A Minimum Description Length Objective Function for Groupwise Non-Rigid Image Registration

Stephen Marsland[01] and Carole Twining[02]

[1]IIST, Massey University, Private Bag 11222, Palmerston North, NZ

[2]ISBE, University of Manchester, Oxford Road, Manchester M13 9PT, UK

Abstract

Groupwise non-rigid registration aims to find a dense correspondence across a set of images, so that analogous structures in the images are aligned. For purely automatic inter-subject registration the meaning of correspondence should be derived purely from the available data (i.e., the full set of images), and can be considered as the problem of learning correspondences given the set of example images. We argue that the Minimum Description Length (MDL) approach is a suitable method of statistical inference for this problem, and we give a brief description of applying the MDL approach to transmitting both single images and sets of images, and show that the concept of a reference image (which is central to defining a *consistent* correspondence across a set of images) appears naturally as a valid model choice in the MDL approach. This paper provides a proof-of-concept for the construction of objective functions for image registration based on the MDL principle.

Keywords: Non-rigid registration, correspondence problem, Minimum Description Length, groupwise registration, objective function

1 Introduction

The non-rigid registration of pairs of medical images is a popular area of research, with many different methods having been proposed. One motivation for bringing images into alignment is to allow statistical analysis of the variation of structures within those images, for example, for disease diagnosis, through analysis of the deformation field. This is impractical based on a set of pairwise registrations, because the choice of reference image is crucial, and the same one-to-one correspondence will not be found across the whole set of images. We propose a method of groupwise image registration based on the Minimum Description Length (MDL) principle, which aims to find a dense one-to-one correspondence across a set of images, so that analogous structures in all of the images in the set are aligned. MDL [1] is an approach to model-selection and statistical inference that does not depend on hypothetical data-generating processes; the MDL principle has previously been shown to give excellent results when applied to the correspondence problem in shape modelling [2].

Algorithms for the automatic non-rigid registration of medical images typically involve two independent choices: the objective function, the extremum of which defines what is meant by the 'best' correspondence between the images, and the representation of the deformation field that defines the dense correspondence between the images. The choice of representation of the deformation field applies implicit constraints on the possible deformations, and hence on the possible correspondences. The objective function is typically a sum of several terms – a voxel-based similarity measure, and terms that assign a cost to each deformation. The problem with this approach is that these terms are incommensurate quantities, so that we have to determine appropriate values for their coefficients.

The fact that the inferences we make about the data do not depend on hypothetical data-generating processes is particularly important when the registration is of images of different subjects (inter-subject). In this case, there is no underlying physical process that generates the data, and so, in the absence of expert anatomical knowledge (i.e., for the case of purely *automatic* registration), the meaning of correspondences should be derived purely from the available data (i.e., the set of images). In intra-subject registration there is often some actual physical process determining the observed deformation, for example, tissue deformation due to patient position, the insertion of an external object such as a needle, or patient and organ motion. Alternatively, the deformation may be caused by atrophy, such as in dementia, or growth, as in a tumour. In either case, the most suitable choice of

[0]Joint first authors. s.r.marsland@massey.ac.nz; carole.twining@man.ac.uk

registration algorithm may well be one that closely models the underlying physical process, leading to physically-based registration algorithms (e.g., [3]), or physically-based models (e.g., [4]) that can be used to evaluate the results of non-rigid registration algorithms.

2 Introduction to the Minimum Description Length (MDL)

The MDL principle states that the best model to represent some given data is the one that gives the smallest stochastic complexity to the data, where the stochastic complexity is the length of the message required to transmit the data to some observer, when the data is encoded using the specified model. In general, a complete message consists of two parts – the parameter values of the model, and the data encoded using the model. Code lengths (the length of the encoded message required to transmit one parameter or one piece of data) are calculated using the fundamental result of Shannon [5] – if there are a set of possible, discrete events $\{i\}$ with associated model probabilities $\{p_i\}$, then the optimum code length required to transmit the occurrence of event i is given by:

$$\mathcal{L}_i = -\ln p_i \text{ nats.} \tag{1}$$

The total message length/description length is then given by the sum of the parameter length and the data length:

$$\mathcal{L} = \mathcal{L}_{\text{para}} + \mathcal{L}_{\text{data}}, \;\; \mathcal{L}_{\text{data}} = \sum_i \mathcal{L}_i, \tag{2}$$

where the parameter length $\mathcal{L}_{\text{para}}$ is the sum of the code lengths for transmitting the set of parameter values of the model. It is trivial to show that the data length $\mathcal{L}_{\text{data}}$ is minimised when the model probabilities $\{p_i\}$ exactly match the empirical distribution of the data. The MDL criterion minimises the description length $\mathcal{L}$, balancing model complexity (as measured by $\mathcal{L}_{\text{para}}$) against the degree of match between the empirical and model distributions.

Consider a positive integer of the form $n = 2^k$, $k \in \mathbb{Z}^+$, which contains k bits, with:

$$\mathcal{L}_{\text{int}}(n) = k \text{ bits} = 1 + \text{int}\left(\log_2 n\right) \text{ bits.} \tag{3}$$

Using natural logarithms rather than base 2, this then gives an approximate message length for transmission of an unsigned integer n of:

$$\mathcal{L}_{\text{int}}(n) \approx \frac{1}{e} + \ln n \text{ nats,} \tag{4}$$

where we have converted from bits to nats (1 nat $\equiv e$ bits), as well as converting to a continuum version of the function.

Similarly, the message length to transmit a real number x to an accuracy $\delta = 2^k$, $k \in \mathbb{Z}$, and to transmit the accuracy δ is given by:

$$\begin{aligned} \mathcal{L}(x;\delta) &= \mathcal{L}_{\text{int}}\left(\frac{x}{\delta}\right) \approx \frac{1}{e} + \ln\left(\frac{x}{\delta}\right) \text{ nats,} & (5) \\ \mathcal{L}(\delta) &= 1 + \text{int}|\log_2(\delta)| \text{ bits} \\ &\approx \frac{1}{e}\left(1 + |\log_2(\delta)|\right) \text{ nats.} & (6) \end{aligned}$$

It is important to note that in this final approximation, $\mathcal{L}(\cdot)$ is *not* a continuous function – it is only strictly defined for arguments $\delta = 2^k$, $k \in \mathbb{Z}$.

3 MDL Encoding of Single Images

Let us consider the simple case of transmitting one-dimensional ordered quantized[1] data $\{\hat{y}_i : i = 1, \ldots N\}$, with quantization parameter Δ. We will suppose that the data is such that the mean is approximately zero, and that the receiver already knows this and the quantization parameter, but that the range of the data is not known *a priori*.

We consider two models in this paper. The first is one of the simplest parameterised models: a Gaussian model, of zero mean and width $\hat{\sigma}$. The width is quantized using a parameter δ_σ, where δ_σ is restricted to the set of values $\{\delta_\sigma = 2^k : k \in \mathbb{Z}, \;\; \delta_\sigma \leq \sigma\}$. Both δ_σ and $\hat{\sigma}$ have to be transmitted. We will consider here just the case where[2] $\hat{\sigma} \gg \Delta$. The full set of model probabilities $\{p(\hat{y}) : \hat{y} = m\Delta, \, m \in \mathbb{Z}\}$ can be approximated by:

$$\begin{aligned} p(\hat{y}) &= \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \int_{\hat{y}-\frac{\Delta}{2}}^{\hat{y}+\frac{\Delta}{2}} \exp\left(-\frac{\hat{y}^2}{2\hat{\sigma}^2}\right) \\ &\approx \frac{\Delta}{\sqrt{2\pi\hat{\sigma}^2}} \exp\left(-\frac{\hat{y}^2}{2\hat{\sigma}^2}\right), & (7) \end{aligned}$$

providing a complete description length of:

$$\begin{aligned} \mathcal{L}_{\text{Gauss}} \;\; & (\{\hat{y}_i\}) = \mathcal{L}_{\text{int}}\left(\frac{\hat{\sigma}}{\delta_\sigma}\right) + \mathcal{L}(\delta_\sigma) \\ - \;\; & \sum_{i=1}^{N} \log_2(p(\hat{y}_i)) \text{ bits} & (8) \\ \approx \;\; & -N\ln\Delta + \frac{N}{2}\ln(2\pi) + \frac{1}{e} + N\ln(\hat{\sigma}) \\ + \;\; & \ln\left(\frac{\hat{\sigma}}{\delta_\sigma}\right) + \mathcal{L}(\delta_\sigma) + \sum_{i=1}^{N} \frac{\hat{y}_i^2}{2\hat{\sigma}^2} \text{ nats.} & (9) \end{aligned}$$

If we treat the quantized variable $\hat{\sigma}$ as a continuous variable σ (i.e., we take the limit $\delta_\sigma \to 0$), then for fixed data the optimum continuum value is given by $\sigma^2_{\text{opt}} = \frac{1}{N+1}\sum_{i=1}^{N} \hat{y}_i^2$.

[1] We will use $\hat{\cdot}$ to denote quantized variables.

[2] The case of Gaussian models where $\hat{\sigma} \approx \Delta$ is dealt with in [2], although only for the case of data where the range is known.

We would also like to be able to estimate the optimum value for δ_σ – that is, we wish to find a continuous function of δ_σ that approximates the discrete function given in (8). There are two types of terms involving δ_σ: the approximation of the $\log_2$ term, and terms arising from the quantization of σ_{opt}. If $\delta_\sigma < 1$ then $\mathcal{L}(\delta_\sigma) \approx \frac{1}{e} - \ln(\delta_\sigma)$ nats. We know that $|\hat{\sigma}_{\text{opt}} - \sigma_{\text{opt}}| \leq \frac{\delta_\sigma}{2}$ and that the data, and hence σ_{opt}, is *fixed*, whilst δ_σ, and hence $\hat{\sigma}_{\text{opt}}$, vary. We therefore take $\hat{\sigma}_{\text{opt}} = \sigma_{\text{opt}} + d\sigma$, where we will assume that $d\sigma$ has a flat distribution within the range $|d\sigma| \leq \frac{\delta_\sigma}{2}$. So, our *estimate* of functions $f(\hat{\sigma}_{\text{opt}})$ is:

$$f(\hat{\sigma}_{\text{opt}}) \approx f(\sigma_{\text{opt}}) + \frac{\delta_\sigma^2}{24} f''(\sigma_{\text{opt}}) + O\left(\delta_\sigma^4\right). \quad (10)$$

Then, using the expansion of $\ln(\hat{\sigma}_{\text{opt}})$ we get a description length of:

$$\mathcal{L}_{\text{Gauss}}(\{\hat{y}_i\}) \approx \frac{2}{e} - N\ln(\Delta) + \frac{N}{2}\ln(2\pi) + \frac{N+3}{2} + (N+1)\ln\sigma_{\text{opt}} - \ln\frac{12\sigma_{\text{opt}}^2}{(N+1)}, \quad (11)$$

where we have used the fact that, to lowest order, the optimum value of the parameter accuracy δ_σ is given by:

$$\delta_\sigma^2 = \frac{12\sigma_{\text{opt}}^2}{(N+1)}. \quad (12)$$

In Figure 1 we compare the exact expression for the description length (8) with the approximate continuous form (11) for 3 datasets of varying variance. Each dataset $\{\hat{y}_i\}$ consists of 50 quantized values randomly selected from a Gaussian distribution so that the mean is precisely zero. In each case, we use the calculated value of $\hat{\sigma}_{\text{opt}}$ or σ_{opt} as appropriate, since it was found that this gives an extremely good estimate of the true optimum value of σ, whatever the value of δ_σ. We can see from the figure that (12) gives a good order-of-magnitude estimate for the optimum value of δ_σ, across a range of values of σ_{opt} that covers 7 orders of magnitude, despite the fact that the approximate continuum expression was derived just for the case $\delta_\sigma < 1$, and the relatively small size of the dataset. The minimum value for the description length given in (11) is also seen to be a reasonable estimate.

The second model we consider consists of simply transmitting the histogram of the data as our model, with the bin widths given by the quantization scale of the data Δ. In terms of parameterised models, this is the most complex, since the model is exactly the empirical distribution of the data. The set of occupied bin positions is given by $\{b_\alpha : b_\alpha = m_\alpha\Delta, m_\alpha \in \mathbb{Z}\}$, with occupancies $\{n_\alpha \geq 1\}$, so the message length for transmitting all the parameters of the histogram is:

$$\mathcal{L}_{\text{hist:param}} \approx \sum_\alpha \left\{ \frac{3}{e} + \ln(1 + |m_\alpha|) + \ln(n_\alpha) \right\} \text{nats}. \quad (13)$$

The full description length is:

$$\mathcal{L}_{\text{hist}} = \mathcal{L}_{\text{hist:param}} + \sum_\alpha n_\alpha \ln\left(\frac{n_\alpha}{N}\right), \quad (14)$$

where the second term is $\mathcal{L}_{\text{hist:data}}$. To transmit data according to its empirical distribution, this is the size of the dataset multiplied by the Shannon entropy of the histogram: $H(\{n_\alpha\}) = -\sum_\alpha \frac{n_\alpha}{N} \ln\left(\frac{n_\alpha}{N}\right)$.

We compared encoding single images using the description lengths from both of these models for a wide variety of images, from images of ordinary objects to medical images and images generated from a set of independent Gaussian random variables. In all cases, the increased parameter length for the model caused by using the empirical distribution is more than compensated for by the exact fit to the data. As well as giving a smaller description length, the encoding according to the empirical distribution potentially offers greater discrimination, since we obtain a wider range of description lengths using this model.

4 MDL Encoding of Sets of Aligned Images

We will now consider transmitting a set of aligned 8-bit greyscale images. The sets of images are created by taking an original 190x190 pixel image and adding Gaussian white noise of varying variance. We will consider two models; sending each image separately, and sending a reference image (here, the mean of the set) plus the set of discrepancy images showing how each image differs from the reference. Images will be sent using the histogram encoding, except that we will now use the fact that the range of the image values is fixed, being $0:255$ for an ordinary greyscale image, and $-255:255$ for a discrepancy image. This means that the expression for $\mathcal{L}_{\text{hist:param}}$ from equation (13) becomes:

$$\mathcal{L}_{\text{hist:param}} = M\ln(R) + \sum_{\alpha=1}^{M}\left(\frac{1}{e} + \ln(n_\alpha)\right), \quad (15)$$

where $M = \sum_\alpha n_\alpha$ and the range R is 256 for greyscale images, and 512 for discrepancy images. As before, the $\{n_\alpha\}$ are the occupancies of the M occupied bins. This is equivalent to taking a flat distribution over the R possible positions for occupied bins.

We first investigate whether sending a reference image and discrepancy images gives an advantage over sending the original images separately. To

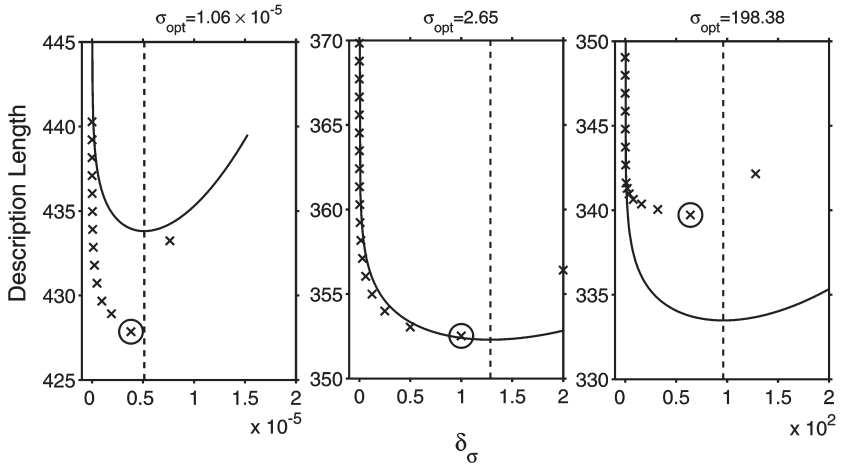

Figure 1: Graphs showing description length as a function of δ_σ for 3 datasets with different variances, with $N = 50$. **Crosses:** the exact description length (equation (8)), with the minimum circled, **Solid line:** the continuum approximation (equation (11)), with the position of the minimum shown by the dashed line.

test this, we took sets of n_s noisy images, with the noise variance fixed. The reference image was taken as the mean of each set, with the same data resolution as the original images (i.e., 8-bit). We then computed the ratio of description lengths for transmission with and without a reference image, as a function of size of the set n_s, and as a function of the noise variance. For all the values of noise variance considered, encoding using an 8-bit reference becomes advantageous (i.e., the ratio of description lengths is less than 1) provided the number of images in the set n_s is large enough. And, as we might have expected, the lower the noise variance, the lower the critical value of n_s.

In this approach, the reference image has to be considered as part of the model we are using to send the *set* of images. As such, the reference image can be considered to consist of the information/structures that are common across the set. It is therefore interesting to ask whether we should use the full 8-bits to describe the reference. The variance of the noise was fixed at 0.2, and the number n_g of quantized grey levels in the reference was varied, whilst the range of the data was maintained. For all set sizes $n_s \geq 3$, there is an advantage to using a reference, provided n_g is chosen with care. Furthermore, as the number of images in the set n_s increases, the optimum value of n_g also increases.

5 An MDL Objective Function for Learning Correspondences

The preceding sections have shown that sending a set of (aligned) images encoded using a reference image produces shorter description lengths that sending each image independently. This allows us to make the critical link between methods of image transmission and correspondence; the reference image, which in some sense contains the structures that are common to the set of *aligned* images, also allows us to define a consistent spatial correspondence across the image set, the generation of which is the aim of automatic non-rigid registration algorithms. The only additional factor that we need to add is the spatial transformation between the original image planes and the reference image plane.

We have a set of images $I_1, \ldots I_{n_s}$ and a reference image I_{ref}. There is also a set of transformations $\{t_i\}$ between the image plane of the reference image and the image plane of each image in the set. It is this set of transformations that defines the dense correspondence across the set of images. Defining a transformation t_i also defines the pullback transformation t_i^{inv}. It is not strictly required that t_i^{inv} is the *exact* inverse of t_i, providing that the transmitter and receiver both use the same algorithm to compute the set $\{t_i^{\text{inv}}\}$ from the set $\{t_i\}$. The set $\{t_i\}$ is enough to define a consistent correspondence across the set, allowing us to find, for each point in the reference, the set of corresponding points across all the images. However, without an *exact* inverse, we cannot find all the points corresponding to a point in image I_i.

Encoding the set of images then proceeds as follows. The transmitter decides on a set of transformations $\{t_i\}$, constructs the set $\{t_i^{\text{inv}}\}$, and maps each image I_i into the plane of the reference. The image values from I_i are resampled onto the regular grid X_{ref} of the reference to give the image $\widetilde{I}_i(X_{\text{ref}})$ (we assume that transmitter and receiver

have previously agreed on a resampling scheme). The set of resampled images in the frame of the reference $\{\widetilde{I}_i(X_{\rm ref})\}$ is then averaged to create the reference image $I_{\rm ref}(X_{\rm ref})$, which is transformed to the image plane of each image I_i in turn, and resampled onto the regular image grid X_i to give the image $\widetilde{I}_{\rm ref}(X_i)$ The discrepancy image between the warped, resampled reference and image I_i is computed, $I_i^{\rm disc}(X_i) = I_i(X_i) - \widetilde{I}_{\rm ref}(X_i)$, and the transmission then comprises the reference image $I_{\rm ref}(X_{\rm ref})$ and the sets of parameterised transformations $\{t_i\}$ and discrepancy images.

To decode the i^{th} image, the receiver decodes the reference image, the transformation, and the discrepancy image. She applies the transformation to the reference image, and resamples it on the regular image grid of image I_i to create the image $\widetilde{I}_{\rm ref}(X_i)$. Adding the discrepancy image allows her to reconstruct the original image $I_i(X_i)$ *exactly*. The description length for this encoding is:

$$\mathcal{L} = \mathcal{L}_{\rm params}\left(\{t_i\}\right) + \mathcal{L}\left(I_{\rm ref}(X_{\rm ref})\right) + \sum_{i=1}^{n_s} \mathcal{L}\left(I_i^{\rm disc}(X_i)\right), \tag{16}$$

where $\mathcal{L}_{\rm params}\left(\{t_i\}\right)$ is the message length for transmitting the set of quantized parameters of the transformations and the set of quantization scales. The only free parameters of the encoding are the set of transformations $\{t_i\}$, which automatically define the correspondence across the set of images. The optimum correspondence is then that given by the set of transformations that minimises this description length.

This enables us to define an objective function for non-rigid registration using this description length. We choose as our parameterised set of transformations the polyharmonic Clamped-Plate splines (CPS) [6] that have been used successfully in non-rigid registration [7]. The CPS interpolates the motion of a set of knotpoints, hence the parameters of a transformation are the initial and final positions of those knotpoints. The boundary conditions on these splines are that the transformation vanishes smoothly on the surface of a ball, which in our case (2D), we take to be the circumcircle of the images.

We establish a spatial reference frame by defining the knotpoint positions $\{x_\alpha^{\rm ref}, y_\alpha^{\rm ref}\}$, $\alpha = 1, \ldots n_k$ on the reference image. The set of transformations is then defined by specifying the knotpoint positions $\{x_\alpha^i, y_\alpha^i\}$, $i = 1, \ldots n_s$ on each image. The description length for the parameters of the set of transformations is then:

$$\mathcal{L}_{\rm params}\left(\{t_i\}\right) = \left[\frac{2}{e} + |\ln(\delta)| + \ln\left(\frac{l}{\delta}\right)\right] + \quad 2(n_s+1)n_k\left[\frac{1}{e} + \ln\left(\frac{2l+1}{\delta}\right)\right] \text{ nats}, \tag{17}$$

where l denotes the range of allowed values of the coordinates and δ the accuracy, with the centre of the image circumcircle being the origin of coordinates. The transformations are given by $t_i = \omega\left(\{x_\alpha^{\rm ref}, y_\alpha^{\rm ref}\} \rightarrow \{x_\alpha^i, y_\alpha^i\}\right)$, where $\omega(\cdot \rightarrow \cdot)$ denotes the CPS interpolant. The inverse can be approximated as $t_i^{\rm inv} = \omega(\{x_\alpha^i, y_\alpha^i\} \rightarrow \{x_\alpha^{\rm ref}, y_\alpha^{\rm ref}\})$. As an example, we take a set of $n_s = 3$ 2D axial T1 MR slices of human brains, which have already been affinely aligned. Following [7], we first generate a set of $n_k = 8$ equi-angularly spaced knotpoints around the skull for each image. We then take the average positions of these points across the set as our knotpoints positions $\{x_\alpha^{\rm ref}, y_\alpha^{\rm ref}\}$. For the purposes of illustration, the image knotpoint positions were initialised to the reference knotpoint positions (as is shown in Figure 2), so that the transformation starts at the identity. The optimisation proceeds by taking each image in turn, and optimising the final position of each knotpoint on that image. We use a fixed position accuracy of $\delta = 0.05$ pixels. As can be seen from the figure, the reference image sharpens – after 6 iterations (that is, 2 passes through each image), we see that the skulls are aligned, apart from at the front of the head. This is because the first image in the set, unlike the other two, does not have a large CSF-filled space at the front, hence the algorithm has aligned the brain surfaces rather than the skulls, which is what caused the ghosting in the reference image in this position.

6 Conclusions & Discussion

This paper has presented a novel approach to constructing objective functions for the registration of groups of images. The objective function is based on computing the minimum description length (MDL) of the set of images, encoded in three parts – (i) a reference image (here, the mean of the images in the set), (ii) a set of transformations between reference and image set that define the correspondence across the set of images, (iii) a set of discrepancy images that show where the aligned reference image differs from the unencoded image, so images are transmitted *exactly*.

The MDL approach allows us to find optimum parameter values (and, in principle, parameter accuracies) for a particular class of encoding model; it also allows us to decide between different classes of model (e.g., where two classes of images may require two separate reference images) – all of this can be done by comparing the appropriate description lengths. We have demonstrated that using a reference image reduces the description length for a set of images, even when the images are noisy,

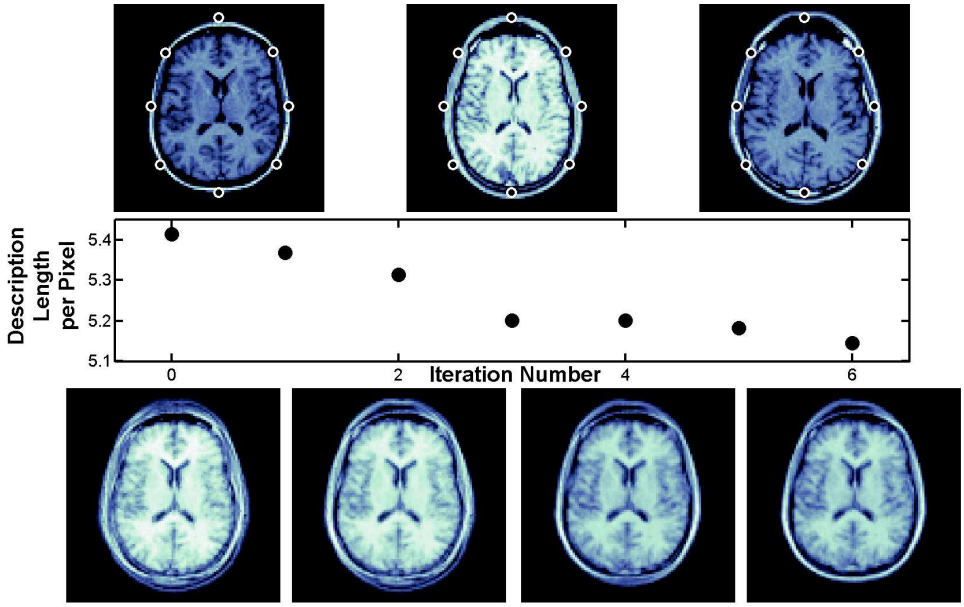

Figure 2: **Top Row:** The group of 3 images to be aligned with the reference image points superimposed, **Second Row:** The description length divided by the total number of pixels in the group of images as a function of iterations, **Bottom Row:** The mean/reference image at the 0^{th}, 2^{nd}, 4^{th}, and 6^{th} iterations.

providing that the set is sufficiently large. We have then shown that an objective function for image registration can be constructed as the description length required to transmit the set of images encoded using the mean of the set and some warp parameters, and have demonstrated this on a small set of 2D MR images of the human brain. The reference, or mean, image has been shown to get sharper as the algorithm proceeds, showing that the images are being brought into alignment. This demonstrates a proof-of-concept on a small group of images with very coarse registration. Demonstrating the method on larger groups of images, with greater number of knotpoints, especially in 3D, is currently under investigation, as is comparing the accuracy of the method to successive pairwise registrations, and performing multimodal registration.

To summarise, the MDL framework allows us to compute optimum parameters, as well as allowing us to choose between different classes of encoding model. So, for example, with regard to the registration results, the next level of comparison would be between sets of transformations with different numbers of knotpoints. Optimising the sets of transformations is the part of the algorithm most closely related to the task of registration, but it should also be possible to optimise the encoding of the discrepancy images. In this paper, each discrepancy image was transmitted separately – a groupwise approach that models the set of discrepancy images (plus further corrections terms between the model representation and the actual discrepancy images) is also possible, and should be simple to compute.

Acknowledgements

This research was supported by the MIAS IRC project, EPSRC grant GR/N14248/01.

References

[1] J. Rissanen, *Stochastic Complexity in Statistical Inquiry*. Singapore: World Scientific, 1989.

[2] R. H. Davies, *et al.*, "3D statistical shape models using direct optimisation of description length," *Lecture Notes in Computer Science*, vol. 2352, pp. 3–20, 2002.

[3] A. Hagemann, *et al.*, "Biomechanical modelling of the human head for physically based, nonrigid registration," *IEEE Transactions on Medical Imaging*, vol. 18, no. 10, pp. 875–884, 1999.

[4] J. A. Schnabel, *et al.*, "Validation of non-rigid registration using finite element methods," *Lecture Notes in Computer Science*, vol. 2082, pp. 344–357, 2001.

[5] C. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423,623–656, 1948.

[6] S. Marsland and C. Twining, "Constructing diffeomorphic representations for the groupwise analysis of non-rigid registrations of medical images," *IEEE Transactions on Medical Imaging*, vol. 23, no. 8, pp. 1006 – 1020, 2004.

[7] S. Marsland and C. J. Twining, "Constructing data-driven optimal representations for iterative pairwise non-rigid registration," *Lecture Notes in Computer Science*, vol. 2717, pp. 50–60, 2003.