A Unified Information-Theoretic Approach to the Correspondence Problem in Image Registration

Carole J. Twining[1], Stephen Marsland[1], and Chris J. Taylor
Imaging Science and Biomedical Engineering
University of Manchester
Oxford Road, Manchester, UK.

Abstract

We consider the correspondence problem associated with the non-rigid registration of a group of images; in particular, the theoretical basis for the derivation of the objective function that defines the 'best' correspondence across a set of images. For intra-subject registration, there is an actual physical deformation process underlying the observed deformation, but for inter-subject registration, there is no such physical process, and hence no hypothetical process that generates the observed data. This leads to the conclusion that our construction should be based on the data alone.

Such a construction is possible using criteria derived from information theory. We show how many commonly used pairwise voxel-based similarity measures can be generated using these criteria, and discuss how this approach can be extended to give a unified theoretical basis for the generation of novel objective functions in the groupwise case, where both image discrepancy and image deformation terms should be included in a principled way.

1 Introduction

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 individual terms are incommensurate quantities, so that we have to determine appropriate values for the coefficients of each term.

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 registration algorithm is one that closely models the underlying physical process, leading to physically-based registration algorithms (e.g., [5, 6]), or physically-based models (e.g., [9]) that can be used to evaluate the results of non-rigid registration algorithms.

However, in inter-subject registration there is no longer a direct underlying physical process that generates the observed data. We therefore contend that 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). Further, any statistical inferences we make about the data should not depend on hypothetical data-generating processes; an assumption that underlies parameter estimation techniques such as maximum likelihood. The Minimum Description Length (MDL) [8] and Minimum Message Length (MML) [13] principles are closely related approaches [1] to model-selection and statistical inference that satisfy these restrictions, and the MDL principle has previously been shown to give excellent results when applied to the correspondence problem in shape modelling [4].

2 Minimum Description Length

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

[1] Joint first authors.
Contact: carole.twining@man.ac.uk, s.r.marsland@massey.ac.nz

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 [10] – 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:

$$\mathcal{L}_i = -\log p_i. \tag{1}$$

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; for details of how to calculate the parameter length, see [8], explicit examples of calculating parameter lengths for the case of Gaussian models are given in [4]. 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.

For the cases of both pairwise and groupwise image registration with a single reference image, the data to be transmitted consists of:

- the N pixel/voxel values of the reference image
- the warp applied to the reference image to register it to each target image in turn
- the pixel-value data for each target image

where the reference image $X = \{x_i : i = 1, \ldots N\}$ only has to be transmitted once and the pixel-value data for each target image may be encoded using the values of the resampled, warped reference image as part of the encoding model.

Given this information, the receiver can then reconstruct each pixelated, quantised target image *exactly*. If, for example, we encode this according to the empirical distribution $p_X(x)$, we obtain a data length of:

$$\begin{aligned} \mathcal{L}_{\text{ref}} &= -\sum_{i=1}^{N} \log p_X(x_i) \\ &= -\sum_x N p_X(x) \log p_X(x) = N H(X), \end{aligned} \tag{3}$$

where $H(X)$ is the Shannon entropy of the reference image.

3 The Pairwise Case

3.1 Transmitting the pixel-value data

We now consider the data lengths $\mathcal{L}_{disc}$ needed to transmit the pixel-value data for the target image based on different modelling choices for the encoding. The warped reference image can be included as part of our encoding model for these values. The target image is denoted by $Y = \{y_i\}$, the warped reference image by $\tilde{X} = \{\tilde{x}_i\}$, and the discrepancy image by $Z = \{z_i\} = \{y_i - \tilde{x}_i\}$. We will restrict ourselves to the case of quantised grayscale images, with $y_i, \tilde{x}_i, z_i \in \mathbb{Z}$.

If we encode Z using a Gaussian model of zero mean and fixed width σ, then to leading order:

$$p(z) = \frac{1}{A(\sigma)} \int_{z-\frac{1}{2}}^{z+\frac{1}{2}} \mathrm{d}t \exp\left(-\frac{t^2}{2\sigma^2}\right) \approx \frac{1}{A(\sigma)} \exp\left(-\frac{z^2}{2\sigma^2}\right), \tag{4}$$

where $A(\sigma)$ is a normalisation factor. From equation (1), this gives a data length:

$$\mathcal{L}_{\text{disc}} = N \log A(\sigma) + \frac{1}{2\sigma^2} \sum_i z_i^2, \tag{5}$$

which is (up to a constant and factors) just the sum-of-squares-difference voxel-based similarity measure.

Similarly, if we instead encode using an exponential distribution $p(z) = \frac{1}{B(\lambda)} \exp(-\lambda|z|)$ then the corresponding data length is:

$$\mathcal{L}_{\text{disc}} = N \log B(\lambda) + \lambda \sum_i |z_i|, \tag{6}$$

which gives the absolute difference similarity measure.

Suppose that we instead transmit the target pixel values directly, rather than the discrepancies. We first separate the pixels in the target image Y into subsets, according to the values of the corresponding pixels in the resampled, warped reference image $\tilde{X}$ and, for each subset, encode using the empirical distribution of values in that subset. If r and t denote pixel values in the warped reference image and target image respectively, then we find the data length for transmitting the target image encoded using the warped reference image is:

$$\mathcal{L}_{Y|\tilde{X}} = -N \sum_{r,t} p_{\tilde{X}}(r) p(t|r) \log p(t|r) = N H(Y|\tilde{X}), \tag{7}$$

where $p(t|r)$ are conditional probabilities, and $H(Y|\tilde{X})$ is the conditional entropy. The mutual information [12] of the target and the warped reference image is given by:

$$I(Y, \tilde{X}) = H(Y) + H(\tilde{X}) - H(Y, \tilde{X}) \equiv H(Y) - H(Y|\tilde{X}). \tag{8}$$

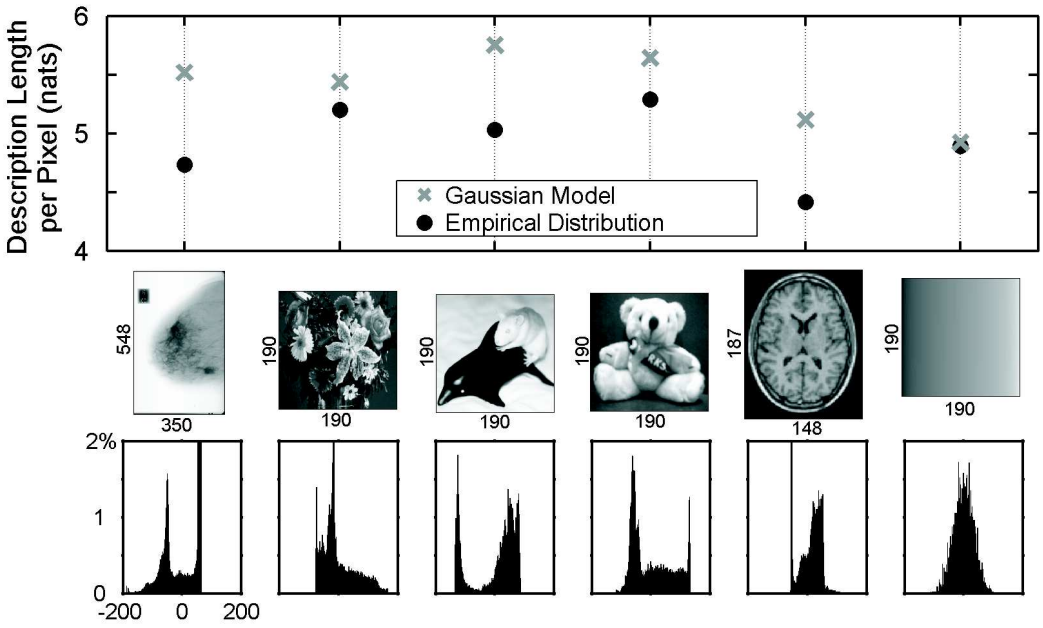

Figure 1. Top row: The Description Lengths per pixel for a set of images, encoded using the 2 different models, Optimised Gaussian: grey crosses, Empirical distribution: black circles. Middle row: Thumbnails of the images with image dimensions in pixels, Bottom row: The centred image histograms, all to the same scale.

If we consider the entropy $H(Y)$ of the target is fixed, then minimising the data length $\mathcal{L}_{Y|\tilde{X}}$ corresponds to maximising the mutual information $I(Y, \tilde{X})$.

The normalised mutual information [11], given by:

$$I_{norm}(Y, \tilde{X}) = \frac{H(Y) + H(\tilde{X})}{H(Y, \tilde{X})} = \frac{\mathcal{L}_{\tilde{X}} + \mathcal{L}_Y}{\mathcal{L}_{\tilde{X}} + \mathcal{L}_{Y|\tilde{X}}}, \quad (9)$$

is just the ratio of the data lengths for transmitting the warped reference and target independently, as opposed to encoding the target using the warped reference.

We see that these commonly-used pairwise voxel-based similarity measures can be related to the data lengths for transmitting the pixel value information of the reference and target. A different choice of similarity measure is then equivalent to a different choice of model used to encode the data. In principle, moving to the MDL framework (i.e., including the parameter lengths for the model) gives us a way of choosing between these different similarity measures; we just compute and compare the description lengths.

Figure 1 shows the description lengths for a version of the Gaussian encoding (equation (5)), where the quantisation parameter δ and the width σ have been optimised for the particular data transmitted. The images used are a set of 8-bit ($= \frac{8}{e} \approx 2.943$ nats) greyscale images, with the data being centred before transmission. The set of images consists of 3 images of ordinary objects, 2 medical images (a mammogram and a slice from a 3D MR image of a normal human brain), and an artificial image constructed from a set of independent Gaussian random variables. We can see that the description length per pixel is of the correct order compared to the greyscale resolution of the original images if we remember that the transmitted data has been centred – this then requires an upper bound of 2×8 bits ≈ 5.886 nats to transmit any such centred image without encoding.

The description lengths of the Gaussian model are compared with those found by using a considerably more complex parameterised model, the empirical distribution described by the histogram of the data. The bin widths for the histogram are given by the quantisation scale of the data, Δ. The set of occupied bin positions is given by $\{b_\alpha : b_\alpha = m_\alpha \Delta, \, m_\alpha \in \mathbb{Z}\}$, with occupancies $\{n_\alpha \geq 1\}$. Hence, the message length for transmitting all the parameters of the histogram is:

$$\begin{aligned} \mathcal{L}_{\text{hist:param}} &= \sum_\alpha \left\{ \frac{1}{e} + \mathcal{L}_{\text{int}}(1 + |m_\alpha|) + \mathcal{L}_{\text{int}}(n_\alpha) \right\} \text{ nats} \\ &\approx \sum_\alpha \left\{ \frac{3}{e} + \ln(1 + |m_\alpha|) + \ln(n_\alpha) \right\} \text{ nats}, \quad (10) \end{aligned}$$

giving a final description length of:

$$\begin{aligned} \mathcal{L}_{\text{hist}} &= \mathcal{L}_{\text{hist:param}} + \mathcal{L}_{\text{hist:data}} \\ &= \mathcal{L}_{\text{hist:param}} - \sum_\alpha n_\alpha \ln\left(\frac{n_\alpha}{N}\right). \quad (11) \end{aligned}$$

Looking again at Figure 1, it can be seem that in all cases, even that of the Gaussian image, the increased parameter length for the model using the empirical distribution is more than compensated for by the exact fit to the data. This discrepancy was not due to any errors in approximating the optimum parameters for the Gaussian, which are very small.

As well as giving a smaller description length, the encoding according to the empirical distribution potentially offers greater discrimination, given that the range of description lengths using this model (0.8743) is greater than the range obtained using the Gaussian model (0.8296). We therefore conclude that the in the MDL framework, the appropriate description length for transmitting a single image is that given by the empirical distribution of the data.

3.2 Transmitting the Deformation Fields

We now consider the message length for transmitting the deformation field information. To obtain a finite message length, we have to quantise our data. However, once this is done, the message lengths for pixel data and deformation field data are commensurate quantities. We do still have an arbitrary parameter (the quantisation scale for the deformation fields), but now this parameter has a clear physical meaning. We reiterate that the receiver can always reconstruct the target image exactly, whatever the quantisation scale.

Consider the pixelwise deformation field $\{\vec{d}_i\}$ with the affine part removed. A naïve quantisation would be to quantise the Cartesian components of each $\vec{d}_i$ using some quantisation parameter. Then, the quantised components could be considered as a set of quantised, pixelated images. We could encode using the empirical distribution (in an analogous manner to equation (3)) to give a deformation field data length of:

$$\mathcal{L}_{\text{deform}} = \sum_{\mu} NH(\{d_i^\mu\}). \tag{12}$$

However, this treats the components of the deformation at a pixel, and the deformation at different pixels, as if they are independent; in practice, representations of the deformation field in non-rigid registration algorithms impose some degree of smoothness on the deformation field, meaning that such an encoding choice will actually grossly over-estimate the stochastic complexity. This is illustrated by the case considered in Figure 2. Here, we generated a set of 150 random diffeomorphic warps of a 190×190 image grid with a fixed number of parameters, so that we would expect the stochastic complexity of these warps to be approximately constant. This will not be the case for the stochastic complexity computed using the Shannon entropy, since, as is shown in the figure, the entropy tends to increase with the mean displacement of the warp.

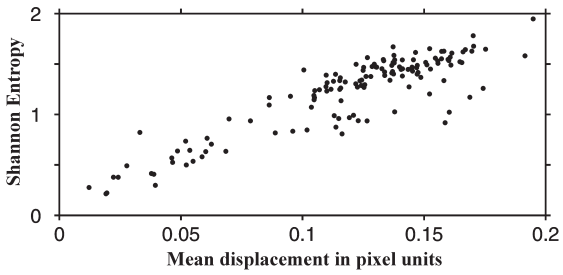

Figure 2. The Shannon entropy of a deformation field versus the mean pixel displacement.

This suggests that for the deformation field in the pairwise case, we should encode using a low-dimensional representation of the deformation field, for example describing the parameters of the warps applied, if they are available.

4 The Groupwise Case

We will now consider the case of groupwise registration to a single reference image. We will begin by discussing some of the points raised by attempting to construct a groupwise objective function within this MDL framework. This will be followed by a simple example showing groupwise *affine* registration using an MDL objective function.

As was noted previously, the reference image is transmitted to the receiver just once. For each target image we transmit the warp of the reference image and the pixel values of the target, encoding the latter using the resampled, warped reference image. As regards the pixel value data to be transmitted, for each target this is the size of the target image, and is in the frame of the target image, *not* the frame of the unwarped reference image. We could encode this in a similar manner to the pairwise case (see equations (5, 6, 7)). However, the sum of pairwise mutual information (8) is no longer strictly suitable in this MDL framework, since it includes the cost of transmitting the reference image multiple times, nor is the normalised mutual information (9), since it is a ratio of data lengths rather than a data length itself. If the target images are all of the same size, and are all affinely aligned (for example, by using the groupwise affine MDL example given in section 4.1), we could instead consider the set of pixel discrepancy images between each target and its respective warped, resampled reference image. We could then transmit the set of pixel discrepancies at a given pixel using, for example, a Gaussian model or a histogram.

Let us consider the simple case where we have chosen to encode both the pixel value data and the deformation field data using Gaussian models. It is shown in [4] that the de-

scription length for data encoded using a simple Gaussian model can, in the limit of a large number of examples, be approximated by a form involving the determinant of the data covariance matrix (as was used by Kotcheff and Taylor [7]). This simplified expression involves the sum of the logs of the eigenvalues of the covariance matrix, and hence can be related to the total variance of the data.

In this approximate form of the MDL objective function, we obtain something analogous to the form of objective function for groupwise registration as used by Cootes et al. [3], where the objective function used a sum of log probabilities, which for a Gaussian model reduces to the sum of variances of the data terms. However, they were required to specify the arbitrary parameter that determined the relative weights of the shape (that is, deformation) and texture (that is, pixel value data) parts.

A similar situation occurs in the formulation of Active Appearance Models (AAM) [2], where the arbitrary parameter is the relative weighting between the shape and texture parts of the model. However, the AAM is explicitly constructed from an annotated training set; the weighting parameter can then be chosen so that the shape and texture parts of the training set have equal total variance. In our case, this would be equivalent to knowing the required dense correspondence for some set of training images.

The role of the AAM weighting parameter is taken up in our method by the relative weighting between the data quantisation parameter for the deformation field (or for the deformation field parameters in some representation), and the pixel quantisation. So, although we have gained nothing in terms of the number of parameters to be determined, our quantisation parameters have a clear physical meaning.

4.1 The Rigid Case

To demonstrate the feasibility of the MDL objective function, we here consider the simplest case of a set of images produced by simply translating and resampling a single image. The extension to groupwise non-rigid registration will be demonstrated in another paper. For the affine case the transformations $\{t_i\}$ are just translations with parameters $\{x_i, y_i\}$. If we suppose that these are transmitted to an accuracy δ, and with some maximum modulus l, then the message length for the transformations is given by:

$$\begin{aligned}\mathcal{L}_{\text{params}}\left(\{t_i\}\right) &= \left(\frac{1}{e} + |\ln(\delta)|\right) + \left(\frac{1}{e} + \ln\left(\frac{l}{\delta}\right)\right) \\ &+ \sum_{i=1}^{n_s} 2\left[\frac{1}{e} + \ln\left(\frac{2l+1}{\delta}\right)\right] \text{ nats.} \quad (13)\end{aligned}$$

The images are then individually transmitting using the histogram encoding described earlier (see equation (11)), except that the range of the data is now known, and so equation (10) is replaced by:

$$\mathcal{L}_{\text{hist:param}} = M\ln\left(R\right) + \sum_{\alpha=1}^{M}\left(\frac{1}{e} + \ln(n_\alpha)\right), \quad (14)$$

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. The only free parameters of the encoding are the set of transformations $\{t_i\}$; a set of such transformations automatically defines the correspondence across the set of images. The optimum correspondence is then that given by the set of transformations that minimises the description length.

From an implementation point of view, it is important to note that we can optimise each transformation t_i individually; varying t_i alters the contribution of the i^{th} image to the mean, which alters the reference image, which hence alters the discrepancy images for all images in the set. So, although we can sequentially optimise the transformations, the effect is actually a fully groupwise one. The results of such an optimisation for a set of $n_s = 5$ images is shown in Figure 3. Each iteration of the algorithm corresponds to optimising just one of the transformations $\{t_i\}$. As we might have expected, the optimisation produces a good result after n_s iterations. It is clear that the final reference image is exactly the generalisation of the image set we would have expected, and that the algorithm converges to it despite the extremely poor quality of the initial reference image.

5 Conclusion

We have shown that many of the commonly-used voxel-based similarity measures can be understood in terms of an MDL framework. This framework also provides a natural way of combining formerly incommensurate shape and texture terms in the same objective function in a principled way; although the MDL formulation will still involve the determination of arbitrary parameters, it has the distinct advantage that these parameters now have a clear physical meaning.

Most importantly, it allows us to compare different classes of encoding model (that is, different objective functions), as is demonstrated in this paper. This paper also demonstrates that an objective function based on the MDL principle allows successful affine registration of 2D brain images. The further theoretical work necessary to extend this method to non-rigid registration is the subject of our current work.

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

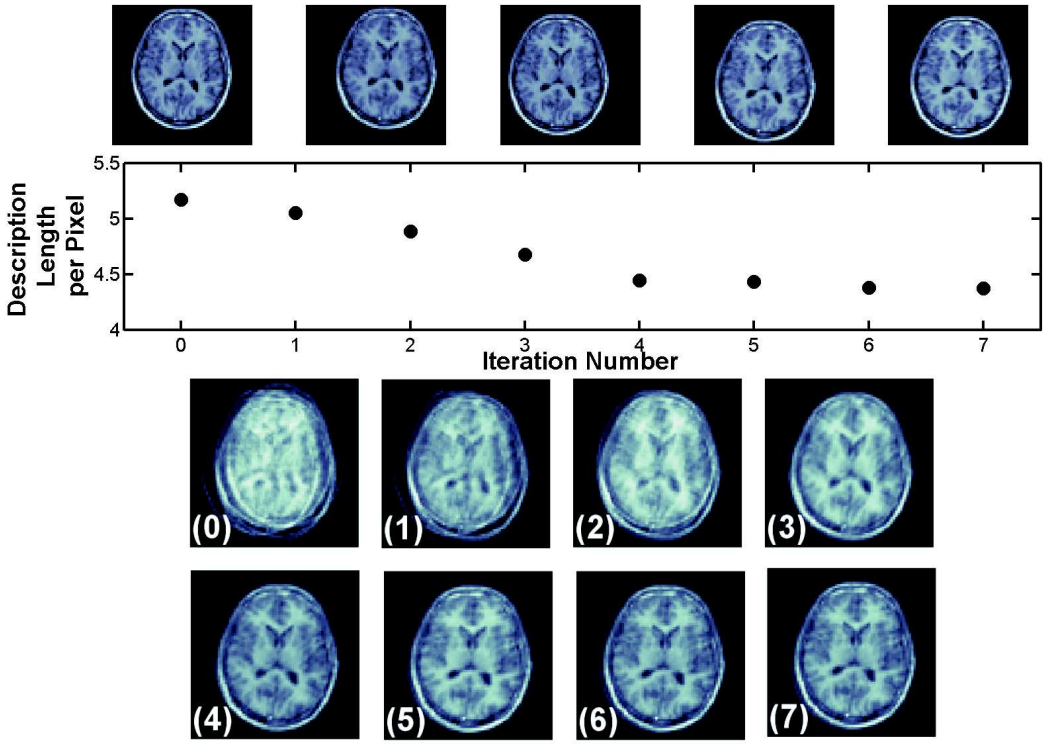

Figure 3. Top Row: The group of 5 images to be aligned (translated versions of the same image). Second Row: The description length divided by the total number of pixels in the group of images as a function of iteration number, Bottom Two Rows: The mean/reference image at each iteration.

References

[1] R. Baxter and J. Oliver. MDL and MML: Similarities and differences. Technical report TR 207, Department of Computer Science, Monash University, Australia, 1994.

[2] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *Lecture Notes in Computer Science*, 1407:484–498, 1998.

[3] T. F. Cootes, S. Marsland, C. J. Twining, K. Smith, and C. J. Taylor. Groupwise diffeomorphic non-rigid registration for automatic model building. In *Proceedings of ECCV*, 2004.

[4] R. H. Davies, C. J. Twining, T. F. Cootes, J. C. Waterton, and C. J. Taylor. 3D statistical shape models using direct optimisation of description length. *Lecture Notes in Computer Science*, 2352:3–20, 2002.

[5] M. Ferrant, S. K. Warfield, C. R. G. Guttmann, R. V. Mulkern, F. A. Jolesz, and R. Kikinis. 3D image matching using a finite element based elastic deformation model. *Lecture Notes in Computer Science*, 1679:202–209, 1999.

[6] A. Hagemann, K. Rohr, H. S. Stiehl, U. Spetzger, and J. M. Gilsbach. Biomechanical modelling of the human head for physically based, nonrigid registration. *IEEE Transactions on Medical Imaging*, 18(10):875–884, 1999.

[7] A. C. W. Kotcheff and C. J. Taylor. Automatic construction of eigenshape models by direct optimization. *Medical Image Analysis*, 2:303–314, 1998.

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

[9] J. A. Schnabel, C. Tanner, A. C. Smith, M. O. Leach, C. Hayes, A. Degenhard, R. Hose, D. L. G. Hill, and D. J. Hawkes. Validation of non-rigid registration using finite element methods. *Lecture Notes in Computer Science*, 2082:344–357, 2001.

[10] C. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423,623–656, 1948.

[11] C. Studholme, D. Hawkes, and D. Hill. An overlap invariant entropy measure of 3D medical image alignment. *Pattern Recognition*, 32(1):71–86, 1999.

[12] P. Viola and W. M. Wells III. Alignement by maximization of mutual information. *International Journal of Computer Vision*, 24(2):137–154, 1997.

[13] C. S. Wallace and P. R. Freeman. Estimation and inference by compact coding. *Journal of the Royal Statistical Society B*, 54(3):240–265, 1987.