Groupwise Construction of Appearance Models using Piece-wise Affine Deformations

T.F. Cootes, C.J. Twining, V.Petrović, R.Schestowitz and C.J. Taylor
Imaging Science and Biomedical Engineering,
The University of Manchester, M13 9PT, UK
t.cootes@manchester.ac.uk

Abstract

We describe an algorithm for obtaining correspondences across a group of images of deformable objects. The approach is to construct a statistical model of appearance which can encode the training images as compactly as possible (a Minimum Description Length framework). Correspondences are defined by piece-wise linear interpolation between a set of control points defined on each image. Given such points a model can be constructed, which can approximate every image in the set. The description length encodes the cost of the model, the parameters and most importantly, the residuals not explained by the model. By modifying the positions of the control points we can optimise the description length, leading to good correspondence. We describe the algorithm in detail and give examples of its application to MR brain images and to faces. We also describe experiments which use a recently-introduced *specificity* measure to evaluate the performance of different components of the algorithm.

1 Introduction

This paper explores a new technique for automatically finding correspondences across a set of images, suitable for constructing statistical models of shape and appearance.

Such methods are important for constructing models from large datasets where manual annotation is too time-consuming to be practical.

We assume that we are supplied with a training set containing numerous images of examples of the structure of interest (eg MR images of the brain), and wish to automatically construct a statistical model capable of generating similar images. If we can compute a set of correspondences across the training set, we can easily build such a model.

A widely used approach is to select one example as a reference template, and then find pair-wise correspondences between this reference and each of the other images in turn (eg [11, 4]). However, this approach does not take account of the importance of the group of images when defining correct correspondence - it is often only by looking at many examples that one can determine which structures should be corresponded and which may be too unreliable.

Inspired by the work on corresponding shapes by minimising a description length (MDL) [5], we seek an MDL approach to corresponding points across whole images. The approach is to construct a statistical appearance model (a model of shape and texture) which can be used to synthesize approximations to the training set images. We can then evaluate the cost of encoding the training images using the model. This involves calculating the cost of the model, the parameters required to approximate each image and the residual differences between the images synthesized by the model and each training image.

We represent dense correspondence fields between images using a piece-wise affine interpolation between a set of control points placed on each image. Given such points on every training image we can build an appearance model and evaluate the description length for the model and training set. We can thus evaluate the effect of modifying the control point positions, leading to an optimisation to find those that give the minimum description length.

In the following we describe the algorithm in detail and describe its application to automatically constructing models of MR images and face images. Any such algorithm is likely to be relatively complex and contain a number of different components. We use a recently-proposed measure of model quality known as *specificity* to evaluate the effects of different components on the results.

2 Background

Finding mappings between structures across a set of images can facilitate many image analysis tasks. One particular area of importance is in medical image interpretation, where image registration can help in tasks as diverse as anatomical atlas matching and labelling, image classification, and data fusion. Many researchers have investigated image registration methods and the use of deformable models, for overviews see for example [12, 9].

Some early work on groupwise image registration based on the discrepancy between the set of images and the reference image has been performed [10], and a groupwise model-matching algorithm that represents image intensities as well as shape has also been proposed [7]. In earlier work we have investigated groupwise registration using compositions of simple warps [3].

The closest work to that proposed here is that by Baker *et al.* [1] who consider building an appearance model as an image coding problem. The model parameters are iteratively re-estimated after fitting the current model to the images, leading to an implicit correspondence defined across the data set. Our work differs from theirs in several respects, including the use of more general assumptions about the form of distributions (non-gaussian) and the measuring of the image residuals in the target frame rather than the reference frame (as required by the MDL formulation).

3 Methods

We seek to create a model of shape and texture which can synthesize the images in the training set as compactly as possible. We treat model building as an optimisation problem in which the objective function is the total cost of synthesizing exact copies of the training set. The basic model will only be able to synthesize approximations to the training images - it is important to evaluate the cost of the residual differences required to generate exact copies (see Figure 1).

The full objective function has the form

$$F_{total} = F_{model} + F_{params} + F_{residuals} \tag{1}$$

where F_{model} is the description length of the model itself, F_{params} is that of the parameters required to synthesize each example and $F_{residuals}$ is the cost of sending the residuals between synthetic and original training images (for more details, see [2]).

This Minimum Description Length (MDL) approach will allow us to perform model selection (for instance, deciding on the best trade-off between warp model complexity and texture model complexity), though that aspect is not explored in this paper.

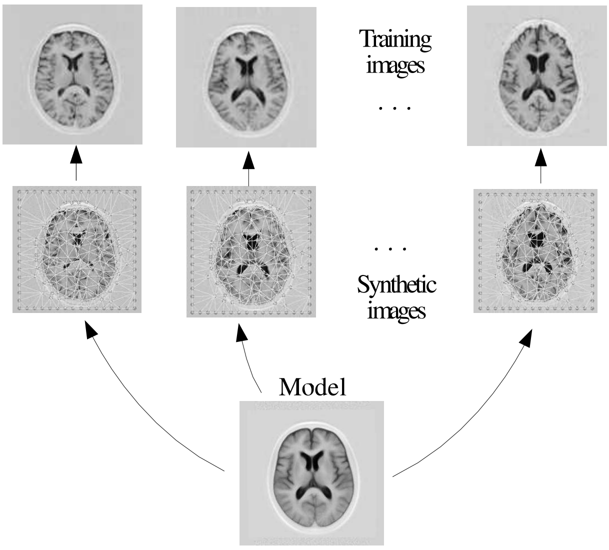


Figure 1: Registration framework. The model generates synthetic images as close as possible to the original training images.

3.1 Synthesizing Images

The model must have a representation of a deformation field, and a method of generating texture in the model frame. We will assume that the deformation field is uniquely defined by the position of a set of control points $\mathbf{x}$, so the deformation is a warping function $\mathbf{y}' = W(\mathbf{y};\mathbf{x})$. The texture in the reference frame is given by some image function $I_r(\mathbf{y};\mathbf{t})$ with texture parameters $\mathbf{t}$.

A new image can be synthesized by the model as

$$I_s(\mathbf{y}) = I_r(W^{-1}(\mathbf{y};\mathbf{x});\mathbf{t}) \tag{2}$$

(for each pixel position in the new image, $\mathbf{y}$, interpolate the reference image at the corresponding point $W^{-1}(\mathbf{y};\mathbf{x})$) to obtain the intensity).

We assume a linear model for the control point positions of the form

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{Pb} + \delta\mathbf{x} \tag{3}$$

where $\bar{\mathbf{x}}$ gives the mean position, $\mathbf{P}$ defines a set of modes with shape parameters $\mathbf{b}$, and $\delta\mathbf{x}$ is a set of residual displacements.

Suppose we wish to use the model to encode a particular training image I. There are four terms to consider:

1. The model texture parameters $\mathbf{t}$
2. The model shape parameters $\mathbf{b}$
3. The residual control point displacements $\delta\mathbf{x}$
4. The set of residual image values $\{r_{\mathbf{y}} = I(\mathbf{y}) - I_s(\mathbf{y})\}$ (one per pixel)

In an optimal coding scheme the cost is related to the log probability of each term,

$$E = \lambda_1 \log p(\mathbf{t}) + \lambda_2 \log p(\mathbf{b}) + \lambda_3 \sum_i \log p(\delta\mathbf{x}_i) + \lambda_4 \sum_{\mathbf{y}} \log p(r_{\mathbf{y}}) \tag{4}$$

where the weights λ_i are chosen to allow for the different units and dimensions of the different terms. Given a training set of images, and the current estimate of control point positions for each, we can construct a model (and appropriate parameter distributions) and use the above equation to estimate the coding cost of the training set using the model. [1]

If we move the control points on one image, we can re-evaluate the total cost, and thus drive an optimisation algorithm.

3.2 Groupwise Registration algorithm

Work on corresponding shapes [5] suggests that it is sufficient to modify the points on one image at a time during the optimisation process. Also, to avoid problems with local minima, when optimising on one image we work with a model built from all images except the current one, and assume the model is fixed during the local optimisation stage.

It is useful to initialise the algorithm by choosing one image as an initial reference frame, then performing pairwise registration onto every other image.

An outline of the groupwise registration algorithm is as follows:

1. Select one image to be used as an initial reference
2. Select control points on this reference image
3. Initial pairwise registration to reference:
    (a) For each image in turn estimate movement of control points to optimise match to reference image
4. Groupwise registration:
    (a) For each image in turn
        i. Compute a shape model from the position of points in other images
        ii. Compute the texture model by warping all other images to the mean shape
        iii. Compute the best-fit of the shape model to the current target image points
        iv. Modify the positions of the points in the target image to minimise the cost of encoding that image using the current model
    (b) Repeat until happy

[1] In the full MDL approach one also includes the model cost including the cost of representing the parameter distributions. For the purposes of this paper we will assume these are fixed.

3.3 Piece-wise Affine Deformation Fields

An important issue for a registration algorithm is the choice of representation for the deformation field. There are three commonly used classes. A dense field representing the movement of every pixel (eg fluid deformation models [6]), a composition of simple warps (eg [8, 3] or fields controlled by a sparse set of control points (eg [11]). In the latter case the control points usually control a set of splines (eg B-splines, thin plate splines or clamped plate splines).

The algorithm described above requires us to be able to invert the deformation efficiently, since during model construction we use $W(\mathbf{y};\mathbf{x})$ to interpolate the training images, and when generating new images we use $W^{-1}(\mathbf{y};\mathbf{x})$ to interpolate the reference frame image. Although we have experimented with smoother interpolating splines, they cannot easily be inverted. Instead we have adopted piece-wise linear interpolation, in which the region of interest is mosaiced by a set of triangles (in 2D) or quadrahedra (in 3D), the control points being the corners (the Delaunay algorithm can be used to obtain this mesh). Within each region we can use an affine approximation of the deformation field, which is easily inverted. Although the resulting piece-wise affine representation is not smooth (the derivatives are not defined at the boundaries), we have found it to produce good results. It is also simple to add constraints to prevent non-invertable mappings.

3.4 Optimising Control Point Positions

The algorithm outlined above involves performing an optimisation of control points, $\mathbf{x}$, on one image at a time. Since for a given set of points $\mathbf{x}$ we can estimate the parameters and residuals, the objective function depends only on $\mathbf{x}$ (call it $E(\mathbf{x})$).

We adopt a coarse to fine regime. Let $\mathbf{x}' = G_n(\mathbf{x};\mathbf{z})$ be a smooth deformation function controlled by the positions of a grid of nodes $\mathbf{z}$ with n points per side, where the total number of nodes is much less than the number of control points $\mathbf{x}$ [4]. We use G to manipulate the control points $\mathbf{x}$, moving sets of them smoothly during the coarse part of the search. Thus rather than modifying the elements of $\mathbf{x}$ independently, we find the (smaller number of) parameters $\mathbf{z}$ which minimize $Q(\mathbf{z}) = E(G_n(\mathbf{x};\mathbf{z}))$. We repeat this with increasing numbers of grid nodes. The finest stage of the search involves directly optimising the values of $\mathbf{x}$.

At each stage we use a simple downhill gradient search, in which we first numerically estimate the gradient, then perform a line search along that direction.

4 Measuring Model Quality

Ideally we would like to evaluate the models by assessing how accurately they compute correspondences across the set. However, this requires correspondences to be known. Though they could be found at a sparse set of points by human annotation, this is both time-consuming and prone to a lack of repeatability.

Alternatively one can generate a synthetic dataset by applying known warps and noise to a single image. Though a useful approach, it can be difficult to apply realistic deformations (because they are often not known) or image degradation.

In the following we use a measure of model *specificity*. This evaluates the similarity between images synthesized by the model (by drawing from a distribution for the parameters) and the original training images. It is an extension of the specificity measures used for evaluating statistical shape models [5], and has recently been shown to be a good surrogate for evaluating correspondence errors.

It is evaluated as follows. Assume that we have a set of N_t training images $\{I_i\}$. By drawing from the appearance model distribution we can generate a set of N_s synthetic images $\{S_j\}$. The specificity is then estimated as

$$E_s = \frac{1}{N_s} \sum_{j=1}^{N_s} \min_i D_s(I_i, S_j) \tag{5}$$

where $D_s(I,S)$ is the mean absolute shuffle distance between images I and S,

$$D_s(I,S) = \frac{1}{n_x n_y} \sum_{\mathbf{x}} \min_{\mathbf{y} \in N(\mathbf{x})} |I(\mathbf{x}) - S(\mathbf{y})| \tag{6}$$

where $N(\mathbf{x})$ is a small neighbourhood around $\mathbf{x}$. In practice we apply our current estimate of the affine transformation between S and I before computing the shuffle distance.

In the experiments below we use $N_s = 1000$ synthetic images. This leads to a standard error on the mean E_s which is small relative to the differences between models.

5 Results of Experiments

5.1 Brain Data Set

We have a dataset of 104 3D MR images of normal brains [2] which have been affine aligned and a single slice at equivalent location extracted from each. Figure 2 shows examples of the extracted slices.

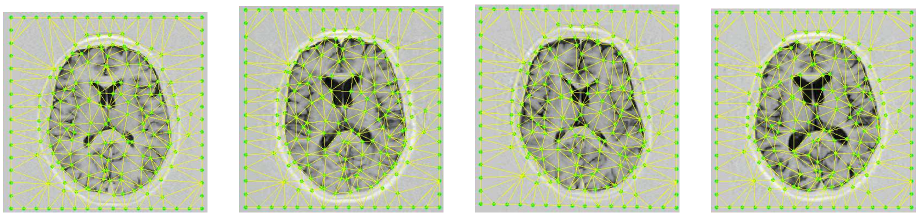
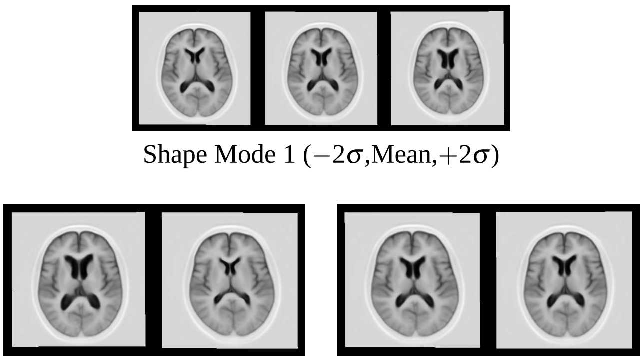


Figure 2: Examples from training set with resulting control points superimposed

Figure 3 shows the first three modes of shape variation of a groupwise model built from all 104 groupwise registered images. The crispness of the mean demonstrates that good correspondence is being achieved.

We have performed a number of experiments which use the specificity measure to compare different algorithms and parameters.

[2] The age matched normals in a dementia study generously provided by P.Bromiley and N.Thacker



Shape Mode 1 (-2σ,Mean,$+2\sigma$)

Shape Mode 2

Shape Mode 3

Figure 3: The mean texture image deformed using the first three models of shape variation (± 2 s.d.)

5.2 Choice of PDF for Residuals

The choice of PDF we use for the distribution of residuals can have an effect on the performance of the registration. If we assume a Gaussian distribution, $p(x) \propto exp(-x^2/(2\sigma^2))$, then the encoding cost is a sum of squares term. If however we assume an exponential distribution, $p(x) \propto exp(-|x|/\sigma))$, the cost is a sum of absolute values, which is inherently more robust to outliers. We have performed groupwise registration using both assumptions. For each model we estimate specificity using Equation 5 using a range of different numbers of modes for the appearance model. The result is plotted in Figure 4. It demonstrates that using the sum of absolutes (exponential distribution) leads to better (lower) values of specificity at a range of modes, suggesting that such a model gives better correspondence. In subsequent experiments we use this exponential distribution.

5.3 Affect of choice of control points on performance

The overall model quality will be affected by the choice of control points used. For instance, we have experimented with three approaches,

1. Placing points on a 16 x 16 grid on the reference frame,
2. Placing points on a grid, but removing those in low variance (flat) regions,
3. As (2) but then moving points to nearby strong edges.

Figure 5 shows plots of specificity vs model modes for each of the three cases. It demonstrates that using control points on strong edges, and ignoring flat regions, seems to give the best performance. This is because it allows more control over the boundary regions in the image.

5.4 Groupwise vs Pairwise

We compare the performance of three different algorithms:

1. simple pairwise registration of the chosen reference image to each training image
2. groupwise registration to a common mean image (the mean texture at the mean shape) with no constraint on shape (assume $p(\mathbf{b})$ and $p(\delta\mathbf{x})$ are flat)

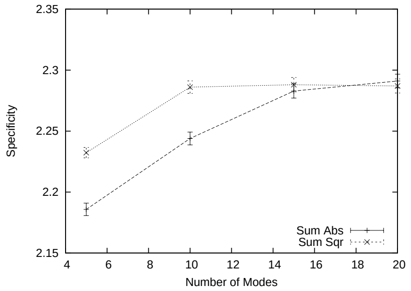


Figure 4: Comparison between different residual objective functions

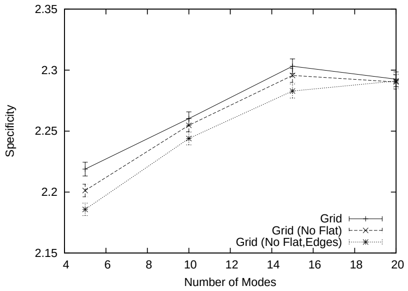


Figure 5: Affect of choice of control points on performance (see text).

3. groupwise registration to a common mean image, assuming $p(\mathbf{b})$ is flat and $p(\delta\mathbf{x})$ is exponential

We make an arbitrary choice for the reference image (the first in the set) and use the exponential distribution for encoding the residuals. The use of the common mean as a reference is the simplest form of model we could use. It is anticipated that using more sophisticated models, such as including the principal modes of variation of shape and texture, would lead to more compact models.

The resulting graph of specificity vs model modes is given in Figure 6. It demonstrates that the groupwise approaches significantly outperform pairwise approaches to an arbitrary reference image. Note that the first groupwise method is equivalent to repeated pairwise registration to the current estimate of the mean image, as used by a number of groups (eg [11]). The second groupwise method, taking advantage of the statistics of the shape across the set, only gives a very slight improvement on this dataset. In relatively clean data such as this the matching is driven by the texture - one would anticipate that the additional shape constraints will only have modest effects.

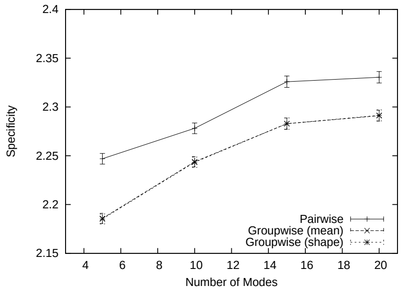


Figure 6: Comparison between different registration algorithms

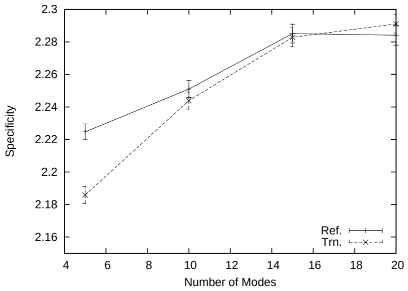


Figure 7: Comparison between evaluating residual in reference frame or image frame

5.5 Frame for Measurement of Residuals

The MDL approach requires that we encode the cost of reconstructing the training images exactly. This means that we must evaluate the residuals in the frame of the training images (ie one per pixel of training images). An alternative approach is to use the current deformation field to pull the training image back into the frame of the reference, and measure the residuals there (ie one per pixel of the reference image). This approach can lead to considerable efficiencies. However, there is a danger that the deformation field can distort to minimise the effect of hard-to-model regions in the training image, leading to a misleading correspondence. We have compared the models constructed when measuring residuals in the training and reference images – see Figure 7. This demonstrates that measuring the residuals in the training image frame (as recommended by the MDL approach) leads to better models.

5.6 Face Data

We applied the model building algorithms to images taken from a sequence of a talking face (every 10^{th} frame, 100 in total). Figure 8 shows examples from the frames together with the resulting control point positions. Note that a fully automatic (but fairly simple) method was used to select control points on the initial reference image (initialising on a grid, then moving each point to a nearby strong edge) - better results would probably be obtained with a more sophisticated choice of points. Figure 9 shows two of the resulting model modes.

As for the MR images, we performed experiments evaluating the effect of different algorithm components, and came to similar conclusions.

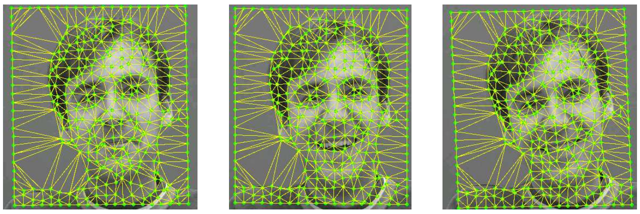


Figure 8: Examples from face training set with resulting control points

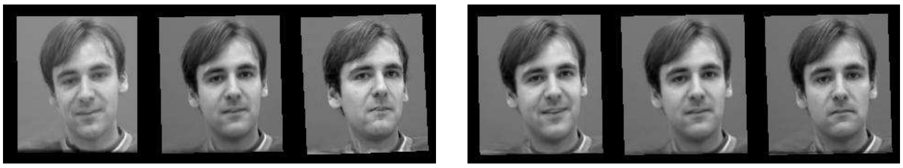


Figure 9: First and third combined face model modes

6 Discussion and Conclusions

We have described a method of computing correspondences across a training set of images which minimises the description length required to encode the data using an appearance model constructed from the correspondences. We have used piece-wise affine deformation fields as they allow simple computation of the inverse, required for efficient implementation of the algorithm. We have applied the technique to two large datasets, and performed experiments which suggest that the use of groupwise registration gives better correspondences than simple pairwise approaches, that the choice of control points can have a significant effect on results and that it is advantageous to estimate residuals in the image frame, rather than the reference frame.

In the future we will extend this framework to use the MDL approach to decide what is the most appropriate tradeoff between the complexity of the shape model and that of the texture model, and to automatically select the most efficient representations for different regions of the image. The algorithms extend naturally to 3D, and we will explore their performance on full 3D MR images.

Techniques such as this will be very important for building statistical appearance models from large datasets, where manual annotation is not practical.

Acknowledgements

The work in this paper was funded by EPSRC and MRC under the MIAS-IRC grant.

References

[1] S. Baker, I. Matthews, and J.Schneider. Automatic construction of active appearance models as an image coding problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(10):1380–84, 2004.

[2] C.J.Twining, T.F.Cootes, S.Marsland, V.Petrovic, R. Schestowitz, and C.J.Taylor. A unified information-theoretic approach to groupwise non-rigid registration and model building. In 19^{th} *Conference on Information Processing in Medical Imaging*, 2005.

[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 8^{th} *European Conference on Computer Vision*, volume 4, pages 316–327. Springer, 2004.

[4] T.F. Cootes, C.J. Twining, and C.J. Taylor. Diffeomorphic statistical shape models. In *15th British Machine Vision Conference*, volume 1, pages 447–456, 2004.

[5] R.H. Davies, C.Twining, T.F. Cootes, and C.J. Taylor. A minimum description length approach to statistical shape modelling. *IEEE Transactions on Medical Imaging*, 21:525–537, 2002.

[6] Michael J. Jones and Tomaso Poggio. Multidimensional morphable models. In 6^{th} *International Conference on Computer Vision*, pages 683–688, 1998.

[7] Michael J. Jones and Tomaso Poggio. Multidimensional morphable models : A framework for representing and matching object classes. *International Journal of Computer Vision*, 2(29):107–131, 1998.

[8] J. Lötjönen and T. Mäkelä. Elastic matching using a deformation sphere. In *MICCAI*, pages 541–548, 2001.

[9] J. B. Antoine Maintz and Max A. Viergever. A survey of medical image registration. *Medical Image Analysis*, 2(1):1–36, 1998.

[10] S. Marsland, C. Twining, and C. Taylor. Groupwise non-rigid registration using polyharmonic clamped-plate splines. In *MICCAI*, Lecture Notes in Computer Science, 2003.

[11] D. Rueckert, A.F. Frangi, and J.A. Schnabel. Automatic construction of 3D statistical deformation models using non-rigid registration. In *MICCAI*, pages 77–84, 2001.

[12] B Zitová and J Flusser. Image registration methods: A survey. *Image and Vision Computing*, 21:977–1000, 2003.