Nonlinear Spatial Normalization using Basis Functions

John Ashburner & Karl J. Friston

The Wellcome Department of Cognitive Neurology, Institute of Neurology, Queen Square, London WC1N 3BG, United Kingdom

Running Title Nonlinear Spatial Normalization

Keywords registration, anatomy, imaging, stereotaxy, basis functions, spatial normalization, PET, MRI, functional mapping.

Address for correspondence
Wellcome Department of Cognitive Neurology,
Functional Imaging Laboratory,
12 Queen Square,
London. WC1N 3BG
U.K.

Tel +44 (0)171 833 7472

Fax +44 (0)171 813 1420

email j.ashburner@fil.ion.ucl.ac.uk

Abstract

We describe a comprehensive framework for performing rapid and automatic non-label based nonlinear spatial normalizations. The approach adopted minimizes the residual squared difference between an image and a template of the same modality. In order to reduce the number of parameters to be fitted, the nonlinear warps are described by a linear combination of low spatial frequency basis functions. The objective is to determine the optimum coefficients for each of the bases by minimizing the sum of squared differences between the image and template, while simultaneously maximizing the smoothness of the transformation using a *maximum a posteriori* (MAP) approach. Most MAP approaches assume that the variance associated with each voxel is already known and that there is no covariance between neighboring voxels. The approach described here attempts to estimate this variance from the data, and also corrects for the correlations between neighboring voxels. This makes the same approach suitable for the spatial normalization of both high quality MR images, and low resolution noisy PET images. A fast algorithm has been developed 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.

1 Background

This paper concerns the problem of nonlinear spatial normalization: Namely how to map a single subject's brain image into a standard space. The solution of this problem allows for a wide range of voxel-based analyses and facilitates the comparison of different subjects and databases. The problem of spatial normalization is not a trivial one; indeed at some anatomical scales it is not clear that a solution even exists.

A fundamental advantage of using spatially normalized images is that activations can be reported according to a set of meaningful Euclidian coordinates within a standard space (Fox, 1995). New results can be readily incorporated into ongoing brain atlas and database projects such as that being developed by the International Consortium for Human Brain Mapping (ICBM) (Mazziotta *et al.*, 1995). The most commonly adopted coordinate system within the brain imaging community is that described by the atlas of Talairach & Tournoux (1988).

When whole brain structural images (typically high resolution MRI) of the subject are available in addition to the functional images, the images can be co-registered using any one of a number of methods for inter-modality registration (Pelizzari *et al.*, 1988; Woods *et al.*, 1992; Studholme *et al.*, 1995; Collignon *et al.*, 1995; Ashburner & Friston, 1997). This allows the spatial transformations that warp the images to the reference space to be determined from the structural images. These warps can then be applied to the functional images. Because there are only six rigid body parameters required to map between the structural and functional images, the co-registration parameters can be determined fairly accurately. The structural images should have higher spatial resolution, less noise and more structural information than the functional images, allowing a more accurate nonlinear registration to be obtained.

However, not every functional imaging unit has ready access to a high quality MR scanner, so for many functional imaging studies there are no structural images of the subject available to the researcher. In this case, it is necessary to determine the required warps based solely

upon the functional images. These images may have a limited field of view, contain very little useful signal or be particularly noisy. An ideal spatial normalization routine would need to be robust enough to cope with this type of data.

Nonlinear spatial transformations 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 as landmarks is that there are very few readily identifiable discrete points in the brain. A similar problem is faced during identification of homologous lines. 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 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* (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 minimizes some index of the difference between an object and a template 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 coefficient 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 gray levels of the different tissue types between the image and template.

There are a number of approaches to non-label based spatial normalization. A potentially enormous number of parameters are required to describe the nonlinear transformations that

warp two images together (ie. the problem is very high dimensional). The forms of spatial normalization tend to differ in how they cope with the large number of parameters.

Some have abandoned conventional optimization approaches, and use viscous fluid models (Christensen *et al.*, 1994; Christensen *et al.*, 1996) to describe the warps. In these models, finite element methods are 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 nonlinear displacements and also ensure that the topology of the warped image is preserved, but they do have the disadvantage that they are computationally expensive. Not every unit in the functional imaging field has the capacity to routinely perform spatial normalizations using these methods.

Others adopt a multi-resolution approach whereby only a few of the parameters are determined at any one time (Collins *et al.*, 1994b). Usually, the entire volume is used to determine parameters that describe global low frequency deformations. The volume is then subdivided, and slightly higher frequency deformations are found for each subvolume. This continues until the desired deformation precision is achieved.

Another approach is to reduce the number of parameters that model the deformations. Some groups simply use only a nine or twelve parameter affine transformation to spatially normalize their images, accounting for differences in position, orientation and overall brain size. Low spatial frequency global variability in head shape can be accommodated by describing deformations by a linear combination of low frequency basis functions (Amit *et al.*, 1991). 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 method described in this paper is one such approach. The rational for adopting a low dimensional approach is that there is not necessarily a one-to-one mapping between any pair of brains. Different subjects have different patterns of gyral convolutions and even if gyral anatomy can be matched exactly, this is no guarantee that areas of functional specialization will be matched in a homologous way. For the purpose of averaging signals from functional images of different subjects, very

high resolution spatial normalization may be unnecessary or unrealistic.

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 nonlinear warping are more arbitrary. Without any constraints it is of course possible to transform any image such that it matches another exactly. The issue is therefore less about the nature of the transformation and more about defining constraints or priors under which a transformation is effected. The validity of a transformation can usually be reduced to the validity of these priors. Priors are normally incorporated using 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 posteriori probability of being correct, and is the estimate that we attempt to obtain in this paper. The MVE is used by Miller *et al.* (1993; 1994), and is the solution that is the conditional mean of the posterior. The MVE is probably more appropriate than the MAP estimate for spatial normalization. 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.

The remainder of this paper is organized as follows: The theory section describes how the registrations are performed, beginning with a description of the Gauss Newton optimization scheme employed. Following this, the paper describes specific implementational details for determining the optimal linear combination of spatial basis functions. This involves using properties of *Kronecker tensor products* for the rapid computation of the curvature matrix used by the optimization. A Bayesian framework using priors based upon membrane energy is then incorporated into the registration model. The second section provides an evaluation focusing on the utility of nonlinear deformations per se, and then the use of priors in a Bayesian framework. There then follows a discussion of the issues raised, and those for future consideration.

2 Theory

There are two steps involved in registering any pair of images together. There is the *registration* itself, whereby the parameters describing a transformation are determined. Then there is the *transformation*, where one of the images is transformed according to the set of parameters. The registration step involves matching the object image to some form of standardized template image. Unlike in the work of Christensen *et al.* (1994; 1996) or the segmentation work by Collins *et al.* (1994a; 1995), spatial normalization requires that the images themselves are transformed to the space of the template, rather than a transformation being determined that transforms the template to the individual images.

The nonlinear spatial normalization approach described here assumes that the images have already been approximately registered with the template according to a nine- (Collins *et al.*, 1994b) or twelve-parameter (Ashburner *et al.*, 1997) affine registration. This section will illustrate how the parameters describing global shape differences between the images and template are determined.

The section begins by introducing a simple method of optimization based upon partial derivatives. Then the parameters describing the spatial transformations are introduced. In the current approach, the nonlinear warps are modeled 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 optimization method is extended to utilize Bayesian statistics in order to obtain a more robust fit. This requires knowledge of the errors associated with the parameter estimates, and also knowledge of the *a priori* distribution from which the parameters are drawn. This distribution is modeled in terms of a cost function based on the *membrane energy* of the deformations. We conclude with some comments on extending the schemes when matching images from different modalities.

2.1 The Basic Optimization Algorithm

The objective of optimization is to determine a set of parameters for which some function is minimized (or maximized). One of the simplest cases involves determining the optimum parameters for a model in order to minimize of the sum of squared differences between the model and a set of real world data (χ^2). Usually there are many parameters in the model, and it is not possible to exhaustively search through the whole parameter space. The usual approach is to make an initial estimate, and to iteratively search from there. At each iteration, the model is evaluated using the current estimates, and χ^2 computed. A judgement is then made about how the parameters should be modified, before continuing on to the next iteration.

The image registration approach described here is essentially an optimization. In the simplest case, one image (the object image) is spatially transformed so that it matches another (the template image), by minimizing χ^2. The parameters that are optimized are those that describe the spatial transformation (although there are often other nuisance parameters required by the model, such as intensity scaling parameters). The algorithm of choice (Friston *et al.*, 1995) is one that is similar to *Gauss-Newton* optimization, and it is illustrated here:

Suppose that $e_i(\mathbf{p})$ is the function describing the difference between the object and template images at voxel i, when the vector of model parameters have values $\mathbf{p}$. For each voxel (i), a first approximation of Taylor's Theorem can be used to estimate the value that this difference will take if the parameters $\mathbf{p}$ are increased by $\mathbf{t}$:

$$e_i(\mathbf{p}+\mathbf{t}) = e_i(\mathbf{p}) + t_1\frac{\partial e_i(\mathbf{p})}{\partial p_1} + t_2\frac{\partial e_i(\mathbf{p})}{\partial p_2}\ldots$$

This allows the construction of a set of simultaneous equations (of the form $\mathbf{Ax} \simeq \mathbf{e}$) for estimating the values that $\mathbf{t}$ should assume to in order to minimize $\sum_i e_i(\mathbf{p}+\mathbf{t})^2$:

$$\begin{pmatrix} \frac{\partial e_1(\mathbf{p})}{\partial p_1} & \frac{\partial e_1(\mathbf{p})}{\partial p_2} & \cdots \\ \frac{\partial e_2(\mathbf{p})}{\partial p_1} & \frac{\partial e_2(\mathbf{p})}{\partial p_2} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \begin{pmatrix} t_1 \\ t_2 \\ \vdots \end{pmatrix} \simeq \begin{pmatrix} e_i(\mathbf{p}) \\ e_i(\mathbf{p}) \\ \vdots \end{pmatrix}$$

From this we can derive an iterative scheme for improving the parameter estimates. For

iteration n, the parameters $\mathbf{p}$ are updated as:

$$\mathbf{p}^{(n+1)} = \mathbf{p}^{(n)} - \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{e} \tag{1}$$

where $\mathbf{A} = \begin{pmatrix} \frac{\partial e_1(\mathbf{p})}{\partial p_1} & \frac{\partial e_1(\mathbf{p})}{\partial p_2} & \cdots \\ \frac{\partial e_2(\mathbf{p})}{\partial p_1} & \frac{\partial e_2(\mathbf{p})}{\partial p_2} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}$ and $\mathbf{e} = \begin{pmatrix} e_i(\mathbf{p}) \\ e_i(\mathbf{p}) \\ \vdots \end{pmatrix}$.

This process is repeated until χ^2 can no longer be decreased - or for a fixed number of iterations. There is no guarantee that the best global solution will be reached, since the algorithm can get caught in a local minimum. The number of potential local minima is decreased by working with smooth images. This also has the effect of making the first order Taylor approximation more accurate for larger displacements.

In practice, $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}^T\mathbf{e}$ from Eqn. 1 are computed 'on the fly' for each iteration. By computing these matrices using only a few rows of $\mathbf{A}$ and $\mathbf{e}$ at a time, much less computer memory is required than is necessary for storing the whole of matrix $\mathbf{A}$. Also, the partial derivatives $\partial e_i(\mathbf{p})/\partial p_j$ can be rapidly computed from the gradients of the images using the chain rule [1]. These calculations will be illustrated more fully in the next section.

2.2 Parameterizing the Spatial Transformations

The spatial transformations are described by a linear combination of smooth basis functions. The choice of basis functions depends partly upon how translations at the boundaries should behave. If points at the boundary 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 boundaries, 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

[1] Tri-linear interpolation of the voxel lattice (rather than the sampling lattice) is used to resample the images at the desired co-ordinates. Gradients of the images are obtained at the same time, using a finite difference method on the same lattice. No assumptions are made about voxel values that lie outside the field of view of image $\mathbf{f}$. Points where $\mathbf{y}_i$ falls outside the domain of $\mathbf{f}$ are not included in the computations.

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

$$\begin{aligned} b_{m,1} &= \frac{1}{\sqrt{M}} && m = 1..M \\ b_{m,j} &= \sqrt{\frac{2}{M}}\cos(\tfrac{\pi(2m-1)(j-1)}{(2M)}) && m = 1..M, j = 2..J \end{aligned} \tag{2}$$

The two dimensional DCT basis functions are shown in Figure 1, and a schematic example of a deformation based upon the DCT is shown in Figure 2.

[Figure 1 about here.]

[Figure 2 about here.]

The optimized parameters can be separated into a number of distinct groups. The most important are those for describing translations in the three orthogonal directions ($\mathbf{t}_1$, $\mathbf{t}_2$ and $\mathbf{t}_3$). The model for defining the nonlinear warps uses deformations consisting of a linear combination of basis functions. In three dimensions, the transformation from coordinates $\mathbf{x}_i$, to coordinates $\mathbf{y}_i$ is:

$$\begin{aligned} y_{1,i} &= x_{1,i} - u_{1,i} = x_{1,i} - \sum_{j=1}^{J} t_{j,1} b_{1,j}(\mathbf{x}_i) \\ y_{2,i} &= x_{2,i} - u_{2,i} = x_{2,i} - \sum_{j=1}^{J} t_{j,2} b_{2,j}(\mathbf{x}_i) \\ y_{3,i} &= x_{3,i} - u_{3,i} = x_{3,i} - \sum_{j=1}^{J} t_{j,3} b_{3,j}(\mathbf{x}_i) \end{aligned}$$

where t_{jd} is the jth coefficient for dimension d, and $b_{jd}(\mathbf{x})$ is the jth basis function at position $\mathbf{x}$ for dimension d.

The optimization involves minimizing the sum of squared differences between the object image ($\mathbf{f}$) and a 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:

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

Each element of vector $\mathbf{e}$ (from the previous section) contains $f(\mathbf{y}_i) - wg(\mathbf{x}_i)$. Derivatives of the function $f(\mathbf{y}_i) - wg(\mathbf{x}_i)$ with respect to each parameter are required in order to compute matrix $\mathbf{A}$. These can be obtained using the chain rule:

$$\begin{aligned}
\frac{\partial f(\mathbf{y}_i)}{\partial t_{j,1}} &= \frac{\partial f(\mathbf{y}_i)}{\partial y_{1,i}} \frac{\partial y_{1,i}}{\partial t_{j,1}} = \frac{\partial f(\mathbf{y}_i)}{\partial y_{1,i}} b_j(\mathbf{x}_i) \\
\frac{\partial f(\mathbf{y}_i)}{\partial t_{j,2}} &= \frac{\partial f(\mathbf{y}_i)}{\partial y_{2,i}} \frac{\partial y_{2,i}}{\partial t_{j,2}} = \frac{\partial f(\mathbf{y}_i)}{\partial y_{2,i}} b_j(\mathbf{x}_i) \\
\frac{\partial f(\mathbf{y}_i)}{\partial t_{j,3}} &= \frac{\partial f(\mathbf{y}_i)}{\partial y_{3,i}} \frac{\partial y_{3,i}}{\partial t_{j,3}} = \frac{\partial f(\mathbf{y}_i)}{\partial y_{3,i}} b_j(\mathbf{x}_i)
\end{aligned}$$

In order to adopt the Gauss-Newton optimization strategy, $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}^T\mathbf{e}$ need to be computed on each iteration. Assuming that the lowest J frequencies of a three dimensional DCT are used to define the warps, and there are I sampled positions in the image, then the theoretical size of the matrix $\mathbf{A}$ is $I \times (3J^3 + 1)$. The straightforward computation of $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}^T\mathbf{e}$ would be very time consuming. We now describe how this can be done in a much more expedient manner.

2.2.1 A Fast Algorithm

The fast algorithm for computing $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}^T\mathbf{e}$ is shown in Figure 3. The remainder of this section explains the matrix terminology used, and why it is so efficient.

[Figure 3 about here.]

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

matrices $\mathbf{F}$ and $\mathbf{G}$ respectively. Row m of $\mathbf{F}$ is denoted by $\mathbf{f}_{m,:}$, and column n by $\mathbf{f}_{:,n}$. The basis functions used by the algorithm are generated from a separable form from matrices $\mathbf{B_1}$ and $\mathbf{B_2}$. By treating the transform coefficients as matrices $\mathbf{T_1}$ and $\mathbf{T_2}$, the deformation fields can be rapidly constructed by computing $\mathbf{B_1 T_1 B_2}^T$ and $\mathbf{B_1 T_2 B_2}^T$.

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 matrixes 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 $diag(-\nabla_\mathbf{1}\mathbf{f}_{:,m})\mathbf{B_1}$ simply means multiplying each element of row i of $\mathbf{B_1}$ by $-\nabla_\mathbf{1}\mathbf{f}_{i,m}$, and the symbol '$\otimes$' refers to the *Kronecker tensor product*. If $\mathbf{B_2}$ is a matrix of order $M \times J$, and $\mathbf{B_1}$ is a second matrix, then:

$$\mathbf{B_2} \otimes \mathbf{B_1} = \begin{pmatrix} b_{211}\mathbf{B_1} & \dots & b_{21J}\mathbf{B_1} \\ \vdots & \ddots & \vdots \\ b_{2M1}\mathbf{B_1} & \dots & b_{2MJ}\mathbf{B_1} \end{pmatrix}$$

The advantage of the algorithm shown in Figure 3 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 performance enhancement results from a reordering of a set of operations like $(\mathbf{B_3} \otimes \mathbf{B_2} \otimes \mathbf{B_1})^T(\mathbf{B_3} \otimes \mathbf{B_2} \otimes \mathbf{B_1})$, to the equivalent $(\mathbf{B_3}^T\mathbf{B_3}) \otimes (\mathbf{B_2}^T\mathbf{B_2}) \otimes (\mathbf{B_1}^T\mathbf{B_1})$. Assuming that the matrices $\mathbf{B_3}$, $\mathbf{B_2}$ and $\mathbf{B_1}$ all have order $M \times J$, then the number of floating point operations is reduced from $M^3J^3(J^3+2)$ to approximately $3M(J^2+J)+J^6$. If M equals 32, and J equals 4, we expect a performance increase of about a factor of 20,000. 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.

2.3 A *Maximum A Posteriori* Solution

Without regularization, 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. Regularization is achieved using Bayesian statistics.

Bayes rule is generally expressed in the continuous form:

$$p(a_{\mathbf{p}}|b) = \frac{p(b|a_{\mathbf{p}})p(a_{\mathbf{p}})}{\int_{\mathbf{q}} p(b|a_{\mathbf{q}})p(a_{\mathbf{q}})d\mathbf{q}}$$

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

When close to the minimum, the optimization becomes almost a linear problem. This allows us to assume that the errors of the fitted parameters ($\mathbf{p}$) can be locally approximated by a multi-normal distribution with covariance matrix $\mathbf{C}$. We assume that the true parameters are drawn from a known underlying multi-normal distribution of zero mean and covariance matrix ($\mathbf{C_0}$). By using the *a priori* probability density function (p.d.f) of the parameters, we can obtain a better estimate of the true parameters by taking a weighted average of $\mathbf{p}$ and zero:

$$\mathbf{p_b} = (\mathbf{C_0}^{-1} + \mathbf{C}^{-1})^{-1}\mathbf{C}^{-1}\mathbf{p} \tag{3}$$

An estimation of $\mathbf{C}$ is required in order to employ this approach. This is the estimated covariance matrix of the standard errors of the fitted parameters, and is derived from the data itself. If the observations are independent, and each has unit standard deviation, then $\mathbf{C}$ is given by $(\mathbf{A}^T\mathbf{A})^{-1}$. In practice, the standard deviations of the observations are unknown, so we assume they are equal for all observations, and estimate this value from the sum of squared differences:

$$\sigma^2 = \sum_{i=1}^{I} e_i(\mathbf{p})^2/\nu \tag{4}$$

where ν refers to the degrees of freedom. This gives a covariance matrix:

$$\mathbf{C} = (\mathbf{A}^T\mathbf{A})^{-1}\sigma^2 \tag{5}$$

If the sampling is sparse relative to the smoothness, then $\nu = I - J$, where I is the number of sampled locations in the images and J is the number of estimated parameters. However this is not usually the case, so an estimate of the effective degrees of freedom is made as described in Ashburner *et al.* (1997): 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 the Gaussian point spread function can be obtained (Poline *et al.*, 1995). To estimate the degrees of freedom, we assume that the residuals are comprised of a number of independent voxels, that have been convolved with a spatially invariant Gaussian kernel. After convolution, each voxel will contain contributions from its neighbors, as well as from itself. The effective number of degrees of freedom is based upon the ratio of the contribution from the central voxel to the contributions from all voxels. We believe that this component is important because it is essential for a proper characterization of the errors on the parameter estimates. This particular approach accounts for spatial correlations in the images that can clearly differ from image to image and among modalities.

As mentioned previously, when the parameter estimates are close to the minimum the registration problem is almost linear. Prior to this, the problem is nonlinear and covariance matrix $\mathbf{C}$ no longer directly reflects the certainties of the parameter estimates. However, it does indicate the certainties of the changes made in the parameter estimates at each iteration, so this information can still be incorporated into the iterative optimization scheme. By combining Eqns. (1), (3) and (5), we obtain the following scheme:

$$\mathbf{p_b}^{(n+1)} = (\mathbf{C_0}^{-1}\sigma^2 + \mathbf{A}^T\mathbf{A})^{-1}(\mathbf{A}^T\mathbf{A}\mathbf{p_b}^{(n)} - \mathbf{A}^T\mathbf{e}) \tag{6}$$

Another way of thinking about this optimization scheme, is that two criteria are simultaneously being minimized. The first is the sum of squared differences between the images, and the second is a squared distance between the parameters and their known expectation

($\mathbf{p_b}^T\mathbf{C_0}^{-1}\mathbf{p_b}$). 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. The effective degrees of freedom also play a role in this, since the residuals are smoother when the estimates are further from the solution. As the parameters converge to their final solution, the amount of regularization decreases accordingly.

2.3.1 The *A Priori* Distribution

The first requirement for a MAP approach is to define some form of prior distribution for the parameters. For a simple linear 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 modeling these priors, each of which refers to some type of 'energy' term. The form of regularization described here is based upon the *membrane energy* or *laplacians* of the deformation field (Amit *et al.*, 1991; Gee *et al.*, 1997). Two other types of linear regularization (*bending energy* and *linear-elastic energy*) are described in the Appendixes. None of these schemes enforce a strict one to one mapping between the object and template images, but this makes little difference for the small deformations that we are interested in here.

In three dimensions, the *membrane energy* of the deformation field $\mathbf{u}$ is:

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

where λ is simply a scaling constant. The membrane energy can be computed from the coefficients of the basis functions by $\mathbf{t_1}^T\mathbf{H}\mathbf{t_1} + \mathbf{t_2}^T\mathbf{H}\mathbf{t_2} + \mathbf{t_3}^T\mathbf{H}\mathbf{t_3}$, where $\mathbf{t_1}$, $\mathbf{t_2}$ and $\mathbf{t_3}$ refer to vectors containing the parameters describing translations in the three dimensions. The matrix $\mathbf{H}$ is defined by:

$$\begin{aligned}
\mathbf{H} = \quad & \lambda \left(\mathbf{B_3}'^T \mathbf{B_3}' \right) \otimes \left(\mathbf{B_2}^T \mathbf{B_2} \right) \otimes \left(\mathbf{B_1}^T \mathbf{B_1} \right) \\
+ \quad & \lambda \left(\mathbf{B_3}^T \mathbf{B_3} \right) \otimes \left(\mathbf{B_2}'^T \mathbf{B_2}' \right) \otimes \left(\mathbf{B_1}^T \mathbf{B_1} \right) \\
+ \quad & \lambda \left(\mathbf{B_3}^T \mathbf{B_3} \right) \otimes \left(\mathbf{B_2}^T \mathbf{B_2} \right) \otimes \left(\mathbf{B_1}'^T \mathbf{B_1}' \right)
\end{aligned}$$

where the notation $\mathbf{B_1}'$ refers to the first derivatives of $\mathbf{B_1}$.

Assuming that the parameters consist of $(\mathbf{t_1}^T\mathbf{t_2}^T\mathbf{t_3}^T w)^T$, matrix $\mathbf{C_0}^{-1}$ from Eqn. 6 can be constructed from $\mathbf{H}$ by:

$$\mathbf{C_0}^{-1} = \begin{pmatrix} \mathbf{H} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{H} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{H} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & 0 \end{pmatrix} \tag{7}$$

$\mathbf{H}$ is all zeros, except for the diagonal. Elements on the diagonal represent the reciprocal of the *a priori* variance of each parameter, and each is given by:

$$h_{j+J(k+J\times l)} = \lambda\pi^2 M^{-2}\left((j-1)^2 + (k-1)^2 + (l-1)^2\right)$$

where M is the dimension of the DCT (see Eq. 2), and J is the number of low frequency coefficients used in any dimension.

Values of λ that are too large will provide too much regularization and result in under-estimated deformations. If the values are too small, there will not be enough regularization and the resulting deformations will include a large amount of noise. There is no simple way of estimating what the best values for these constants should be.

In the absence of known priors, the membrane energy provides a useful model in which we assume that the probability of a particular set of parameters arising is inversely related to the membrane energy associated with that set. Clearly this model is somewhat ad hoc, but is a useful and sensible one. If the true prior distribution of the parameters is known (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.

2.4 Templates and Intensity Transformations

So far, only a single intensity scaling parameter (w) has been considered. This is most effective when there is a linear relation between the images. However, for spatially normalizing some images, it is necessary to include other parameters describing intensity transformations.

The optimization can be assumed to minimize two sets of parameters: those that describe spatial transformations ($\mathbf{p_t}$), and those for describing intensity transformations ($\mathbf{p_w}$). This means that the difference function can be expressed in the generic form:

$$e_i(\mathbf{p}) = f(\mathbf{t}(\mathbf{x}_i, \mathbf{p_t})) - w(\mathbf{x}_i, \mathbf{p_w})$$

where $\mathbf{f}$ is the object image, $\mathbf{t}()$ is a vector function describing the spatial transformations based upon parameters $\mathbf{p_t}$ and $w()$ is a scalar function describing intensity transformations based on parameters $\mathbf{p_w}$. $\mathbf{x}_i$ represents the coordinates of the ith sampled point.

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:

$$\sum_i \left(f(\mathbf{x}_i, \mathbf{p_t}) - (p_{w1} g(\mathbf{x}_i) + p_{w2} x_{1i} g(\mathbf{x}_i) + p_{w3} x_{2i} g(\mathbf{x}_i) + p_{w4} x_{3i} g(\mathbf{x}_i))\right)^2$$

More complex variations could be included by modulating with other basis functions (such as the DCT basis function set described earlier) (Friston *et al.*, 1995). Information on the smoothness of the inhomogeneities could be incorporated by appropriate modifications to the matrix $\mathbf{C_0}^{-1}$.

Another 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:

$$\sum_i \left(f(\mathbf{M}\mathbf{x}_i) - (p_{w1}g_1(\mathbf{x}_i) + p_{w2}g_2(\mathbf{x}_i))\right)^2$$

Again, some form of model describing the likely *a priori* distributions of the parameters could be included.

3 Evaluation

This section provides an anecdotal evaluation of the techniques presented in the previous section. We compare spatial normalization both with and without nonlinear deformations, and compare nonlinear deformations with and without the use of Bayesian priors.

T1 weighted MR images of 12 subjects were spatially normalized to the same anatomical space. The normalizations were performed twice, first using only 12 parameter affine transformations and then using affine transformations followed by nonlinear transformations. The nonlinear transformation used 392 parameters to describe deformations in each of the directions, and four parameters to model a linear scaling and simple linear image intensity inhomogenaeties (making a total of 1180 parameters in all). The basis functions were those of a three dimensional DCT, and the regularization minimized the membrane energy of the deformation fields. Twelve iterations of the nonlinear registration algorithm were performed.

Figure 4 shows pixel by pixel means and standard deviations of the normalized images. The mean image from the nonlinear normalization shows more contrast and has edges that are slightly sharper. The standard deviation image for the nonlinear normalization 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.

[Figure 4 about here.]

This evaluation should illustrate the fact that the nonlinear normalization 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 regularization. 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 favorable results for less heavily regularized algorithms. With less regularization, the optimum solution is based more upon minimizing 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 5, where the distortions of gyral anatomy clearly have a very low face validity (lower right panel).

[Figure 5 about here.]

4 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 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 normalization (either within or between subjects) in functional mapping studies is to maximize the sensitivity to neurophysiological 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.

4.1 Limitations of the Nonlinear Registration

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 modeled, since 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 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 (about 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 regularization, this constraint is rarely broken.

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). Other methods which

search for more parameters should be used when more precision is required. 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).

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. These methods are very good at registering grey matter with gray 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 studies prior to performing the 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 registering 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. Lesioned areas could be assigned lower weights, so that they have much less

influence on the final solution.

4.2 Conclusions

Consider the deformation fields required to map brain images to a common stereotactic space. Fourier transforming the fields reveal that most of the variance is low frequency - even when the deformations have been determined using good high dimensional nonlinear registration methods. Therefore, an efficient representation of the fields can be obtained from the low frequency coefficients of the transform. The current approach to spatial normalization utilizes this compression. Rather than estimating warps based upon literally millions of parameters, only a few hundred parameters are used to represent the deformations as a linear combination of a few low frequency basis functions.

The method we have developed is automatic and non-label based. A *maximum a posteriori* (MAP) approach is used to regularize the optimization. However, the main difference between this and other MAP approaches is that an estimate of the errors is derived from the data itself. This estimate also includes a correction for local correlations between voxels. An implication of this is that the approach is suitable for spatially normalizing a wide range of different image modalities. High quality MR images, and also low resolution noisy PET images can be treated the same way.

The spatial normalization converges rapidly, because it uses an optimization strategy with fast local convergence properties. Each iteration of the algorithm requires the computation of a Hessian matrix ($\mathbf{A}^T\mathbf{A}$). The straightforward computation of this matrix would be prohibitively time consuming. However, this problem has been solved by developing an extremely fast method of computing this matrix that relies on the separable properties of the basis functions. A performance increase of several orders of magnitude is achieved in this way.

Acknowledgements

This work was supported by the Wellcome Trust.

References

Amit, Y., Grenander, U., & Piccioni, M. 1991. Structural Image Restoration through Derormable Templates. *Journal of the American Statistical Association*, **86**, 376–387.

Ashburner, J., & Friston, K. J. 1997. Multimodal Image Coregistration and Partitioning - a Unified Framework. *to appear in NeuroImage.*

Ashburner, J., Neelin, P., Collins, D. L., Evans, A. C., & Friston, K. J. 1997. Incorporating Prior Knowledge into Image Registration. *to appear in NeuroImage.*

Bookstein, F. L. 1989. Principal Warps: Thin-plate splines and the decomposition of deformations. *IEEE Trans Pattern Anal Machine Intelligence*, **11(6)**, 567–585.

Bookstein, F. L. 1997a. Landmark Methods for Forms Without Landmarks: Morphometrics of Group Differences in Outline Shape. *Medical Image Analysis*, **1**(3), 225–243.

Bookstein, F. L. 1997b. Quadratic Variation of Deformations. *Pages 15–28 of: Information Processing in Medical Imaging.*

Christensen, G. E. 1994. Deformable Shape Models for Anatomy. *Doctoral Thesis.*

Christensen, G. E., Rabbitt, R. D., & Miller, M. I. 1994. 3D Brain Mapping Using Using a Deformable Neuroanatomy. *Physics in Medicine and Biology*, **39**, 609–618.

Christensen, G. E., Rabbitt, R. D., & Miller, M. I. 1996. Deformable Templates Using Large Deformation Kinematics. *IEEE Transactions on Image Processing*, **5**, 1435–1447.

Collignon, A., Maes, F., Delaere, D., Vandermeulen, D., Suetens, P., & Marchal, G. 1995. Automated Multi-Modality Image Registration Based on Information Theory. *Pages 263–274 of: Information Processing in Medical Imaging.*

Collins, D. L., Peters, T. M., & Evans, A. C. 1994a. An automated 3D Non-linear image deformation procedure for determination of gross morphometric variability in human brain. *Proc. Conference on Visualisation in Biomedical Computing*, 180–190.

Collins, D. L., Neelin, P., Peters, T. M., & Evans, A. C. 1994b. Automatic 3D intersubject registration of MR volumetric data in standardized Taliarach space. *J. Comput. Assist. Tomogr.*, **18**, 192–205.

Collins, D. L., Evans, A. C., Holmes, C., & Peters, T. M. 1995. Automatic 3D Segmentation of Neuro-Anatomical Structures from MRI. *Pages 139–152 of: Information Processing in Medical Imaging.*

Fox, P. T. 1995. Spatial Normalization Origins: Objectives, Applications, and Alternatives. *Human Brain Mapping*, **3**, 161–164.

Friston, K. J., Ashburner, J., Frith, C. D., Poline, J.-B., Heather, J. D., & Frackowiak, R. S. J. 1995. Spatial Registration and Normalization of Images. *Human Brain Mapping*, **2**, 165–189.

Gee, J. C., Haynor, D. R., Briquer, L. Le, & Bajcsy, R. K. 1997. Advances in Elastic Matching Theory and its Implementation. *In: CVRMed-MRCAS'97.* Springer-Verlag.

Mazziotta, J. C., Toga, A. W., Evans, A., Fox, P., & Lancaster, J. 1995. A Probablistic Atlas of the Human Brain: Theory and Rationale for Its Development. *NeuroImage*, **2**, 89–101.

Miller, M. I., Christensen, G. E., Amit, Y., & Grenander, U. 1993. Mathematical Textbook of Deformable Neuroanatomies. *Proc. Natl. Acad. Sci.*, **90**, 11944–11948.

Pelizzari, C. A., Chen, G. T. Y., Spelbring, D. R., Weichselbaum, R. R., & Chen, C. T. 1988. Accurate three-dimensional registration of CT, PET and MR images of the brain. *J. Comput. Assist. Tomogr.*, **13**, 20–26.

Poline, J.-B., Friston, K. J., Worsley, K. J., & Frackowiak, R. S. J. 1995. Estimating Smoothness in Statistical Parametric Maps: Confidence Intervals on p-Values. *J. Comput. Assist. Tomogr.*, **19(5)**, 788–796.

Studholme, C., Hill, D. L. G., & Hawkes, D. J. 1995. Multiresolution Voxel Similarity Measures for MR-PET Coregistration. *Pages 287–298 of: Information Processing in Medical Imaging.*

Talairach, J., & Tournoux. 1988. *Coplanar stereotaxic atlas of the human brain.* New York: Thieme Medical.

Thompson, P. M., & Toga, A. W. 1996. Visualization and Mapping of Anatomic Abnormalities using a Probablistic Brain Atlas Based on Random Fluid Transformations. *Pages 383–392 of: Proceedings of the International Conference on Visualization in Biomedical Computing.*

Woods, R. P., Cherry, S. R., & Mazziotta, J. C. 1992. Rapid automated algorithm for aligning and reslicing PET images. *J. Comput. Assist. Tomogr.*, **16**, 620–633.

A Other Linear Priors

As an alternative to the *membrane energy* prior already discussed, we now show how two other priors can easily be incorporated into the model. For simplicity, they will only be shown for the two dimensional case.

A.1 Bending Energy

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

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

This can be computed by:

$$\begin{aligned} \lambda \mathbf{t_1}^T (\mathbf{B_2}'' \otimes \mathbf{B_1})^T (\mathbf{B_2}'' \otimes \mathbf{B_1}) \mathbf{t_1} \quad &+ \quad \lambda \mathbf{t_1}^T (\mathbf{B_2} \otimes \mathbf{B_1}'')^T (\mathbf{B_2} \otimes \mathbf{B_1}'') \mathbf{t_1} + \\ 2\lambda \mathbf{t_1}^T (\mathbf{B_2}' \otimes \mathbf{B_1}')^T (\mathbf{B_2}' \otimes \mathbf{B_1}') \mathbf{t_1} \quad &+ \quad \lambda \mathbf{t_2}^T (\mathbf{B_2}'' \otimes \mathbf{B_1})^T (\mathbf{B_2}'' \otimes \mathbf{B_1}) \mathbf{t_2} + \\ \lambda \mathbf{t_2}^T (\mathbf{B_2} \otimes \mathbf{B_1}'')^T (\mathbf{B_2} \otimes \mathbf{B_1}'') \mathbf{t_2} \quad &+ \quad 2\lambda \mathbf{t_2}^T (\mathbf{B_2}' \otimes \mathbf{B_1}')^T (\mathbf{B_2}' \otimes \mathbf{B_1}') \mathbf{t_2} \end{aligned}$$

where the notation $\mathbf{B_1}'$ and $\mathbf{B_1}''$ refer to the first and second derivatives of $\mathbf{B_1}$. This is simplified to $\mathbf{t_1}^T \mathbf{H} \mathbf{t_1} + \mathbf{t_2}^T \mathbf{H} \mathbf{t_2}$ where:

$$\mathbf{H} = \lambda \left(\left(\mathbf{B_2}''^T \mathbf{B_2}'' \right) \otimes \left(\mathbf{B_1}^T \mathbf{B_1} \right) + \left(\mathbf{B_2}^T \mathbf{B_2} \right) \otimes \left(\mathbf{B_1}''^T \mathbf{B_1}'' \right) + 2 \left(\mathbf{B_2}'^T \mathbf{B_2}' \right) \otimes \left(\mathbf{B_1}'^T \mathbf{B_1}' \right) \right)$$

Matrix $\mathbf{C_0}^{-1}$ from Eqn. 6 can be constructed from $\mathbf{H}$ as described by Eqn. 7, but this time values on the diagonals are given by:

$$h_{j+k \times J} = \lambda \left(\left(\frac{\pi(j-1)}{M} \right)^4 + \left(\frac{\pi(k-1)}{M} \right)^4 + 2 \left(\frac{\pi(j-1)}{M} \right)^2 \left(\frac{\pi(k-1)}{M} \right)^2 \right)$$

A.2 Linear Elasticity

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

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

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

$$\begin{aligned}
&(\mu+\lambda/2)\mathbf{t_1}^T(\mathbf{B_2}\otimes\mathbf{B_1}')^T(\mathbf{B_2}\otimes\mathbf{B_1}')\mathbf{t_1}+(\mu+\lambda/2)\mathbf{t_2}^T(\mathbf{B_2}'\otimes\mathbf{B_1})^T(\mathbf{B_2}'\otimes\mathbf{B_1})\mathbf{t_2}\\
&+\mu/2\mathbf{t_1}^T(\mathbf{B_2}'\otimes\mathbf{B_1})^T(\mathbf{B_2}'\otimes\mathbf{B_1})\mathbf{t_1}+\mu/2\mathbf{t_2}^T(\mathbf{B_2}\otimes\mathbf{B_1}')^T(\mathbf{B_2}\otimes\mathbf{B_1}')\mathbf{t_2}\\
&+\mu/2\mathbf{t_1}^T(\mathbf{B_2}'\otimes\mathbf{B_1})^T(\mathbf{B_2}\otimes\mathbf{B_1}')\mathbf{t_2}+\mu/2\mathbf{t_2}^T(\mathbf{B_2}\otimes\mathbf{B_1}')^T(\mathbf{B_2}'\otimes\mathbf{B_1})\mathbf{t_1}\\
&+\lambda/2\mathbf{t_1}^T(\mathbf{B_2}\otimes\mathbf{B_1}')^T(\mathbf{B_2}'\otimes\mathbf{B_1})\mathbf{t_2}+\lambda/2\mathbf{t_2}^T(\mathbf{B_2}'\otimes\mathbf{B_1})^T(\mathbf{B_2}\otimes\mathbf{B_1}')\mathbf{t_1}
\end{aligned}$$

A regularization 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:

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

where:

$$\begin{aligned}
\mathbf{H_1} &= (\mu+\lambda/2)(\mathbf{B_2}^T\mathbf{B_2})\otimes(\mathbf{B_1}'^T\mathbf{B_1}')+\mu/2(\mathbf{B_2}'^T\mathbf{B_2}')\otimes(\mathbf{B_1}^T\mathbf{B_1})\\
\mathbf{H_2} &= (\mu+\lambda/2)(\mathbf{B_2}'^T\mathbf{B_2}')\otimes(\mathbf{B_1}^T\mathbf{B_1})+\mu/2(\mathbf{B_2}^T\mathbf{B_2})\otimes(\mathbf{B_1}'^T\mathbf{B_1}')\\
\mathbf{H_3} &= \lambda/2(\mathbf{B_2}^T\mathbf{B_2}')\otimes(\mathbf{B_1}'^T\mathbf{B_1})+\mu/2(\mathbf{B_2}'^T\mathbf{B_2})\otimes(\mathbf{B_1}^T\mathbf{B_1}')
\end{aligned}$$

List of Figures

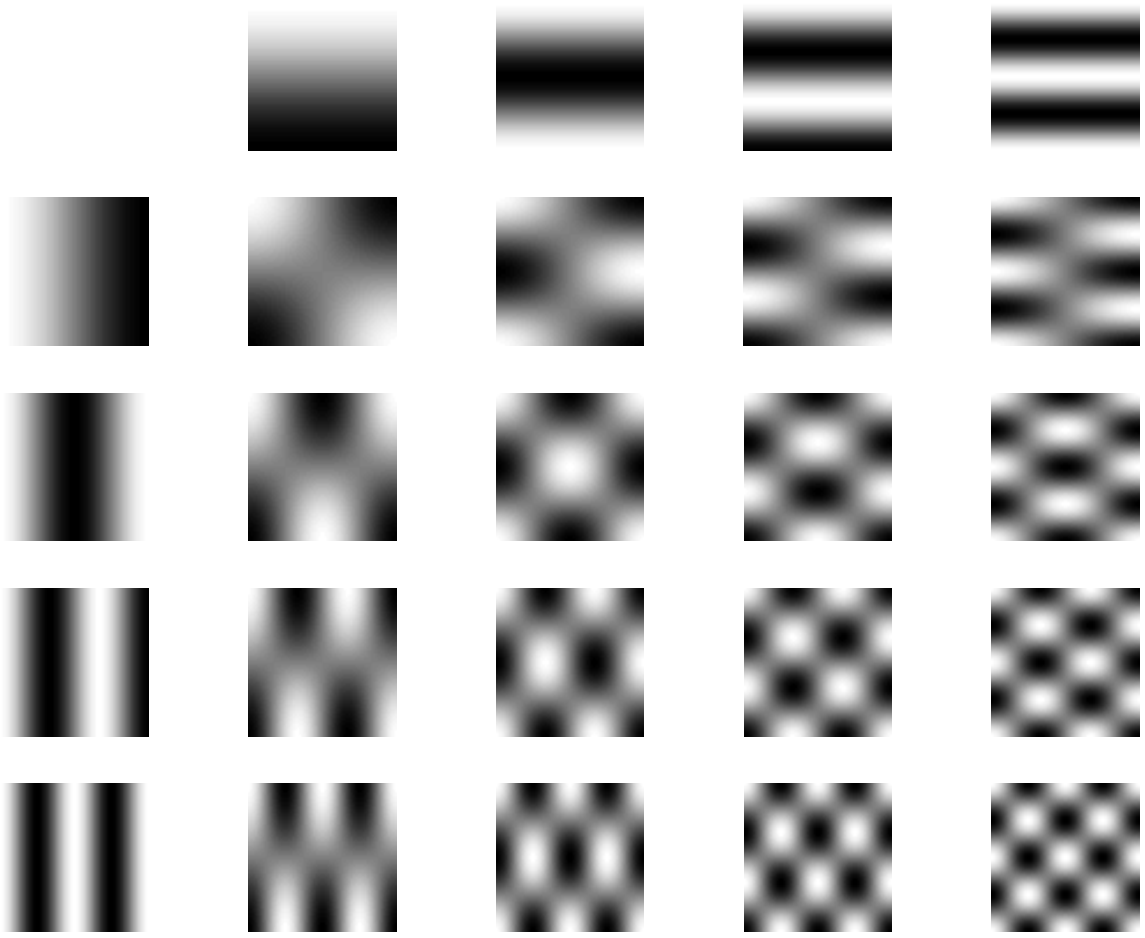

Figure 1: The lowest frequency basis functions of a two dimensional Discrete Cosine Transform.

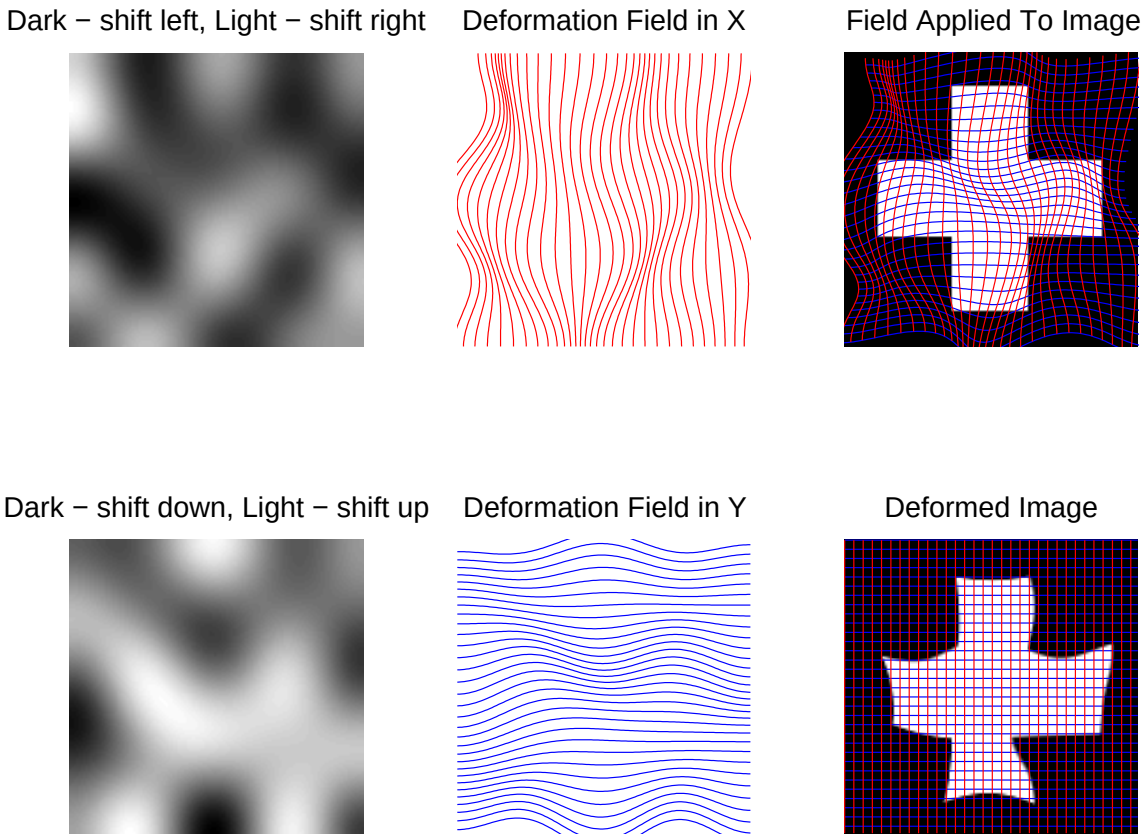

Figure 2: For the two dimensional case, the deformation field consists of two scalar fields. One for horizontal deformations, and the other for vertical deformations. The images on the left show the deformation fields as a linear combination of the basis images (see Figure 1). The center column shows the deformations in a more intuitive sense. The deformation field is applied by overlaying it on the object image, and re-sampling (right).

$$\alpha = (\,\mathbf{0}\,)$$
$$\beta = (\,\mathbf{0}\,)$$

$$for \quad m = 1 \ldots M$$
$$\mathbf{C} = \mathbf{b_2}_{m,:}{}^T \mathbf{b_2}_{m,:}$$
$$\mathbf{E_1} = diag(-\nabla_1 \mathbf{f}_{:,m})\mathbf{B_1}$$
$$\mathbf{E_2} = diag(-\nabla_2 \mathbf{f}_{:,m})\mathbf{B_1}$$

$$\alpha = \alpha + \begin{pmatrix} \mathbf{C} \otimes (\mathbf{E_1}{}^T \mathbf{E_1}) & \mathbf{C} \otimes (\mathbf{E_1}{}^T \mathbf{E_2}) & \mathbf{b_2}_{m,:}{}^T \otimes (\mathbf{E_1}{}^T \mathbf{g}_{:,m}) \\ (\mathbf{C} \otimes (\mathbf{E_1}{}^T \mathbf{E_2}))^T & \mathbf{C} \otimes (\mathbf{E_2}{}^T \mathbf{E_2}) & \mathbf{b_2}_{m,:}{}^T \otimes (\mathbf{E_2}{}^T \mathbf{g}_{:,m}) \\ (\mathbf{b_2}_{m,:}{}^T \otimes (\mathbf{E_1}{}^T \mathbf{g}_{:,m}))^T & (\mathbf{b_2}_{m,:}{}^T \otimes (\mathbf{E_1}{}^T \mathbf{g}_{:,m}))^T & \mathbf{g}_{:,m}{}^T \mathbf{g}_{:,m} \end{pmatrix}$$

$$\beta = \beta + \begin{pmatrix} \mathbf{b_2}_{m,:}{}^T \otimes (\mathbf{E_1}{}^T (\mathbf{f}_{:,m} - w\mathbf{g}_{:,m})) \\ \mathbf{b_2}_{m,:}{}^T \otimes (\mathbf{E_2}{}^T (\mathbf{f}_{:,m} - w\mathbf{g}_{:,m})) \\ \mathbf{g}_{:,m}{}^T (\mathbf{f}_{:,m} - w\mathbf{g}_{:,m}) \end{pmatrix}$$

$$end$$

Figure 3: A two dimensional illustration of the fast algorithm for computing $\mathbf{A}^T\mathbf{A}$ (α) and $\mathbf{A}^T\mathbf{e}$ (β).

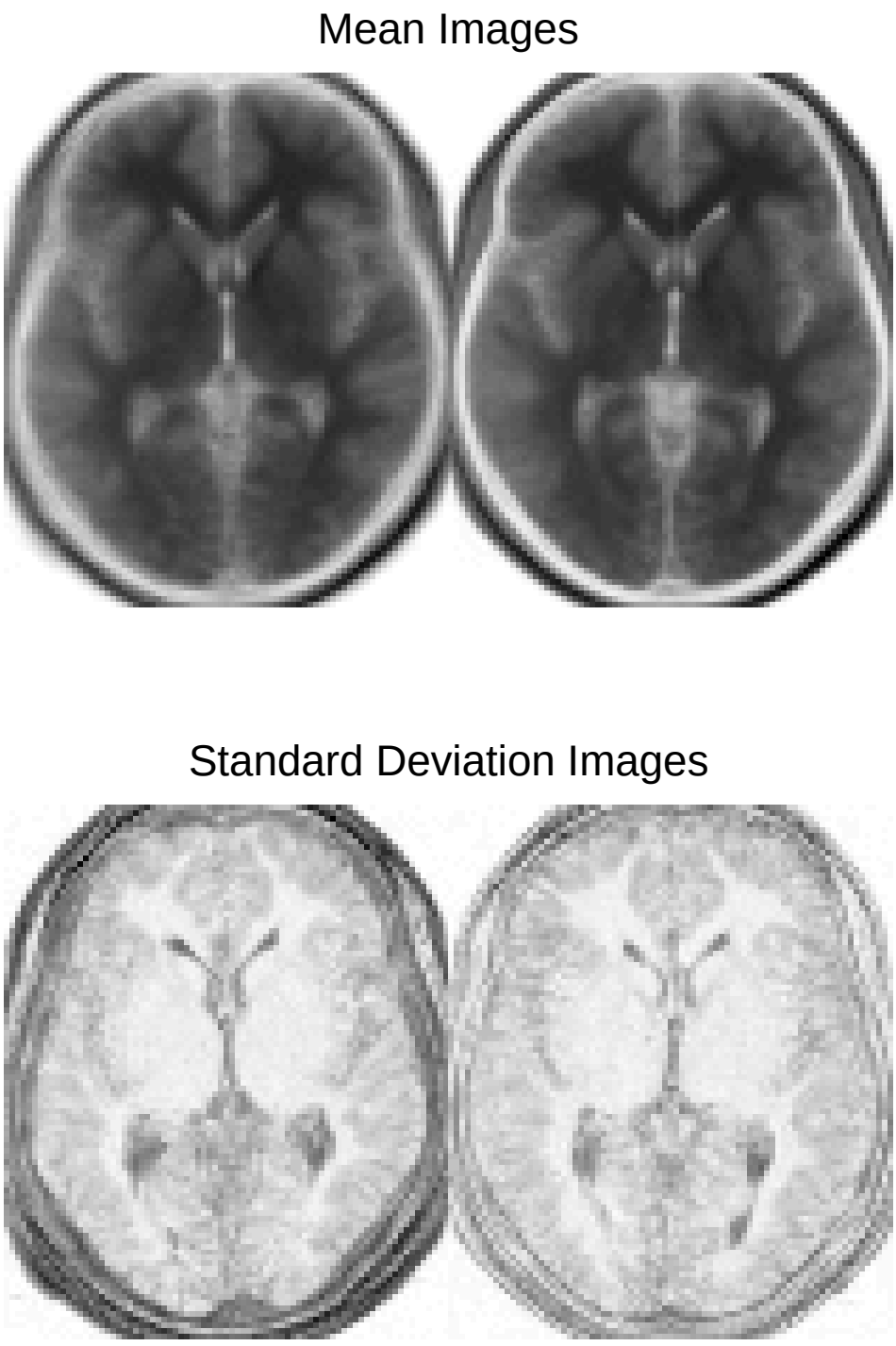

Figure 4: Means and standard deviations of spatially normalized T1 weighted images from 12 subjects. The images on the left were derived using only affine registration. Those on the right used nonlinear registration in addition to the affine registration.

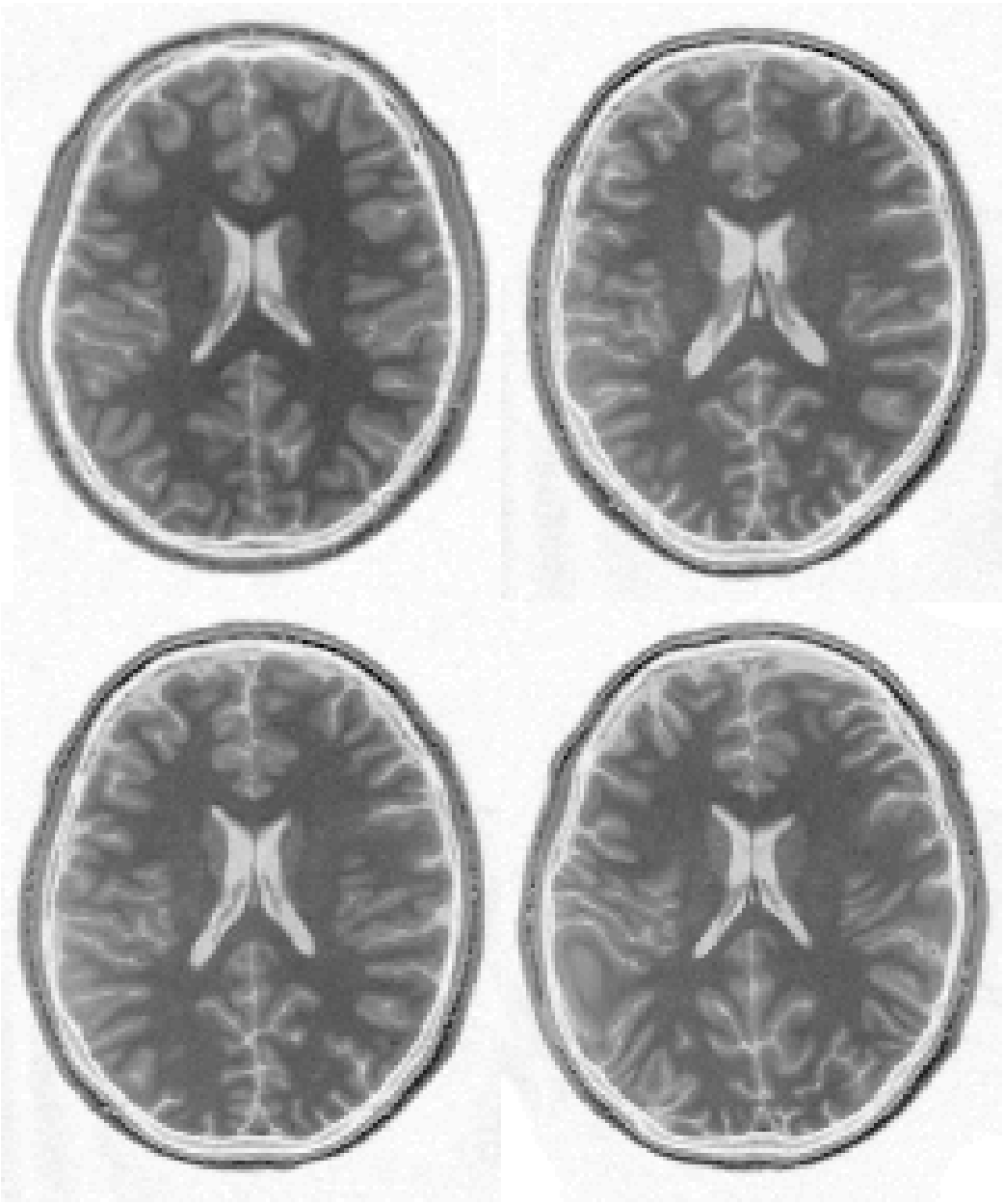

Figure 5: The image shown at the top-left is the object or template image. At the top-right is an image that has been registered with it using a 12-parameter affine registration. The image at the bottom-left is the same image registered using the 12-parameter affine registration, followed by a regularized global nonlinear registration (using 1180 parameters and 12 iterations). 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.