Chapter 3

Image Warping using Basis Functions

Contents

3.1 Introduction

Methods of registering images can be broadly divided into *label based* and *non-label based*. Label based techniques identify homologous features (labels) in the image and template and find the transformations that best superpose them. The labels can be points, lines or surfaces. Homologous features are often identified manually, but this process is time consuming and subjective. Another disadvantage of using points or lines as landmarks is that there are very few readily identifiable discrete points or lines in the brain. However, surfaces are more readily identified, and in many instances they can be extracted automatically (or at least semi-automatically). Once they are identified, the spatial transformation is effected by bringing the homologies together. If the labels are points, then the required transformations at each of those points is known. Between the points, the deforming behaviour is not known, so it is forced to be as 'smooth' as possible. There are a number of methods for modelling this smoothness. The simplest models include fitting splines through the points in order to minimise *bending energy* (Bookstein, 1997a; Bookstein,

1989). More complex forms of interpolation are often used when the labels are surfaces. For example Thompson *et al.*(1996) map surfaces together using a fluid model.

Non-label based approaches identify a spatial transformation that minimises some index of the difference between a source and a template image, where both are treated as unlabeled continuous processes. The matching criterion is usually based upon minimising the sum of squared differences or maximising the correlation between the images. For this criterion to be successful, it requires the template to appear like a warped version of the image. In other words, there must be correspondence in the grey levels of the different tissue types between the source image and template.

A potentially enormous number of parameters are required to describe the nonlinear transformations that warp two images together (i.e., the problem is very high dimensional). However, much of the spatial variability can be captured using just a few parameters. Low spatial frequency global variability of head shape can be accommodated by describing deformations by a linear combination of low frequency basis functions. One very widely used basis function registration method is part of the AIR package (Woods *et al.*, 1998a; Woods *et al.*, 1998b), which uses polynomial basis functions to model shape variability. For example, a two dimensional third order polynomial basis function mapping can be defined something like:

$$\begin{aligned} y_1 = & q_1 + q_2 x_1 + q_3 x_1^2 + q_4 x_1^3 + \\ & q_5 x_2 + q_6 x_1 x_2 + q_7 x_1^2 x_2 + \\ & q_8 x_2^2 + q_9 x_1 x_2^2 + \\ & q_{10} x_2^3 \\ y_2 = & q_{11} + q_{12} x_1 + q_{13} x_1^2 + q_{14} x_1^3 + \\ & q_{15} x_2 + q_{16} x_1 x_2 + q_{17} x_1^2 x_2 + \\ & q_{18} x_2^2 + q_{19} x_1 x_2^2 + \\ & q_{20} x_2^3 \end{aligned} \tag{3.1}$$

The small number of parameters will not allow every feature to be matched exactly, but it will permit the global head shape to be modelled. The method of nonlinear registration described in this chapter is a similar approach, but uses discrete cosine transform basis functions instead of polynomials. The rational for adopting the low dimensional approach is that it allows rapid modelling of the global brain shape.

The deformations required to transform images to the same space are not clearly defined. Unlike rigid body transformations, where the constraints are explicit, those for warping are more arbitrary. Regularisation schemes are therefore necessary when attempting image registration with many parameters, thus ensuring that voxels remain close to their neighbours. Regularisation is normally incorporated by some form of Bayesian scheme, using estimators such as the *maximum a posteriori* (MAP) estimate or the *minimum variance estimate* (MVE). The MAP estimate is the single solution that has the highest posterior probability of being correct, and is the estimate used for the fully automatic non-label based spatial normalisation method described in this chapter.

3.2 Methods

This section begins by introducing a modification to the optimisation method described in Section 2.4, such that more robust *maximum a posteriori* (MAP) parameter estimates can be obtained. The first step in registering images from different subjects involves determining the optimum 12 parameter affine transformation. A procedure for doing this using the MAP optimisation scheme is described. Because the variability of head sizes is known *a priori*, the registration can be made more robust by incorporating this knowledge. The next part describes nonlinear registration for correcting gross differences in head shapes that can not be accounted for by the affine normalisation alone. The nonlinear warps are modelled by linear combinations of smooth basis functions, and a fast algorithm for determining the optimum combination of basis functions is described. For speed and simplicity, a relatively small number of parameters (approximately 1000) are used to describe the nonlinear components of the registration. The MAP scheme requires some form of prior distribution for the basis function coefficients, so a number of different forms for this distribution are then presented. The last part of this section describes a variety of possible models for intensity transforms. In addition to spatial transformations, it is sometimes desirable to also include intensity transforms in the registration model, as one image may not look exactly like a spatially transformed version of the other.

3.2.1 A *Maximum A Posteriori* Solution

A Bayesian registration scheme is used in order to obtain a *maximum a posteriori* estimate of the registration parameters. Given some prior knowledge of the variability of brain shapes and sizes that may be encountered, a MAP registration scheme is able to give a more accurate (although biased) estimate of the true shapes of the brains. This is illustrated by a very simple one dimensional example in Figure 3.1. The use of a MAP parameter estimate reduces any potential over-fitting of the data, which may lead to unnecessary deformations that only reduce the residual variance by a tiny amount. It also makes the registration scheme more robust by reducing the search space of the algorithm, and therefore the number of potential local minima.

Bayes' rule can be expressed as:

$$p(\mathbf{q}|\mathbf{b}) \propto p(\mathbf{b}|\mathbf{q})p(\mathbf{q}) \tag{3.2}$$

where $p(\mathbf{q})$ is the prior probability of parameters $\mathbf{q}$, $p(\mathbf{b}|\mathbf{q})$ is the conditional probability that $\mathbf{b}$ is observed given $\mathbf{q}$ and $p(\mathbf{q}|\mathbf{b})$ is the posterior probability of $\mathbf{q}$, given that measurement $\mathbf{b}$ has been made. The *maximum a posteriori* (MAP) estimate for parameters $\mathbf{q}$ is the mode of $p(\mathbf{q}|\mathbf{b})$. The *maximum likelihood* (ML) estimate is a special case of the MAP estimate, in which $p(\mathbf{q})$ is uniform over all values of $\mathbf{q}$. For our purposes, $p(\mathbf{q})$ represents a known prior probability distribution from which the parameters are drawn, $p(\mathbf{b}|\mathbf{q})$ is the likelihood of obtaining the data $\mathbf{b}$ given the parameters, and $p(\mathbf{q}|\mathbf{b})$ is the function to be maximised. The optimisation can be simplified by assuming that all probability distributions can be approximated by multi-normal (multidimensional and normal) distributions, and can therefore be described by a mean vector and a covariance matrix.

A probability is related to its Gibbs form by $p(a) \propto e^{-H(a)}$. Therefore the posterior probability is maximised when its Gibbs form is minimised. This is equivalent to minimising $H(\mathbf{b}|\mathbf{q}) +$

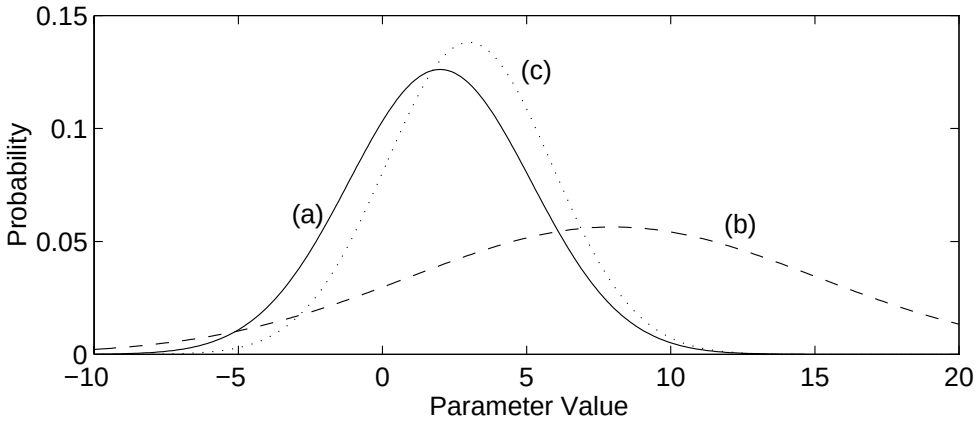

Figure 3.1: This figure illustrates a hypothetical example with one parameter, where the prior probability distribution is better described than the likelihood. The solid Gaussian curve (a) represents the prior probability distribution (p.d.f), and the dashed curve (b) represents a maximum likelihood parameter estimate (from fitting to observed data) with its associated certainty. The true parameter is known to be drawn from distribution (a), but it can be estimated with the certainty described by distribution (b). Without the MAP scheme, a more precise estimate would probably be obtained for the true parameter by taking the most likely *a priori* value, rather than the value obtained from a maximum likelihood fit to the data. This would be analogous to cases where the number of parameters is reduced in a maximum likelihood registration model in order to achieve a better solution (e.g., see page 45). The dotted line (c) shows the posterior p.d.f obtained using Bayesian statistics. The maximum value of (c) falls at the MAP estimate. It combines previously known information with that from the data to give a more accurate estimate.

$H(\mathbf{q})$ (the posterior potential). In this expression, $H(\mathbf{b}|\mathbf{q})$ (the likelihood potential) is related to the residual sum of squares. If the parameters are assumed to be drawn from a multi-normal distribution described by a mean vector $\mathbf{q_0}$ and covariance matrix $\mathbf{C_0}$, then $H(\mathbf{q})$ (the prior potential) is simply given by:

$$H(\mathbf{q}) = (\mathbf{q} - \mathbf{q_0})^T \, {\mathbf{C_0}}^{-1} \, (\mathbf{q} - \mathbf{q_0}) \tag{3.3}$$

Eqn. 2.22 gives the following maximum likelihood updating rule for the parameter estimation:

$${\mathbf{q_{ML}}}^{(n+1)} = \mathbf{q}^{(n)} - \left(\mathbf{A}^T\mathbf{A}\right)^{-1} \mathbf{A}^T\mathbf{b} \tag{3.4}$$

Assuming equal variance for each observation (σ^2) and ignoring covariances among them, the formal covariance matrix of the fit on the assumption of normally distributed errors is given by $\sigma^2 \left(\mathbf{A}^T\mathbf{A}\right)^{-1}$. When the distributions are normal, the MAP estimate is simply the average of the prior and likelihood estimates, weighted by the inverses of their respective covariance matrices:

$$\mathbf{q}^{(n+1)} = \left({\mathbf{C_0}}^{-1} + \mathbf{A}^T\mathbf{A}/\sigma^2\right)^{-1} \left({\mathbf{C_0}}^{-1}\mathbf{q_0} + \mathbf{A}^T\mathbf{A}/\sigma^2 {\mathbf{q_{ML}}}^{(n+1)}\right) \tag{3.5}$$

The MAP optimisation scheme is obtained by combining Eqns. 3.4 and 3.5.

$$\mathbf{q}^{(n+1)} = \left({\mathbf{C_0}}^{-1} + \mathbf{A}^T\mathbf{A}/\sigma^2\right)^{-1} \left({\mathbf{C_0}}^{-1}\mathbf{q_0} + \mathbf{A}^T\mathbf{A}\mathbf{q}^{(n)}/\sigma^2 - \mathbf{A}^T\mathbf{b}/\sigma^2\right) \tag{3.6}$$

For the sake of the registration, it is assumed that the exact form for the *a priori* probability distribution ($\mathbf{q_0}$ and $\mathbf{C_0}$) is known. However, because the registration may need to be done on a wide range of different image modalities, with differing contrasts and signal to noise ratios, it is not possible to easily and automatically know what value to use for σ^2. In practice, σ^2 is assumed to be the same for all observations, and is estimated from the sum of squared differences from the current iteration:

$$\sigma^2 = \sum_{i=1}^{I} b_i(\mathbf{q})^2/\nu \tag{3.7}$$

where ν refers to the degrees of freedom. If the sampling is sparse relative to the smoothness, then $\nu \simeq I - J$, where I is the number of sampled locations in the images and J is the number of estimated parameters [1].

However, complications arise because the images are smooth, resulting in the observations not being independent, and a reduction in the effective number of degrees of freedom. The degrees of freedom are corrected using the principles described by Friston (1995a) [although this approach is not strictly correct (Worsley & Friston, 1995), it gives an estimate that is close enough for these purposes]. The effective degrees of freedom are estimated by assuming that the difference between $\mathbf{f}$ and $\mathbf{g}$ approximates a continuous, zero-mean, homogeneous, smoothed *Gaussian random field*. The approximate parameter of a Gaussian point spread function describing the smoothness in direction k (assuming that the axes of the Gaussian are aligned with the axes of the image co-ordinate system) can be obtained by (Poline *et al.*, 1995):

$$w_k = \sqrt{\frac{\sum_{i=1}^{I} b_i(\mathbf{q})^2}{2\sum_{i=1}^{I} (\nabla_k b_i(\mathbf{q}))^2}} \tag{3.8}$$

[1] Strictly speaking, the computation of the degrees of freedom should be more complicated than this, as this simple model does not account for the regularisation (See Section 7.3).

Multiplying w_k by $\sqrt{8\log_e(2)}$ produces an estimate of the full width at half maximum of the Gaussian. If the images are sampled on a regular grid where the spacing in each direction is s_k, the number of effective degrees of freedom[2] becomes approximately:

$$\nu = (I - J) \prod_k \frac{s_k}{w_k (2\pi)^{1/2}} \tag{3.9}$$

This is essentially a scaling of $I - J$ by the number of resolution elements per voxel.

This approach has the advantage that when the parameter estimates are far from the solution, σ^2 is large, so the problem becomes more heavily regularised with more emphasis being placed on the prior information. For nonlinear warping, this is analogous to a coarse to fine registration scheme. The penalty against higher frequency warps is greater than that for those of low frequency (see Section 3.2.4). In the early iterations, the estimated σ^2 is higher leading to a heavy penalty against all warps, but with more against those of higher frequency. The algorithm does not fit much of the high frequency information until σ^2 has been reduced. In addition to a gradual reduction in σ^2 due to the decreasing residual squared difference, σ^2 is also reduced because the estimated smoothness is decreased, leading to more effective degrees of freedom. Both these factors are influential in making the registration scheme more robust to local minima.

3.2.2 Affine Registration

Almost all between subject co-registration or spatial normalisation methods for brain images begin by determining the optimal nine or twelve parameter affine transformation that registers the images together. This step is normally performed automatically by minimising (or maximising) some mutual function of the images. The objective of affine registration is to fit the source image $\mathbf{f}$ to a template image $\mathbf{g}$, using a twelve parameter affine transformation (via a matrix $\mathbf{M}$ generated from parameters q_1 to q_{12}). The images may be scaled quite differently, so an additional intensity scaling parameter (q_{13}) is included in the model. The objective function that is minimised is therefore:

$$\sum_i \left(f(\mathbf{M}\mathbf{x}_i) - q_{13} g(\mathbf{x}_i)\right)^2 \tag{3.10}$$

Without constraints and with poor data, simple ML parameter optimisation (similar to that described in Section 2.5) can produce some extremely unlikely transformations. For example, when there are only a few slices in the image, it is not possible for the algorithms to determine an accurate zoom in the out of plane direction. Any estimate of this value is likely to have very large errors. When a regularised approach is not used, it may be better to assign a fixed value for this difficult-to-determine parameter, and simply fit for the remaining ones.

By incorporating prior information into the optimisation procedure, a smooth transition between fixed and fitted parameters can be achieved. When the error for a particular fitted parameter is known to be large, then that parameter will be based more upon the prior information. In order to adopt this approach, the prior distribution of the parameters should be known. A suitable *a priori* distribution of the parameters ($\mathbf{q_0}$ and $\mathbf{C_0}$ from Eqn. 3.6) was determined from

[2]Note that this only applies when $s_k < w_k(2\pi)^{1/2}$, otherwise $\nu = I - J$. Alternatively, to circumvent this problem the degrees of freedom can be better estimated by $(I - J)\prod_k \mathrm{erf}(2^{-3/2} s_k / w_k)$. This gives a similar result to the approximation by Friston (1995a) for smooth images, but never allows the computed value to exceed $I - J$.

affine transformations estimated from 51 high resolution T_1 weighted brain MR images using the basic least squares optimisation algorithm. The transformation matrices were defined by $\mathbf{M} = \mathbf{M_f}^{-1}\mathbf{M_a}^{-1}\mathbf{M_g}$, where $\mathbf{M_a}$ (refer back to Sections 2.2.1 and 2.2.2) is constructed from parameters $\mathbf{q}$:

$$\mathbf{M_a} = \begin{bmatrix} 1 & 0 & 0 & q_1 \\ 0 & 1 & 0 & q_2 \\ 0 & 0 & 1 & q_3 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & cos(q_4) & sin(q_4) & 0 \\ 0 & -sin(q_4) & cos(q_4) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} cos(q_5) & 0 & sin(q_5) & 0 \\ 0 & 0 & & 0 \\ -sin(q_5) & 0 & cos(q_5) & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \times \ldots$$
$$\begin{bmatrix} cos(q_6) & sin(q_6) & 0 & 0 \\ -sin(q_6) & cos(q_6) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} q_7 & 0 & 0 & 0 \\ 0 & q_8 & 0 & 0 \\ 0 & 0 & q_9 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & q_{10} & q_{11} & 0 \\ 0 & 1 & q_{12} & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (3.11)$$

The results for the translation and rotation parameters (q_1 to q_6) are ignored, since these depend only on the positioning of the subjects in the scanner, and do not reflect variability of head shape and size.

The mean zooms required to fit the individual brains to the space of the template (parameters q_7 to q_9) were 1.10, 1.05 and 1.17 in the left-right, posterior-anterior and inferior-superior directions respectively, reflecting the fact that the template was larger than a typical head. The variance-covariance matrix for these parameters was:

$$\begin{bmatrix} 0.00210 & 0.00094 & 0.00134 \\ 0.00094 & 0.00307 & 0.00143 \\ 0.00134 & 0.00143 & 0.00242 \end{bmatrix}$$

giving a correlation coefficient matrix of:

$$\begin{bmatrix} 1.00 & 0.37 & 0.59 \\ 0.37 & 1.00 & 0.52 \\ 0.59 & 0.52 & 1.00 \end{bmatrix}$$

As expected, these parameters are correlated, since larger brains are generally larger in all dimensions. This allows partial prediction of the optimal zoom in one direction given the zooms in the others, a fact that is useful for spatially normalising images containing a limited number of transverse slices. The means of the parameters defining shear were close to zero (-0.0024, 0.0006 and -0.0107 for q_{10}, q_{11} and q_{12} respectively). The variances of the parameters are 0.000184, 0.000112 and 0.001786, with very little covariance.

A number of affine registrations were evaluated both with and without incorporating the MAP scheme. This evaluation is described in Section 3.3.1.

3.2.3 Nonlinear Registration

The nonlinear spatial normalisation approach described here assumes that the image has already been approximately registered with the template according to a twelve-parameter affine registration. This section illustrates how the parameters describing global shape differences (not accounted for by affine registration) between an image and template can be determined.

The model for defining nonlinear warps uses deformations consisting of a linear combination of low-frequency periodic basis functions. The spatial transformation from co-ordinates $\mathbf{x}_i$, to co-ordinates $\mathbf{y}_i$ is:

$$\begin{aligned} y_{1i} &= x_{1i} + u_{1i} = x_{1i} + \sum_j q_{j1} d_j(\mathbf{x}_i) \\ y_{2i} &= x_{2i} + u_{2i} = x_{2i} + \sum_j q_{j2} d_j(\mathbf{x}_i) \\ y_{3i} &= x_{3i} + u_{3i} = x_{3i} + \sum_j q_{j3} d_j(\mathbf{x}_i) \end{aligned} \tag{3.12}$$

where q_{jk} is the jth coefficient for dimension k, and $d_j(\mathbf{x})$ is the jth basis function at position $\mathbf{x}$.

The choice of basis functions depend upon the distribution of warps likely to be required, and also upon how translations at borders should behave. If points at the borders over which the transform is computed are not required to move in any direction, then the basis functions should consist of the lowest frequencies of the three dimensional discrete sine transform (DST). If there are to be no constraints at the borders, then a three dimensional discrete cosine transform (DCT) is more appropriate. Both of these transforms use the same set of basis functions to represent warps in each of the directions. Alternatively, a mixture of DCT and DST basis functions can be used to constrain translations at the surfaces of the volume to be parallel to the surface only (*sliding* boundary conditions). By using a different combination of DCT and DST basis functions, the corners of the volume can be fixed and the remaining points on the surface can be free to move in all directions (*bending* boundary conditions) (Christensen, 1994). These various boundary conditions are illustrated in Figure 3.2.

The basis functions used here are the lowest frequency components of the three (or two) dimensional DCT. In one dimension, the DCT of a function is generated by pre-multiplication with the matrix $\mathbf{D}^T$, where the elements of the $I \times M$ matrix $\mathbf{D}$ are defined by:

$$\begin{aligned} d_{i1} &= \frac{1}{\sqrt{I}} \quad i = 1..I \\ d_{im} &= \sqrt{\frac{2}{I}} \cos\left(\frac{\pi(2i-1)(m-1)}{2I}\right) \quad i = 1..I, m = 2..M \end{aligned} \tag{3.13}$$

A set of low frequency two dimensional DCT basis functions are shown in Figure 3.3, and a schematic example of a two dimensional deformation based upon the DCT is shown in Figure 3.4.

As for affine registration, the optimisation involves minimising the sum of squared differences between a source ($\mathbf{f}$) and template image ($\mathbf{g}$). The images may be scaled differently, so an additional parameter (w) is needed to accommodate this difference. The minimised function is then:

$$\sum_i \left(f(\mathbf{y}_i) - w g(\mathbf{x}_i)\right)^2 \tag{3.14}$$

The approach described in Section 2.4 is used to optimise the parameters $\mathbf{q_1}$, $\mathbf{q_2}$, $\mathbf{q_3}$ and w, and requires derivatives of the function $f(\mathbf{y}_i) - w g(\mathbf{x}_i)$ with respect to each parameter. These can be obtained using the chain rule:

$$\frac{\partial f(\mathbf{y}_i)}{\partial q_{j1}} = \frac{\partial f(\mathbf{y}_i)}{\partial y_{1i}} \frac{\partial y_{1i}}{\partial q_{j1}} = \frac{\partial f(\mathbf{y}_i)}{\partial y_{1i}} d_j(\mathbf{x}_i)$$

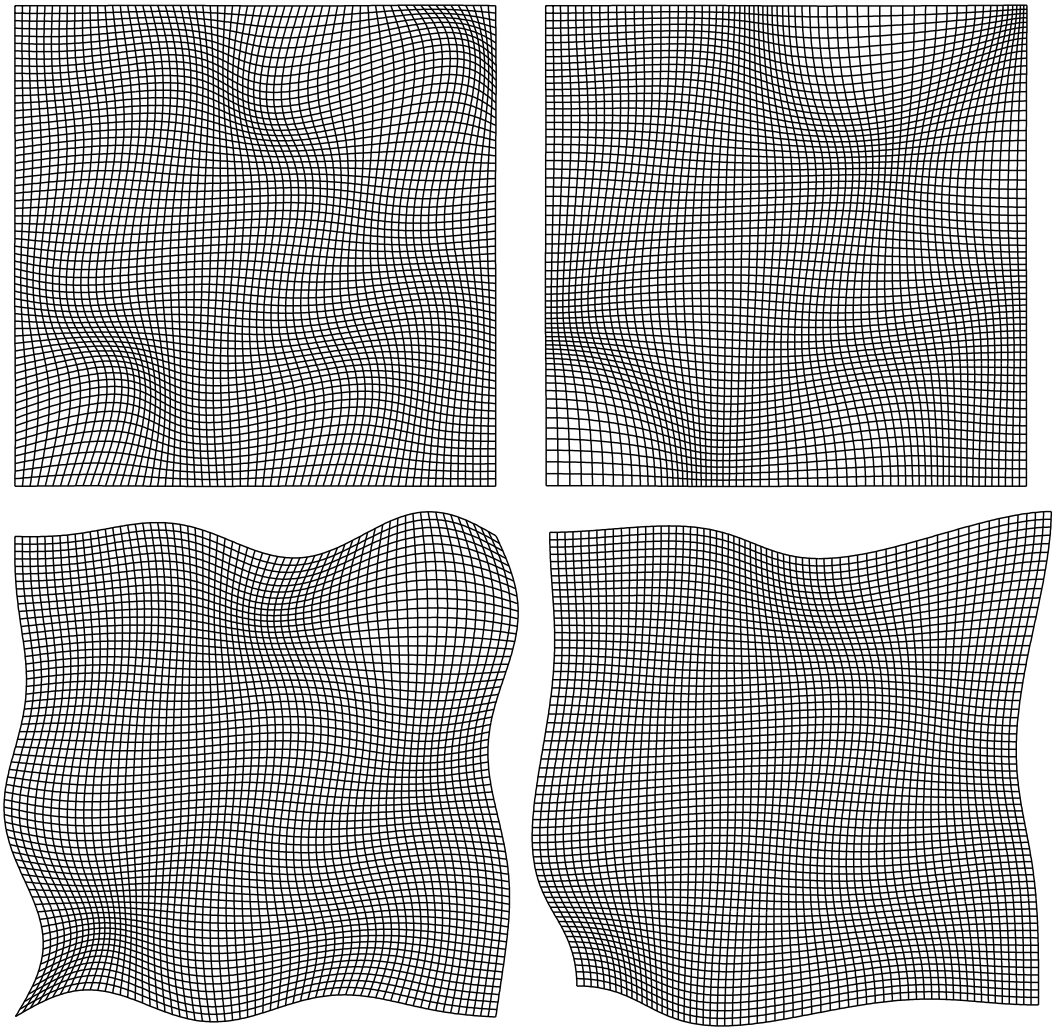

Figure 3.2: Different boundary conditions. Above left: fixed boundaries (generated purely from DST basis functions). Above right: sliding boundaries (from a mixture of DCT and DST basis functions). Below left: bending boundaries (from a different mixture of DCT and DST basis functions). Below right: free boundary conditions (purely from DCT basis functions). These deformation fields were all computed using the same 4×4 randomly generated coefficients (normal distribution of unit variance), and are assumed to cover a unit square.

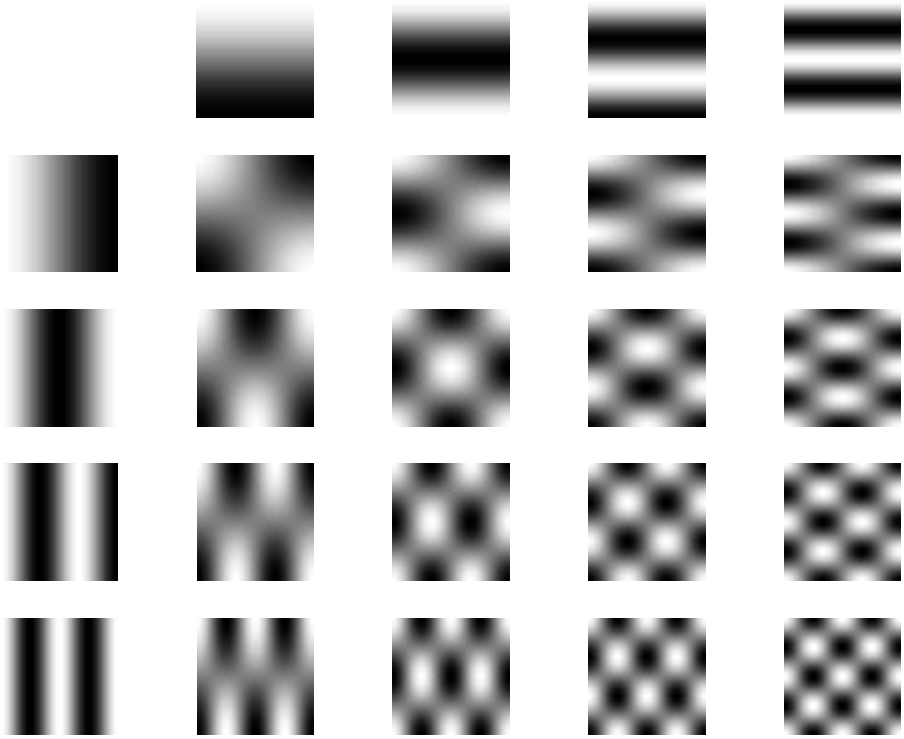

Figure 3.3: The lowest frequency basis functions of a two dimensional Discrete Cosine Transform.

$$\begin{aligned}\frac{\partial f(\mathbf{y}_i)}{\partial q_{j2}} &= \frac{\partial f(\mathbf{y}_i)}{\partial y_{2i}}\frac{\partial y_{2i}}{\partial q_{j2}} = \frac{\partial f(\mathbf{y}_i)}{\partial y_{2i}} d_j(\mathbf{x}_i) \\ \frac{\partial f(\mathbf{y}_i)}{\partial q_{j3}} &= \frac{\partial f(\mathbf{y}_i)}{\partial y_{3i}}\frac{\partial y_{3i}}{\partial q_{j3}} = \frac{\partial f(\mathbf{y}_i)}{\partial y_{3i}} d_j(\mathbf{x}_i)\end{aligned} \tag{3.15}$$

The approach involves iteratively computing $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}^T\mathbf{b}$. However, because there are many parameters to optimise, these computations can be very time consuming. There now follows a description of a very efficient way of computing these matrices.

A Fast Algorithm

A fast algorithm for computing $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}^T\mathbf{b}$ is shown in Figure 3.5. The remainder of this section explains the matrix terminology used, and why it is so efficient.

For simplicity, the algorithm is only illustrated in two dimensions, although it has been implemented to estimate warps in three dimensions. Images $\mathbf{f}$ and $\mathbf{g}$ are considered as $I \times J$ matrices $\mathbf{F}$ and $\mathbf{G}$ respectively. Row i of $\mathbf{F}$ is denoted by $\mathbf{f}_{i,:}$, and column j by $\mathbf{f}_{:,j}$. The basis functions used by the algorithm are generated from a separable form from matrices $\mathbf{D_1}$ and $\mathbf{D_2}$, with dimensions $I \times M$ and $J \times N$ respectively. By treating the transform coefficients as $M \times N$ matrices $\mathbf{Q_1}$ and $\mathbf{Q_2}$, the deformation fields can be rapidly constructed by computing $\mathbf{D_1 Q_1 D_2}^T$ and $\mathbf{D_1 Q_2 D_2}^T$.

Between each iteration, image $\mathbf{F}$ is resampled according to the latest parameter estimates. The derivatives of $\mathbf{F}$ are also resampled to give $\nabla_\mathbf{1}\mathbf{F}$ and $\nabla_\mathbf{2}\mathbf{F}$. The ith element of each of these matrices contain $f(\mathbf{y}_i)$, $\partial f(\mathbf{y}_i)/\partial y_{1i}$ and $\partial f(\mathbf{y}_i)/\partial y_{2i}$ respectively.

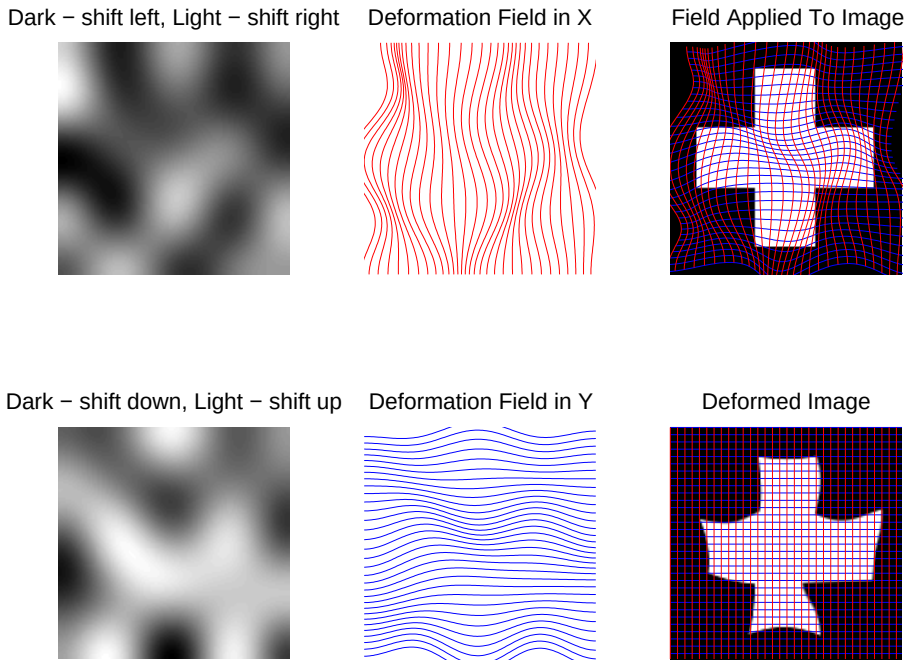

Figure 3.4: In two dimensions, a deformation field consists of two scalar fields. One for horizontal deformations, and the other for vertical deformations. The images on the left show deformations as a linear combination of basis images (see Figure 3.3). The centre column shows the same deformations in a more intuitive sense. The deformation is applied by overlaying it on a source image, and re-sampling (right).

$$\alpha = \begin{bmatrix} \mathbf{0} \end{bmatrix}$$
$$\beta = \begin{bmatrix} \mathbf{0} \end{bmatrix}$$

$$for \quad j = 1 \ldots J$$
$$\mathbf{C} = \mathbf{d_2}_{j,:}{}^T \mathbf{d_2}_{j,:}$$
$$\mathbf{E_1} = diag(\nabla_1 \mathbf{f}_{:,j}) \mathbf{D_1}$$
$$\mathbf{E_2} = diag(\nabla_2 \mathbf{f}_{:,j}) \mathbf{D_1}$$

$$\alpha = \alpha + \begin{bmatrix} \mathbf{C} \otimes (\mathbf{E_1}{}^T \mathbf{E_1}) & \mathbf{C} \otimes (\mathbf{E_1}{}^T \mathbf{E_2}) & -\mathbf{d_2}_{j,:}{}^T \otimes (\mathbf{E_1}{}^T \mathbf{g}_{:,j}) \\ (\mathbf{C} \otimes (\mathbf{E_1}{}^T \mathbf{E_2}))^T & \mathbf{C} \otimes (\mathbf{E_2}{}^T \mathbf{E_2}) & -\mathbf{d_2}_{j,:}{}^T \otimes (\mathbf{E_2}{}^T \mathbf{g}_{:,j}) \\ (-\mathbf{d_2}_{j,:}{}^T \otimes (\mathbf{E_1}{}^T \mathbf{g}_{:,j}))^T & (-\mathbf{d_2}_{j,:}{}^T \otimes (\mathbf{E_1}{}^T \mathbf{g}_{:,j}))^T & \mathbf{g}_{:,j}{}^T \mathbf{g}_{:,j} \end{bmatrix}$$

$$\beta = \beta + \begin{bmatrix} \mathbf{d_2}_{j,:}{}^T \otimes (\mathbf{E_1}{}^T (\mathbf{f}_{:,j} - w\mathbf{g}_{:,j})) \\ \mathbf{d_2}_{j,:}{}^T \otimes (\mathbf{E_2}{}^T (\mathbf{f}_{:,j} - w\mathbf{g}_{:,j})) \\ \mathbf{g}_{:,j}{}^T (\mathbf{f}_{:,j} - w\mathbf{g}_{:,j}) \end{bmatrix}$$

$$end$$

Figure 3.5: A two dimensional illustration of the fast algorithm for computing $\mathbf{A}^T\mathbf{A}$ (α) and $\mathbf{A}^T\mathbf{b}$ (β).

The notation $diag(\nabla_{\mathbf{1}}\mathbf{f}_{:,j})\mathbf{D_1}$ simply means multiplying each element of row i of $\mathbf{D_1}$ by $\nabla_{\mathbf{1}}\mathbf{f}_{i,j}$, and the symbol '$\otimes$' refers to the *Kronecker tensor product*. If $\mathbf{D_2}$ is a matrix of order $J \times N$, and $\mathbf{D_1}$ is a second matrix, then:

$$\mathbf{D_2} \otimes \mathbf{D_1} = \begin{bmatrix} d_{\mathbf{2}11}\mathbf{D_1} & \dots & d_{\mathbf{2}1N}\mathbf{D_1} \\ \vdots & \ddots & \vdots \\ d_{\mathbf{2}J1}\mathbf{D_1} & \dots & d_{\mathbf{2}JN}\mathbf{D_1} \end{bmatrix} \tag{3.16}$$

The advantage of the algorithm shown in Figure 3.5 is that it utilises some of the useful properties of Kronecker tensor products. This is especially important when the algorithm is implemented in three dimensions. The performance enhancement results from a reordering of a set of operations like $(\mathbf{D_3}\otimes\mathbf{D_2}\otimes\mathbf{D_1})^T(\mathbf{D_3}\otimes\mathbf{D_2}\otimes\mathbf{D_1})$, to the equivalent $(\mathbf{D_3}^T\mathbf{D_3})\otimes(\mathbf{D_2}^T\mathbf{D_2})\otimes(\mathbf{D_1}^T\mathbf{D_1})$. Assuming that the matrices $\mathbf{D_3}$, $\mathbf{D_2}$ and $\mathbf{D_1}$ all have order $I \times M$, then the number of floating point operations is reduced from $I^3M^3(M^3+2)$ to approximately $3I(M^2+M)+M^6$. If I equals 32, and M equals 4, then a performance increase of about a factor of 20,000 would be expected. The limiting factor to the algorithm is no longer the time taken to create the curvature matrix ($\mathbf{A}^T\mathbf{A}$), but is now the amount of memory required to store it and the time taken to invert it.

3.2.4 Linear Regularisation for Nonlinear Registration

Without regularisation in the nonlinear registration, it is possible to introduce unnecessary deformations that only reduce the residual sum of squares by a tiny amount. This could potentially make the algorithm very unstable. Regularisation is achieved by minimising the sum of squared difference between the template and the warped image, while simultaneously minimising some function of the deformation field. The principles are Bayesian and make use of the MAP scheme described in Section 3.2.1.

The first requirement for a MAP approach is to define some form of prior distribution for the parameters. For a simple linear[3] approach, the priors consist of an *a priori* estimate of the mean of the parameters (assumed to be zero), and also a covariance matrix describing the distribution of the parameters about this mean. There are many possible forms for these priors, each of which describes some form of 'energy' term. If the true prior distribution of the parameters is known (somehow derived from a large number of subjects), then $\mathbf{C_0}$ could be an empirically determined covariance matrix describing this distribution. This approach would have the advantage that the resulting deformations are more typically "brain like", and so increase the face validity of the approach.

The three distinct forms of linear regularisation that will now be described are based upon *membrane energy*, *bending energy* and *linear-elastic energy*. None of these schemes enforce a strict one to one mapping between the source and template images, but this makes little difference for the small deformations required here. Each of these models needs some form of elasticity constants (λ and sometimes μ). Values of these constants that are too large will provide too much regularisation and result in greatly underestimated deformations. If the values are too

[3] Although the cost function associated with these priors is quadratic, the priors are linear in the sense that they minimise the sum of squares of a linear combination of the model parameters. This is analogous to solving a set of linear equations by minimising a quadratic cost function.

small, there will not be enough regularisation and the resulting deformations will overfit the data. Section 7.3 will introduce one possible method of estimating what the best values for these constants should be.

Membrane Energy

The simplest model used for linear regularisation is based upon minimising the *membrane energy* of the deformation field $\mathbf{u}$ (Amit *et al.*, 1991; Gee *et al.*, 1997). By summing over i points in three dimensions, the membrane energy of $\mathbf{u}$ is given by:

$$\sum_i \sum_{j=1}^{3} \sum_{k=1}^{3} \lambda \left(\frac{\partial u_{ji}}{\partial x_{ki}} \right)^2 \tag{3.17}$$

where λ is simply a scaling constant. The membrane energy can be computed from the coefficients of the basis functions by $\mathbf{q_1}^T\mathbf{H}\mathbf{q_1} + \mathbf{q_2}^T\mathbf{H}\mathbf{q_2} + \mathbf{q_3}^T\mathbf{H}\mathbf{q_3}$, where $\mathbf{q_1}$, $\mathbf{q_2}$ and $\mathbf{q_3}$ refer to vectors containing the parameters describing translations in the three dimensions. The matrix $\mathbf{H}$ is defined by:

$$\begin{aligned} \mathbf{H} = \quad & \lambda \left(\dot{\mathbf{D}}_\mathbf{3}^T \dot{\mathbf{D}}_\mathbf{3} \right) \otimes \left(\mathbf{D_2}^T \mathbf{D_2} \right) \otimes \left(\mathbf{D_1}^T \mathbf{D_1} \right) \\ + \quad & \lambda \left(\mathbf{D_3}^T \mathbf{D_3} \right) \otimes \left(\dot{\mathbf{D}}_\mathbf{2}^T \dot{\mathbf{D}}_\mathbf{2} \right) \otimes \left(\mathbf{D_1}^T \mathbf{D_1} \right) \\ + \quad & \lambda \left(\mathbf{D_3}^T \mathbf{D_3} \right) \otimes \left(\mathbf{D_2}^T \mathbf{D_2} \right) \otimes \left(\dot{\mathbf{D}}_\mathbf{1}^T \dot{\mathbf{D}}_\mathbf{1} \right) \end{aligned} \tag{3.18}$$

where the notation $\dot{\mathbf{D}}_\mathbf{1}$ refers to the first derivatives of $\mathbf{D_1}$.

Assuming that the parameters consist of $\left[\mathbf{q_1}^T \mathbf{q_2}^T \mathbf{q_3}^T w\right]^T$, matrix $\mathbf{C_0}^{-1}$ from Eqn. 3.6 can be constructed from $\mathbf{H}$ by:

$$\mathbf{C_0}^{-1} = \begin{bmatrix} \mathbf{H} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & 0 \end{bmatrix} \tag{3.19}$$

$\mathbf{H}$ is all zeros, except for the diagonal. Elements on the diagonal represent the reciprocal of the *a priori* variance of each parameter. If all the DCT matrices are $I \times M$, then each diagonal element is given by:

$$\begin{aligned} h_{j+M(k-1+M(l-1))} &= \lambda \pi^2 I^{-2} \left((j-1)^2 + (k-1)^2 + (l-1)^2 \right) \\ &\text{over } j = 1 \ldots M,\ k = 1 \ldots M \text{ and } l = 1 \ldots M. \end{aligned} \tag{3.20}$$

The nonlinear registration algorithm described here is implemented in three dimensions using membrane energy as the cost functions. For completeness, the other two commonly used cost functions will now be described, but only for the two dimensional case.

Bending Energy

Bookstein's thin plate splines (1997b; 1997a) minimise the *bending energy* of deformations. For a two dimensional deformation, the bending energy is defined by:

$$\lambda \sum_i \left(\left(\frac{\partial^2 u_{1i}}{\partial x_{1i}^2} \right)^2 + \left(\frac{\partial^2 u_{1i}}{\partial x_{2i}^2} \right)^2 + 2 \left(\frac{\partial^2 u_{1i}}{\partial x_{1i} \partial x_{2i}} \right)^2 \right) \quad +$$

$$\lambda \sum_i \left(\left(\frac{\partial^2 u_{2i}}{\partial x_{1i}^2} \right)^2 + \left(\frac{\partial^2 u_{2i}}{\partial x_{2i}^2} \right)^2 + 2 \left(\frac{\partial^2 u_{2i}}{\partial x_{1i} \partial x_{2i}} \right)^2 \right) \tag{3.21}$$

This can be computed by:

$$\begin{aligned} \lambda \mathbf{q_1}^T (\ddot{\mathbf{D}}_\mathbf{2} \otimes \mathbf{D_1})^T (\ddot{\mathbf{D}}_\mathbf{2} \otimes \mathbf{D_1}) \mathbf{q_1} &+ \lambda \mathbf{q_1}^T (\mathbf{D_2} \otimes \ddot{\mathbf{D}}_\mathbf{1})^T (\mathbf{D_2} \otimes \ddot{\mathbf{D}}_\mathbf{1}) \mathbf{q_1} + \\ 2\lambda \mathbf{q_1}^T (\dot{\mathbf{D}}_\mathbf{2} \otimes \dot{\mathbf{D}}_\mathbf{1})^T (\dot{\mathbf{D}}_\mathbf{2} \otimes \dot{\mathbf{D}}_\mathbf{1}) \mathbf{q_1} &+ \lambda \mathbf{q_2}^T (\ddot{\mathbf{D}}_\mathbf{2} \otimes \mathbf{D_1})^T (\ddot{\mathbf{D}}_\mathbf{2} \otimes \mathbf{D_1}) \mathbf{q_2} + \\ \lambda \mathbf{q_2}^T (\mathbf{D_2} \otimes \ddot{\mathbf{D}}_\mathbf{1})^T (\mathbf{D_2} \otimes \ddot{\mathbf{D}}_\mathbf{1}) \mathbf{q_2} &+ 2\lambda \mathbf{q_2}^T (\dot{\mathbf{D}}_\mathbf{2} \otimes \dot{\mathbf{D}}_\mathbf{1})^T (\dot{\mathbf{D}}_\mathbf{2} \otimes \dot{\mathbf{D}}_\mathbf{1}) \mathbf{q_2} \end{aligned} \tag{3.22}$$

where the notation $\dot{\mathbf{D}}_\mathbf{1}$ and $\ddot{\mathbf{D}}_\mathbf{1}$ refer to the column-wise first and second derivatives of $\mathbf{D_1}$. This is simplified to $\mathbf{q_1}^T \mathbf{H} \mathbf{q_1} + \mathbf{q_2}^T \mathbf{H} \mathbf{q_2}$ where:

$$\mathbf{H} = \lambda \left(\left(\ddot{\mathbf{D}}_\mathbf{2}^T \ddot{\mathbf{D}}_\mathbf{2} \right) \otimes \left(\mathbf{D_1}^T \mathbf{D_1} \right) + \left(\mathbf{D_2}^T \mathbf{D_2} \right) \otimes \left(\ddot{\mathbf{D}}_\mathbf{1}^T \ddot{\mathbf{D}}_\mathbf{1} \right) + 2 \left(\dot{\mathbf{D}}_\mathbf{2}^T \dot{\mathbf{D}}_\mathbf{2} \right) \otimes \left(\dot{\mathbf{D}}_\mathbf{1}^T \dot{\mathbf{D}}_\mathbf{1} \right) \right) \tag{3.23}$$

Matrix $\mathbf{C_0}^{-1}$ from Eqn. 3.6 can be constructed from $\mathbf{H}$ as:

$$\mathbf{C_0}^{-1} = \begin{bmatrix} \mathbf{H} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & 0 \end{bmatrix} \tag{3.24}$$

with values on the diagonals of $\mathbf{H}$ given by:

$$h_{j+(k-1)\times M} = \lambda \left(\left(\frac{\pi(j-1)}{I} \right)^4 + \left(\frac{\pi(k-1)}{I} \right)^4 + 2 \left(\frac{\pi(j-1)}{I} \right)^2 \left(\frac{\pi(k-1)}{I} \right)^2 \right) \quad \text{over } j = 1 \ldots M \text{ and } k = 1 \ldots M \tag{3.25}$$

Linear-Elastic Energy

The *linear-elastic* energy (Miller *et al.*, 1993) of a two dimensional deformation field is:

$$\sum_{j=1}^{2} \sum_{k=1}^{2} \sum_i \frac{\lambda}{2} \left(\frac{\partial u_{ji}}{\partial x_{ji}} \right) \left(\frac{\partial u_{ki}}{\partial x_{ki}} \right) + \frac{\mu}{4} \left(\frac{\partial u_{ji}}{\partial x_{ki}} + \frac{\partial u_{ki}}{\partial x_{ji}} \right)^2 \tag{3.26}$$

where λ and μ are the *Lamé* elasticity constants. The elastic energy of the deformations can be computed by:

$$\begin{aligned} (\mu + \lambda/2) \mathbf{q_1}^T (\mathbf{D_2} \otimes \dot{\mathbf{D}}_\mathbf{1})^T (\mathbf{D_2} \otimes \dot{\mathbf{D}}_\mathbf{1}) \mathbf{q_1} + (\mu + \lambda/2) \mathbf{q_2}^T (\dot{\mathbf{D}}_\mathbf{2} \otimes \mathbf{D_1})^T (\dot{\mathbf{D}}_\mathbf{2} \otimes \mathbf{D_1}) \mathbf{q_2} \\ + \mu/2 \mathbf{q_1}^T (\dot{\mathbf{D}}_\mathbf{2} \otimes \mathbf{D_1})^T (\dot{\mathbf{D}}_\mathbf{2} \otimes \mathbf{D_1}) \mathbf{q_1} + \mu/2 \mathbf{q_2}^T (\mathbf{D_2} \otimes \dot{\mathbf{D}}_\mathbf{1})^T (\mathbf{D_2} \otimes \dot{\mathbf{D}}_\mathbf{1}) \mathbf{q_2} \\ + \mu/2 \mathbf{q_1}^T (\dot{\mathbf{D}}_\mathbf{2} \otimes \mathbf{D_1})^T (\mathbf{D_2} \otimes \dot{\mathbf{D}}_\mathbf{1}) \mathbf{q_2} + \mu/2 \mathbf{q_2}^T (\mathbf{D_2} \otimes \dot{\mathbf{D}}_\mathbf{1})^T (\dot{\mathbf{D}}_\mathbf{2} \otimes \mathbf{D_1}) \mathbf{q_1} \\ + \lambda/2 \mathbf{q_1}^T (\mathbf{D_2} \otimes \dot{\mathbf{D}}_\mathbf{1})^T (\dot{\mathbf{D}}_\mathbf{2} \otimes \mathbf{D_1}) \mathbf{q_2} + \lambda/2 \mathbf{q_2}^T (\dot{\mathbf{D}}_\mathbf{2} \otimes \mathbf{D_1})^T (\mathbf{D_2} \otimes \dot{\mathbf{D}}_\mathbf{1}) \mathbf{q_1} \end{aligned} \tag{3.27}$$

A regularisation based upon this model requires an inverse covariance matrix ($\mathbf{C_0}^{-1}$) that is not a simple diagonal matrix. This matrix is constructed as follows:

$$\mathbf{C_0}^{-1} = \begin{bmatrix} \mathbf{H_1} & \mathbf{H_3} & \mathbf{0} \\ \mathbf{H_3}^T & \mathbf{H_2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & 0 \end{bmatrix} \tag{3.28}$$

where:

$$\begin{aligned}
\mathbf{H_1} = &\quad (\mu+\lambda/2)(\mathbf{D_2}^T\mathbf{D_2}) \otimes (\mathbf{D_1}^T\mathbf{D_1}) + \mu/2(\mathbf{D_2}^T\mathbf{D_2}) \otimes (\mathbf{D_1}^T\mathbf{D_1}) \\
\mathbf{H_2} = &\quad (\mu+\lambda/2)(\mathbf{D_2}^T\mathbf{D_2}) \otimes (\mathbf{D_1}^T\mathbf{D_1}) + \mu/2(\mathbf{D_2}^T\mathbf{D_2}) \otimes (\mathbf{D_1}^T\mathbf{D_1}) \\
\mathbf{H_3} = &\quad \lambda/2(\mathbf{D_2}^T\mathbf{D_2}) \otimes (\mathbf{D_1}^T\mathbf{D_1}) + \mu/2(\mathbf{D_2}^T\mathbf{D_2}) \otimes (\mathbf{D_1}^T\mathbf{D_1})
\end{aligned} \tag{3.29}$$

3.2.5 Templates and Intensity Transformations

Sections 3.2.2 and 3.2.3 have modelled a single intensity scaling parameter (q_{13} and w respectively), but more generally, the optimisation can be assumed to minimise two sets of parameters: those that describe spatial transformations ($\mathbf{q_s}$), and those for describing intensity transformations ($\mathbf{q_t}$). This means that the difference function can be expressed in the generic form:

$$b_i(\mathbf{q}) = f(\mathbf{s}(\mathbf{x}_i, \mathbf{q_s})) - t(\mathbf{x}_i, \mathbf{q_t}) \tag{3.30}$$

where $\mathbf{f}$ is the source image, $\mathbf{s}()$ is a vector function describing the spatial transformations based upon parameters $\mathbf{q_s}$ and $t()$ is a scalar function describing intensity transformations based on parameters $\mathbf{q_t}$. $\mathbf{x}_i$ represents the co-ordinates of the ith sampled point.

The previous subsections simply considered matching one image to a scaled version of another, in order to minimise the sum of squared differences between them. For this case, $t(\mathbf{x}_i, \mathbf{q_t})$ is simply equal to $q_{t1}g(\mathbf{x}_i)$, where q_{t1} is a simple scaling parameter and $\mathbf{g}$ is a template image. This is most effective when there is a linear relation between the image intensities. Typically, the template images used for spatial normalisation will be similar to those shown in the top row of Figure 3.6. The simplest least squares fitting method is not optimal when there is not a linear relationship between the images. Examples of nonlinear relationships are illustrated in Figure 3.7, which shows histograms (scatter-plots) of image intensities plotted against each other.

An important idea is that a given image can be matched not to one reference image, but to a series of images that all conform to the same space. The idea here is that (ignoring the spatial differences) any given image can be expressed as a linear combination of a set of reference images. For example these reference images might include different modalities (e.g., PET, SPECT, ^{18}F-DOPA, ^{18}F-deoxy-glucose, T_1-weighted MRI T_2^*-weighted MRI .. etc.) or different anatomical tissues (e.g., grey matter, white matter, and CSF segmented from the same T_1-weighted MRI) or different anatomical regions (e.g., cortical grey matter, sub-cortical grey mater, cerebellum ... etc.) or finally any combination of the above. Any given image, irrespective of its modality could be approximated with a function of these images. A simple example using two images would be:

$$\sum_i \left(f(\mathbf{M}\mathbf{x}_i) - (q_{t1}g_1(\mathbf{x}_i) + q_{t2}g_2(\mathbf{x}_i))\right)^2 \tag{3.31}$$

In Figure 3.8, a plane of a T1 weighted MRI is modelled by a linear combination of the five other template images shown in Figure 3.6. Similar models were used to simulate T2 and PD weighted MR images. The linearity of the scatter-plots (compared to those in Figure 3.7) shows that MR images of a wide range of different contrasts can be modelled by a linear combination of a limited number of template images. Visual inspection shows that the simulated images are very similar to those shown in Figure 3.6.

Alternatively, the intensities could vary spatially (for example due to inhomogeneities in the MRI scanner). Linear variations in intensity over the field of view can be accounted for by

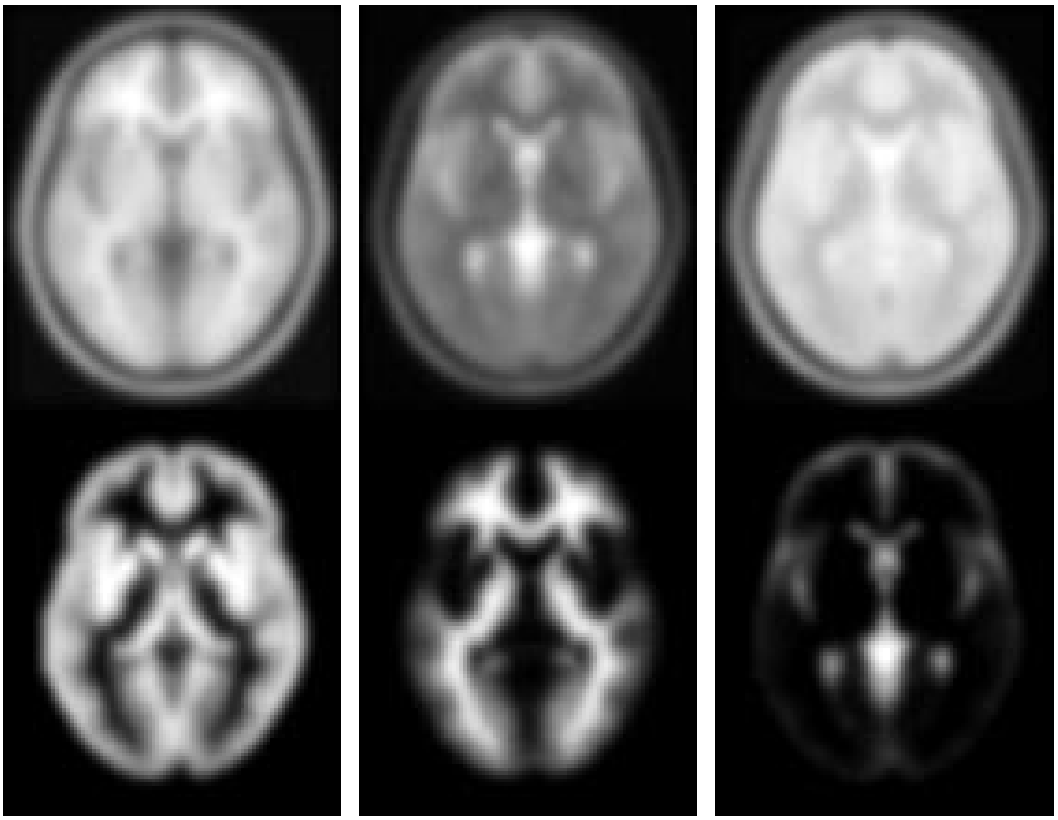

Figure 3.6: Example template images. Above: T1 weighted MRI, T2 weighted MRI and PD weighted MRI. Below: Grey matter probability distribution, White matter probability distribution and CSF probability distribution. All the data were generated at the McConnel Brain Imaging Centre, Montréal Neurological Institute at McGill University, and are based on the averages of about 150 normal brains. The original images were reduced to 2mm resolution and convolved with an 8mm FWHM Gaussian kernel to be used as templates for spatial normalisation.

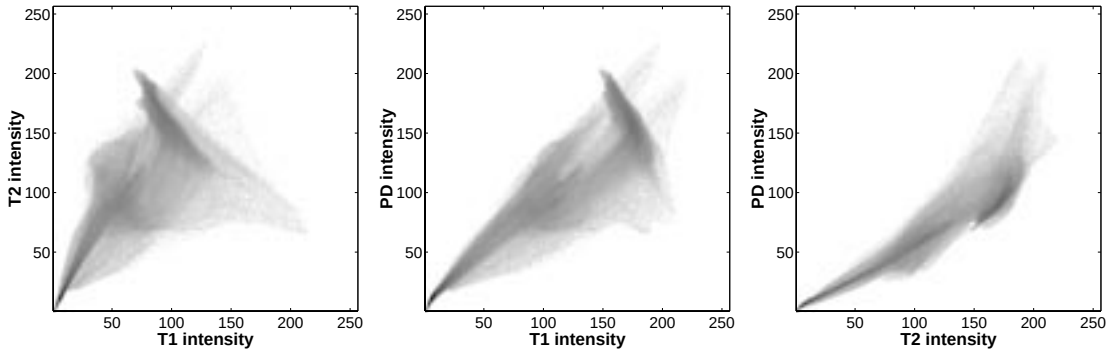

Figure 3.7: Two dimensional histograms of template images (intensities shown as $log(1+n)$, where n is the value in each bin). The histograms were based on the whole volumes of the template images shown in the top row of Figure 3.6.

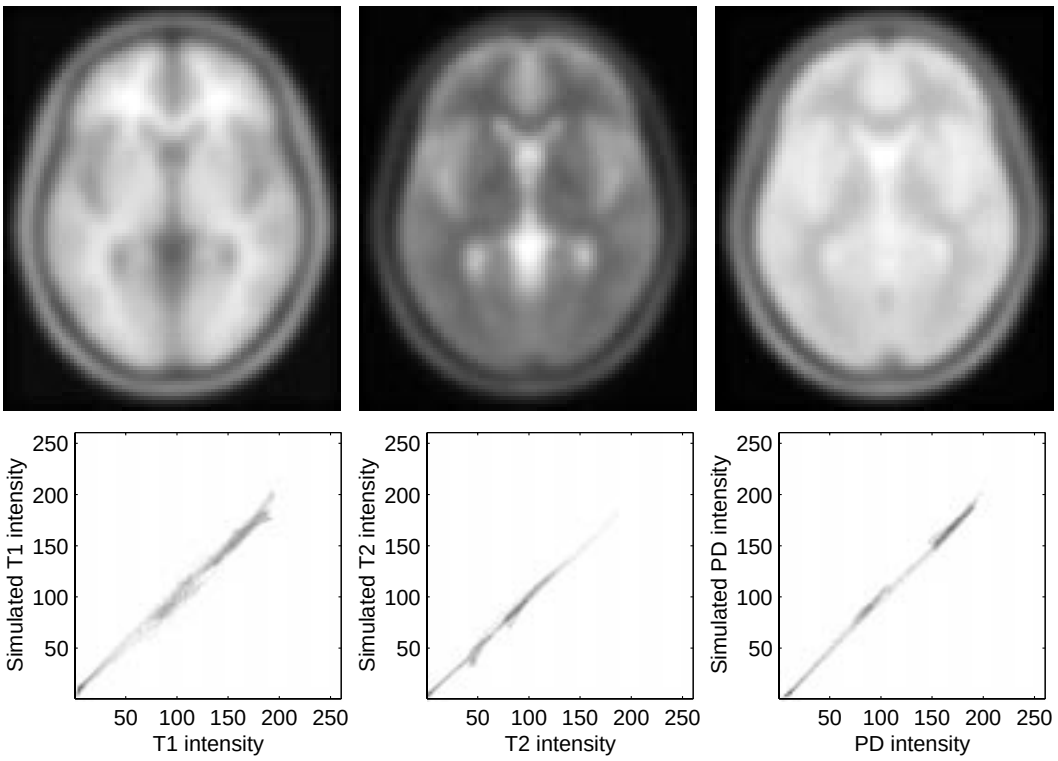

Figure 3.8: Simulated images of T1, T2 and PD weighted images, and histograms of the real images versus the simulated images.

optimising a function of the form:

$$\sum_i \left(f(\mathbf{x}_i, \mathbf{q_s}) - (q_{t1}g(\mathbf{x}_i) + q_{t2}x_{1i}g(\mathbf{x}_i) + q_{t3}x_{2i}g(\mathbf{x}_i) + q_{t4}x_{3i}g(\mathbf{x}_i))\right)^2 \tag{3.32}$$

More complex variations could be included by modulating with other basis functions (such as the DCT basis function set described in Section 3.2.3) (Friston *et al.*, 1995c). The examples shown so far have been linear in their parameters describing intensity transformations. A simple example of an intensity transformation that is nonlinear would be:

$$\sum_i \left(f(\mathbf{x}_i, \mathbf{q_s}) - q_{t1}g(\mathbf{x}_i)^{q_{t2}}\right)^2 \tag{3.33}$$

Collins *et al.*(1994b) suggested that – rather than matching the image itself to the template – some function of the image should be matched to a template image transformed in the same way. He found that the use of gradient magnitude transformations lead to more robust solutions, especially in cases of partial volume coverage or intensity inhomogeneity artifacts (in MR images). Other spatially invariant moments may also contain other useful matching information. The algorithms described here perform most efficiently with smooth images. Much of the high frequency information in the images is lost in the smoothing step, but information about important image features may be retained in separate (smoothed) moment images. Simultaneous registrations (comparable to those in the previous chapter that matched grey matter with grey matter, and white matter with white matter) using these extracted features may be a useful technique for preserving information, while still retaining the advantages of using smooth images in the registration.

Another idea for introducing more accuracy would be to simultaneously spatially normalise co-registered images to corresponding templates. For example, by simultaneously matching a PET image to a PET template, at the same time as matching a structural MR image to a corresponding MR template, more accuracy could be obtained than by matching the images individually. Section 2.6 described a method of between modality registration where the first step involves simultaneous affine registration of a pair of images to a pair of templates in order to extract the rigid body transformation that maps the images together. There is no reason why nonlinear warping can not also be included in this model to further increase the accuracy of the rigid registration, while also improving the spatial normalisation.

3.3 Evaluation

The criteria for 'good' spatial transformations can be framed in terms of validity, reliability and computational efficiency. The validity of a particular transformation device is not easy to define or measure and indeed varies with the application. For example a rigid body transformation may be perfectly valid for realignment but not for spatial normalisation of an arbitrary brain into a standard stereotactic space. Generally the sorts of validity that are important in spatial transformations can be divided into (i) *Face validity*, established by demonstrating the transformation does what it is supposed to and (ii) *Construct validity*, assessed by comparison with other techniques or constructs. Face validity is a complex issue in functional mapping. At first glance, face validity might be equated with the co-registration of anatomical homologues in two images. This would be complete and appropriate if the biological question referred to structural differences or modes of variation. In other circumstances however this definition of face validity is not appropriate. For example, the purpose of spatial normalisation (either within or between subjects) in functional mapping studies is to maximise the sensitivity to neurophysiological change elicited by experimental manipulation of sensorimotor or cognitive state. In this case a better definition of a valid normalisation is that which maximises condition-dependent effects with respect to error (and if relevant inter-subject) effects. This will probably be effected when functional anatomy is congruent. This may or may not be the same as registering structural anatomy.

One approach is illustrated for assessing validity by comparing affine registrations both with and without the incorporation of the MAP scheme. It was found that the affine transformations derived using the Bayesian scheme are much more robust, and that the rate of convergence is greater. The final part of the evaluations illustrate that the nonlinear registration reduces structural variability on a global scale.

3.3.1 Evaluation of the MAP Scheme for Affine Registration

The affine registration scheme relies on optimising a set of 12 parameters that define the overall size and position of the head. The optimisation searches for the closest local minimum to the initial estimates, so it relies on these starting estimates being reasonably close to the optimum solution. In practice, this should not be problem. For PET images of the brain, the position of the subject within the scanner should not vary greatly. Also, the images are almost exclusively reconstructed in the same transverse orientation. Once a suitable set of starting estimates for one subject has been determined, it should be possible to use the same one for all subsequent

subjects. The situation is slightly more complicated for MR images, where the images can be acquired in any orientation. However, in most medical image format standards, the orientation and position of the images is stored within the headers. This information can be automatically extracted and used as starting estimates for the registration.

The MAP optimisation scheme from Section 3.2.1 was evaluated for affine registrations with respect to the same optimisations performed without using the MAP scheme. It was found to converge more rapidly to a good solution, and also give much more robust and reliable results for limited data. These evaluations are detailed below.

Plots of convergence – with and without Bayesian extension

The affine registration algorithm was applied to 100 T1 weighted images, in order to match the images to a T1 template image. All images were smoothed with a Gaussian kernel of 8mm full width at half maximum. The voxels were reduced to $2 \times 2 \times 4$mm with a field of view of $256 \times 256 \times 128$mm in x, y and z respectively, in order to facilitate faster computations.

The optimisations were performed three times: (A) Without the Bayesian scheme, for a 12 parameter affine transformation. (B) With the Bayesian scheme, for a 12 parameter affine transformation. (C) Without the Bayesian scheme, for a six parameter rigid body transformation (to demonstrate that the Bayesian scheme is not simply optimising a rigid body transformation).

During the optimisation procedure, the images were sampled approximately every 8mm. 32 iterations were used, and the residual variance (χ^2) recorded for each iteration. 50 of the subjects were given good starting estimates (i), and 50 were given starting estimates that deviated from the optimal solution by about 10cm (ii).

There were 2 cases from (ii) in which the starting estimates were insufficiently close to the solution, for either (A) or (B) to converge satisfactorily. These cases have been excluded from the results.

Figure 3.9 shows the average χ^2 for all images plotted against iteration number. As can be seen from these plots, (B) leads to a more rapid estimation of the optimal parameters, even though convergence appears faster at the start of (A). The plot of convergence for (C) illustrates the point that the Bayesian method is not over-constrained and simply optimising a set of rigid body parameters.

Figure 3.10 compares the number of iterations required by (A) and (B) in order to reduce the χ^2 to within 1% of the minimum of both schemes. In several cases of (A), the optimisation had not converged within the 32 iterations. There were only 5 cases where (B) does not obtain a value for χ^2 that is as low as that from (A). In two of the cases, the results from (A) were very close to those from (B). However, in the other three cases, examination of the parameter estimates from scheme (A) showed that it had found a minimum that was clearly not a proper solution. The zooms determined, after 32 iterations, were (0.96,0.98,0.11), (2.10,0.72,0.0003) and (1.09,0.24,0.02). These are clearly not correct!

The algorithm requires relatively few iterations to reach convergence. The speed of each iteration for the affine normalisation depends upon the number of sampled voxels. On a SPARC Ultra 2, an iteration takes one second when about 26000 points (and their gradients) are sampled using tri-linear interpolation.

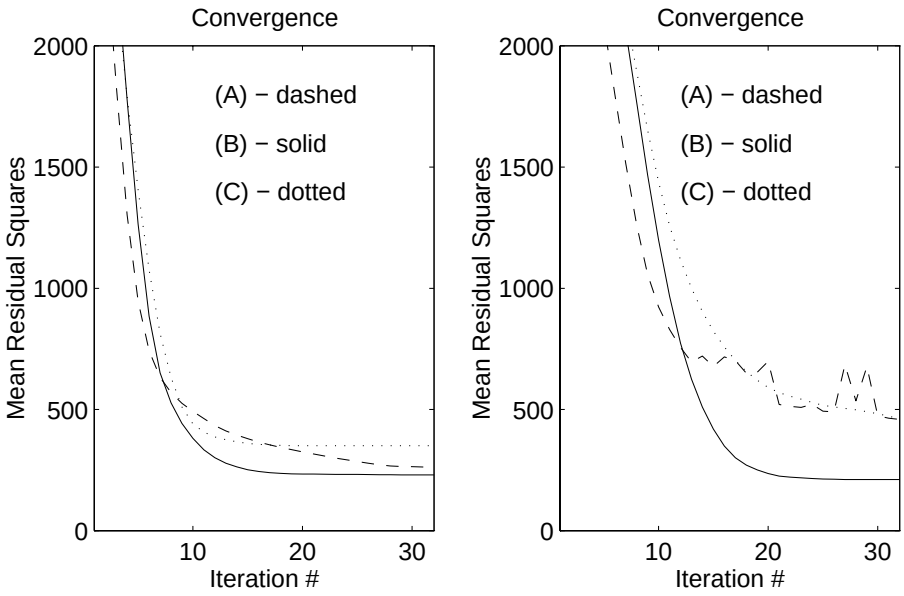

Figure 3.9: The average χ^2 for the images plotted against iteration number. Left: given good starting estimates (i). Right: given poor starting estimates (ii). The dashed lines (A) show convergence for a 12 parameter affine transformation without using the Bayesian scheme. The solid lines (B) show the same, but with the Bayesian scheme. Convergence for a six parameter rigid body transformation (C) is shown in the dotted lines.

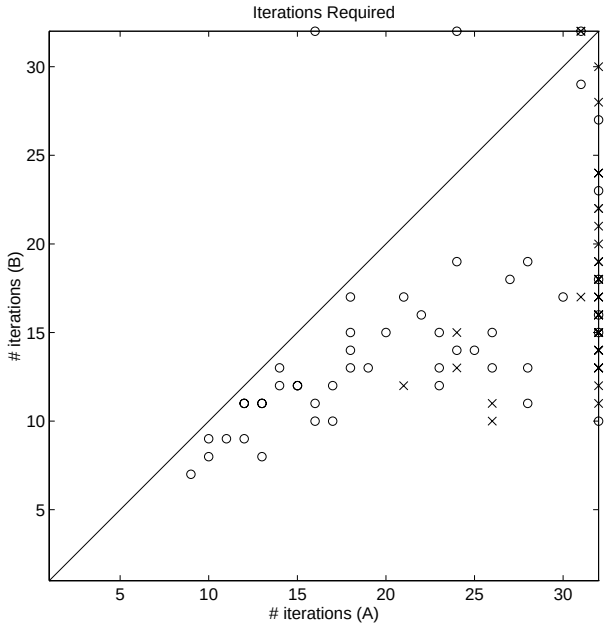

Figure 3.10: The number of iterations in which convergence to within 1% of the minimum mean residual sum of squares had not been reached. The non-Bayesian scheme (A) is on the X axis, with the Bayesian scheme (B) on the Y axis. Results from optimisations given good starting estimates are shown as circles, whereas those with bad starting estimates are shown as crosses.

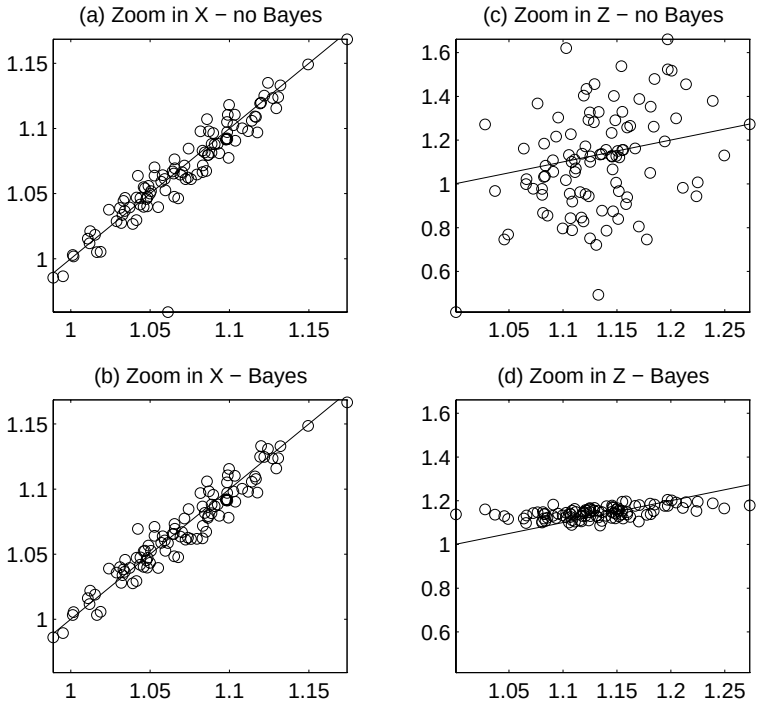

Figure 3.11: Plots of the parameter estimates from reduced data, against estimates using the complete data. As expected, the Bayesian scheme makes little difference for the estimates of the zoom in the X direction [(a) and (b)], whereas the Bayesian scheme heavily biases the zoom in Z towards the mean of the prior distribution [(c) and (d)].

Comparisons of Affine Normalisation with Limited Data

Occasionally the image that is to be spatially normalised is of poor quality. It may have a low signal to noise ratio, or it may contain only a limited number of slices. When this is the case, the parameter estimates for the spatial normalisation are likely to be unreliable. Here, a further comparison of affine registrations, with and without the incorporation of prior information [(E) and (D) respectively], is presented. This time, only four planes from the images were sampled, to simulate an effective field of view of 16 mm. The optimisations were given good initial parameter estimates, and the results compared with those obtained using the complete data.

The resulting parameter estimates from (D) and (E) are plotted against those from (B) in Figure 3.11. As can be seen from the plots, where the parameters can be estimated accurately, the results from (D) and (E) are similar. However, where there is not enough information in the images to determine an accurate parameter estimate, the results of (E) are properly biased towards the prior estimate.

3.3.2 Comparing Spatial Normalisation both With and Without Nonlinear Deformations

This section provides an anecdotal evaluation of the nonlinear warping techniques. Spatial normalisation is compared both with and without nonlinear deformations, and nonlinear deforma-

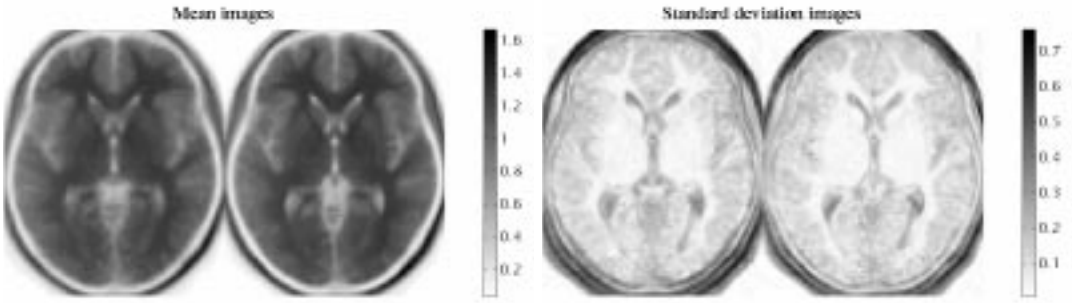

Figure 3.12: Means and standard deviations of spatially normalised T1 weighted images from 12 subjects. The images on the left of each pair were derived using only affine registration. Those on the right used nonlinear registration in addition to the affine registration.

tions compared with and without using Bayesian priors.

T1 weighted MR images of 12 subjects were spatially normalised to the same anatomical space. The normalisations were performed twice, first using only 12 parameter affine transformations and then using affine transformations followed by nonlinear warps. In both cases, the registration was weighted using the image shown in Figure 6.5 (page 129), so that any confounding effects of skull and scalp differences could be discounted. The nonlinear transformation used 392 ($7\times8\times7$) parameters to describe deformations in each of the directions, and four parameters to model a linear scaling and simple linear image intensity inhomogeneities (making a total of 1180 parameters in all). The basis functions were those of a three dimensional DCT, and the regularisation minimised the membrane energy of the deformation fields (using a value of 0.01 for λ). Twelve iterations of the nonlinear registration algorithm were performed.

Figure 3.12 shows pixel by pixel means and standard deviations of the normalised images. In order to create these mean and standard deviation images, the spatially normalised images were first scaled such that each of their weighted mean intensities was unity, where the weighting was done using the image in Figure 6.5. The mean image from the nonlinear normalisation shows more contrast and has edges that are slightly sharper. The standard deviation image for the nonlinear normalisation shows decreased intensities, demonstrating that the intensity differences between the images have been reduced. However, the differences tend to reflect changes in the global shape of the heads, rather than differences between the cortical anatomy. More examples of affine versus nonlinearly warped images are shown in Figures 4.15 and 4.16 of the next chapter. The average weighted residual squared difference between the normalised images and the mean image of the group was computed. Again, the weighting was done so that the residual squared differences were derived primarily from voxels in the brain. The average squared difference was 0.0237 for the affine only normalised images and 0.0187 for those that had also been nonlinearly warped. This shows that a 20% reduction of residual variance can be achieved by following an affine registration by the nonlinear warping described here.

This evaluation should illustrate the fact that nonlinear normalisation clearly reduces the sum of squared intensity differences between the images. The amount of residual variance could have been reduced further by decreasing the amount of regularisation. This however, may lead to some very un-natural looking distortions being introduced, due to an over-estimation of the *a priori* variability.

Evaluations like this tend to show more favourable results for less heavily regularised algorithms. With less regularisation, the optimum solution is based more upon minimising the difference between the images, and less upon knowledge of the *a priori* distribution of the parameters. This is illustrated for a single subject in Figure 3.13, where the distortions of gyral anatomy clearly have a very low face validity (lower right panel).

3.4 Discussion

Because the deformations are only defined by a few hundred parameters, the nonlinear registration method described here does not have the potential precision of some other methods. High frequency deformations can not be modelled because the deformations are restricted to the lowest spatial frequencies of the basis functions. This means that the current approach is unsuitable for attempting exact matches between fine cortical structures.

The current method is relatively fast, (taking in the order of 30 seconds per iteration – depending upon the number of basis functions used). The speed is partly a result of the small number of parameters involved, and the simple optimisation algorithm that assumes an almost quadratic error surface. Because the images are first matched using a simple affine transformation, there is less 'work' for the algorithm to do, and a good registration can be achieved with only a few iterations (less than 20). The method does not rigorously enforce a one-to-one match between the brains being registered. However, by estimating only the lowest frequency deformations and by using appropriate regularisation, this constraint is rarely broken.

The approach in this chapter searches for a MAP estimate of the parameters defining the warps. However, optimisation problems for complex nonlinear models such as those used for image registration can easily get caught in local minima, so there is no guarantee that the estimate determined by the algorithm is globally optimum. Even if the best MAP estimate is achieved, there will be many other potential solutions that have similar probabilities of being correct. A further complication arises from the fact that there is no one-to-one match between the small structures (especially gyral and sulcal patterns) of any two brains. This means that it is not possible to obtain a single objective high frequency match however good an algorithm is for determining the best MAP estimate. Because of these issues, registration using the minimum variance estimate (MVE) may be more appropriate. Rather than searching for the single most probable solution, the MVE is the average of all possible solutions, weighted by their individual probabilities of being correct. Although useful approximations have been devised (Miller *et al.*, 1993; Christensen, 1994), this estimate is still difficult to achieve in practice because of the enormous amount of computing power required. The MVE is probably more appropriate than the MAP estimate for spatial normalisation, as it is (on average) closer to the "true" solution. However, if the errors associated with the parameter estimates and also the priors are normally distributed, then the MVE and the MAP estimate are identical. This is partially satisfied by smoothing the images before registering them.

When higher spatial frequency warps are to be fitted, more DCT coefficients are required to describe the deformations. There are practical problems that occur when more than about the $8 \times 8 \times 8$ lowest frequency DCT components are used. One of these is the problem of storing and inverting the curvature matrix ($\mathbf{A}^T\mathbf{A}$). Even with deformations limited to $8 \times 8 \times 8$ coefficients,

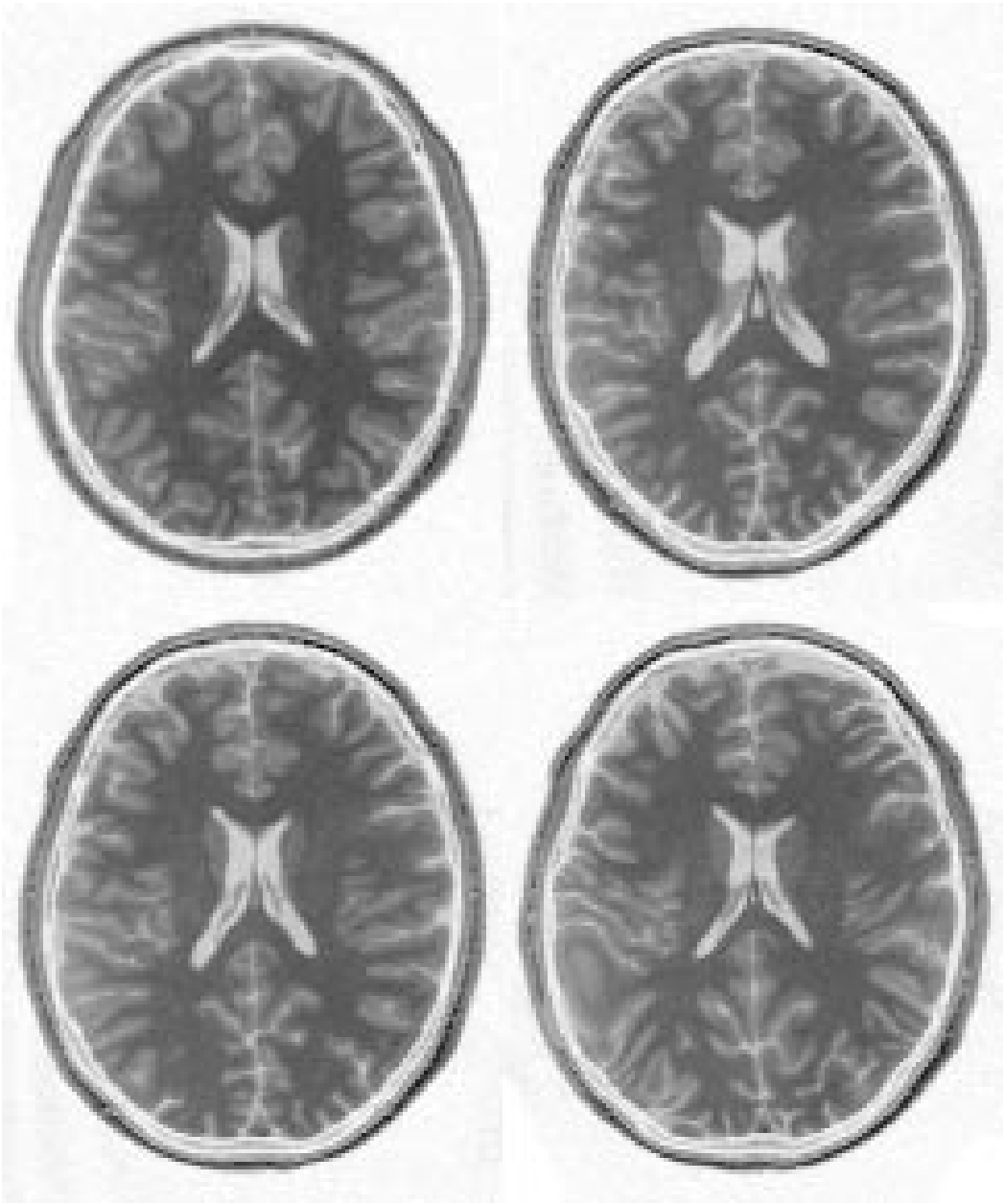

Figure 3.13: The image shown at the top-left is the template image. At the top-right is an image that has been registered with it using a 12-parameter affine registration. The image at the bottom-left is the same image registered using the 12-parameter affine registration, followed by a regularised global nonlinear registration (using 1180 parameters, 12 iterations and a λ of 0.01). It should be clear that the shape of the image approaches that of the template much better after nonlinear registration. At the bottom right is the image after the same affine transformation and nonlinear registration, but this time without using any regularisation. The mean squared difference between the image and template after the affine registration was 472.1. After the regularised nonlinear registration this was reduced to 302.7. Without regularisation, a mean squared difference of 287.3 is achieved, but this is at the expense of introducing a lot of unnecessary warping.

there are at least 1537 unknown parameters, requiring a curvature matrix of about 18Mbytes (using double precision floating point arithmetic). High-dimensional registration methods that search for more parameters should be used when more precision is required in the deformations. These include the method of Collins *et al.*(1994a), high dimensional linear-elasticity model (Miller *et al.*, 1993) and the viscous fluid models (Christensen *et al.*, 1996; Thompson & Toga, 1996). The next chapter also describes one such method.

In practice however, it may be meaningless to even attempt an exact match between brains beyond a certain resolution. There is not a one-to-one relationship between the cortical structures of one brain and those of another, so any method that attempts to match brains exactly must be folding the brain to create sulci and gyri that do not really exist. Even if an exact match is possible, because the registration problem is not convex, the solutions obtained by high dimensional warping techniques may not be truly optimum. High-dimensional registrations methods are often very good at registering grey matter with grey matter (for example), but there is no guarantee that the registered grey matter arises from homologous cortical features.

Also, structure and function are not always tightly linked. Even if structurally equivalent regions can be brought into exact register, it does not mean that the same is true for regions that perform the same or similar functions. For inter-subject averaging, an assumption is made that functionally equivalent regions lie in approximately the same parts of the brain. This leads to the current rationale for smoothing images from multi-subject functional imaging studies prior to performing statistical analyses. Constructive interference of the smeared activation signals then has the effect of producing a signal that is roughly in an average location. In order to account for substantial fine scale warps in a spatial normalisation, it is necessary for some voxels to increase their volumes considerably, and for others to shrink to an almost negligible size. The contribution of the shrunken regions to the smoothed images is tiny, and the sensitivity of the tests for detecting activations in these regions is reduced. This is another argument in favour of registering only on a global scale.

The constrained normalisation described here assumes that the template resembles a warped version of the image. Modifications are required in order to apply the method to diseased or lesioned brains. One possible approach is to assume different weights for different brain regions. Lesioned areas can be assigned lower weights, so that they have much less influence on the final solution.

The registration scheme described in this chapter is constrained to describe warps with a few hundred parameters. More powerful and less expensive computers are rapidly evolving, so algorithms that are currently applicable will become increasingly redundant as it becomes feasible to attempt more precise registrations. Scanning hardware is also improving, leading to improvements in the quality and diversity of images that can be obtained. Currently, most registration algorithms only use the information from a single image from each subject. This is typically a T1 MR image, which provides limited information that simply delineates grey and white matter. For example, further information that is not available in the more conventional sequences could be obtained from diffusion weighted imaging. Knowledge of major white matter tracts should provide structural information more directly related to connectivity and implicitly function, possibly leading to to improved registration of functionally specialised areas.