Optimisation in Robust Linear Registration of Brain Images

FMRIB Technical Report TR00MJ2
(A related paper has been submitted to Medical Image Analysis)

Mark Jenkinson and Stephen Smith

Oxford Centre for Functional Magnetic Resonance Imaging of the Brain (FMRIB),
Department of Clinical Neurology, University of Oxford, John Radcliffe Hospital,
Headley Way, Headington, Oxford, UK

Abstract

Registration is an important component of medical image analysis and for analysing large amounts of data it is desirable to have fully automatic registration methods. Many different automatic registration methods have been proposed to date, and almost all share a common mathematical framework — one of optimising a cost function. To date little attention has been focused on the optimisation method itself, as opposed to other aspects of the problem like defining suitable cost functions. However, the success of most registration methods hinges on optimisation method. This report examines the assumptions underlying the registration problem and shows that the use of local optimisation methods together with the standard multi-resolution approach is *not* sufficient to reliably find the global minimum. In addition, a global optimisation method is proposed that is specifically tailored to this registration problem. A full discussion of all the necessary implementation details is included as this forms an important aspect of any practical method. Furthermore, results are presented that show that the proposed method is more reliable at finding the global minimum than several of the currently available registration packages in common usage.

Keywords: affine transformation, global optimisation, multimodal registration, multi-resolution search, robustness

1 Introduction

Registration is an important component in many medical image analysis applications. It is used in motion correction, multi-modal fusion, mapping to Talairach space and many other tasks. Furthermore, when analysing large quantities of data, such as in a clinical study or within a busy imaging unit, it is desirable to have fully automatic registration methods. Such methods offer reliability and repeatability as well as minimising user interaction.

A standard method of solving the registration problem is to treat it as a mathematical optimisation, using a cost (or similarity) function to quantify the quality of the alignment of the two images for any given transformation. In practice, this formulation relies on the use of a global optimisation method. This optimisation method is crucial for obtaining good registrations. However, to date most attention has been focused on other aspects of the problem, such as defining cost functions, rather than on the optimisation method.

Most of the mathematical optimisation methods that exist are only suitable for local optimisation and therefore will not find the global solution in general. These methods include gradient descent, Powell's method, simplex methods and so on (Press et al., 1995). Furthermore, the global optimisation methods that do exist often require a very large number of iterations to satisfy convergence criteria (Geman and Geman, 1984; Ingber, 1989), which is impractical for volumetric registration where evaluating the cost function for many values of registration parameters is particularly time consuming.

In an attempt to solve the global optimisation problem in a reasonable amount of time, many registration methods rely on a multi-resolution approach. This is hoped to enable local optimisation methods to reach the global minimum. The assumption is that it is easier to find the global minimum when using large sub-sampling, since coarser transformation steps can be taken, which should reduce the chance of there being a local minimum between the initial starting position and the global minimum. This global minimum is then tracked across different resolutions by successive local optimisations at a series of sub-samplings.

This report examines the assumptions underlying the registration problem (sections 2 and 3) and shows that the use of local optimisation methods together with the standard multi-resolution approach is *not* sufficient to reliably find the global minimum. In addition, a global optimisation method is proposed (section 4) that is specifically tailored to this registration problem, together with a full discussion of all the necessary implementation details (section 5). Finally, results are presented (section 6) that show that the proposed method is more reliable at finding the global minimum than several of the currently available registration packages in common usage.

2 Mathematical Formulation

The standard registration problem is to find a transformation that best aligns a reference image I^r to another (floating) image I^f. In this context, an image defines the intensity for each spatial location. That is, it is a mapping of positions to intensity values; $I : \mathcal{R}^3 \rightarrow \mathcal{R}$. Given this definition, registration is formulated as a mathematical problem by taking a cost function, $C(I_1, I_2)$, that quantifies the quality of the registration and then finding the transformation T^* which gives the minimum cost:

$$T^* = \arg \min_{T \in S_T} C(I^r, I^f \circ T) \tag{1}$$

where S_T is the space of allowable transformations.

2.1 Transformation Space

It is necessary when beginning to formulate the registration problem to decide what will be the space of allowable transformations, S_T. The most general class of transformations are the local, non-linear transformations. These allow each voxel to be moved separately, leading to potentially millions of Degrees Of Freedom (DOF), although in practice, some constraints are necessary such as requiring that topology be preserved.

One basic class of transformations are the linear transformations where all voxels are constrained to move according to a global, linear relationship. That is,

$$\begin{bmatrix} x' \\ y' \\ z' \\ 1 \end{bmatrix} = \begin{bmatrix} A_{11} & A_{12} & A_{13} & A_{14} \\ A_{21} & A_{22} & A_{23} & A_{24} \\ A_{31} & A_{32} & A_{33} & A_{34} \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \\ 1 \end{bmatrix} \tag{2}$$

or $X' = AX$ using homogeneous coordinates.

The most general of these transformations is the affine transformation with 12 DOF ($A_{11}, \ldots, A_{34}$), which includes rigid body (6 DOF) and similarity (7 DOF) transformations as particular cases. These transformations are also interesting physically because they represent the transformations that physical, rigid objects undergo. However, even for these simple linear cases, with a maximum of 12 DOF for the entire volume, the registration problem is still difficult. Therefore this report examines these affine transformations because they are the simplest, most common transformations used and, furthermore, many non-linear methods rely on having an initial linear fit as a preprocessing step.

2.2 Interpolation

For discrete data, the intensity is normally only defined on a grid, G, of discrete locations or lattice sites. That is, the data is stored as $I_{ijk} = I(x_i, y_j, z_k)$ where I_{ijk} represents the discrete data and $(x_i, y_j, z_k) \in G$ the coordinates of the lattice sites , with $I(\cdot)$ the underlying but unobservable continuous image. However, when the lattices for I^r and I^f are not perfectly aligned, it is necessary to evaluate the intensity at points in between the lattice sites. This is common because virtually any transformation that is applied to I^f will cause the lattices to be out of alignment.

To evaluate the intensity at intermediate locations requires interpolation. The interpolation can be viewed as reconstructing a full continuous image from the discrete points, although to evaluate the cost function it is usually only necessary to know the intensity at the corresponding lattice sites. That is, if (x_i, y_j, z_k) represent the lattice sites for I^r then it is usually only necessary to know the value of $I^f(T(x_i, y_j, z_k))$. Typically, interpolation methods are based on a convolution of the discrete data with some continuous kernel such as trilinear, spline and (windowed) sinc kernels.

One major effect that the choice of interpolation has is to what degree the cost function becomes continuous or discontinuous. This is also affected by the boundary conditions used, such as padding with zeros or only

using the overlapping volume. Studying the precise effects of interpolation is an active research area (Hajnal et al., 1995; Pluim et al., 2000; Thacker et al., 1999) but is beyond the scope of this report. Therefore, in this report trilinear interpolation is used on the overlapping volume. These choices require no additional parameters to be set and were motivated largely by experience.

2.3 Cost Functions

Cost functions can use either direct intensity information or landmark information. For each type of cost function there are two main categories: intra-modal (for example, sum of absolute differences) and inter-modal (for example, mutual information). Not surprisingly, it is significantly harder to find functions with desirable properties for inter-modal registration and consequently much research has been concerned with finding such functions. For this report it is the intensity-based inter-modal cost functions that are investigated: for example, Mutual Information, Correlation Ratio, *etc.* These are more difficult to optimise, since there are many non-linear, potentially discontinuous terms involved that result in functions that are less smooth and regular. In general, though, all cost functions (except some landmark-based cost functions which are convex) require global optimisation.

2.4 Optimisation

The theoretical registration problem, as posed above, is fully specified by a transformation space, an interpolation method and a cost function. However, in practice, an optimisation method is required to find the transformation that minimises the cost function. This, therefore, is an implementational issue rather than one that relates purely to the theory. Furthermore, it is necessary that the method be a *global* optimisation method rather than one of the more common *local* methods (for example, Powell's Method, Gradient Descent, *etc.*).

Although there do exist general global optimisation methods (Simulated Annealing being a notable example), it has been shown (in the No Free Lunch theorem (Wolpert and Macready, 1996)) that there is no general method that is superior for all problems (when taken on average). Therefore the method used should ideally be tuned for the particular problem at hand — in this case, volumetric registration.

The performance of the optimisation method is important since evaluating the cost function requires a large amount of computation. This, together with the general combinatorial explosion which occurs for higher dimensional spaces (for example, $\mathcal{R}^{12}$), makes any simple search algorithm impractical. In addition, it is also difficult to satisfy the convergence criteria in reasonable amounts of time for statistical optimisation methods such as Simulated Annealing (even in the fast version (Ingber, 1989)). Consequently, a compromise is made by shortening the schedule in order to achieve a solution within the time available.

To overcome these problems, one common tactic is to adopt a multi-resolution approach in conjunction with a local optimisation method (Woods et al., 1993; Studholme et al., 1996). The multi-resolution approach involves computing the cost function for various sub-samplings of the volumes. For instance, let I_n^r denote the reference volume sampled with cubic voxels of side length n mm, and similarly for the floating volume. This creates a set of scales at which the problem can be solved. That is,

$$T_n^* = \arg\min_{T \in S_T} C(I_n^r, I_n^f \circ T). \tag{3}$$

Then, an initial solution is found for large n, and progressively refined by lowering n since

$$T^* = \lim_{n \to 0} T_n^*. \tag{4}$$

Furthermore, denote the local optimisation as a function that takes an initial transformation and returns a new transformation, which is an estimate of a nearby local minima. That is, $\mathrm{Op}_n : S_T \to S_T$ so that

$$T_{\mathrm{new}} = \mathrm{Op}_n(T_{\mathrm{old}}). \tag{5}$$

Note that the optimisation function also depends on the sub-sampling, n, via the cost function since this is based on the sub-sampled images and so changes with the sub-sampling. Therefore, the multi-resolution approach begins with an initial transformation estimate, T_0, usually taken to be the identity transformation, and proceeds iteratively as:

$$\begin{aligned} T_1 &= \mathrm{Op}_{n_0}(T_0) \\ T_2 &= \mathrm{Op}_{n_1}(T_1) \\ &\vdots \quad\quad \vdots \\ T_P &= \mathrm{Op}_{n_{P-1}}(T_{P-1}) \end{aligned} \tag{6}$$

where a typical set of values would be $P = 4$ with $n_0 = 8$, $n_1 = 4$, $n_2 = 2$, $n_3 = 1$ in mm.

The advantage of this multi-resolution approach is that the initial optimisation, at large n, has a dramatically reduced computational load, since the number of sample points is substantially less. In addition, for large sub-samplings it should be the gross features of the image which dominate, and so the overall alignment should be easier to find. This belief is equivalent to saying that there are less local minima for the optimisation method to get caught in, although, as shown in the next section, this is not necessarily true.

2.5 Difficulties

The standard formulation described above is based, in most cases, on the following assumptions:

1. the location of the global minimum of the cost function, T^*, corresponds to the desired solution,
2. at the maximum sub-sampling $n = n_0$, the global minimum is the nearest minimum to the starting position and can be found by local optimisation: $T^*_{n_0} \approx \mathrm{Op}_{n_0}(T_0)$,
3. the location of the global minimum found using one sub-sampling, n_1 is inside the basin of attraction (as defined by the optimisation method) of the global minimum for the next sub-sampling, n_2: $T^*_{n_2} \approx \mathrm{Op}_{n_2}(T_{n_1})$.

Here the basin of attraction for some minimum is defined as the set of initial transformations that, after successive applications of the local optimisation algorithm, converge to that minimum. That is: $B(T) = \{ T_0 \mid \mathrm{Op}_n^M(T_0) \to T \text{ as } M \to \infty \}$. This set depends on the precise optimisation algorithm used and should be empty for any T that is not a minimum. Furthermore, such sets have very complicated boundaries, often with a fractal nature.

However, these assumptions do not always hold. The following describes some cases where they are not true.

- If the cost function for some extreme transformation gives a low value then the global minimum will be (at least degenerately) given by this limiting case. For example, large scalings can create low cost values even though the registration is poor. Furthermore, limiting the domain is not a general solution to this as it would be necessary to guarantee that the cost at the edge of the domain was higher than the global minimum value which is unknown.
- Sub-sampling to lower resolutions may not reduce the number of local minima sufficiently. In fact, some work on interpolation (Pluim et al., 2000) has shown how it can actually create additional local minima.
- Minima will move in scale-space, which is a well known phenomenon, so that the location of a minimum for some sub-sampling may fall inside the basin of attraction of a different minimum for another sub-sampling.

To develop a reliable, automatic registration method it is necessary to examine these assumptions and tailor the method to this problem. In order to do this the characteristics of a typical cost function needs to be understood, and these are examined in the next section.

3 Characterisation of Cost Functions

3.1 Computational Complexity

In order to understand the computational difficulty of the problem it is helpful to consider the amount of calculations required. For instance, take a typical field of view (FOV) for the brain as 200mm cubed, and let each voxel be a cube of length d mm. Therefore, there are $(200/d)^3$ voxels in the volume.

To interpolate this volume requires, generally, one evaluation of the interpolation function (acting on the floating image) for each voxel location in the reference image. For trilinear interpolation there are at least 8 floating point operations required for each interpolation evaluation. Therefore, the total cost of interpolation is at least $8 \times (200/d)^3$ operations.

In addition, to calculate the cost function typically requires at least one operation (to estimate variance) per voxel, making the number of operations required per voxel at least 9. Therefore, on a machine that could achieve 500 MFLOPS, the time required for one cost function evaluation, with voxel size of 1mm would be at least $9 \times 200^3/(5 \times 10^8) = 0.144$ seconds. However, in practice there are some additional calculations required as well as memory access overheads, and so our implementation (on a Pentium3 500MHz) takes approximately 1 second for this evaluation.

Now consider implementing a search through the transformation space by constructing a grid with N different values for each parameter, so that there are a total of N^D elements (where D is the Degree Of Freedom — 12 for

N	D	d	No. of Operations	Time (sec)
10	12	1	7.2×10^{19}	1.44×10^{11}
10	12	2	9.0×10^{18}	1.8×10^{10}
10	12	4	1.13×10^{18}	2.3×10^{9}
10	12	8	1.4×10^{17}	2.8×10^{8}
10	6	1	7.2×10^{13}	1.44×10^{5}
10	6	2	9.0×10^{12}	18000
10	6	4	1.13×10^{12}	2250
10	6	8	1.4×10^{15}	281
3	12	1	3.8×10^{13}	76528
3	12	2	4.8×10^{12}	9566
3	12	4	6.0×10^{11}	1196
3	12	8	7.5×10^{10}	149

Table 1: Some typical computational requirements for implementing a simple grid-based search strategy (see text for explanation of terms). The times are based on an idealised 500 MFLOP performance with no memory overhead. In practice our implementation was approximately 7 times slower than this.

affine). The total number of calculations required to evaluate the cost function at each point is $7.2 \times 10^7 (N^D/d^3)$ which would ideally take $0.144 \times (N^D/d^3)$ seconds at 500 MFLOPS. Table 1 shows the times required for a range of moderate parameters. From this table it can be seen that evaluations for 12 DOF using 1mm voxels are prohibitive, taking over 4000 years for a grid with $N = 10$ and over 21 hours for a simple neighbourhood $N = 3$. However, using 8mm cubed voxels improves the execution time although the biggest improvement comes from restricting the dimensionality of the transformation space (that is, the DOF).

For a simple, but usable, search routine it would be necessary to search with a grid of at least $N = 100$, in which case even for 6 DOF and 8mm voxels the time taken would exceed 7 hours. This would be unacceptable for most users, especially given that our implementation would be likely to take over 50 hours. However, in combination with a local optimisation method, a coarse search method using 8mm voxels can be used to create a useful optimisation method, which is pursued in section 4.3.

3.2 Asymptotic Behaviour

Consider two types of extreme (or asymptotic) registrations:

- large translations — such that the volume overlap is minimal.
- large scale disparity — such that a small portion of the floating volume is stretched to cover the entire reference volume.

Both of these cases represent poor registrations and it is therefore important that the cost values associated with them are high. If this is not the case then the global minimum may be associated with these transformations rather than the desired one, violating the first assumption outlined in section 2.5. Furthermore, limiting the domain to exclude these transformations will only eliminate the problem if the cost at the domain boundary is guaranteed to be larger than the global minimum, which is generally difficult or impossible to do as the value at the global minimum is unknown.

In examining the asymptotic behaviour it is important to define what the boundary conditions are. That is, what is done for points in space which do not lie in one or other of the volume domains. One option is to (conceptually) pad the volumes with zeros to give them infinite domain. However, this creates artificial intensity boundaries when the FOV only includes part of the head (for example, when the top few mm of the head/brain are not included the intensity suddenly drops from that of brain matter to zero) — which is relatively common. These artificial boundaries would then bias the registration, which is undesirable. Therefore, a better alternative is to only calculate the cost function for the region where the volumes overlap. This usually requires some extra calculations in practice, as the normalisation now depends on the overlap volume, which depends on the transformation, but the resulting registration is, in theory, unbiased.

The cost functions that will be compared here are: the Woods function (Woods et al., 1993), Correlation Ratio (Roche et al., 1998), Joint Entropy (Studholme et al., 1995; Collignon et al., 1995), Mutual Information (Viola and Wells, 1997; Maes et al., 1997), and Normalised Mutual Information (Studholme et al., 1999;

Maes, 1998). Denoting the two images by X (typically I^r) and Y (typically $I^f \circ T$), the respective *cost* functions[1] are defined as:

$$C^W = \sum_i \frac{n_i}{N} \frac{\sqrt{\mathrm{Var}(Y_i)}}{\mu(Y_i)} \tag{7}$$

$$C^{CR} = \frac{1}{\mathrm{Var}(Y)} \sum_i \frac{n_i}{N} \mathrm{Var}(Y_i) \tag{8}$$

$$C^{JE} = H(X,Y) \tag{9}$$

$$C^{MI} = H(X,Y) - H(X) - H(Y) \tag{10}$$

$$C^{NMI} = \frac{H(X,Y)}{H(X)+H(Y)}. \tag{11}$$

Here the quantities X and Y represent the images as the set of intensities evaluated at the discrete, valid grid points. That is, $X = \{X(q) \mid q \in \mathrm{Dom}(X) \cap G\}$ where G represents the discrete grid and $\mathrm{Dom}(X)$ the continuous, valid domain (that is, the FOV). Also:

- $\mu(A)$ is the mean of set A
- $\mathrm{Var}(A)$ is the variance of the set A
- Y_i is the ith iso-set defined by X as $Y_i = \{Y(q) \,|\, I_i < X(q) < I_{i+1},\, q \in \mathrm{Dom}(X) \cap \mathrm{Dom}(Y) \cap G\}$
- $n_i = \mathrm{Card}(Y_i)$ (number of elements in the set Y_i) such that $N = \sum_i n_i$
- $H(X,Y) = -\sum_{ij} p_{ij} \log p_{ij}$ is the standard entropy definition where $p_{ij} = \mathrm{Card}(\{q \,|\, I_i < X(q) < I_{i+1}$ and $I_j < Y(q) < I_{i+1}\})/N$ represents the probability of the (i,j) joint histogram bin, and similarly for the marginals, $H(X)$ and $H(Y)$.

These definition require the specification of a partition of the intensities: $\{I_0, I_1, \ldots, I_M\}$. This partition is used to define the various histograms and iso-sets required for the calculation of the cost functions. In particular, a discrete bin (or iso-set) number is calculated for each voxel using the intensity at that voxel. For example, at location q in image X the intensity is $X(q)$ and the bin number is k if $I_k < X(q) < I_{k+1}$. Then, given this bin number, iso-sets or the joint histogram are easily determined.

That is, the kth iso-set, Y_k, is the set of Y intensities where the corresponding voxel in X has a bin number k — the intensity $X(q)$ is between I_k and I_{k+1}. The joint histogram is composed of a number of elements, where the element $b(i,j)$ represents the number of voxels where the intensity of X is in the ith bin, (I_i, I_{i+1}), *and* the intensity of Y for the same (corresponding) voxel is in the jth bin, (I_j, I_{j+1}). The probability associated with this element p_{ij} is then simply the value of the element divided by the sum of all the elements.

Furthermore, for each of these cost functions the range can easily be determined, and is:

$$0 \leq C^W \leq \infty \tag{12}$$

$$0 \leq C^{CR} \leq 1 \tag{13}$$

$$0 \leq C^{JE} \leq \infty \tag{14}$$

$$-\infty \leq C^{MI} \leq 0 \tag{15}$$

$$0 \leq C^{NMI} \leq 1. \tag{16}$$

3.2.1 Large Translation

In this case the intersection of the volume domains becomes small and generally any tissue in one volume corresponds only to background in the other volume. More specifically, consider the case where there is a single, small amount of tissue contained in either volume, with the proportion of voxels containing tissue to total number of voxels in the intersection volume being χ_1 for X and χ_2 for Y (see also figure 1). Furthermore, without loss of generality, take the background intensity as zero and the tissue intensity as B. The cost functions

[1] To form cost functions where low values represent good registrations, many definitions are reversed like $C^{MI} = -$Mutual Information.

are then given by:

$$
\begin{aligned}
C^W &\approx (1-\chi_1)\frac{\sqrt{\frac{\chi_2}{1-\chi_1}B^2 - (\frac{\chi_2}{1-\chi_1}B)^2}}{\frac{\chi_2}{1-\chi_1}B} \\
&\to 0 \quad \text{as} \quad \chi_1 \to 0 \qquad (17)\\
&\to \infty \quad \text{as} \quad \chi_2 \to 0 \qquad (18)\\
C^{CR} &\approx (1-\chi_1)\left(\frac{\frac{\chi_2}{1-\chi_1}B^2 - (\frac{\chi_2}{1-\chi_1}B)^2}{\chi_2 B^2 - \chi_2^2 B^2}\right) \to 1 \\
&\quad \text{as} \quad \chi_1, \chi_2 \to 0 \qquad (19)\\
C^{JE} &= H(X,Y) \approx -\chi_1 \log(\chi_1) - \chi_2 \log(\chi_2) \\
&\quad -(1-\chi_1-\chi_2)\log(1-\chi_1-\chi_2) \to 0 \\
&\quad \text{as} \quad \chi_1, \chi_2 \to 0 \qquad (20)\\
C^{MI} &= H(X,Y) - H(X) - H(Y) \\
&\approx -\chi_1 \log(\chi_1) - \chi_2 \log(\chi_2) \\
&\quad -(1-\chi_1-\chi_2)\log(1-\chi_1-\chi_2) \\
&\quad +\chi_1 \log(\chi_1) + (1-\chi_1)\log(1-\chi_1) \\
&\quad +\chi_2 \log(\chi_2) + (1-\chi_2)\log(1-\chi_2) \to 0 \\
&\quad \text{as} \quad \chi_1, \chi_2 \to 0 \qquad (21)\\
C^{NMI} &= \frac{H(X,Y)}{H(X)+H(Y)} \\
&\to \frac{-\chi_1 \log(\chi_1) - \chi_2 \log(\chi_2)}{-\chi_1 \log(\chi_1) - \chi_2 \log(\chi_2)} = 1 \\
&\quad \text{as} \quad \chi_1, \chi_2 \to 0. \qquad (22)
\end{aligned}
$$

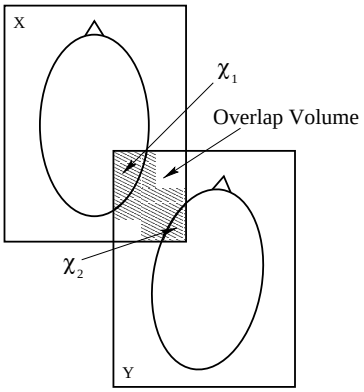


Figure 1: Example of an asymptotically large transformation between images X and Y where the overlap volume becomes small. The proportion of non-background voxels in the overlap volume (shaded) is denoted by χ_1 for X and χ_2 for Y.

Thus the Woods function (as $\chi_1 \to 0$) and Joint Entropy actually approach the minimum values which indicates, erroneously, that the registration is good. The other three cost functions approach their maximum values, indicating poor registrations.

3.2.2 Large Scale Disparity

In this case a small portion of the floating volume is stretched to cover the reference volume.[2] Therefore the floating volume (Y) is approximately constant over the volume of overlap. Consequently:

$$C^W \approx \sum_i \frac{n_i}{N} \frac{\sqrt{\mathrm{Var}(Y_i)}}{\bar{Y}} \to 0 \tag{23}$$

$$C^{CR} \approx \frac{1}{\mathrm{Var}(Y)} \sum_i \frac{n_i}{N} \mathrm{Var}(Y_i)$$

$$\approx \frac{1}{\mathrm{Var}(Y)} \mathrm{Var}(Y) = 1 \tag{24}$$

$$C^{JE} = H(X,Y) \approx H(X) \tag{25}$$

$$C^{MI} \approx H(X) - H(X) - 0 = 0 \tag{26}$$

$$C^{NMI} \approx \frac{H(X)}{H(X)+0} = 1. \tag{27}$$

where the joint histogram becomes compressed into a small horizontal band (minimal variation in Y) and it is assumed that the spatial variation of Y is uncorrelated with the iso-sets of X, such that $\sum_i \frac{n_i}{N} \mathrm{Var}(Y_i) \approx \mathrm{Var}(Y)$.

Once again, the Woods function and Joint Entropy give low cost values indicating a good registration, and so are likely to violate the assumption (stated in section 2.5) that the global minimum is the desired solution. Consequently, these functions should not be used. Note that although the Joint Entropy does not approach zero, the value is $H(X) = H(X,X)$ which is the same as it would be for a *perfect* registration of the volume with itself.

Overall, only the Correlation Ratio, Mutual Information and Normalised Mutual Information display the correct asymptotic behaviour and therefore these are the only suitable cost functions (from this selection) to use.

3.3 Extent of Minima

The number and location of local minima for a given cost function can only be determined empirically, as otherwise the registration problem could be solved analytically. However, the typical extent of local minima, the size of their basins of attraction and their inherent smoothness are important for designing effective optimisation strategies.

One approach for examining the cost function is to look at how it changes as individual parameter values vary about some fixed point. That is, to plot the function values versus one parameter value, with all other parameters held constant. This also avoids the problem of trying to visualise functions in high-dimensional spaces (for example, $\mathcal{R}^{12}$ for affine transformations).

As it is the global minimum that is of most interest, the fixed point is chosen to be the global minimum itself. Figure 2 shows the cost function plots for each parameter (3 rotation angles, 3 translations, 3 scalings and 3 skews) using the images shown in figure 3. This image pairing was chosen as it had been found to be a difficult pair to register successfully. However, despite this fact, the global minimum is still quite broad and well defined.

The plots shown in figure 2 give an idea of the gross behaviour of the cost function about the global minimum. However, on a small scale there exist many small, local minima, as shown in figure 4a. In addition, there also exist larger local minima as shown in figure 4b, where the central point is no longer the global minimum, but a transformation more typical of an initial alignment. Consequently a successful optimisation method must be able to cope with both small, densely packed local minima and much larger, and widely separated, local minima. These problems are respectively dealt with in section 4 by (1) an adaptive step local minimisation method (section 4.2) and (2) an initial search stage (section 4.3).

3.4 Local Continuity

Ideally the cost function should be continuous with respect to the transformation. However, by working with discrete data (to reduce computational load) some discontinuities are usually introduced, as can be seen in figure 4a. The exact nature of the discontinuities depend on both the data and the interpolation method. Such

[2] Similar results are obtained when the floating volume is compressed, although the analysis is slightly different as the overlap volume varies with scale.

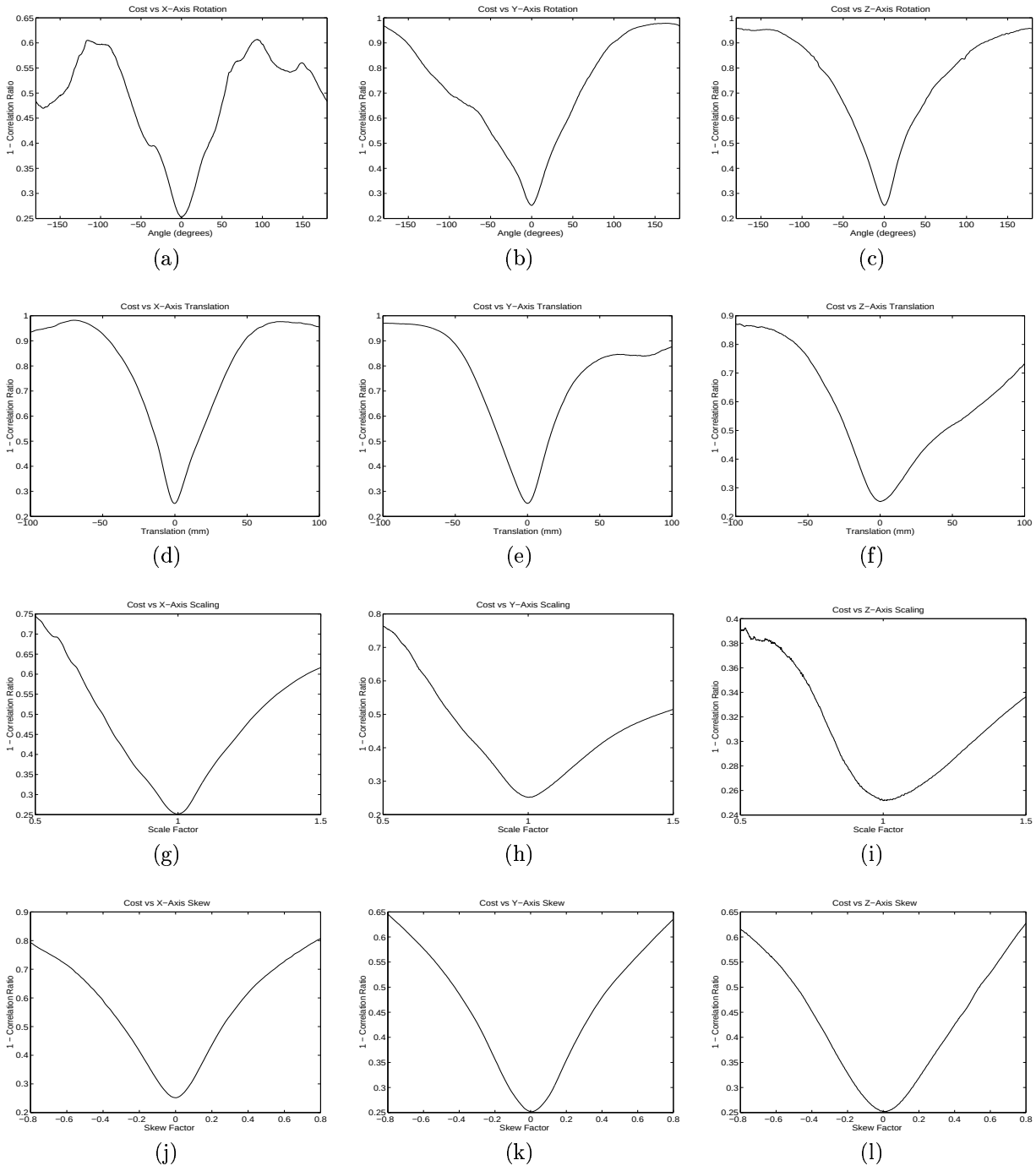
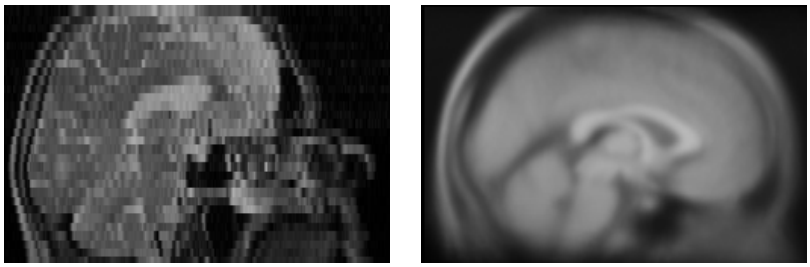
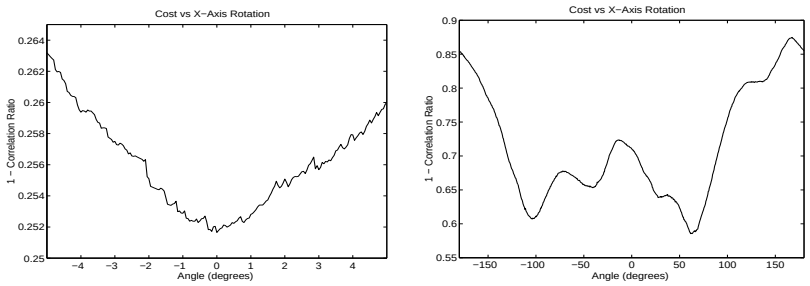


Figure 2: Plots of the Correlation Ratio cost function versus individual parameter values. In each plot, a single parameter is varied over a large range while the others are kept fixed. The central point about which the parameters varied was the global minimum. All cost function calculations were done using the image pair shown in figure 3 with a resampling (including appropriate pre-blurring) to cubic voxels with 8mm side-length.



(a) (b)

Figure 3: Example slices from the volumes used for plotting the cost function. Figure (a) shows a T2 weighted image of a subject while figure (b) shows the MNI 305 image (Collins et al., 1994). The voxel dimensions for these images were $0.93 \times 0.93 \times 5$ mm for (a), and $1 \times 1 \times 1$ mm for (b).



(a) (b)

Figure 4: Plots of the cost function versus rotation about the x-axis, showing the presence of local minima. Figure (a) shows an expansion of the plot shown in figure 2a, over a small range about the global minimum. It can be seen that several small, local minima exist, generated by small fluctuations in the cost function. In figure (b) the cost function is plotted about a different central point — one that is offset from the global minimum. This produces significant changes in the cost function, showing the presence of large local minima.

discontinuities can be reduced by methods such as Partial Volume Interpolation (Maes et al., 1997), but these are not well adapted to certain transformations such as scaling changes.

Although it should be possible to place bounds on the difference between discrete cost evaluations for small changes in transformation, it is a difficult analysis and is left as a topic for further investigation. Moreover, in practice, the optimisation methods usually converge satisfactorily and simple re-samplings can avoid problems associated with grid alignment (Pluim et al., 2000).

4 Global Optimisation Method

This section presents a global optimisation method that is specifically tailored for volume registration of brain images. It uses the Correlation Ratio as a cost function and trilinear interpolation, but is general and can be easily adapted to most cost functions, provided that they have the correct asymptotic behaviour.

4.1 Overview

The fundamental idea of this method is to combine a local optimisation method (Powell's (Press et al., 1995)) with an initial search. Furthermore, the search is tuned so that it is computationally feasible. In particular, our implementation of the whole optimisation method executes in approximately 40 minutes on a modern PC.

There are four different sub-sampling levels that are used by the optimisation algorithm: $n_0 = 8, n_1 = 4, n_2 = 2, n_3 = 1$. Initially, 8mm cubed voxels are used and a full search is conducted over the rotation angles. Following this, various local optimisation are performed with a variety of starting points in the local neighbourhood of the best points identified in the search. These local optimisations are done using 4mm, 2mm and finally 1mm cubed voxels. Each of these stages is described below in more detail.

4.2 Resolution Adapted Powell's Method

A major component in the overall optimisation process is the local optimisation method. This local optimisation is called repeatedly and so should be efficient. Here Powell's method was chosen since it is a commonly used and efficient local optimisation method (Press et al., 1995).

The changes in transformation used by the optimisation method should also be adapted to the resolution of the volumes. For example, computing very small changes in transformation when the voxel size is large corresponds to very small sub-voxel changes and is a waste of computation. Consequently, a lower limit for the transformation step size is enforced. This step size is easily calculated by imposing the condition that the maximum voxel position shift should be less than half a voxel. More precisely, with a brain of radius R, and an origin located at the centre of mass for the brain, the maximum shifts are:

$$\text{Translation}: \quad \Delta x_{max} = \Delta t \tag{28}$$

$$\text{Rotation}: \quad \Delta x_{max} = 2R\cos\left(\frac{\Delta\theta}{2}\right) \tag{29}$$

$$\text{Scale}: \quad \Delta x_{max} = R|\Delta s| \tag{30}$$

$$\text{Skew}: \quad \Delta x_{max} = R|\Delta k|. \tag{31}$$

For cubic voxels of side-length n mm the constraint becomes $x_{max} \leq \frac{n}{2}$ which gives:

$$\Delta t \leq \frac{n}{2} \tag{32}$$

$$\Delta\theta \leq \frac{n}{2R} \tag{33}$$

$$\Delta s \leq \frac{n}{2R} \tag{34}$$

$$\Delta k \leq \frac{n}{2R}. \tag{35}$$

For $n = 1$ mm this gives $\Delta t \leq 0.5$ mm, $\Delta\theta \leq 0.3°$, $\Delta s \leq 0.005$ and $\Delta k \leq 0.005$. However, since the voxel shift due to translation is constant across the volume, whereas all others are less towards the centre, it pays to be more conservative for Δt.

Given these lower limits on the parameter steps, Powell's algorithm is modified by changing the termination conditions so that when the 1D optimisation (Brent's method) has bounded the minimum within an interval that is less than one parameter step, then it returns the mid-point of the current interval. In this way, significant savings in computation can be had without any sacrifice in accuracy. Moreover, accuracy is largely determined on the final pass, which can have even more conservative bounds if better accuracy is required.

4.3 Initial Search

As shown in section 3.1, the amount of time required to perform all but the most perfunctory search is prohibitive with any sub-sampling where the voxels are smaller than 8mm cubed. However, in order to find the global minima reliably, some search must be performed.

To allow a useful search to be conducted in under 20 minutes, a multi-stage search was devised. This search is based on the observation that finding the correct orientation, or rotation, is the most difficult task since the three rotation parameters are highly coupled and most erroneous registrations have an incorrect orientation. Therefore, the search concentrates on the rotational part of the transformation space. The outline is:

1. Initially form a coarse grid over the 3 rotation parameters (Euler angles), with M angles per dimension (total of M^3 grid points).
2. At each point in the grid, perform a 4 DOF *local optimisation* to find the translation and (global) scale — using initial values of translation = 0 and scale = 1.
3. Calculate the median scaling from the optimised results.
4. Form a finer grid over the rotation angles, with N angles per dimension.
5. Provide initial parameter estimates for translation at each fine grid point by interpolating the optimised translation values found at the coarse grid points.
6. Evaluate the cost function (no optimisation) at all points on the fine grid with initial scale set to the previously calculated median scale.
7. Find all points that have lower cost than their neighbours (local minima) and perform a 7 DOF optimisation for each of these points, storing the results of the transformations (and costs) both before and after optimisation.

Ideally, by choosing a large value for M, the initial stage (steps 1 and 2) of the search would be sufficient. However, as it typically takes about 2 seconds to perform each 4 DOF optimisation in our implementation, only a small value for M can be chosen if the search is to be carried out in a limited amount of time. Therefore another two stages are introduced into the search to allow a finer grid ($N > M$) to be used.

To find the cost at each point on a finer grid (where $N > M$) only a single evaluation, rather than an optimisation, is done. However, in order that the cost value is close to the optimal value it is necessary to ensure than they are not significantly influenced by translation and scale differences. In order to achieve this, good initial estimates for these parameters are required. The initial estimates used are obtained from the parameters stored after optimisation on the coarse grid. Since the translation is strongly coupled to the rotation (that is, the correct translation is significantly different for different rotation values) the initial translation parameters are obtained by interpolating the values from the coarse grid. However, the scaling parameter is not strongly coupled with rotation, and so a single, robust estimate (the median) is taken from all the coarse grid values.

In practice, our implementation takes about 2 seconds for each 4 DOF optimisation, since the extra overhead is proportionally higher at the 8mm resolution. Therefore, even lower values for M and N are required than the values given in section 3.1 would suggest. However, it has been found that values as low as $M = 6$ and $N = 20$ still give good results.

This method relies on certain assumptions, namely:

1. at least one point will occur in the global minimum's basin of attraction, and be lower than its neighbours.
2. the initial estimates for translation and scale, provided by the coarse grid 4 DOF optimisation, are reasonable.
3. the typical anisotropic scaling and skews are sub-voxel with 8mm cubed voxels, and can be ignored.

The first two assumptions can be partly justified by the results shown in section 3.3. Firstly, it can be seen that the typical size of the global minimum's basin of attraction is large enough that a suitably fine grid should manage to have at least one point within the basin of attraction. Secondly, the cost varies more slowly, and more smoothly, with translation and scale than it does with rotation. Consequently, it is more likely that the change in rotation between grid points will be the dominant cause of changes in the cost function, rather than small errors in the initial estimation of scale and translation.

4.3.1 Re-optimising with Perturbations

Following the search with 8mm cubed voxels, a progressive refinement is required, starting with 4mm voxel size.

Since the 8mm solution may not lie inside the 4mm basin of attraction, a 7 DOF optimisation is run at 4mm on the optimised solution found in 8mm. Furthermore, several perturbations of the *un-optimised* 8mm solution are also treated similarly. These perturbations attempt to correct for typical registration errors, with the perturbations being: $\pm\frac{1}{2}\Delta\theta_{fine}$ in each rotation angle and $\pm\Delta s, \pm 2\Delta s$ in scale, giving a total of 10 different perturbations, where $\Delta\theta$ and Δs are the step sizes evaluated for the 8mm resolution. It has been found that this simple, relatively inexpensive procedure, corrects for the majority of mis-registrations.

It has also been found that, on occasion, the relative rankings of corresponding minima change between sub-samplings. This is especially true when the cost values are very similar for two competing minima. Therefore, the global minimum for the 4mm case may not correspond to the global minimum for the 8mm case. To correct for this change in minima ranking, the best 3 solutions from the 8mm case are used, rather than just the single best solution. This has been found, empirically, to compensate for the next major case of mis-registration.

Overall, therefore, a total of 33 optimisations with 7 DOF are performed at the 4mm resolution. This acts as a local search about each of the most promising local minima found by the previous (8mm) stage. The time taken to complete this stage is typically less than 10 minutes but it is a critical stage in eliminating many common mis-registrations.

4.3.2 Refining the Transformation

Once the solution has been found for the 4mm resolution, it is assumed that this corresponds to the desired solution and simple re-optimisation occurs at the 2mm and 1mm resolutions.

For the 2mm case, the optimisation is initially run with 7 then 9 then 12 DOF in order to iteratively improve the fit. Note that it is only at this resolution that anisotropic scalings and skews are included in the optimisation.

In the 1mm case, the computational burden is high, taking over 3 seconds for each cost function evaluation (compare this to around 50ms for a single evaluation at 8mm). Therefore, only a single optimisation pass with 12 DOF is run, in order to make minor adjustments in the transformation. Even so, this stage typically takes 10 minutes to run which is approximately 25% of the entire execution time.

4.3.3 Summary

No guarantee of finding the global minimum is available using this method. However, this is true for most global optimisation methods which at most provide only statistical guarantees for convergence which can never be met in practice (as they require infinite time in theory). The timing for this method, on the other hand, is more or less guaranteed to be less than an hour for our implementation.

Some assumptions in the standard multi-resolution approach have been relaxed here, while others are retained. In particular, the assumptions that are being used are:

- The global minimum for the 1mm resolution represents the desired solution.
- Significant change in the position of corresponding minima only occurs between the 8mm and 4mm resolutions.
- Anisotropic scaling and skew are minimal and can be ignored for the 8mm and 4mm resolutions.
- The search grid will evaluate one point in the global minimum's basin of attraction for the 8mm resolution and that this will be a local minima in the fine grid.
- Initial estimates of the translation and scale, as found from the optimised transformations at each point in the coarse grid, are sufficiently accurate for fine grid cost evaluations.

Of these, some assumptions can be further relaxed, but at the expense of computation time which can easily become excessive.

Much is based on empirical observations; however, the performance will not be worse than using the standard (no search) multi-resolution approach as this is equivalent to having $M = N = 1$ and using no perturbations.

5 Implementation

In almost all methods, there are several choices which need to be made at the implementation stage. These choices are important and a re-implementation of a method will be unlikely to give similar results unless these details are treated the same way.

5.1 Parameterisation

The first implementation choice is the way in which the transformation is represented. A simple, but poor, choice is to use one parameter for each entry in the first three rows of the 4×4 affine matrix. However, this highly couples the parameters which makes it very difficult for Powell's method to work efficiently. In addition, the matrix elements themselves are not particularly meaningful. Therefore, the best parameterisation is one that decouples the parameters in a sensible way.

One standard decomposition of an affine matrix, which decouples the parameters, is to split the transformation into 3 rotation, 3 translation, 3 scale and 3 skew parameters. That is:

$$\begin{bmatrix} R_{11} & R_{12} & R_{13} & t_x \\ R_{21} & R_{22} & R_{23} & t_y \\ R_{31} & R_{32} & R_{33} & t_z \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} s_x & 0 & 0 & 0 \\ 0 & s_y & 0 & 0 \\ 0 & 0 & s_z & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & k_1 & k_2 & 0 \\ 0 & 1 & k_3 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{36}$$

where the product is the 4×4 affine matrix, R_{ij} represents the 3×3 rotation matrix, t_i represents a 3×1 translation vector, s_i are the scale parameters, and k_i are the skew parameters. It is also necessary to further decompose the rotation matrix into three independent parameters such as Euler angles which represent a rotation by three independent rotations, each about one of the coordinate axes: $R = R(\theta_x)R(\theta_y)R(\theta_z)$.

This fully accounts for the 12 Degrees Of Freedom and allows simple restrictions to be placed on the parameters, such as $s_x = s_y = s_z$. Consequently, the reduced 9, 7 and 4 DOF optimisations required by the method outlined in the previous section are straightforward to implement.

These parameters, however, are still not fully specified as the decomposition can be performed about any given origin. This origin functions as the centre of rotations, scales and skews, and affects the coupling between these parameters and the translation parameters. By choosing the origin to be the *Centre of Mass* for the volume, the coupling is minimised in the sense that applying a rotation to some hypothesized alignment results in a transformation where the translation is still approximately correct, as opposed to choosing the origin at the corner of the image where applying a rotation will significantly affect the overlap.

The Centre of Mass (COM) is defined by taking the intensity as a mass density (or the intensity less the minimum intensity in order to assure that it is positive) so that:

$$COM = \frac{\sum_i \vec{x}_i I(\vec{x}_i)}{\sum_i I(\vec{x}_i)} \tag{37}$$

where $\vec{x}_i$ are the 3D spatial locations of the lattice sites.

For example, consider an ellipsoid in the reference and floating image that is almost aligned correctly, except for a small rotation, see figure 5. Applying the correct adjustment in rotation but using an origin at the corner of the volume will cause the orientation of the two ellipsoids to be correct but will result in a translational error. Therefore, to find the correct transformation using this parameterisation requires first one rotation parameter to be adjusted and then all three translation parameters. However, because the cost function will be greatly affected by the amount of overlap of the images, the minima of cost as the rotation is varied is largely going to reflect the change in position (translation) rather than the change in orientation (rotation). On the other hand, by having the origin at the COM, the rotation change does not affect the translation greatly, and a search along each parameter independently is much more likely to find the correct solution.

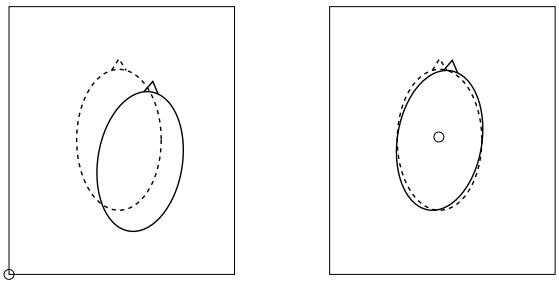
Even with this choice of origin there is coupling between the translation and remaining parameters. Furthermore, as Powell's method is capable of estimating linear parameter couplings it might seem that decoupling the parameters initially is unimportant. However, these couplings are non-linear and therefore beyond the scope of Powell's method to estimate.

In addition, by taking the origin at the COM, the identity transformation will align the COM of the two volumes, which makes this a sensible choice for any initial transformation.

5.2 Histogram Bin Size

Another important choice to make for most cost functions is the histogram bin size, or equivalently, number of bins. All the cost functions investigated in section 3.2 require some form of intensity binning. For the Correlation Ratio and Woods function the intensity bins are only used for the reference volume in order to determine the iso-sets, whereas for the entropy-based functions, intensity binning occurs for both images.

The number of intensity bins determines two things: (1) how good the statistics will be in reflecting the ideal, continuous distribution and (2) the effective fidelity of the intensities (that is, using 256 bins is equivalent



Origin at corner Origin at COM

Figure 5: Example comparing the same small rotation about the corner of the volume with the rotation about the Centre of Mass. In each case the original image is shown with dashed lines and the rotated version with solid lines, while the origin as shown as a small circle. Note that the overlap of the two brain areas is significantly diminished when the origin is at the corner of the volume since the translation is coupled with the rotation.

to that image having an 8 bit intensity range). In most cases it is the first point that is of most interest, as the Signal to Noise Ratio in a typical image limits the maximum practical fidelity.

Determining how many histogram bins should be used for estimating distributions is a problem in non-parametric statistics, although histogram-based methods are not the only form of distribution estimation (for an overview see (Izenman, 1991)). However, histogram-based methods are the most practical as other methods usually involve too much computational overhead to be useful for this problem.

It has been shown (Scott, 1979) that the optimal histogram bin size, which provides the most efficient, unbiased estimation of the probability density function, is achieved when:

$$W = 3.49\sigma N^{-1/3} \tag{38}$$

where W is the width of the histogram bin, σ is the standard deviation of the distribution and N is the number of available samples. In practice, the estimated standard deviation, s, must be used. A similar, but more robust, result was also obtained by Freedman and Diaconis (summarised in (Izenman, 1991)), which gives the bin width as:

$$W = 2(IQR)N^{-1/3} \tag{39}$$

where IQR is the interquartile range (the 75th percentile minus the 25th percentile).

In both formulations the width is proportional to $N^{-1/3}$. However, it is common practice for the number of histogram bins (and hence the size) to be set as a pre-defined constant in a registration method. This results in the statistics becoming less precise as the sub-sampling is increased. For example, with 8mm cubed voxels and a 200mm cubed FOV, there are 15625 voxels in the image, and if the expected number of samples per bin in a joint histogram is to be at least 10, then there must be 1562 bins or less, or less than 40 intensity bins per image. However, it is not unusual to find methods that use 256 bins for this data, which results in an average of 0.2 samples per bin which will result in very poor statistics, and therefore less smooth cost functions.

For the method proposed here, the histogram bin size is set proportional to $N^{-1/3}$. This means that at a resolution with voxel size of n mm cubed, the number of voxels, or samples, is $N = (200/n)^3$. Therefore, the histogram bin size is given by $W = n(IQR)/100$. However, this is only true for 1D histograms, where intensity bins are only used for a single image, such as for Correlation Ratio and the Woods function. For the entropy-based functions, a 2D joint histogram is required and so the result must be generalised.

The two easiest ways of generalising the above result are (1) to keep the bin width the same or (2) to keep the average number of samples per bin the same. Given that the number of bins within the IQR is now $(IQR)^2/W^2$ and the estimated number of samples per bin is $NW^2/(2IQR^3)$, the two generalisations give bin widths of $W_1 = n(IQR)/100$ and $W_2 = \sqrt{n}(IQR)/10$ respectively. However, this gives a very large bin width for the latter case, so that for the 8mm resolution, $W_1 = (IQR)/12.5$ as compared to $W_2 \approx (IQR)/3.5$, whereas the expected number of samples per bin are 50 and 625 respectively. Therefore, the former generalisation ensures a sufficient number of samples per bin whilst providing a better fidelity, and so this is the one that is chosen.

In practice, the number of intensity bins used per image at resolution n is $256/n$. This is equivalent to estimating the IQR to 39% of the total intensity range. In addition, this gives the same number of bins at a 1mm resolution (that is, 256) as is common in many other methods, while maintaining near optimal sampling requirements and a reasonable level of fidelity, in keeping with the expected Signal to Noise Ratio.

5.3 Sampling and Sub-Sampling

The precise method of sampling and sub-sampling the volumes to calculate the cost function is also important. In the method proposed here the reference volume is initially re-sampled to an isotropic grid with voxel size 1mm cubed. This is done by interpolating the values available in the original volume which usually has anisotropic voxel dimensions. Once this isotropic 1mm resolution reference volume has been obtained, the 2mm, 4mm and 8mm sub-sampled versions are created.

Sub-sampling the reference volume by a factor of two is done by first blurring the intensities using a convolution with a discrete, 3D Gaussian kernel where: $FWHM = n$ mm (or $\sigma = 0.425n$ mm), with n being the size of the required sub-sampling (that is, 2, 4 or 8). This blurring is done so that all points on the lattice contribute equally to the sub-sampled version. The sub-sampling then simply takes every nth point on the lattice in each direction. Therefore, the new volume contains $1/n^3$ as many points as the original and so the total storage for all four volumes (1mm, 2mm, 4mm and 8mm resolutions) is just 14% more than for the 1mm resolution volume alone.

To evaluate the cost function at a resolution of n mm requires the intensities at the isotropic lattice sites to be known for both the reference and floating volumes. The reference volume intensities are already known, having been calculated and stored as described above. Therefore, it is only necessary to calculate the floating volume intensities. As various transformations are applied to the floating volume during the optimisation procedure, the interpolated values are not stored but just calculated as required by the cost function. However, before the optimisation an initial blurring is applied to the stored intensities in the floating volume, as is done for the reference volume. In this case though, the volume usually has an anisotropic voxel size, and so the discrete Gaussian kernel used is also anisotropic, reflecting the unequal sampling of the continuous, isotropic Gaussian kernel.

6 Results

The registration method described above in sections 4 and 5 has been implemented in C++ and is called FLIRT — FMRIB's Linear Image Registration Tool. This program has undergone extensive trials over several months, being used by various researchers including trained neurologists, psychologists and physiologists. During this time it has been used to perform many thousands of registrations in the context of FMRI analysis and structural studies (Smith et al., 2000).

Feedback from the users has been positive, with the vast majority of registrations producing acceptable results and only a few cases of failure, or visually unacceptable registrations. Of the failures reported it was found that the other registration methods available (see below) also failed to find an acceptable registration. Furthermore, it was found that these volumes were usually difficult to manually register as they often had unusual features such as particularly enlarged ventricles. Consequently, these cases are ones where the main problem appears to be the transformation space. That is, a good *affine* registration does not exist. It would require the use of higher order transformations (non-linear warpings) to achieve a good solution.

One aspect that was highlighted during these trials was that, as reported before (Woods et al., 1993), in cases where the scalp or skull appeared in one image but not the other, better results were sometimes obtained by "stripping" the scalp and skull from the image. Without doing this stripping a typical registration error that occurred was that the outer brain surface would be aligned with the scalp/skull surface in the other image. That is, the scale was incorrectly set. This is a direct result of the cost function having similar values for this situation and the "correct" scale, since in one case the gap between the scalp/skull and brain is matched poorly, whilst in the other case the scalp/skull itself is matched poorly. Therefore, it should be kept in mind that such skull/scalp stripping should still be performed, either manually or automatically, in order to get the best results.

6.1 Consistency Test

The results presented above are purely qualitative, based on the subjective assessment of many individuals. However, as is generally the case for many registration problems in practice, there was no ground truth available to test the registration against. This makes the area of quantitative assessment of methods quite difficult.

In order to test the method more quantitatively, a comparative consistency test was performed. This test aims to measure the robustness, rather than the accuracy (West et al., 1997) of the registration method. Robustness is defined here as the ability to get close to the global minimum on all trials, whereas accuracy is the ability to precisely locate a (possibly local) minimum of the cost function. For example, one method might always (say in over 99.99% of cases) be between 0.2mm and 0.6mm from the best possible solution compared to another method that was often less than 0.1mm from the best solution but would sometimes (say in 5% of

cases) fail to find the global minimum and get trapped in a local minimum which could be in excess of 10mm from the best solution. In this case the former method would be considered more robust than the latter while the latter method would be more accurate but less robust. Ideally a registration method should be both.

The consistency test is designed to assess one *necessary, but not sufficient*, aspect of robustness. That is, the ability to find the same solution regardless of the initial position. Any robust method, which always finds the global minimum, will give the same solution each time whereas a non-robust method which can be trapped by a local minimum is likely to give different solutions depending on the initial position. However, this condition is not sufficient in determining robustness as the same, consistent, solution may just be a large local minimum, rather than the global minimum. Therefore it is also necessary to check that the registration solution is acceptable to someone trained in neuroanatomy. This aspect was addressed in the trials described above.

More specifically, the consistency test for an individual image I involved taking the image and applying several pre-determined affine transformations, A_j to it. All these images (both transformed and un-transformed) were registered to a given reference image, I^r, giving transformations T_j. If the method was consistent the composite transformations $T_j \circ A_j$ should all be the same, which is illustrated in figure 6. Moreover, an RMS deviation between the composite registration and the registration from the un-transformed case allows quantification of the consistency.

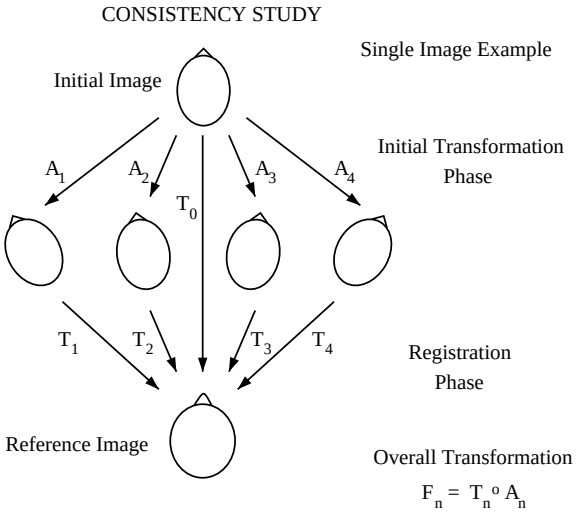


Figure 6: Illustration of the consistency test for a single image. An image (top) has a number of initial affine transformations A_j (rotations are used in this study) applied to it. The resulting images (middle) are then registered to the reference image (bottom), giving transformations T_j. Therefore, the overall transformation from the initial image to the reference image is $F_j = T_j \circ A_j$, and these are compared with T_0 which is the registration of the initial image directly to the reference image. For a consistent method, all the transformations, F_j, should be the same as T_0.

The particular test used (which is also described in (Jenkinson and Smith, 1999)) used 18 different images as the floating images (like the one shown in figure 3a), all with the MNI 305 brain (Collins et al., 1994) as the reference image. The 18 images were all $256 \times 256 \times 30$, T2 weighted MRI images with voxel dimensions of 0.93mm by 0.93 mm by 5mm, while the MNI 305 template is a $172 \times 220 \times 156$, T1 weighted MRI images with voxel dimensions of 1mm by 1 mm by 1mm.

In addition to FLIRT, several other registration packages were tested. These were AIR (Woods et al., 1993), SPM (Friston et al., 1995), UMDS (Studholme et al., 1996) and MRITOTAL (Collins et al., 1994). These methods were chosen because the authors' implementations were available, and so this constituted a fair test as opposed to a re-implementation of a method described in a paper, where often the lack of precise implementation details makes it difficult to produce a good working method.

The results of such a test, using six different rotations about the Anterior-Posterior axis, are shown in figure 7. It can be seen that only FLIRT and MRITOTAL were consistent with this set of images. This indicates that the other methods, AIR, SPM and UMDS, get trapped in local minima more easily, and are not as robust. In particular, rotations of only 0.5° sometimes resulted in large differences in the final registrations, showing how sensitive the methods are to initial position.

A further consistency test was then performed comparing only MRITOTAL and FLIRT. This test used initial scalings rather than rotations. The reason that this is important is that MRITOTAL uses a purely local

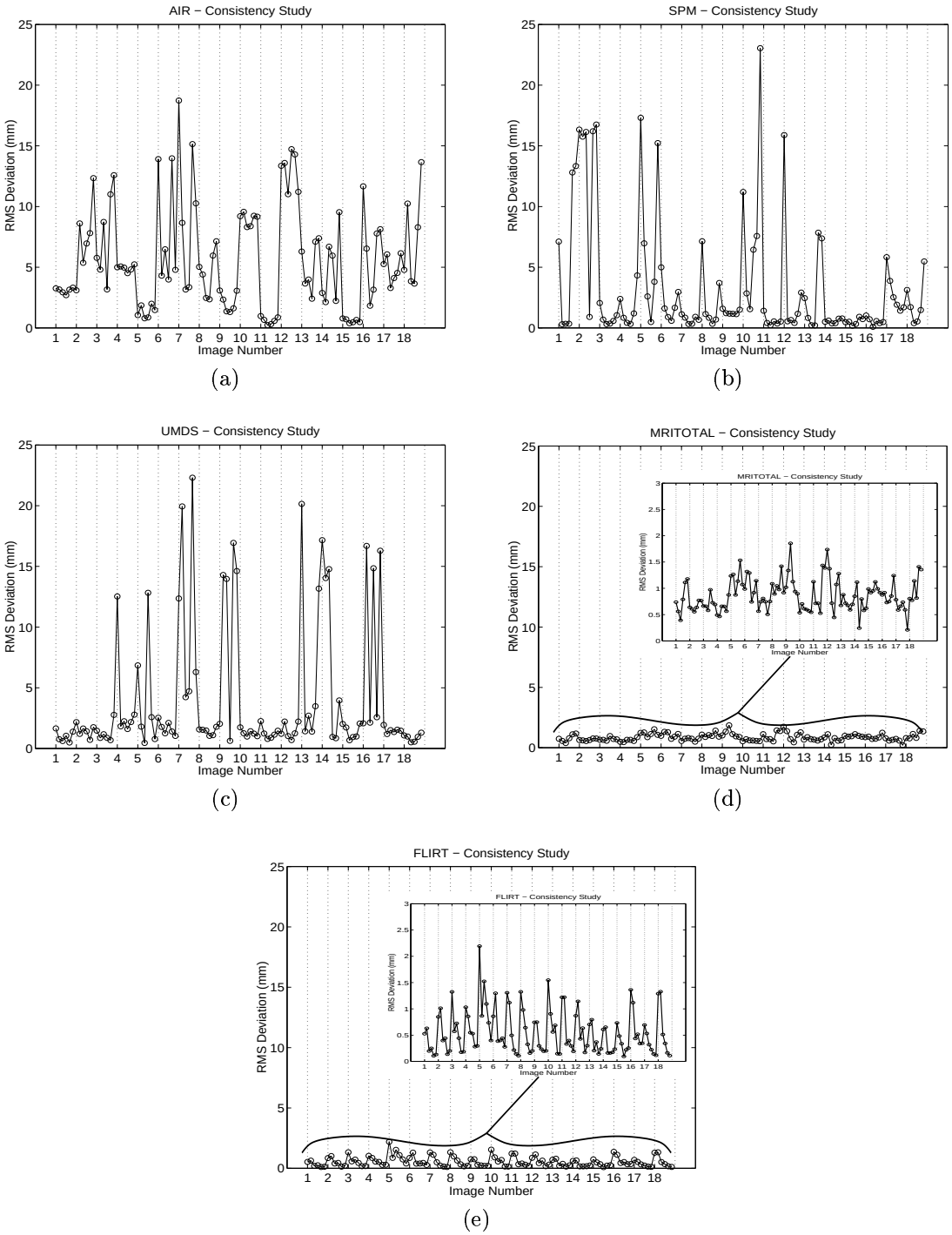
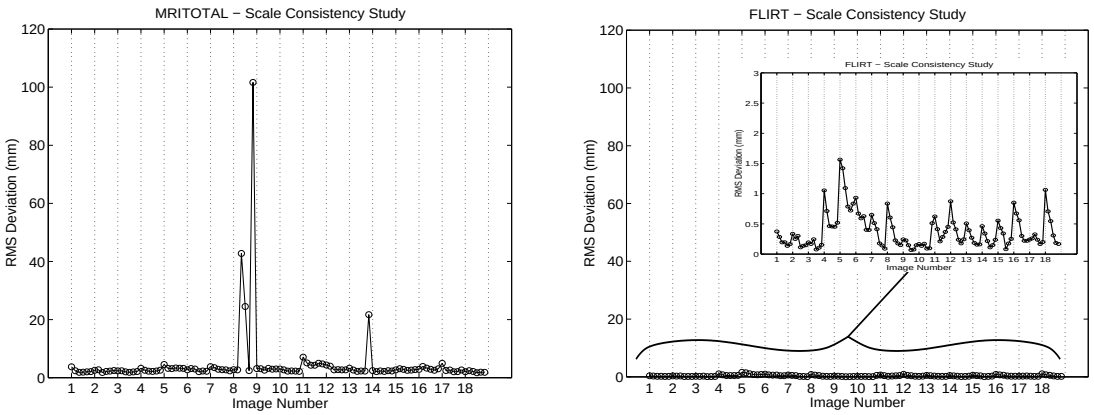


Figure 7: Results of the consistency study, plotting RMS deviation (in mm) versus image number for (a) AIR, (b) SPM, (c) UMDS, (d) MRITOTAL and (e) FLIRT. For each of the 18 source images (T2 weighted MRI images with voxel dimensions of 0.93mm by 0.93 mm by 5mm) there are 6 results corresponding to initial starting rotations of -10,-2,-0.5,0.5,2, and 10 degrees about the y-axis (anterior-posterior axis). All of the methods, except FLIRT and MRITOTAL, show large deviations and are therefore inconsistent and non-robust.

optimisation method (Gradient Descent) but relies on initial pre-processing to provide a good starting position. This pre-processing is done by finding the principle axes of both volumes and initially aligning them. However, this initial alignment does not give any information about scaling and is dependent on the FOV, since when the edges of the volume truncate the image it can have a significant impact on the principle axes that are computed.

The results of the scaling consistency test are shown in figure 8. It can be seen that, although generally consistent, in three cases MRITOTAL produces registrations that deviate by more than 20mm (RMS) from each other. In contrast, FLIRT was consistent (less than 2mm RMS) for all images.



(a) (b)

Figure 8: Results of the scale consistency study, plotting RMS deviation (in mm) versus image number for (a) MRITOTAL and (b) FLIRT. For each of the 18 source images (T2 weighted MRI images with voxel dimensions of 0.93mm by 0.93 mm by 5mm) there are 6 results corresponding to initial scalings of 0.7, 0.8, 0.9, 1.1, 1.2 and 1.3 about the Centre of Mass. In three cases MRITOTAL shows large deviations and is therefore inconsistent and non-robust.

7 Discussion

In this report the problem of global optimisation for brain image registration was examined. Only affine registration was examined as although this is a much easier problem than general, non-linear transformations, finding the global minimum is still difficult. Furthermore, many non-linear methods rely on an initial affine registration to find a good starting position, and so having a good method of affine registration is important.

The standard mathematical formulation of the registration problem was detailed and its assumptions examined. This included an analysis of the asymptotic behaviour of several multi-modal cost functions and showed that only the Correlation Ratio, Mutual Information and Normalised Mutual Information correctly yielded high costs for these poor registrations. Consequently, only these functions (out of those considered) are suitable for the global optimisation formulation. In addition, the distribution of local minima was sampled empirically for large sub-sampling (8mm cubed voxels) which demonstrated that there were still many local minima, so that a local optimisation method would not be sufficient to reliably find the global minimum.

A global optimisation method, which was tailored to this registration problem, was then proposed together with a complete discussion of the implementation details required to turn the method into a practical registration tool. This method uses the Correlation Ratio (although any suitable cost function could be used) and combines a common local optimisation method (Powell's method — adapted for the voxel size to improve the efficiency) with several search strategies. The main search is a global search through the transformation space using the most sub-sampled volumes, with smaller neighbourhood-based searches taking place as the sub-sampling decreases. In such methods, the major constraint is the amount of time that is considered reasonable. In the design of this method the search time was limited to 20 minutes using our implementation. Even with this constraint it was still possible to perform a full search and identify the global minimum more reliably.

Evaluation of this registration method was done in two ways. Firstly, the software has been used routinely in the FMRIB Centre as an experimental trial, allowing many people to use and comment on the performance. The qualitative feedback was positive in its own right and in comparison with other available methods. This method has now been used to satisfactorily solve thousands of registration problems.

Secondly, quantitative results were found for a consistency test. This test is designed to examine the

robustness of a registration method. Results showed that the method was highly consistent on a set of difficult images. Furthermore, several other available packages were tested on the same set of images and did not achieve the same level of consistency, sometimes demonstrating substantial inconsistencies.

The global optimisation method proposed here does not, however, guarantee finding the global minimum. This is typical though, as even methods such as Simulated Annealing only provide a statistical guarantee which cannot be met in practice. The results, though, are encouraging, and by using finer search grids, the likelihood of finding the global minimum can be increased. This requires that there be sufficient time at hand, or a sufficiently fast computer. However, even with modest resources this method can find the global minimum (solving the registration problem) within one hour, more reliably than the other methods tested.

Optimisation is only one aspect of the registration problem, although it is practically a very important one. Other aspects such as interpolation, alternative cost functions and understanding the properties of existing cost functions remain important areas for further work. In addition, for higher dimensional transformations, the optimisation problem becomes even more difficult, and so this too is an important area for future research.

Finally, as stated before, it is important to be precise about the implementational details of such methods. This allows (1) the methods to be more easily re-implemented by others, (2) the various methods to be compared fairly (by using the author's parameters) and (3) the results to be repeated. The alternative is finding the best value for various implementation parameters by trial and error which is extremely tedious and error prone. In addition to including the implementation details in this report the source code for the FLIRT package is available for downloading from `www.fmrib.ox.ac.uk`. This should avoid others needing to re-implement the method and facilitate the evaluation of the method, hopefully leading to further improvements.

8 Acknowledgements

The authors wish to thank the Medical Research Council and the European MICRODAB project for supporting this work.

References

Collignon, A., Vandermeulen, D., Suetens, P., and Marchal, G. 1995. 3D multi-modality medical image registration using feature space clustering. In *Proceedings of the 1st International Conference on Computer Vision, Virtual Reality and Robotics in Medicine*, pp. 195–204, Nice, France.

Collins, D., Neelin, P., Peters, T., and Evans, A. 1994. Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. *Journal of Computer Assisted Tomography*, **18(2):**192–205.

Friston, K., Ashburner, J., Frith, C., Poline, J.-B., Heather, J., and Frackowiak, R. 1995. Spatial registration and normalization of images. *Human Brain Mapping*, **2:**165–189.

Geman, S. and Geman, D. 1984. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, **6(6):**721–741.

Hajnal, J., Saeed, N., Soar, E., Oatridge, A., Young, I., and Bydder, G. 1995. A registration and interpolation procedure for subvoxel matching of serially acquired MR images. *Journal of Computer Assisted Tomography*, **19(2):**289–296.

Ingber, A. L. 1989. Very fast simulated re-annealing. *Jounral of Mathematical and Computational Modelling*, **12:**967–973.

Izenman, A. J. 1991. Recent developments in nonparametric density estimation. *Journal of the American Statistical Association*, **86(413):**205–224.

Jenkinson, M. and Smith, S. 1999. An investigation of the robustness of registration methods. In *Int. Workshop on Biomedical Image Registration*, pp. 200–210.

Maes, F. 1998. *Segmentation and Registration of Multimodal Medical Images: from Theory, Implementation and Validation to a Useful Tool in Clinical Practice.* PhD thesis, Catholic University of Leuven, Belgium.

Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., and Suetens, P. 1997. Multimodality image registration by maximisation of mutual information. *IEEE Trans. on Medical Imaging*, **16(2):**187–198.

Pluim, J. P. W., Maintz, J. A., and Viergever, M. A. 2000. Interpolation artefacts in mutual information based image registration. *Computer Vision and Image Understanding*, **77(2):**211–232.

Press, W., Teukolsky, S., Vetterling, W., and Flannery, B. 1995. *Numerical Recipies in C*. Cambridge University Press, second edition.

Roche, A., Malandain, G., Pennec, X., and Ayache, N. 1998. Multimodal image registration by maximization of the correlation ratio. Technical Report 3378, INRIA Sophia-Antipolis.

Scott, D. 1979. On optimal and data-based histograms. *Biometrika*, **66:**605–610.

Smith, S., De Stefano, N., Jenkinson, M., and Matthews, P. 2000. Normalised accurate measurement of longitudinal brain change. *Journal of Computer Assisted Tomography*, . submitted.

Studholme, C., Hill, D., and Hawkes, D. 1995. Multiresolution voxel similarity measures for MR-PET registration. In *Proceedings of Information Processing in Medical Imaging*, pp. 287–298, Brest, France.

Studholme, C., Hill, D., and Hawkes, D. 1996. Automated 3D registration of MR and CT images of the head. *Medical Image Analysis*, **1(2):**163–175.

Studholme, C., Hill, D., and Hawkes, D. 1999. An overlap invariant entropy measure of 3D medical image alignment. *Pattern Recognition*, **32:**71–86.

Thacker, N. A., Jackson, A., and Moriarty, D. 1999. Improved quality of re-sliced MR images using re-normalized sinc interpolation. *Journal of Magnetic Resonance Imaging*, **10:**582–588.

Viola, P. and Wells, W. 1997. Alignment by maximization of mutual information. *International Journal of Computer Vision*, **24(2):**137–154.

West et al., J. 1997. Comparison and evaluation of retrospective intermodality brain image registration techniques. *Journal of Computer Assisted Tomography*, **21(4):**554–566.

Wolpert, D. H. and Macready, W. G. 1996. No free lunch theorems for search. Technical Report SFI-TR-95-02-010, The Santa Fe Insititute.

Woods, R., Mazziotta, J., and Cherry, S. 1993. MRI–PET registration with automated algorithm. *Journal of Computer Assisted Tomography*, **17(4):**536–546.