Chapter 3

Spatial Normalization using Basis Functions

John Ashburner & Karl J. Friston
The Wellcome Dept. of Imaging Neuroscience,
12 Queen Square, London WC1N 3BG, UK.

Contents

Abstract

This chapter describes the steps involved in registering images of different subjects into roughly the same co-ordinate system, where the co-ordinate system is defined by a template image (or series of images). The method only uses up to a few hundred parameters, so can only model global brain shape. It works by estimating the optimum coefficients for a set of bases, by minimizing the sum of squared differences between the template and source image, while simultaneously maximizing the smoothness of the transformation using a *maximum a posteriori* (MAP) approach.

3.1 Introduction

Sometimes it is desirable to warp images from a number of individuals into roughly the same standard space to allow signal averaging across subjects. This procedure is known as spatial normalization. In functional imaging studies, spatial normalization of the images is useful for determining what happens generically over individuals. A further advantage of using spatially normalized images is that activation sites can be reported according to their Euclidean co-ordinates within a standard space [21]. The most commonly adopted co-ordinate system within the brain imaging community is that described by [32], although new standards are now emerging that are based on digital atlases [18, 19, 27].

Methods of registering images can be broadly divided into *label based* and *intensity based*. Label based techniques identify homologous features (labels) in the source and reference images 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 as landmarks is that there are very few readily identifiable discrete points in the brain. Lines and 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 behavior is not known, so it is forced to be as 'smooth' as possible. There are a number of methods for modeling this smoothness. The simplest models include fitting splines through the points in order to minimize *bending energy* [5, 4]. More complex forms of interpolation, such as viscous fluid models, are often used when the labels are surfaces [35, 16].

Intensity (non-label) based approaches identify a spatial transformation that optimizes some voxel-similarity measure between a source and reference image, where both are treated as unlabeled continuous processes. The matching criterion is usually based upon minimizing the sum of squared differences or maximizing the correlation between the images. For this criterion to be successful, it requires the reference to appear like a warped version of the source image. In other words, there must be correspondence in the grey levels of the different tissue types between the source and reference images. In order to warp together images of different modalities, a few intensity based methods have been devised that involve optimizing an information theoretic measure [31, 33]. Intensity matching methods are usually very susceptible to poor starting estimates, so more recently a number of hybrid approaches have emerged that combine intensity based methods with matching user defined features (typically sulci).

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. Some research groups use only a nine- or twelve-parameter affine transformation to approximately register images of different subjects, accounting for differences in position, orientation and overall brain dimensions. Low spatial frequency global variability of head shape can be accommodated by describing deformations by a linear combination of low frequency basis functions. One widely used basis function registration method is part of the AIR package [36, 37], which uses polynomial basis functions to model shape variability. For example, a two dimensional third order polynomial basis function

mapping can be defined:

$$\begin{aligned}
y_1 = & q_1 & +q_2x_1 & +q_3x_1^2 & +q_4x_1^3+ \\
& q_5x_2 & +q_6x_1x_2 & +q_7x_1^2x_2 & + \\
& q_8x_2^2 & +q_9x_1x_2^2 & + & \\
& q_{10}x_2^3 & & & \\
y_2 = & q_{11} & +q_{12}x_1 & +q_{13}x_1^2 & +q_{14}x_1^3+ \\
& q_{15}x_2 & +q_{16}x_1x_2 & +q_{17}x_1^2x_2 & + \\
& q_{18}x_2^2 & +q_{19}x_1x_2^2 & + & \\
& q_{20}x_2^3 & & &
\end{aligned}$$

Other low-dimensional registration methods may employ a number of other forms of basis function to parameterize the warps. These include Fourier bases [10], sine and cosine transform basis functions [9, 3], B-splines [31, 33], and piecewise affine or trilinear basis functions (see [25] for a review). The small number of parameters will not allow every feature to be matched exactly, but it will permit the global head shape to be modeled. The rationale for adopting a low dimensional approach is that it allows rapid modeling of 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. Regularization schemes are therefore necessary when attempting image registration with many parameters, thus ensuring that voxels remain close to their neighbors. Regularization is often incorporated by some form of Bayesian scheme, using estimators such as the *maximum a posteriori* (MAP) estimate or the *minimum variance estimate* (MVE). Often, the prior probability distributions used by registration schemes are linear, and include minimizing the *membrane energy* of the deformation field [1, 24], the *bending energy* [5] or the *linear-elastic energy* [28, 16]. None of these linear penalties explicitly preserve the topology[1] of the warped images, although cost functions that incorporate this constraint have been devised [17, 2]. A number of methods involve repeated Gaussian smoothing of the estimated deformation fields [14]. These methods can be classed among the elastic registration methods because convolving a deformation field is a form of linear regularization [8].

An alternative, to using a Bayesian scheme incorporating some form of elastic prior, could be to use a viscous fluid model [11, 12, 8, 34] to estimate the warps. In these models, finite difference methods are often used to solve the partial differential equations that model one image as it "flows" to the same shape as the other. The major advantage of these methods is that they are able to account for large deformations and also ensure that the topology of the warped image is preserved. Viscous fluid models are almost able to warp any image so that it looks like any other image, while still preserving the original topology. These methods can be classed as "plastic" as it is not the deformation fields themselves that are regularized, but rather the increments to the deformations at each iteration.

3.2 Method

This chapter describes the steps involved in registering images of different subjects into roughly the same co-ordinate system, where the co-ordinate system is defined by a template image (or series of images).

[1] The word "topology" is used in the same sense as in "Topological Properties of Smooth Anatomical Maps" [13]. If spatial transformations are not one-to-one and continuous, then the topological properties of different structures can change.

This section begins by introducing a modification to the optimization method described in Section **??**, such that more robust *maximum a posteriori* (MAP) parameter estimates can be obtained. It works by estimating the optimum coefficients for a set of bases, by minimizing the sum of squared differences between the template and source image, while simultaneously minimizing the deviation of the transformation from its expected value. In order to adopt the MAP approach, it is necessary to have estimates of the likelihood of obtaining the fit given the data, which requires prior knowledge of spatial variability, and also knowledge of the variance associated with each observation. True Bayesian approaches assume that the variance associated with each voxel is already known, whereas the approach described here is a type of Empirical Bayesian method, which attempts to estimate this variance from the residual errors (see Chapter 13). Because the registration is based on smooth images, correlations between neighboring voxels are considered when estimating the variance. This makes the same approach suitable for the spatial normalization of both high quality MR images, and low resolution noisy PET images.

The first step in registering images from different subjects involves determining the optimum 12 parameter affine transformation. Unlike Chapter 2 – where the images to be matched together are from the same subject – zooms and shears are needed to register heads of different shapes and sizes. Prior knowledge of the variability of head sizes is included within a Bayesian framework in order to increase the robustness and accuracy of the method.

The next part describes nonlinear registration for correcting gross differences in head shapes that cannot be accounted for by the affine normalization alone. The nonlinear warps are modeled by linear combinations of smooth discrete cosine transform basis functions. A fast algorithm is described that utilizes Taylor's Theorem and the separable nature of the basis functions, meaning that most of the nonlinear spatial variability between images can be automatically corrected within a few minutes. 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:

$$p(\mathbf{q}|\mathbf{b}) \propto p(\mathbf{b}|\mathbf{q})p(\mathbf{q}) \tag{3.1}$$

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

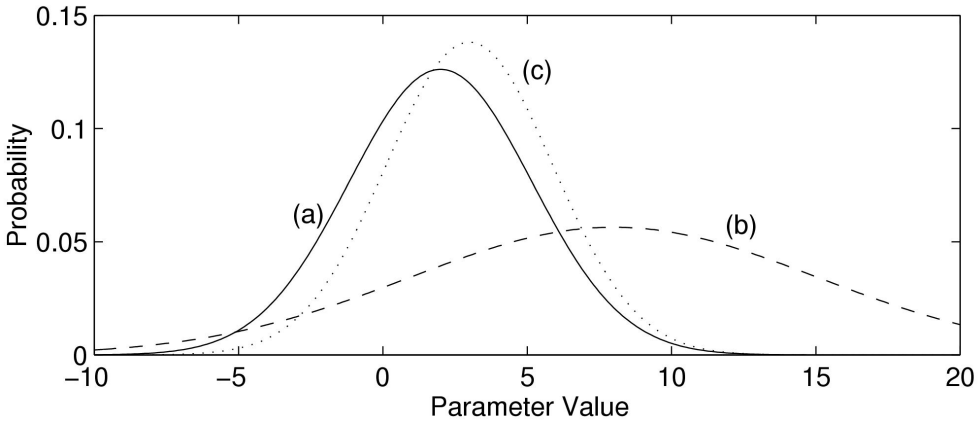

Figure 3.1: This figure illustrates a hypothetical example with one parameter, where the prior probability distribution is better described than the likelihood. The solid Gaussian curve (a) represents the prior probability distribution (p.d.f), and the dashed curve (b) represents a maximum likelihood 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 7). The dotted line (c) shows the posterior p.d.f obtained using Bayesian statistics. The maximum value of (c) is the MAP estimate. It combines previously known information with that from the data to give a more accurate estimate.

$p(\mathbf{q})$ 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 maximized. The optimization 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 maximized when its Gibbs form is minimized. This is equivalent to minimizing $H(\mathbf{b}|\mathbf{q}) + H(\mathbf{q})$ (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:

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

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

$${\mathbf{q_{ML}}}^{(n+1)} = \mathbf{q}^{(n)} - \left(\mathbf{A}^T\mathbf{A}\right)^{-1} \mathbf{A}^T\mathbf{b} \tag{3.2}$$

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:

$$\mathbf{q}^{(n+1)} = \left({\mathbf{C_0}}^{-1} + \mathbf{A}^T\mathbf{A}/\sigma^2\right)^{-1} \left({\mathbf{C_0}}^{-1}\mathbf{q_0} + \mathbf{A}^T\mathbf{A}/\sigma^2 {\mathbf{q_{ML}}}^{(n+1)}\right) \tag{3.3}$$

The MAP optimization scheme is obtained by combining Eqns. 3.2 and 3.3:

$$\mathbf{q}^{(n+1)} = \left({\mathbf{C_0}}^{-1} + \mathbf{A}^T\mathbf{A}/\sigma^2\right)^{-1} \left({\mathbf{C_0}}^{-1}\mathbf{q_0} + \mathbf{A}^T\mathbf{A}\mathbf{q}^{(n)}/\sigma^2 - \mathbf{A}^T\mathbf{b}/\sigma^2\right) \tag{3.4}$$

For the sake of the registration, it is assumed that the exact form for the prior probability distribution ($N(\mathbf{q_0}, \mathbf{C_0})$) 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:

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

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 [2].

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 [23] (although this approach is not strictly correct [38], 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

[2] Strictly speaking, the computation of the degrees of freedom should be more complicated than this, as this simple model does not account for the regularization.

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 [29]:

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

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 [3] becomes approximately:

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

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 regularized 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 normalization 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 minimizing (or maximizing) 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. The images may be scaled quite differently, so an additional intensity scaling parameter is included in the model.

Without constraints and with poor data, simple ML parameter optimization (similar to that described in Section **??**) 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 regularized 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 optimization 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. This can be derived from the zooms and shears determined by registering a large number of brain images to the template.

[3] 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 [23] for smooth images, but never allows the computed value to exceed $I - J$.

3.2.3 Nonlinear Registration

The nonlinear spatial normalization 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:

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

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) [9]. 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:

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

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 optimization involves minimizing 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 minimized function is then:

$$\sum_i (f(\mathbf{y}_i) - wg(\mathbf{x}_i))^2$$

The approach described in Section **??** is used to optimize the parameters $\mathbf{q_1}$, $\mathbf{q_2}$, $\mathbf{q_3}$ and w, and requires derivatives of the function $f(\mathbf{y}_i) - wg(\mathbf{x}_i)$ with respect to each parameter. These

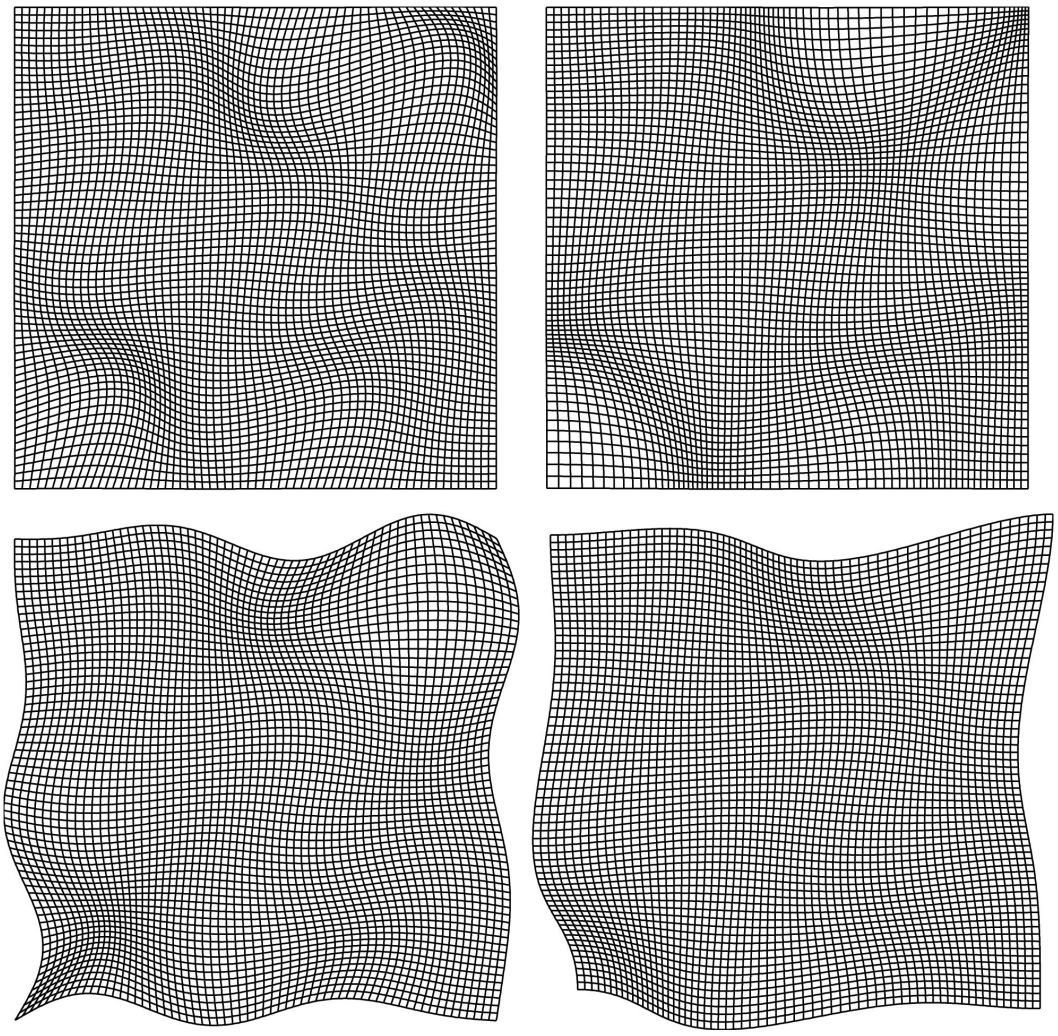

Figure 3.2: Different boundary conditions. Above left: fixed boundaries (generated purely from DST basis functions). Above right: sliding boundaries (from a mixture of DCT and DST basis functions). Below left: bending boundaries (from a different mixture of DCT and DST basis functions). Below right: free boundary conditions (purely from DCT basis functions).

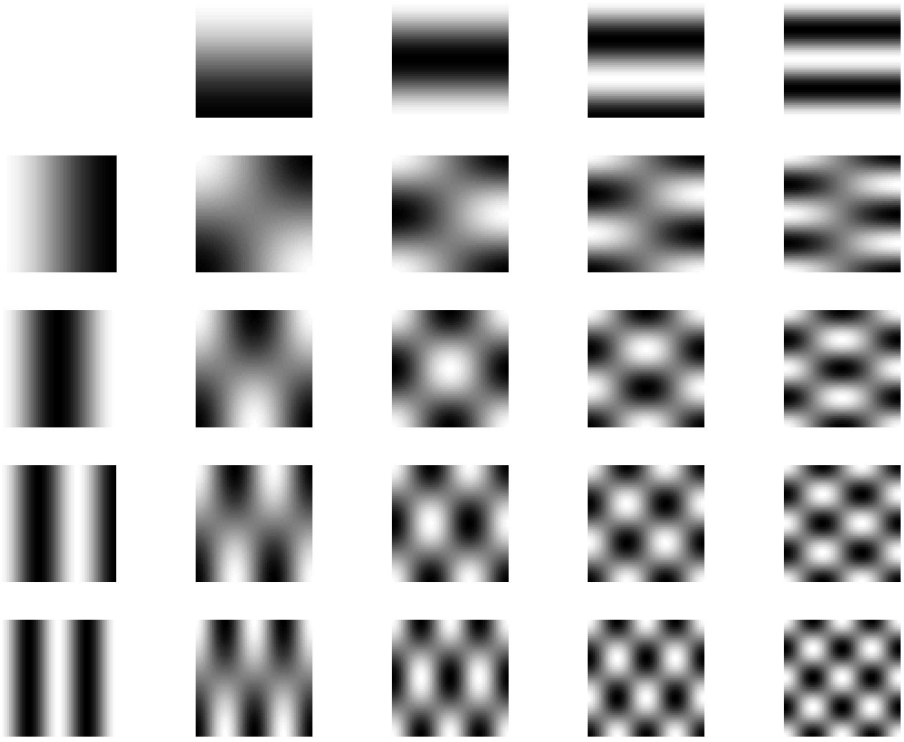

Figure 3.3: The lowest frequency basis functions of a two dimensional Discrete Cosine Transform.

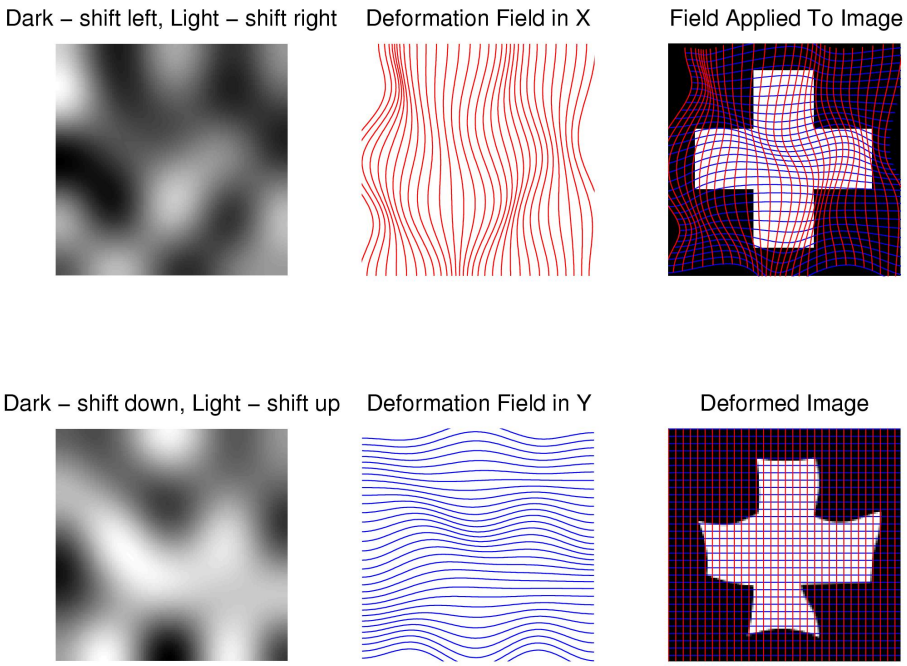

Figure 3.4: In two dimensions, a deformation field consists of two scalar fields. One for horizontal deformations, and the other for vertical deformations. The images on the left show deformations as a linear combination of basis images (see Figure 3.3). The center column shows the same deformations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

$\alpha = [\mathbf{0}]$

$\beta = [\mathbf{0}]$

$for \quad j = 1 \ldots J$

$\mathbf{C} = \mathbf{d_2}_{j,:}{}^T \mathbf{d_2}_{j,:}$

$\mathbf{E_1} = diag(\nabla_\mathbf{1} \mathbf{f}_{:,j}) \mathbf{D_1}$

$\mathbf{E_2} = diag(\nabla_\mathbf{2} \mathbf{f}_{:,j}) \mathbf{D_1}$

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

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

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).

can be obtained using the chain rule:

$$\begin{aligned} \frac{\partial f(\mathbf{y}_i)}{\partial q_{j1}} &= \frac{\partial f(\mathbf{y}_i)}{\partial y_{1i}} \frac{\partial y_{1i}}{\partial q_{j1}} = \frac{\partial f(\mathbf{y}_i)}{\partial y_{1i}} d_j(\mathbf{x}_i) \\ \frac{\partial f(\mathbf{y}_i)}{\partial q_{j2}} &= \frac{\partial f(\mathbf{y}_i)}{\partial y_{2i}} \frac{\partial y_{2i}}{\partial q_{j2}} = \frac{\partial f(\mathbf{y}_i)}{\partial y_{2i}} d_j(\mathbf{x}_i) \\ \frac{\partial f(\mathbf{y}_i)}{\partial q_{j3}} &= \frac{\partial f(\mathbf{y}_i)}{\partial y_{3i}} \frac{\partial y_{3i}}{\partial q_{j3}} = \frac{\partial f(\mathbf{y}_i)}{\partial y_{3i}} d_j(\mathbf{x}_i) \end{aligned}$$

The approach involves iteratively computing $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}^T\mathbf{b}$. However, because there are many parameters to optimize, 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. 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 $\mathbf{D_1}\mathbf{Q_1}\mathbf{D_2}^T$ and $\mathbf{D_1}\mathbf{Q_2}\mathbf{D_2}^T$.

Between each iteration, image $\mathbf{F}$ is re-sampled according to the latest parameter estimates. The derivatives of $\mathbf{F}$ are also re-sampled 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.

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:

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

The advantage of the algorithm shown in Figure 3.5 is that it utilizes some of the useful properties of Kronecker tensor products. This is especially important when the algorithm is implemented in three dimensions. 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 Regularization for Nonlinear Registration

Without regularization in the nonlinear registration, it is possible to introduce unnecessary deformations that only reduce the residual sum of squares by a tiny amount (see Figure 3.6). This could potentially make the algorithm very unstable. Regularization is achieved by minimizing the sum of squared difference between the template and the warped image, while simultaneously minimizing 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[4] 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 regularization 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 regularization and result in greatly underestimated deformations. If the values are too small, there will not be enough regularization and the resulting deformations will over-fit the data.

Membrane Energy

The simplest model used for linear regularization is based upon minimizing the *membrane energy* of the deformation field $\mathbf{u}$ [1, 24]. By summing over i points in three dimensions, the membrane energy of $\mathbf{u}$ is given by:

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

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

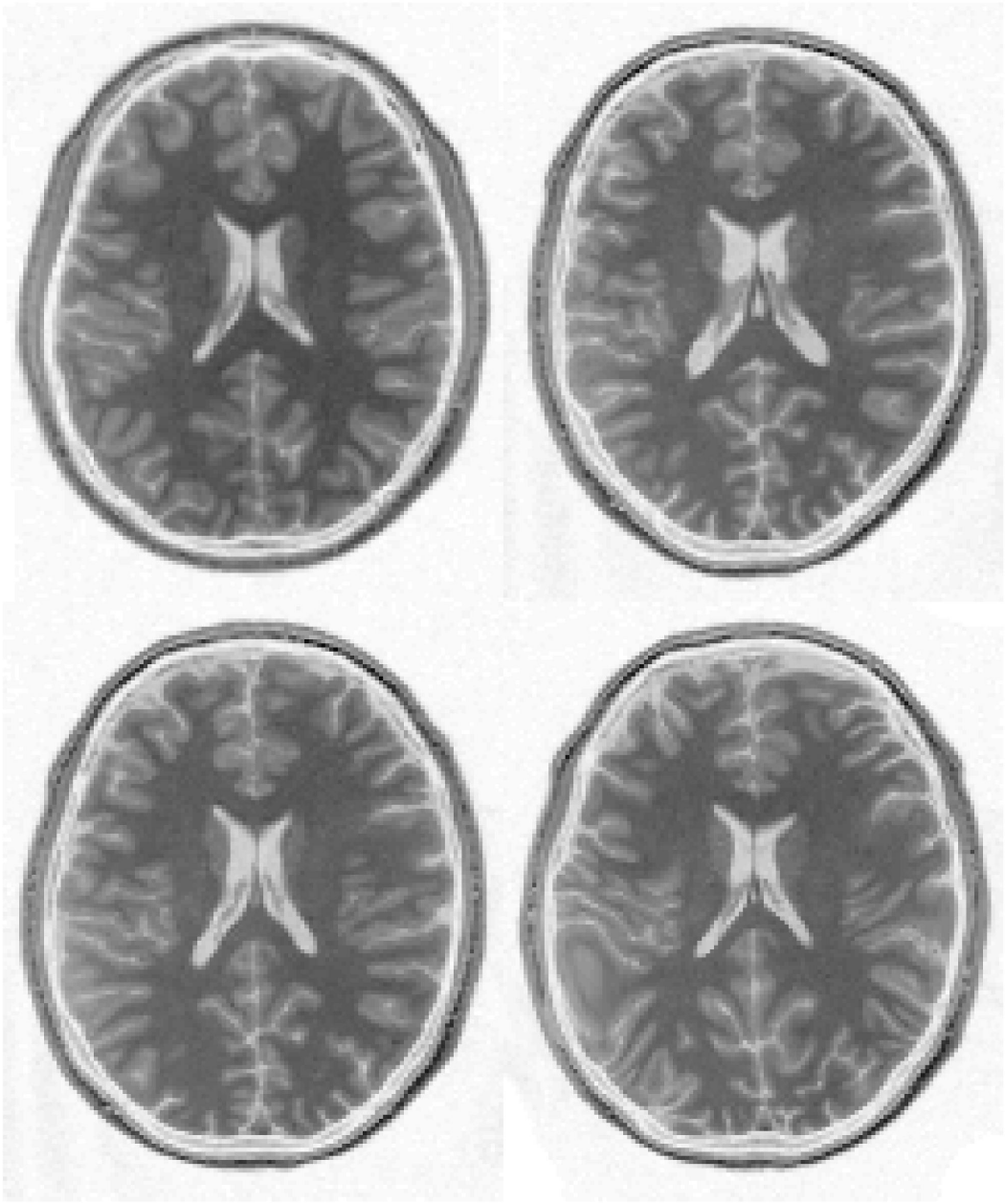

Figure 3.6: The image shown at the top-left is the template image. At the top-right is an image that has been registered with it using a 12-parameter affine registration. The image at the bottom-left is the same image registered using the 12-parameter affine registration, followed by a regularized global nonlinear registration. 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 regularization. The mean squared difference between the image and template after the affine registration was 472.1. After the regularized nonlinear registration this was reduced to 302.7. Without regularization, a mean squared difference of 287.3 is achieved, but this is at the expense of introducing a lot of unnecessary warping.

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:

$$\begin{aligned}\mathbf{H} = &\quad \lambda\left(\dot{\mathbf{D_3}}^T\dot{\mathbf{D_3}}\right)\otimes\left(\mathbf{D_2}^T\mathbf{D_2}\right)\otimes\left(\mathbf{D_1}^T\mathbf{D_1}\right)\\ &+\quad \lambda\left(\mathbf{D_3}^T\mathbf{D_3}\right)\otimes\left(\dot{\mathbf{D_2}}^T\dot{\mathbf{D_2}}\right)\otimes\left(\mathbf{D_1}^T\mathbf{D_1}\right)\\ &+\quad \lambda\left(\mathbf{D_3}^T\mathbf{D_3}\right)\otimes\left(\mathbf{D_2}^T\mathbf{D_2}\right)\otimes\left(\dot{\mathbf{D_1}}^T\dot{\mathbf{D_1}}\right)\end{aligned}$$

where the notation $\dot{\mathbf{D_1}}$ 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. 3.4 can be constructed from $\mathbf{H}$ by:

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

$\mathbf{H}$ 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:

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

Bending Energy

Bookstein's thin plate splines [6, 5] minimize the *bending energy* of deformations. For a two dimensional deformation, the bending energy is defined by:

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

This can be computed by:

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

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:

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

Matrix $\mathbf{C_0}^{-1}$ from Eqn. 3.4 can be constructed from $\mathbf{H}$ as:

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

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

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

Linear-Elastic Energy

The *linear-elastic* energy [28] of a two dimensional deformation field is:

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

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

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

A regularization based upon this model requires an inverse covariance matrix that is not a simple diagonal matrix. This matrix is constructed as follows:

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

where:

$$\begin{aligned}\mathbf{H_1} = &\ (\mu+\lambda/2)(\mathbf{D_2}^T\mathbf{D_2})\otimes(\dot{\mathbf{D}}_\mathbf{1}^T\dot{\mathbf{D}}_\mathbf{1}) + \mu/2(\dot{\mathbf{D}}_\mathbf{2}^T\dot{\mathbf{D}}_\mathbf{2})\otimes(\mathbf{D_1}^T\mathbf{D_1})\\ \mathbf{H_2} = &\ (\mu+\lambda/2)(\dot{\mathbf{D}}_\mathbf{2}^T\dot{\mathbf{D}}_\mathbf{2})\otimes(\mathbf{D_1}^T\mathbf{D_1}) + \mu/2(\mathbf{D_2}^T\mathbf{D_2})\otimes(\dot{\mathbf{D}}_\mathbf{1}^T\dot{\mathbf{D}}_\mathbf{1})\\ \mathbf{H_3} = &\ \lambda/2(\mathbf{D_2}^T\dot{\mathbf{D}}_\mathbf{2})\otimes(\dot{\mathbf{D}}_\mathbf{1}^T\mathbf{D_1}) + \mu/2(\dot{\mathbf{D}}_\mathbf{2}^T\mathbf{D_2})\otimes(\mathbf{D_1}^T\dot{\mathbf{D}}_\mathbf{1})\end{aligned}$$

3.2.5 Templates and Intensity Transformations

Sections 3.2.2 and 3.2.3 have modeled a single intensity scaling parameter (q_{13} and w respectively), but more generally, the optimization can be assumed to minimize 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:

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

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 minimize the sum of squared differences between them. For this case, $t(\mathbf{x}_i, \mathbf{q_t})$ is simply

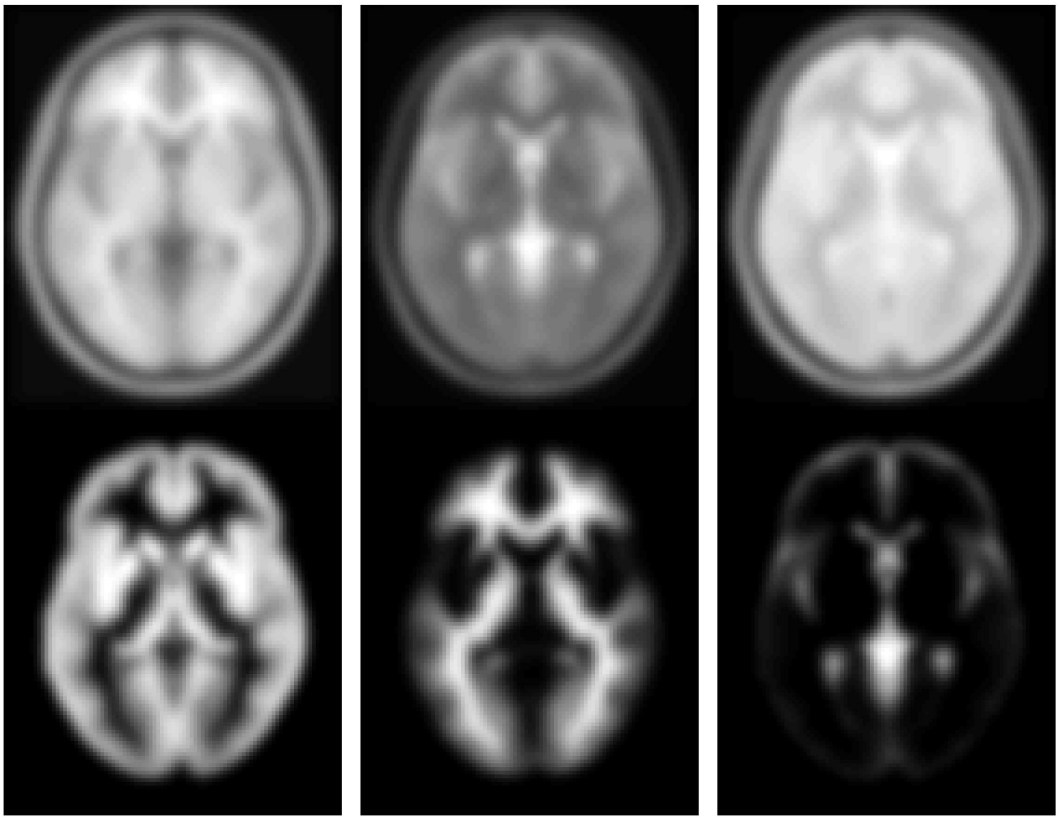

Figure 3.7: Example template images. Above: T1 weighted MRI, T2 weighted MRI and PD weighted MRI. Below: Grey matter probability distribution, White matter probability distribution and CSF probability distribution. All the data were generated at the McConnel Brain Imaging Center, Montréal Neurological 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 normalization.

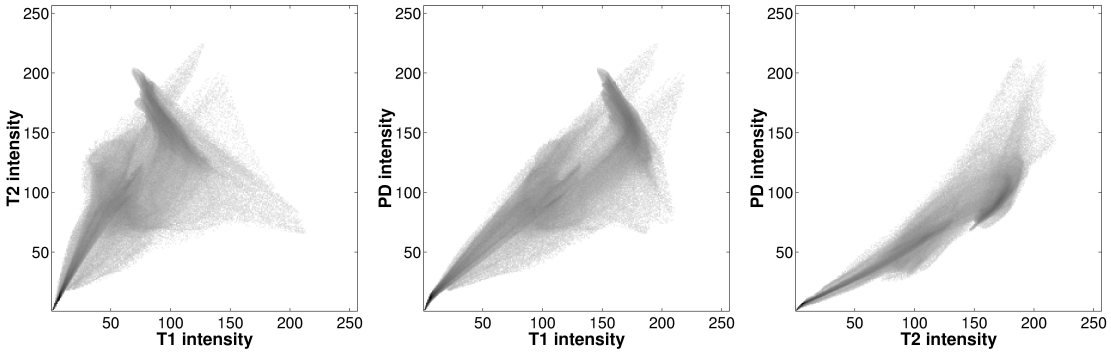

Figure 3.8: Two dimensional histograms of template images (intensities shown as $log(1+n)$, where n is the value in each bin). The histograms were based on the whole volumes of the template images shown in the top row of Figure 3.7.

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 normalization will be similar to those shown in the top row of Figure 3.7. 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 3.8, 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:

$$b_i(\mathbf{q}) = f(\mathbf{s}(\mathbf{x}_i, \mathbf{q_s})) - (q_{t1}g_1(\mathbf{x}_i) + q_{t2}g_2(\mathbf{x}_i))$$

In Figure 3.9, a plane of a T1 weighted MRI is modeled by a linear combination of the five other template images shown in Figure 3.7. Similar models were used to simulate T2 and PD weighted MR images. The linearity of the scatter-plots (compared to those in Figure 3.8) shows that MR images of a wide range of different contrasts can be modeled 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 3.7.

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 optimizing a function of the form:

$$b_i(\mathbf{q}) = f(\mathbf{s}(\mathbf{x}_i, \mathbf{q_s})) - (q_{t1}g(\mathbf{x}_i) + q_{t2}x_{1i}g(\mathbf{x}_i) + q_{t3}x_{2i}g(\mathbf{x}_i) + q_{t4}x_{3i}g(\mathbf{x}_i))$$

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) [22]. 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:

$$b_i(\mathbf{q}) = f(\mathbf{s}(\mathbf{x}_i, \mathbf{q_s})) - q_{t1}g(\mathbf{x}_i)^{q_{t2}}$$

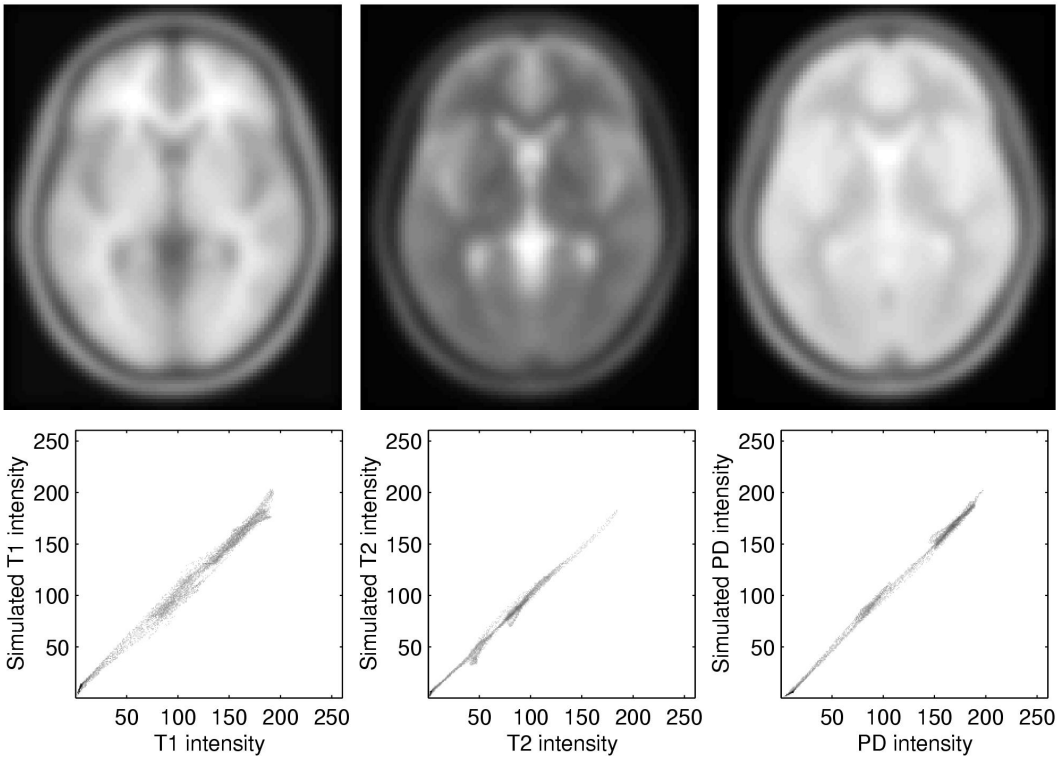

Figure 3.9: Simulated images of T1, T2 and PD weighted images, and histograms of the real images versus the simulated images.

[15] 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 limited brain coverage or intensity inhomogeneity artifacts (in MR images). Other rotationally invariant moments also contain useful matching information [30]. 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 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 by making use of internal consistency would be to simultaneously spatially normalize 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. A similar approach could be devised for simultaneously matching different tissue types from classified images together [26], although a more powerful approach is to incorporate tissue classification and registration into the same Bayesian model [20].

3.3 Discussion

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 normalization of an arbitrary brain into a standard stereotaxic 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 normalization (either within or between subjects) in functional mapping studies is to maximize the sensitivity to neuro-physiological change elicited by experimental manipulation of sensorimotor or cognitive state. In this case a better definition of a valid normalization is that which maximizes 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.

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 cannot be modeled 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 (see Figures 3.10 and 3.11).

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 optimization 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

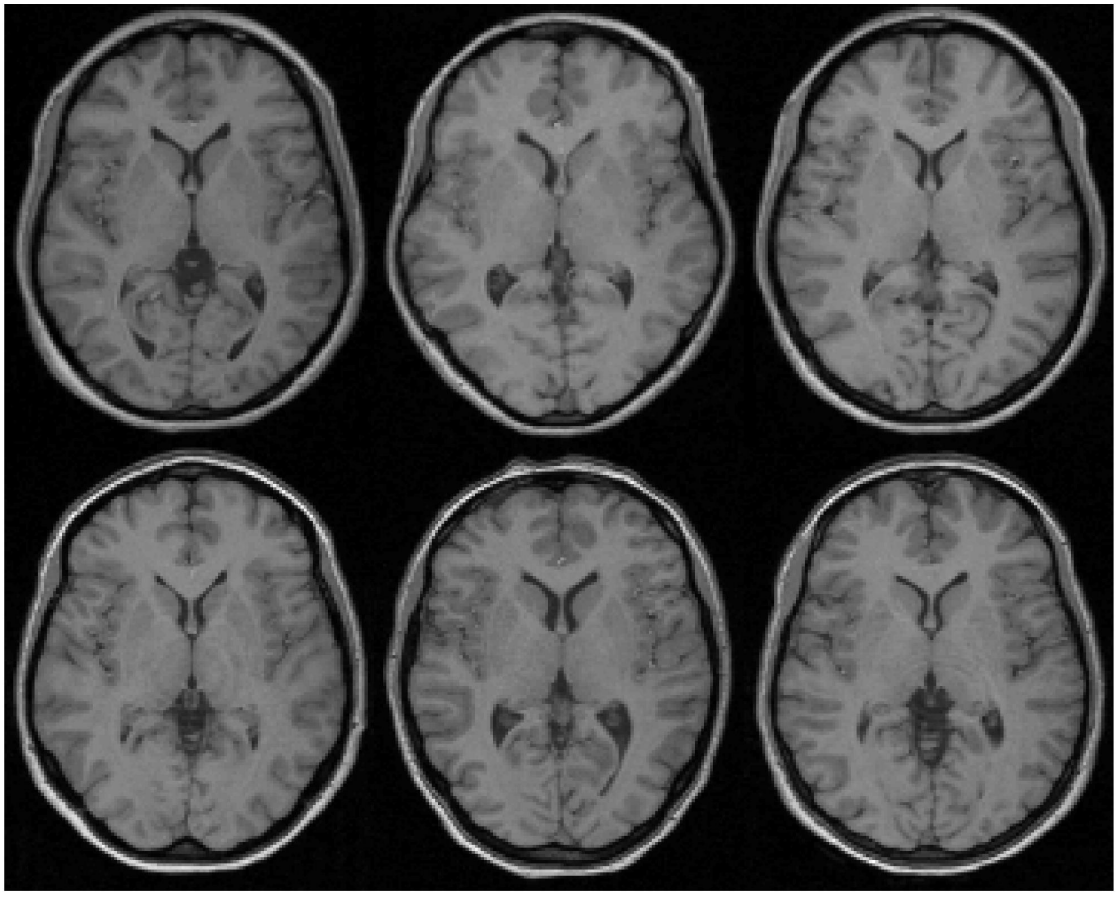

Figure 3.10: Images of six subjects registered using a 12-parameter affine registration (see also Figure 3.11). The affine registration matches the positions and sizes of the images.

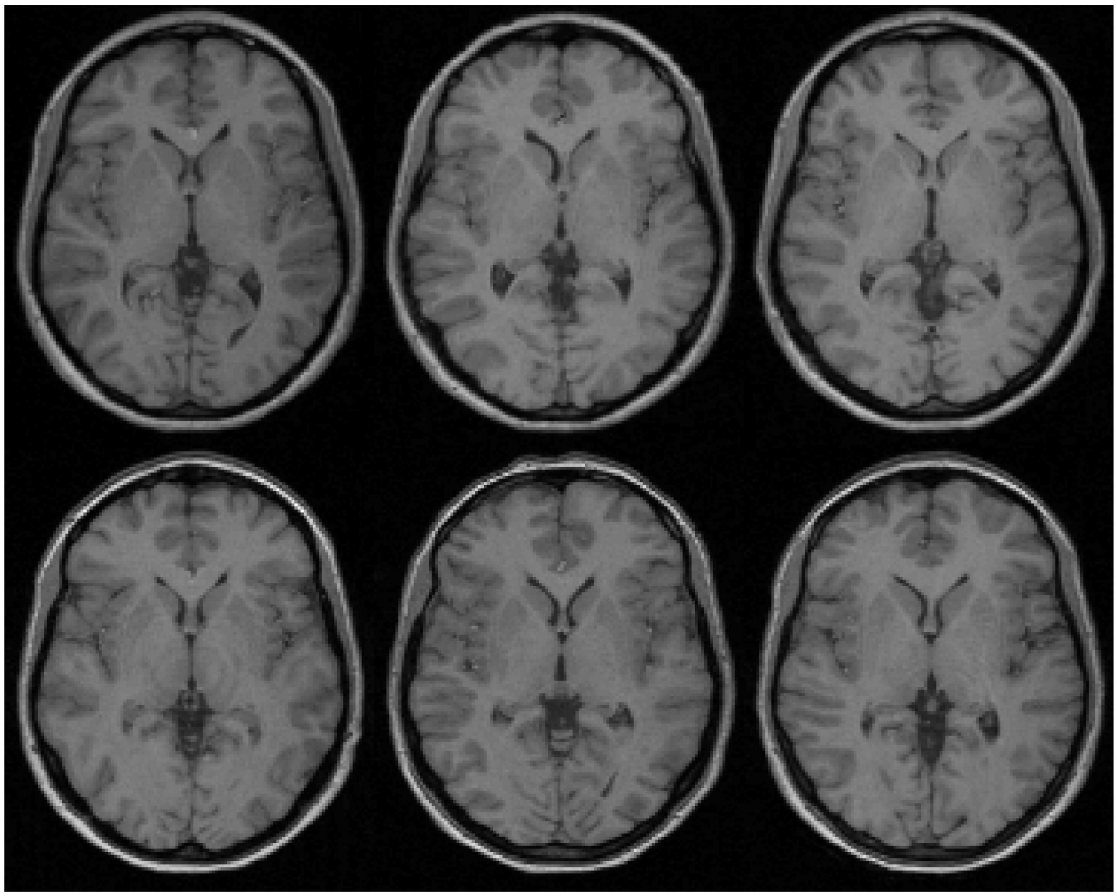

Figure 3.11: Six subjects brains registered with both affine and basis function registration (see also Figure 3.10). The basis function registration estimates the global shapes of the brains, but is not able to account for high spatial frequency warps.

the brains being registered. However, by estimating only the lowest frequency deformations and by using appropriate regularization, this constraint is rarely broken.

The approach in this chapter searches for a MAP estimate of the parameters defining the warps. However, optimization 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 posterior probabilities. Although useful approximations have been devised [28, 9], 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 normalization, 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, 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.

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 normalization, 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 favor of spatially normalizing only on a global scale.

The constrained normalization 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

[7]. 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 improved registration of functionally specialized areas.

Bibliography

[1] Y. Amit, U. Grenander, and M. Piccioni. Structural image restoration through deformable templates. *Journal of the American Statistical Association*, 86:376–387, 1991.

[2] J. Ashburner and K. J. Friston. High-dimensional nonlinear image registration. *NeuroImage*, 7(4):S737, 1998.

[3] J. Ashburner and K. J. Friston. Nonlinear spatial normalization using basis functions. *Human Brain Mapping*, 7(4):254–266, 1999.

[4] F. L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(6):567–585, 1989.

[5] F. L. Bookstein. Landmark methods for forms without landmarks: Morphometrics of group differences in outline shape. *Medical Image Analysis*, 1(3):225–243, 1997.

[6] F. L. Bookstein. Quadratic variation of deformations. In J. Duncan and G. Gindi, editors, *Proc. Information Processing in Medical Imaging*, pages 15–28, Berlin, Heidelberg, New York, 1997. Springer-Verlag.

[7] M. Brett, A. P. Leff, C. Rorden, and J. Ashburner. Spatial normalization of brain images with focal lesions using cost function masking. *NeuroImage*, 14(2):486–500, 2001.

[8] M. Bro-Nielsen and C. Gramkow. Fast fluid registration of medical images. *Lecture Notes in Computer Science*, 1131:267–276, 1996.

[9] G. E. Christensen. Deformable shape models for anatomy. Doctoral thesis, Washington University, Sever Institute of Technology, 1994.

[10] G. E. Christensen. Consistent linear elastic transformations for image matching. In A. Kuba et al., editor, *Proc. Information Processing in Medical Imaging*, pages 224–237, Berlin, Heidelberg, 1999. Springer-Verlag.

[11] G. E. Christensen, R. D. Rabbitt, and M. I. Miller. 3D brain mapping using using a deformable neuroanatomy. *Physics in Medicine and Biology*, 39:609–618, 1994.

[12] G. E. Christensen, R. D. Rabbitt, and M. I. Miller. Deformable templates using large deformation kinematics. *IEEE Transactions on Image Processing*, 5:1435–1447, 1996.

[13] G. E. Christensen, R. D. Rabbitt, M. I. Miller, S. C. Joshi, U. Grenander, T. A. Coogan, and D. C. Van Essen. Topological properties of smooth anatomic maps. In Y. Bizais, C. Barillot, and R. Di Paola, editors, *Proc. Information Processing in Medical Imaging*, pages 101–112, Dordrecht, The Netherlands, 1995. Kluwer Academic Publishers.

[14] D. L. Collins, A. C. Evans, C. Holmes, and T. M. Peters. Automatic 3D segmentation of neuro-anatomical structures from MRI. In Y. Bizais, C. Barillot, and R. Di Paola, editors, *Proc. Information Processing in Medical Imaging*, pages 139–152, Dordrecht, The Netherlands, 1995. Kluwer Academic Publishers.

[15] D. L. Collins, P. Neelin, T. M. Peters, and A. C. Evans. Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. *Journal of Computer Assisted Tomography*, 18:192–205, 1994.

[16] C. Davatzikos. Spatial normalization of 3D images using deformable models. *Journal of Computer Assisted Tomography*, 20(4):656–665, 1996.

[17] P. J. Edwards, D. L. G. Hill, and D. J. Hawkes. Image guided interventions using a three component tissue deformation model. In *Proc. Medical Image Understanding and Analysis*, 1997.

[18] A. C. Evans, D. L. Collins, S. R. Mills, E. D. Brown, R. L. Kelly, and T. M. Peters. 3D statistical neuroanatomical models from 305 MRI volumes. In *Proc. IEEE-Nuclear Science Symposium and Medical Imaging Conference*, pages 1813–1817, 1993.

[19] A. C. Evans, M. Kamber, D. L. Collins, and D. Macdonald. An MRI-based probabilistic atlas of neuroanatomy. In S. Shorvon, D. Fish, F. Andermann, G. M. Bydder, and Stefan H, editors, *Magnetic Resonance Scanning and Epilepsy*, volume 264 of *NATO ASI Series A, Life Sciences*, pages 263–274. Plenum Press, 1994.

[20] B. Fischl, D. H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. van der Kouwe, R. Killiany, D. Kennedy, S. Klaveness, A. Montillo, N. Makris, B. Rosen, and A. M. Dale. Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33:341–355, 2002.

[21] P. T. Fox. Spatial normalization origins: Objectives, applications, and alternatives. *Human Brain Mapping*, 3:161–164, 1995.

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

[23] K. J. Friston, A. P. Holmes, J.-B. Poline, P. J. Grasby, S. C. R. Williams, R. S. J. Frackowiak, and R. Turner. Analysis of fMRI time series revisited. *NeuroImage*, 2:45–53, 1995.

[24] J. C. Gee, D. R. Haynor, L. Le Briquer, and R. K. Bajcsy. Advances in elastic matching theory and its implementation. In P. Cinquin, R. Kikinis, and S. Lavallee, editors, *Proc. CVRMed-MRCAS'97*, Heidelberg, 1997. Springer-Verlag.

[25] C. A. Glasbey and K. V. Mardia. A review of image warping methods. *Journal of Applied Statistics*, 25:155–171, 1998.

[26] C. D. Good, I. S. Johnsrude, J. Ashburner, R. N. A Henson, K. J. Friston, and R. S. J. Frackowiak. *NeuroImage*, 14:21–36, 2001.

[27] J. C. Mazziotta, A. W. Toga, A. Evans, P. Fox, and J. Lancaster. A probablistic atlas of the human brain: Theory and rationale for its development. *NeuroImage*, 2:89–101, 1995.

[28] M. I. Miller, G. E. Christensen, Y. Amit, and U. Grenander. Mathematical textbook of deformable neuroanatomies. *Proc. National Academy of Sciences*, 90:11944–11948, 1993.

[29] J.-B. Poline, K. J. Friston, K. J. Worsley, and R. S. J. Frackowiak. Estimating smoothness in statistical parametric maps: Confidence intervals on p-values. *Journal of Computer Assisted Tomography*, 19(5):788–796, 1995.

[30] D. Shen and C. Davatzikos. HAMMER: Hierarchical attribute matching mechanism for elastic registration. *IEEE Transactions on Medical Imaging*, 21(11):1421–1439, 2002.

[31] C. Studholme, R. T. Constable, and J. S. Duncan. Accurate alignment of functional EPI data to anatomical MRI using a physics-based distortion model. *IEEE Transactions on Medical Imaging*, 19(11):1115–1127, 2000.

[32] J. Talairach and P. Tournoux. *Coplanar stereotaxic atlas of the human brain.* Thieme Medical, New York, 1988.

[33] P. Thévenaz and M. Unser. Optimization of mutual information for multiresolution image registration. *IEEE Transactions on Image Processing*, 9(12):2083–2099, 2000.

[34] J.-P. Thirion. Fast non-rigid matching of 3D medical images. Technical Report 2547, Institut National de Recherche en Informatique et en Automatique, May 1995. Available from http://www.inria.fr/RRRT/RR-2547.html.

[35] P. M. Thompson and A. W. Toga. Visualization and mapping of anatomic abnormalities using a probablistic brain atlas based on random fluid transformations. In *Proc. Visualization in Biomedical Computing*, pages 383–392, 1996.

[36] R. P. Woods, S. T. Grafton, C. J. Holmes, S. R. Cherry, and J. C. Mazziotta. Automated image registration: I. General methods and intrasubject, intramodality validation. *Journal of Computer Assisted Tomography*, 22(1):139–152, 1998.

[37] R. P. Woods, S. T. Grafton, J. D. G. Watson, N. L. Sicotte, and J. C. Mazziotta. Automated image registration: II. Intersubject validation of linear and nonlinear models. *Journal of Computer Assisted Tomography*, 22(1):153–165, 1998.

[38] K. J. Worsley and K. J. Friston. Analysis of fMRI time-series revisited - again. *NeuroImage*, 2:173–181, 1995.