entropy integral

$$h(z) \approx -E_z[\ln P^*(z)] = -\int_{-\infty}^{\infty} P^*(z) \ln P^*(z)\, \mathrm{d}z$$

is difficult if not impossible. This integral can, however, be approximated as a sample mean:

$$h(z) \approx -\frac{1}{N_{\mathrm{B}}} \sum_{z_i \in B} \ln P^*(z_i)\,, \tag{4}$$

where N_{B} is the size of a second sample B. The sample mean converges toward the true expectation at a rate proportional to $1/\sqrt{N_{\mathrm{B}}}$.

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

$$h(z) \approx h^*(z) \equiv \frac{-1}{N_{\mathrm{B}}} \sum_{z_i \in B} \ln \frac{1}{N_{\mathrm{A}}} \sum_{z_j \in A} G_\psi(z_i - z_j)\,. \tag{5}$$

To reiterate, two samples can be used to estimate the entropy of a density: the first is used to estimate the density, the second is used to estimate the entropy[a].

Next we examine the entropy of $v(T(x))$, which is a function of the transformation T. In order to find a maximum of entropy or mutual information, we may ascend the gradient with respect to the transformation T. After some manipulation, the derivative of the entropy may be written as follows,

$$\frac{\mathrm{d}}{\mathrm{d}T} h^*(v(T(x))) = \frac{1}{N_{\mathrm{B}}} \sum_{x_i \in B} \sum_{x_j \in A} W_v(v_i, v_j)(v_i - v_j)^{\mathrm{T}} \psi^{-1} \frac{\mathrm{d}}{\mathrm{d}T}(v_i - v_j)\,, \tag{6}$$

using the following definitions:

$$v_i \equiv v(T(x_i)), \qquad v_j \equiv v(T(x_j)), \qquad v_k \equiv v(T(x_k)),$$

and

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

[a]Using a procedure akin to leave-one-out cross-validation a single sample can be used for both purposes.

The weighting factor $W_v(v_i, v_j)$ takes on values between zero and one. It will approach one if v_i is significantly closer to v_j than it is to any other element of A. It will be near zero if some v_k is significantly closer to v_i than v_j. Distance is interpreted with respect to the squared Mahalanobis distance (see Duda and Hart, 1973) $D_{\psi_v}(v) \equiv v^{\mathrm{T}} \psi_v^{-1} v$. Thus, $W_v(v_i, v_j)$ 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 (Bridle, 1989) on the negative of the Mahalanobis distance to indicate correspondence between v_i and elements of A.

The summand in Equation (6) may also be written as:

$$W_v(v_i, v_j)\, \frac{\mathrm{d}}{\mathrm{d}T} \frac{1}{2} D_{\psi_v}(v_i - v_j)\,.$$

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 of v which W indicates are nearby, i.e. clusters should be tightened.

2.3. Estimation of the derivatives of mutual information

The entropy approximation described in Equation (5) may now be used to evaluate the mutual information between the reference volume and the test volume [Equation (2)]. In order to seek a maximum of the mutual information, we will calculate an approximation to its derivative,

$$\frac{\mathrm{d}}{\mathrm{d}T} I(T) \approx \frac{\mathrm{d}}{\mathrm{d}T} h^*(u(x)) + \frac{\mathrm{d}}{\mathrm{d}T} h^*(v(T(x))) - \frac{\mathrm{d}}{\mathrm{d}T} h^*(u(x), v(T(x)))\,.$$

Recall that the reference volume is not a function of the transformation. As a result its derivative is zero. The remaining two terms are computed using Equation (6). The entropy of the test volume is dependent on the variance of the window functions, ψ_v[b]. The joint entropy of the reference and test volumes is computed using the multi-dimensional generalization of the entropy estimate. In general the joint entropy of two random variables, $h(u(x), v(T(x)))$, can be evaluated by constructing the vector random variable, $w = [u(x), v(T(x))]^{\mathrm{T}}$ and evaluating $h(w)$. The estimate of this entropy will be dependent on the covariance ψ_w of the multi-dimensional Parzen window functions that are used in the density estimator for w. We will assume that this covariance matrix is diagonal: $\psi_w = \mathrm{DIAG}(\psi_{uu}, \psi_{vv})$.

[b]Note: this is not variance of the signal, $v(T(x))$, but the chosen width of the Parzen window functions. A principled scheme for selecting these widths is described in a later section.

Given these definitions we can obtain an estimate for the derivative of the mutual information as follows:

$$\widehat{\frac{\mathrm{d}I}{\mathrm{d}T}} = \frac{1}{N_\mathrm{B}} \sum_{x_i \in B} \sum_{x_j \in A} (v_i - v_j)^\mathrm{T} \times \left[W_v(v_i, v_j) \psi_v^{-1} - W_w(w_i, w_j) \psi_{vv}^{-1} \right] \frac{\mathrm{d}}{\mathrm{d}T} (v_i - v_j).$$

The weighting factors are defined as

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

$$W_w(w_i, w_j) \equiv \frac{G_{\psi_w}(w_i - w_j)}{\sum_{x_k \in A} G_{\psi_w}(w_i - w_k)},$$

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

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

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 $\frac{\mathrm{d}}{\mathrm{d}T}(v_i - v_j)$ will generally involve gradients of the test volume intensities, and the derivative of transformed coordinates with respect to the transformation.

2.4. Stochastic maximization of mutual information

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:

$$\begin{aligned} A &\leftarrow \{\text{sample of size } N_\mathrm{A} \text{ drawn from } x\} \\ B &\leftarrow \{\text{sample of size } N_\mathrm{B} \text{ drawn from } x\} \\ T &\leftarrow T + \lambda \widehat{\frac{\mathrm{d}I}{\mathrm{d}T}} \end{aligned}$$

The parameter λ is called the *learning rate*. The above procedure is repeated a fixed number of times or until convergence is detected. When using this procedure, some care must be taken to ensure that the parameters of transformation remain valid. For example, we may wish to find the best rotation transformation using a matrix representation for T. If the derivatives are with respect to the matrix entries then $T + \lambda \widehat{\frac{\mathrm{d}I}{\mathrm{d}T}}$ may no longer be a rotation matrix (for discussions of such issues see Paul, 1981; Ayache, 1991).

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 estimates instead of the true derivative for optimizing a function (see Widrow and Hoff, 1960; Ljung and Söderström, 1983; Haykin, 1994). 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 $N_\mathrm{A} = N_\mathrm{B} = 50$. We have proven that the technique will always converge to a transformation estimate that is close to locally optimal (Viola, 1995).

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.

2.5. Estimating the covariance

In addition to the learning rate λ, the covariance matrices of the Parzen window functions are important parameters of this technique. We have found that it is not difficult to determine suitable values for these parameters by empirical adjustment, and that is the method we usually use.

An automated method for determining these parameters has been described (Viola, 1995); we outline that approach here. Referring back to Equation (3), ψ should be chosen so that $P^*(z)$ provides the best estimate for $p(z)$. In other words ψ is chosen so that a sample B has the maximum possible likelihood. Assuming that the trials in B are chosen independently, the log likelihood of ψ is:

$$\ln \prod_{z_i \in B} P^*(z_i) = \sum_{z_i \in B} \ln P^*(z_i). \tag{7}$$

This equation bears a striking resemblance to Equation (4), and in fact the log likelihood of ψ is maximized precisely

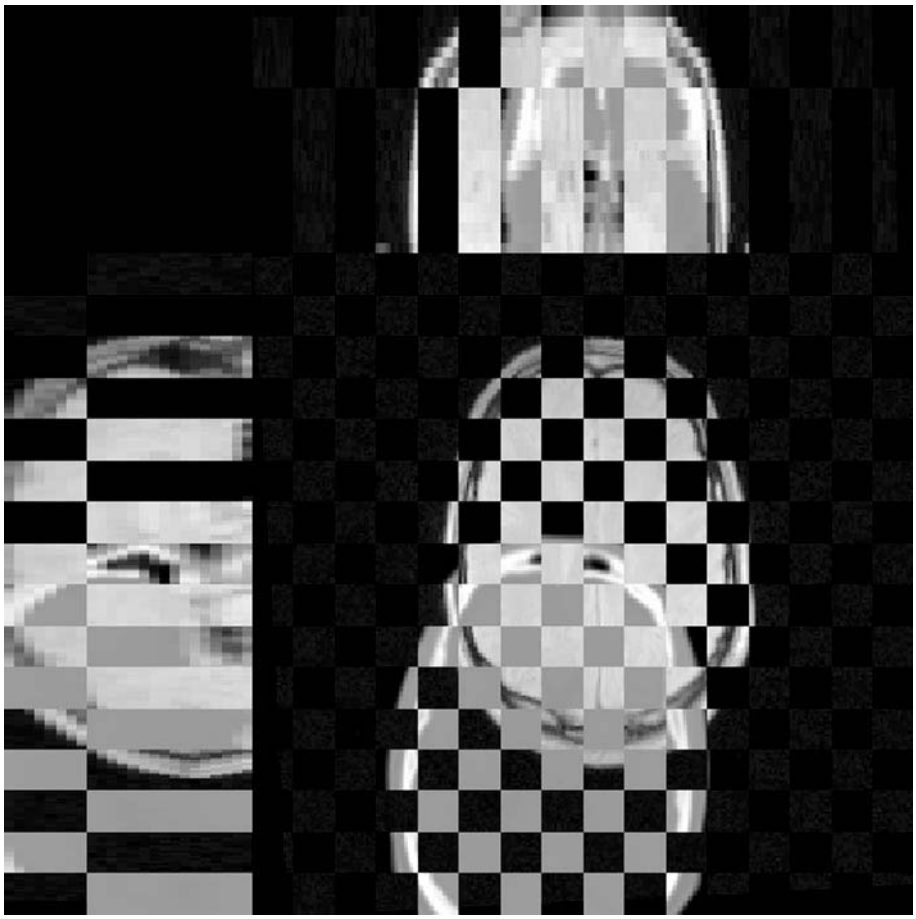


Figure 1. An initial condition for MR–CT registration by maximization of mutual information displayed as a checkerboard composite of the three orthogonal central slices.

when the entropy estimator $h^*(z)$ is minimized.

For simplicity, we assume that the covariance matrices are diagonal,

$$\psi = \mathrm{DIAG}(\sigma_1^2, \sigma_2^2, \ldots) . \tag{8}$$

Following a derivation almost identical to the one described above we can derive an equation analogous to Equation (6),

$$\frac{\mathrm{d}}{\mathrm{d}\sigma_k} h^*(z) = \frac{1}{N_b} \sum_{z_b \in b} \sum_{z_a \in a} W_z(z_b, z_a) \left(\frac{1}{\sigma_k}\right) \left(\frac{[z]_k^2}{\sigma_k^2} - 1\right) \tag{9}$$

where $[z]_k$ is the zth component of the vector $\boldsymbol{z}$. In practice both the transformation T and the covariance ψ can be adjusted simultaneously; so while T is adjusted to maximize the mutual information, $I(u(x), v(T(x)))$, ψ is adjusted to minimize $h^*(v(T(x)))$.

3. EXPERIMENTS

3.1. MRI–CT registration[a]

In this section we describe a series of experiments where the method was used to register MR images and CT images from the same person. Figures 1, 2 and 3 illustrate the data, initial configuration and final configuration for a representative MR–CT registration.

The MRI data consisted of 24 proton-density cross sections of 256×256 pixels each. The pixel dimensions were 1.25 mm squared and the slice spacing was 4 mm. The CT data were

[a]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.

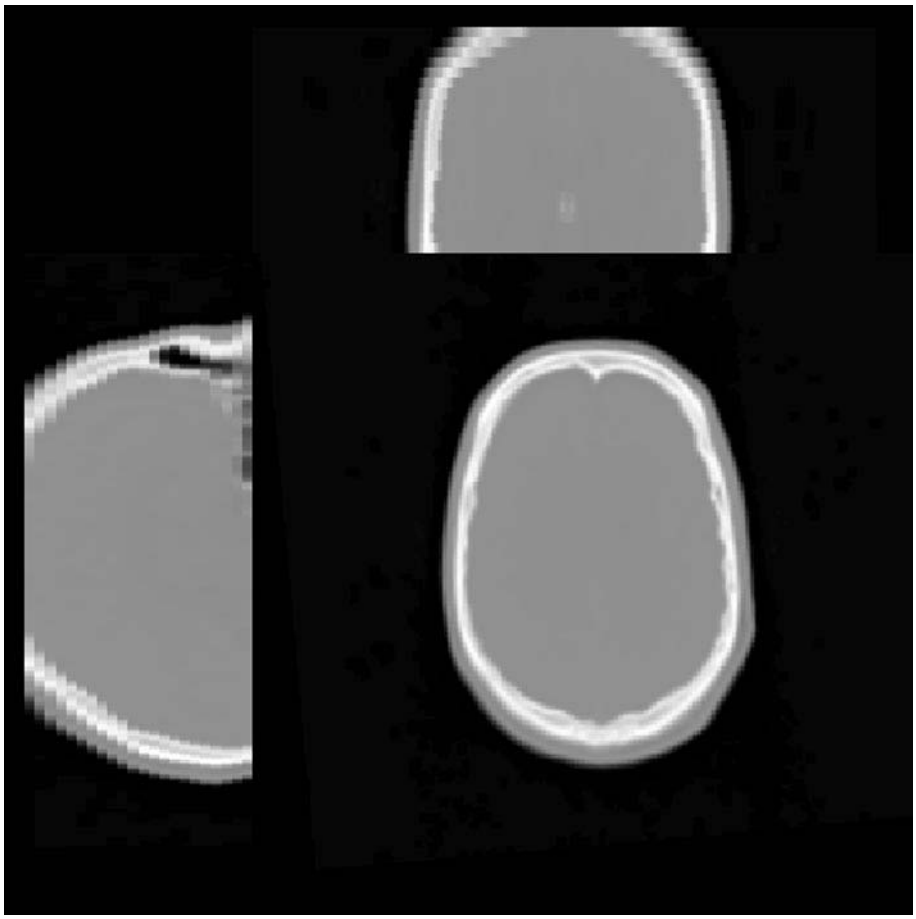


Figure 2. The three orthogonal central slices of the CT data used in the MR–CT experiments.

29 slices of 512 × 512, the pixel dimensions were 0.65 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. Since in theory, mutual information is a symmetrical measure, the assignment of test and reference volumes should be of little importance. However, in our implementation the details of the sampling are not symmetrical. While we do not believe this is an important factor here, we have not fully explored this issue experimentally.

The registration was performed in a coarse-to-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 is less tendency to become trapped in local minima, but the resulting accuracy is reduced.

Smoothing 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 (Burt and Adelson, 1983).

Rigid transformations were used; they were represented by displacement vectors and quaternions. At each 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. At each iteration the quaternions were normalized in order to avoid numerical drift in their magnitude.

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

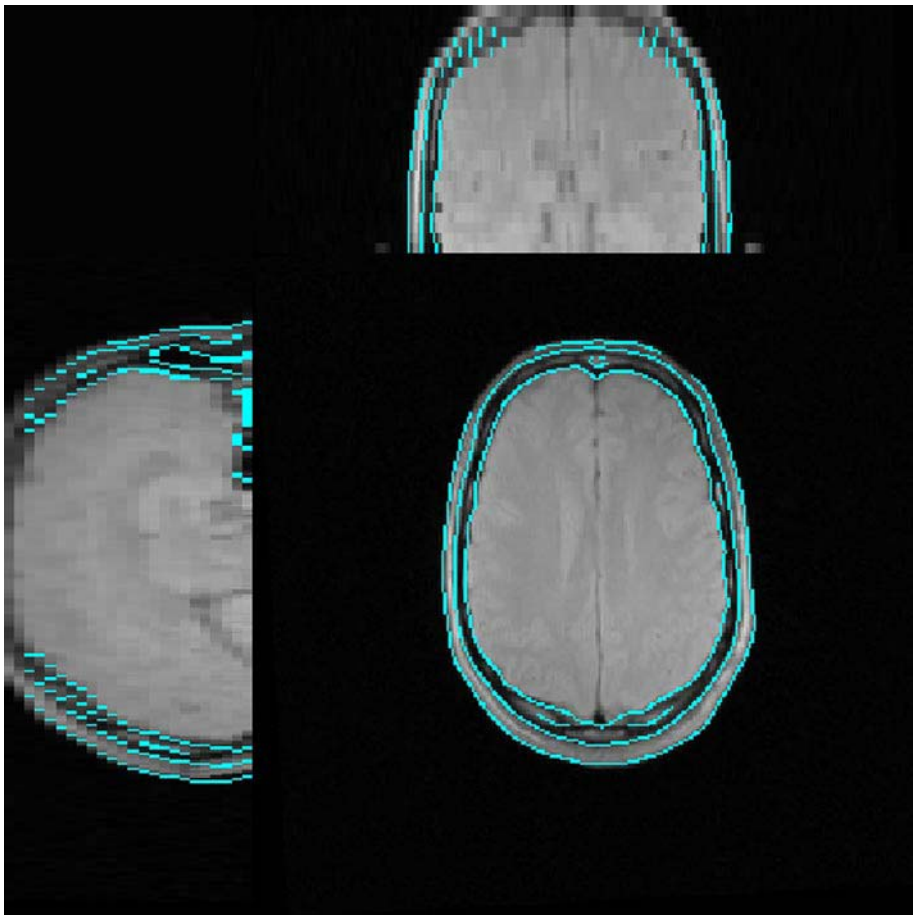


Figure 3. A final configuration for MR–CT registration by maximization of mutual information. The three orthogonal central slices of the MRI data are shown with the edges from the registered and reformatted CT data overlaid.

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 signal 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. These parameters were determined empirically in an effort to obtain good capture range and final accuracy.

Table 2 summarizes a series of randomized experiments that were performed to gain an indication of the reliability, accuracy and repeatability of the registration. Running time for each full registration was approximately 6 min on a Digital Equipment Corporation Alpha 3000/600. Video clip 1 illustrates a coarse-to-fine convergence of MR–CT registration.

3.2. MRI–PET registration[b]

An experiment was performed to investigate the utility of the method for the registration of MR images with PET images. The PET data consisted of seven slices of 256 × 256 pixels each, the interslice spacing was 12 mm, while the pixel size was 1 mm square.

The MRI data consisted of 120 slices of 256 × 256 pixels

[b]Images are courtesy of Dr Jael Travere of Cyceron Center (CEA, Caen, France).

Table 1. Parameters used in hierarchical MR–CT registration.

	XY reduction		*Z* reduction							
Level	MR	CT	MR	CT	Iterations	σ_{uu}	σ_{vv}	σ_v	λ_T	λ_R
1	8:1	16:1	1:1	1:1	10 000	2.0	2.0	4.0	1	0.0001
2	4:1	8:1	1:1	1:1	5 000	2.0	2.0	4.0	0.2	0.000 05
3	2:1	4:1	1:1	1:1	5 000	2.0	2.0	4.0	0.1	0.000 02
4	1:1	2:1	1:1	1:1	5 000	2.0	2.0	4.0	0.05	0.000 01
5	1:1	2:1	1:1	1:1	5 000	2.0	2.0	4.0	0.02	0.000 005

Differing amounts of in-slice (*XY*) and across slices (*Z*) smoothing and reduction were used in order to approach isotropy of voxel dimensions at the smoothest level. The variables σ_{uu}, σ_{vv} and σ_v denote the standard deviations used in the Parzen density approximators, which are the square roots of ψ_{uu}, ψ_{vv} and ψ_v, respectively. The translational and rotational learning rates are λ_T and λ_R, respectively.

Table 2. MR–CT registration results table.

		Initial				Final					
ΔT *XYZ* (±mm)	$\Delta\theta$ (deg)	σ_X	σ_Y (mm)	σ_Z	$\overline{\lvert\Delta\theta\rvert}$ (deg)	σ_X	σ_Y (mm)	σ_Z	$\overline{\lvert\Delta\theta\rvert}$ (deg)	Trials	Success (%)
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	0.98	2.42	70	68
10	10	5.63	5.90	5.89	5.11	1.44	2.05	1.12	3.18	20	100

From a known position and orientation, a random offset uniformly selected from the interval $\pm\Delta T$ was added to each translational axis after the reference volume had been rotated about a randomly selected axis by a random angle uniformly selected from the interval $\pm\Delta\theta$. 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 ($\overline{|\Delta\theta|}$). Finally, the number of trials that succeeded in converging to near the correct solution (by visual inspection) is reported. The final statistics were evaluated only over the successful trials.

Table 3. MR–PET registration parameter table.

	XY reduction		*Z* reduction							
Level	MR	PET	MR	PET	Iterations	σ_{uu}	σ_{vv}	σ_v	λ_T	λ_R
1	8:1	8:1	8:1	1:1	10 000	2.0	2.0	4.0	0.1	0.000 01
2	8:1	4:1	8:1	1:1	10 000	2.0	2.0	4.0	0.05	0.000 005
3	4:1	2:1	4:1	1:1	5 000	2.0	2.0	4.0	0.02	0.000 002
4	2:1	1:1	2:1	1:1	5 000	2.0	2.0	4.0	0.01	0.000 001

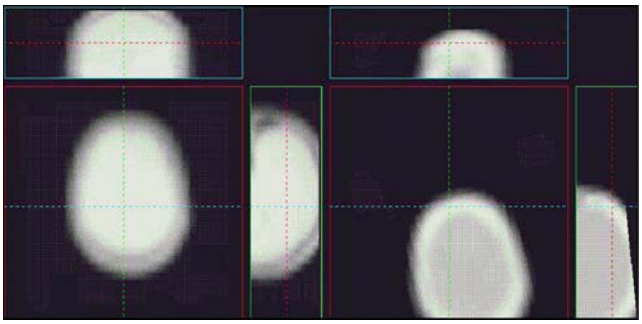
See Table 1 for explanation.

each, the voxels measured 1.3 mm cubed.

The experiments closely followed the procedures described above. The MR image served as the test 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, 5 and 6 which illustrate the data and the initial and final configurations of 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



Video clip 1. This is the first image of a video sequence that illustrates the registration of MR to CT. The sequence is divided into four parts, each part at a different spatial resolution going from coarse to fine.

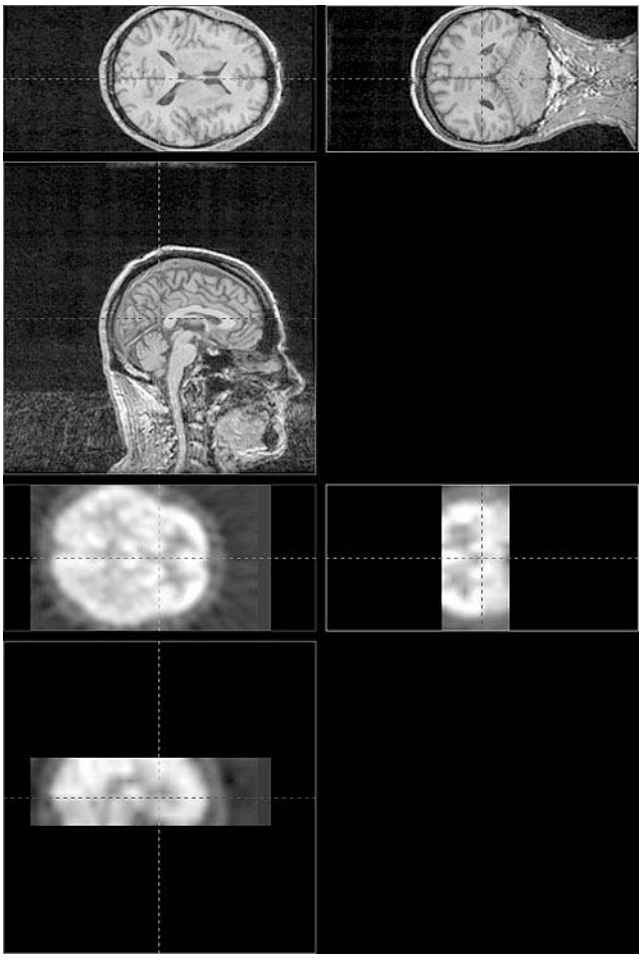


Figure 4. An initial configuration for MR–PET registration. Three orthogonal central slices are shown MR above and PET below. The PET data have been shifted posteriorly.

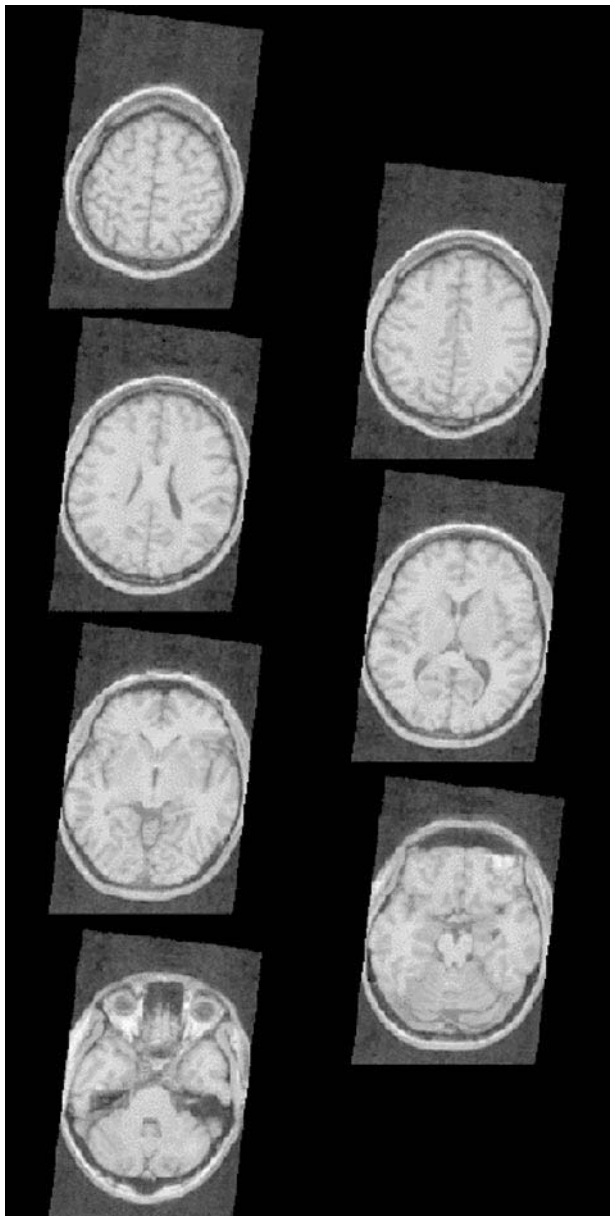


Figure 5. Registered MR data reformatted into the lattice of the PET data.

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 the PET activity was attracted 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

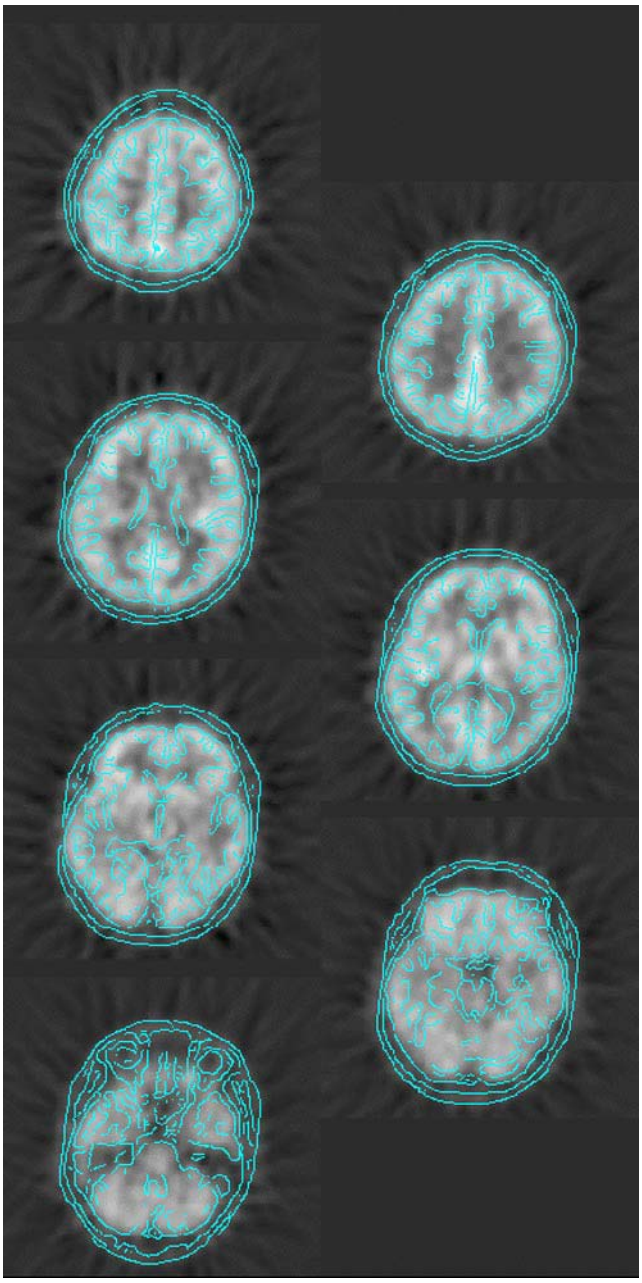


Figure 6. A final configuration for MR–PET registration by maximization of mutual information. The original PET slices are shown along with edges derived from the the MRI data after reformatting into the lattice of the PET data at the final pose.

doing this are available (Cline *et al.*, 1990).

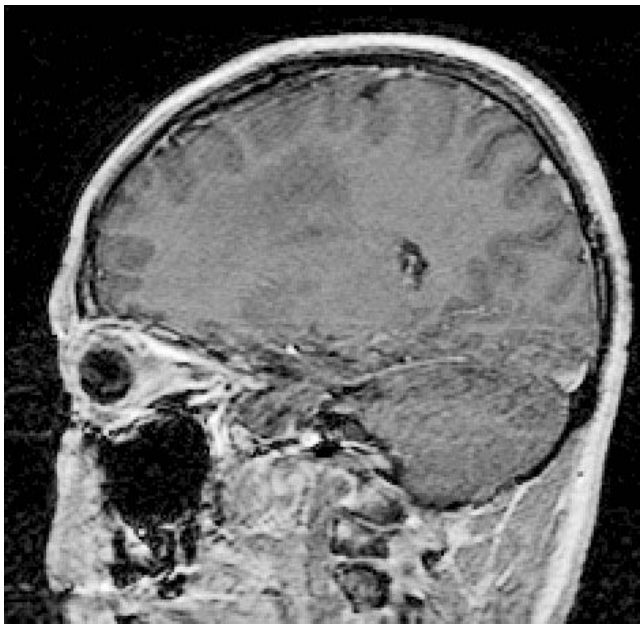


Figure 7. The tumor did not enhance well in the post-contrast SPGR MR images.

4. SURGICAL APPLICATIONS

One of the primary motivations for this research has been the integration of information from differing medical images for surgical exploitation. In this section we describe two examples in which the registration method was utilized in neurosurgical applications.

4.1. Case 1

Radiological examinations of several MRI acquisitions indicated that the patient had a tumor of the frontal lobes bilaterally. While providing good anatomical information, post-contrast gradient-echo (SPGR) MR images did not visualize the tumor well (Figure 7). The tumor was, however, evident with good contrast in a T2-weighted acquisition (see Figure 9). These two scans were registered using the method described above in order to facilitate the construction of 3-D models of the anatomy and pathology for surgical planning and visualization (Kikinis *et al.*, 1996).

The original SPGR MR images were 1.5 mm thick sagittal images, and the original T2-weighted MR images were 5.0 mm thick, 1.0 mm spacing axial images. The results of the registration are illustrated in Figure 8. After registration, the T2-weighted images were reformatted into the lattice of the SPGR images.

Three-dimensional models of the skin, brain, vessels and ventricles were generated from the SPGR MR images, and 3-

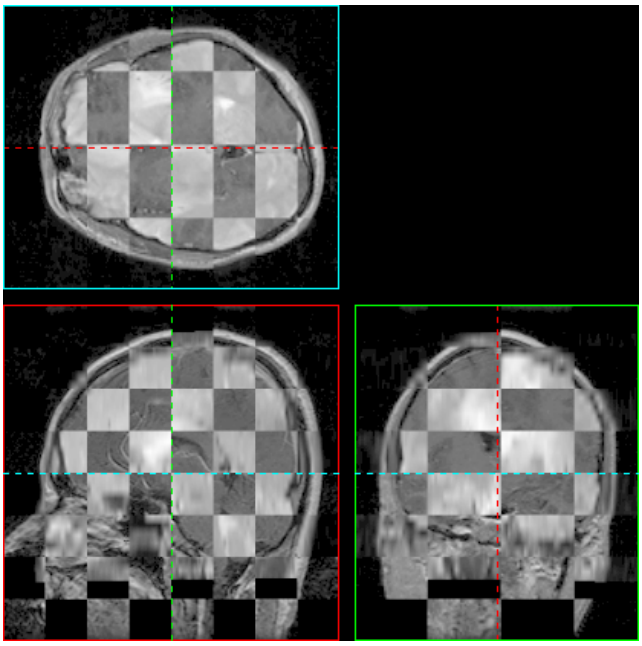


Figure 8. Registration of the SPGR and T2 images is illustrated in composite axial (left upper), sagittal (left lower) and coronal (right lower) images.

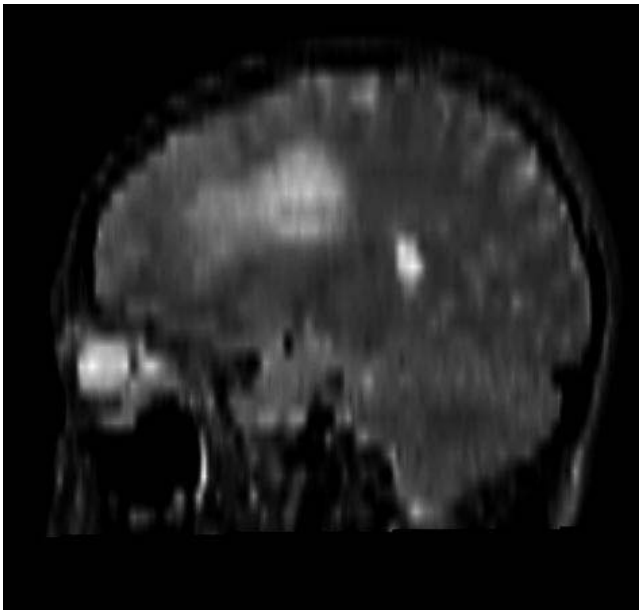


Figure 9. The reformatted T2-weighted images visualized the tumor in the frontal lobe. 3-D models of the tumor and the surrounding edema were extracted from these reformatted T2-weighted images.

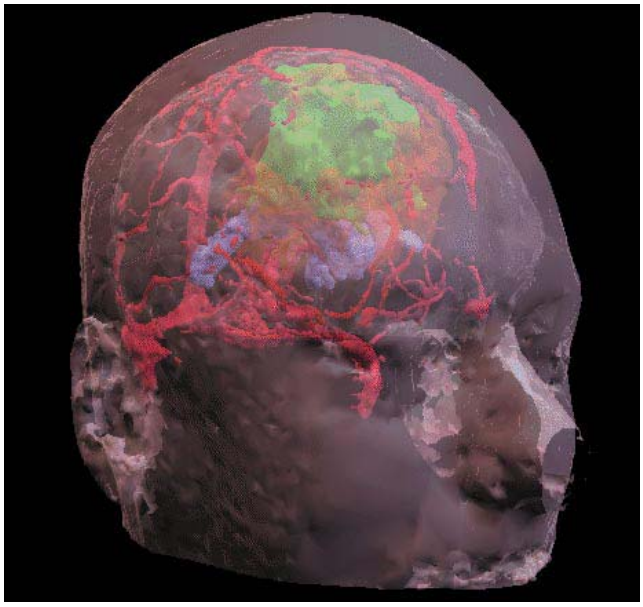


Figure 10. A rendering of the 3-D models constructed from the registered MR images. Models of the skin and the brain were generated from the SPGR MR images, and are rendered as translucent models. The vessels (red) and the ventricles (blue) were also generated from the SPGR images, while the tumor (green) and the surrounding edema (orange) were generated from the reformatted T2-weighted images.

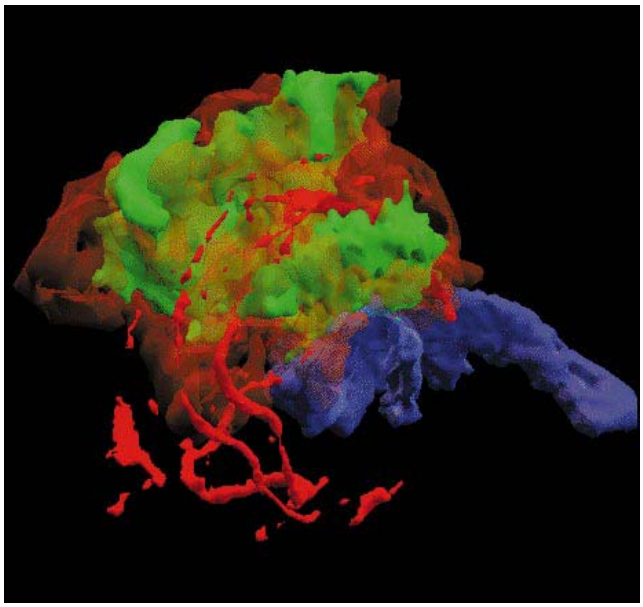
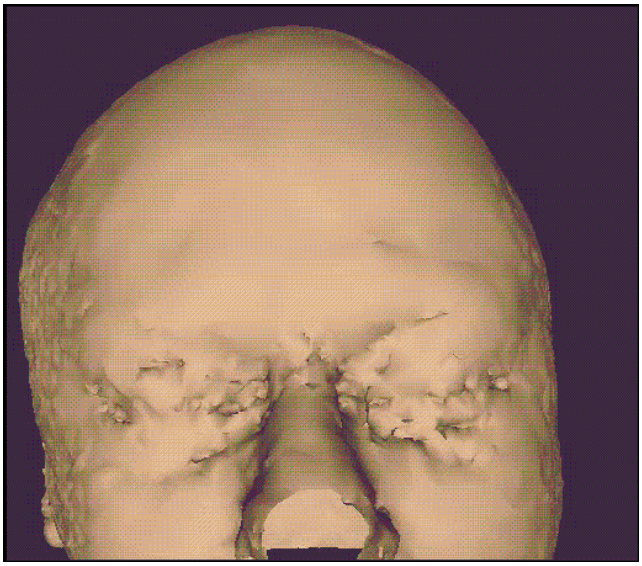


Figure 11. This rendering shows a left upper frontal view. The skin and brain models are suppressed for clarity. The anterior cerebral arteries are seen overriding the tumor, which is consistent with the radiological diagnosis.



Video clip 2. This is the first image of a video sequence that illustrates the results of the registration for the SPGR and T2 weighted scans. The brain is gray, the vascular structure red, the tumor green and the ventricles blue.

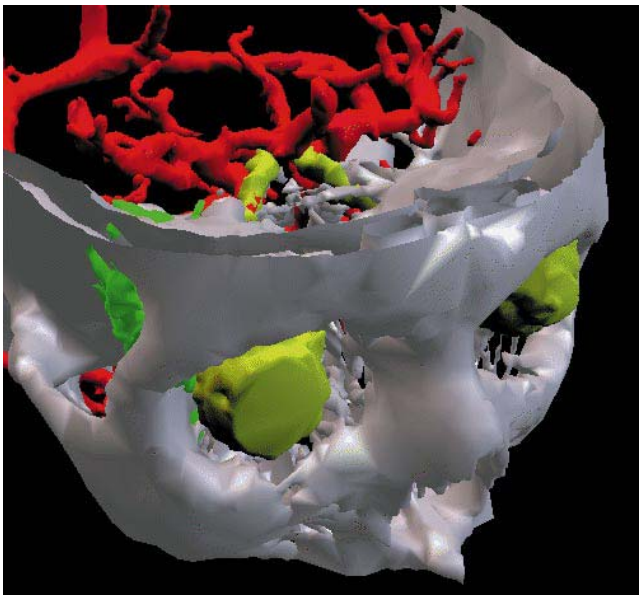


Figure 12. Right frontal view of the 3-D model. The skull (colored white) is derived from CT images, while the vascular tree (red) was derived from an MR angiogram. Models of the tumor (green) and optic nerve with the ocular bulb (yellow) were derived from an SPGR MR sequence.

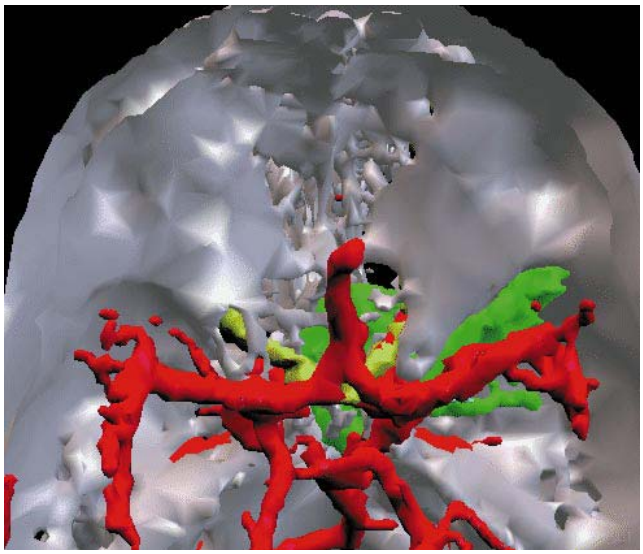


Figure 13. Top view of the 3-D models for case two, showing the spatial relationship of the circle of Willis (a part of the intracranial vascular tree), the optic nerves and the bony structure. The tumor (colored green) is in the middle cranial fossa and in the sphenoid sinus (one of the paranasal sinuses). This spatial information was found useful during surgical planning.

D models of the tumor and surrounding edema were generated from the reformatted T2-weighted MR images (Figure 10).

The radiological diagnosis suggested that parts of the tumor were present in both frontal lobes and that these parts were connected through the corpus callosum (the so-called butterfly configuration), so that the anterior cerebral arteries should override the tumor. This spatial relationship was consistent with the 3-D reconstructions from the registered SPGR and T2-weighted MR images (Figure 11 and Video clip 2).

The surgery was performed in an open-configuration MR unit (Schenk *et al.*, 1995). The abnormal area was biopsied, and proved to be a diffuse infiltrating glioma.

4.2. Case 2

This patient had a skull base meningioma which consisted of both intra- and extra-cranial parts. The 3-D models were made from CT images and two sequences of MR images (see Figure 12). In this 3-D model, the skull was constructed from CT images, the tumor and the optic nerves from SPGR MR images and the vascular tree from MR angiograms. The tumor is seen in extra-cranial and intra-orbital areas as well as in the middle cranial fossa. A top view of the anatomical structures appears in Figure 13.

Table 4. Idealized example: joint distribution on anatomy under transformation, $p(A(x), A(T(x)))$, at alignment.

		$A(x)$					
		AIR	BONE	WM	GM	FAT	Row sum constraint
$A(T(x))$	AIR	P_A	0	0	0	0	P_A
	BONE	0	P_B	0	0	0	P_B
	WM	0	0	P_W	0	0	P_W
	GM	0	0	0	P_G	0	P_G
	FAT	0	0	0	0	P_F	P_F
Column sum constraint		P_A	P_B	P_W	P_G	P_F	

5. DISCUSSION

5.1. Analysis of an idealized example

This section analyzes an idealization of a medical registration problem in order to clarify our registration approach and to suggest its effectiveness. The example used is a simplification of the situation that occurs during the registration of volumetric data of the head from MRI and CT. We will show that under certain reasonable assumptions, the joint signal entropy (an important component of mutual information) is at a local minimum at alignment.

Let us suppose that the anatomy is characterized by a function mapping from locations in space to the following tissue types: air, bone, white matter (WM), gray matter (GM) and fat,

$$A(x) \in \{\text{AIR, BONE, WM, GM, FAT}\},$$

and that the overall probabilities of the various tissues occuring in the volume are the non-zero values P_A, P_B, P_W, P_G, and P_F respectively.

Suppose that there are two observations of the anatomy, $A(x)$ and $A(T(x))$, with the second observed through a coordinate transformation T. We assume that T is a volume-preserving transformation such as rigid-body motion, and that the volume boundary conventions are defined such that the marginal distributions, $p(A(x))$ and $p(A(T(x)))$, are not a function of T. This allows us to ignore the marginal entropy terms of mutual information and to focus solely on the joint entropy term.

Since the joint entropy is a property of the joint distribution let us examine the joint distribution on anatomy under transformation, $p(A(x), A(T(x)))$. This distribution is tabulated in Table 4 for the particular case in which the two signals are properly aligned, e.g. $T(x) = x$. When this holds, the distribution is diagonal. The joint distribution is subject to the constraints that the marginal distributions equal the overall tissue probabilities, this leads to the row- and column-sum constraints that are listed in the table. These constraints hold independently of T. We assume that the joint distribution $p(A(x), A(T(x)))$ departs from being diagonal when T departs from the null transformation. In other words, when the anatomy is compared with itself under a non-null transformation, some mixing of the tissue compartments will occur (otherwise the anatomy is degenerate with respect to the transformations induced by T).

Table 5. Idealized example: contrast properties of hypothetical imaging modalities F and G. Note that modality F does not separate the soft tissues, and modality G does not separate bone from air.

Tissue	F(tissue)	G(tissue)
AIR	F_1	G_1
BONE	F_2	G_1
WM	F_3	G_2
GM	F_3	G_3
FAT	F_3	G_4

Now let us introduce two imaging modalities, F and G, whose contrast properties are described in Table 5. F and G are intended to be analogous to CT and MRI, respectively. We assume that modality F observes the anatomy through a transformation, $T(x)$, with respect to the coordinates of modality G, and define the signals in the following way,

$$u(x) \equiv G(A(x)) \quad \text{and} \quad v(x) \equiv F(A(x)).$$

The joint signal distribution, $p(u(x), v(T(x)))$, is shown in Table 6, for the case that the two signals are properly aligned. Guided by the tissue contrast properties (shown in Table 5), this distribution is easily constructed from the one shown in Table 4, by merging the probabilities in the first two columns and the last three rows. From the definition of the entropy of

Table 6. Idealized example: distribution on joint signal, $p(u(x), v(T(x)))$, at alignment.

		$u(x) = G(A(x))$				
		G_1	G_2	G_3	G_4	Row sum constraint
$v(T(x)) = F(A(T(x)))$	F_1	P_{A}	0	0	0	P_{A}
	F_2	P_{B}	0	0	0	P_{B}
	F_3	0	P_{W}	P_{G}	P_{F}	$P_{\mathrm{W}} + P_{\mathrm{G}} + P_{\mathrm{F}}$
Column sum constraint		$P_{\mathrm{A}} + P_{\mathrm{B}}$	P_{W}	P_{G}	P_{F}	

a discrete probability distribution[a] we see that *in this example* the entropy of the joint signal, *when properly aligned*, is the same as the entropy of the anatomy,

$$\begin{aligned} H(u(x), v(T(x))) &= H(A(x)) = -P_{\mathrm{A}} \ln P_{\mathrm{A}} - P_{\mathrm{B}} \ln P_{\mathrm{B}} \\ &\quad -P_{\mathrm{W}} \ln P_{\mathrm{W}} - P_{\mathrm{G}} \ln P_{\mathrm{G}} - P_{\mathrm{F}} \ln P_{\mathrm{F}}. \end{aligned}$$

We consider a differential change of the transformation away from the null transformation, and assume that this will induce a mixing of tissue compartments that is observable in the joint signal, for example if misalignment causes the air and fat tissue structures to overlap, then the upper-right and/or lower-left entries in the joint signal distribution will become non-zero. Note that any change in the distribution of the joint signal due to misalignment will require that some zero-probability entries become non-zero. This is because, at alignment, the non-zero values are maximal due to the marginal constraints. We assume that a differential change in alignment induces a differential change in the joint signal distribution, thus the effect on the distribution is that some zero-probability entries receive positive differential increments, while some non-zero entries receive negative differential increments.

The partial derivative of the entropy of a distribution with respect to the probability of a particular event is $\frac{\partial H}{\partial p_i} = -(1 + \ln p_i)$. This partial derivative is finite for non-zero probability, and approaches positive infinity as the probability of the event approaches zero. Because of this, the change in the joint signal entropy due to the probability changes described above will be positive, since the effect of the zero-probability entries will dominate. Thus the joint signal entropy at alignment is a local minimum.

This idealization has modeled the imaging modalities as having a few discrete values, while conventional medical imaging modalities typically take on many values, and are more conveniently modeled as continuous intensities. The discrete values used in the modeling here will correspond to specific clusters in conventional data, and these clusters will have some variance due to partial-volume effects and any spatial smoothing that is used.

While not all of the above assumptions are met in real applications, conventional medical image registration problems often have the property that spurious clusters appear in the joint signal under misalignment, due to the simultaneous observation of differing tissues. This may be the cause of the increase in entropy we have observed in the joint signal when misaligned.

5.2. Discussion: correlation and occlusion

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.

Conventional correlation may be seen to align two signals by minimizing a summed quadratic penalty in the difference between their intensities. For the sake of example, let us consider two hypothetical signals that can be aligned well by traditional correlation, i.e. at alignment their intensities are in good agreement. If we then negate the intensity of one of these signals, their intensities will no longer agree, and their alignment by correlation will most likely fail. It is easy to see that the mutual information formulation of alignment is insensitive to, and in fact not affected by, the negation of either of the signals. Similar robustness with respect to other transformations is described in Viola and Wells (1995).

Mutual information also has attractive robustness with respect to occlusions of one of the signals, while traditional correlation is often significantly disturbed by occlusions, since they lead to substantial penalties for disagreement of intensities. In medical imagery, the effect of occlusions on the joint signal is frequently manifested by the appearance of additional intensity clusters where the valid part of one signal is in registration with a background (occluded) value for the other signal. While such additional clusters will typically reduce the mutual information at alignment, we have found that there can still be a good

[a] $H(x) \equiv -\sum_i p(x_i) \ln (p(x_i))$

maximum at alignment, i.e. that the mutual information measure degrades gracefully when subject to partially occluded imagery.

6. RELATED WORK

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 (Van den Elsen, 1993).

Pelizzari *et al*. (1989) have used surface-based methods to register PET and MRI imagery. Jiang *et al*. (1992) have applied a robust variant of chamfer matching to register surfaces from multi-modal medical images. Malandain *et al*. (1995) have described a physically based method for registration of medical images, including PET to MR, that uses potentials of attraction. Grimson *et al*. (1994) have used surface-based methods to register MRI to laser measurements of the skin, as well as to register MRI to MRI. 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 investigated (Hill *et al*., 1994) for the alignment of MRI, CT and other medical image modalities. They showed interesting scatter-plots of the joint data as the registration is disturbed, and used third-order moments of the joint histogram, as well as other measures to characterize the clustering of the joint data.

The use of joint entropy as a criterion for registration of CT and MRI data has been explored (Collignon *et al.*, 1995b). They graphically demonstrated a good minimum by probing the criterion, but no search techniques were described. They also described the use of Parzen density estimation for computing entropy, and their graphs illustrate a reduction in ripple artifacts when Parzen windowing is used.

The use of mutual information as a registration method and the stochastic search technique we use appeared in Viola and Wells (1995). The experiments there were primarily registration of video images to 3-D object models. A simplified medical image problem was described: that of 2-D registration of the two components of a dual-echo MRI slice.

Several researchers have investigated the use of joint entropy to characterize the proper registration of medical imagery (Collignon *et al.*, 1995a; Studholme *et al*., 1995b), and found that it is not a robust measure of registration, with Collignon (1995a) describing difficulties associated with partial overlap of the data. Collignon *et al*. (1995a) and Studholme *et al*. (1995a) found registration based on mutual information to be an attractive approach, with Collignon *et al*. (1995a) describing the use of Powell's optimization method.

In a previous report of this research (Wells *et al.*, 1995), mutual information combined with stochastic search was shown to be a robust approach for the registration of medical imagery.

We believe that mutual information provides some advantage over joint entropy by providing larger capture range—this behavior was apparent in the experiments we have 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 volume is transformed.

Woods *et al.* (1993) 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 (Viola, 1995). 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 are 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 Viola (1995).

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 techniques 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 (1986), Becker and Hinton (1992) and Bell and Sejnowski (1994).

7. SUMMARY AND CONCLUSIONS

The registration of volumetric data from sources such as MR, CT or PET, is of importance for surgical planning, diagnosis and medical research. While there are many existing approaches based on alignment of surfaces, these techniques are dependent on the *a priori* quality of the available segmentations. Alternatively, intensity-based techniques can work directly with the volumetric data. In the past these techniques have relied on somewhat *ad hoc* assumptions about the nature of the signals involved.

We have presented a technique based on mutual information that requires neither a segmentation nor any *ad hoc* assumptions about the nature of the imaging modalities. In addition to being effective and efficient, the technique is quite general. It shows promise in many application domains.

In related work we have shown that the same formalism can be used to register 3-D volumetric information directly to video images of patients (Viola and Wells, 1995). We are currently constructing a unified registration system that can accommodate various planar and volumetric images. In addition we hope to extend these techniques so that they can be used in domains where the correct registration may not be rigid.

ACKNOWLEDGEMENTS

We were partially inspired by the work of Hill and Hawkes on registration of medical images. The experiments reported here occurred while W. M. Wells was visiting INRIA[a] Sophia-Antipolis. He thanks Nicholas Ayache for the opportunity to interact with project Epidaure, and Grégoire Malandain and Xavier Pennec for fruitful discussions and contributions the the success of our experiments. The biopsy referred to in section 4.1 was performed by Drs P. McL. Black, E. Alexander III and T. Moriarty from the Neurosurgery Department of Brigham and Womens' Hospital and Harvard Medical School. The open-configuration MR unit is a joint research project among the MR division of the Radiology Department of Brigham and Womens' Hospital and Harvard Medical School, and General Electric Medical Systems. P. Viola would like to thank the following sources for their support of this research: USAF ASSERT program, Parent Grant#:F49620-93-1-0263 and Howard Hughes Medical Institute. In addition Viola would like to thank Terrence J. Sejnowski of the Salk Institute, San Diego, CA for providing space and resources during part of the preparation of this research. We thank the anonymous reviewers for their constructive comments.

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

REFERENCES

Ayache, N. (1991) *Artificial Vision for Mobile Robots–Stereo-Vision and Multisensor Perception*. MIT Press, Cambridge, MA.

Becker, S. and Hinton, G. E. (1992) Learning to make coherent predictions in domains with discontinuities. In Moody, J. E., Hanson, S. J. and Lippmann, R. P. (eds), *Advances in Neural Information Processing Systems*, vol 4, Denver 1991. Morgan Kaufmann, San Mateo, CA.

Bell, A. (1995) An information-maximisation approach to blind separation. In *Advances in Neural Information Processing Systems*, vol 7, Denver 1994. Morgan Kaufmann, San Francisco, CA.

Bridle, J. (1989) Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters. In Touretzky, D. S. (ed.), *Advances in Neural Information Processing 2*, pp. 211–217. Morgan Kaufman, San Mateo, CA.

Burt, P. and Adelson, E. (1983) The Laplacian pyramid as a compact image code. *IEEE Trans. Communications*, 31, 532–540.

Cline, H., Lorensen, W., Kikinis, R. and Jolesz, F. (1990) Three-dimensional segmentation of MR images of the head using probability and connectivity. *JCAT*, 14, 1037–1045.

Collignon, A. *et al.* (1995a) Automated multi-modality image registration based on information theory. In Bizais, Y. (ed.), *Proc. Information Processing in Medical Imaging Conf.*, pp. 263–274. Kluwer Academic Publishers, Dordrecht.

Collignon, A., Vandermuelen, D., Suetens, P. and Marchal, G. (1995b) 3d multi-modality medical image registration using feature space clustering. In Ayache, N. (ed.), *Computer Vision, Virtual Reality and Robotics in Medicine*, pp.195–204. Springer Verlag, Berlin.

Duda, R. and Hart, P. (1973) *Pattern Classification and Scene Analysis*. John Wiley and Sons, New York.

Grimson, W. *et al*. (1994) An automatic registration method for frameless stereotaxy, image guided surgery, and enhanced reality visualization. In *Proc. Computer Society Conf. on Computer Vision and Pattern Recognition (Seattle, WA)*. IEEE.

Hill, D., Studholme, C. and Hawkes, D. (1994) Voxel Similarity Measures for Automated Image Registration. In *Proc. Third Conf. Visualization in Biomedical Computing*, pp. 205–216. SPIE.

Haykin, S. (1994) *Neural Networks: A Comprehensive Foundation*. Macmillan College Publishing.

Jiang, H., Robb, R. and Holton, K. (1992) A new approach to 3-D registration of multimodality medical images by surface matching. In *Visualization in Biomedical Computing*, pp. 196–213. SPIE.

Kikinis, R. *et al.* (1996) Computer assisted interactive three-dimensional planning for neurosurgical procedures. *Neurosurgery*, in press.

Linsker, R. (1986) From basic network principles to neural architecture. *Proc. Natl Acad. Sci. USA*, 83, 7508–7512, 8390–8394, 8779–8783.

Ljung, L. and Söderström, T. (1983) *Theory and Practice of*

Recursive Identification. MIT Press, Cambridge, MA.

Malandain, G., Fernandez-Vidal, S. and Rocchisani, J. (1995) 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, France.

Papoulis, A. (1991) *Probability, Random Variables, and Stochastic Processes* (3rd edn) McGraw-Hill, Inc.

Paul, R. (1981) *Robot Manipulators: Mathematics, Programming, and Control*. MIT Press, Cambridge, MA.

Pelizzari, C., Chen, G., Spelbring, D., Weichselbaum, R. and Chen, C. (1989) Accurate three dimensional registration of CT, PET and/or MR Images of the brain. *J. Comp. Assis. Tomogr.*, 13, 20–26.

Schenk, J. *et al.* (1995) Superconducting open-configuration MR imaging system for image-guided therapy. *Radiology*, 195, 805–814.

Studholme, C., Hill, D. and Hawkes, D. (1995a) Automated 3D registration of truncated MR and CT images of the head. In Pycock, D. (ed.), *Proc. British Machine Vision Conf. (BMVC'95)*, pp. 27–36. British Machine Vision Association.

Studholme, C., Hill, D., and Hawkes, D. (1995b) Multiresolution voxel similarity measures for mr-pet registration. In Bizais, Y. (ed.), *Proc. Information Processing in Medical Imaging Conf.*, Kluwer Academic Publishers, Dordrecht.

Van den Elsen, P. (1993) *Multimodality Matching of Brain Images*. *Ph.D. Thesis*, Utrecht University.

Viola, P. (1995) *Alignment by Maximization of Mutual Information. Ph.D. Thesis*, Massachusetts Institute of Technology.

Viola, P. and Wells, W. (1995) Alignment by maximization of mutual information. In *Proc. 5th Int. Conf. Computer Vision*, Boston, IEEE.

Wells, W., Viola, P. and Kikinis, R. (1995) Multi-modal volume registration by maximization of mutual information. In *Proc. Second Int. Symp. on Medical Robotics and Computer Assisted Surgery*, pp. 55–62. John Wiley and Sons, New York.

Widrow, B. and Hoff, M. (1960) Adaptive switching circuits. In *1960 IRE WESCON Convention Record*, vol. 4, pp. 96–104. IRE, New York.

Woods, R., Mazziotta, J. and Cherry, S. (1993) MRI-PET registration with automated algorithm. *J. Comp. Assis. Tomogr.*, 17, 536–546.