An Information Theoretic Approach to Statistical Shape Modelling

Rhodri H. Davies, Tim F. Cootes, Carole J. Twining and Chris J. Taylor
Imaging Science, Stopford Building
University of Manchester, Manchester, M13 9PT, UK
rhodri.h.davies@stud.man.ac.uk

Abstract

Statistical shape models have been used widely as a basis for segmenting and interpreting images. A major drawback of the approach is the need to establish a set of dense correspondences across a training set of segmented shapes. By posing the problem as one of minimising the *description length* of the model, we develop an efficient method that *automatically* defines correspondences across a set of shapes. Results are given for several different training sets of shapes, showing that the automatic method constructs significantly better models than those built by hand - the current gold standard.

1 Introduction

Statistical models of shape have been used widely as a basis for segmenting and interpreting images [4]. The basic idea is to establish, from a training set, the pattern of 'legal' variation in the shapes and spatial relationships of structures in a given class of images. Statistical analysis is used to give an efficient parameterisation of this variability, providing a compact representation of shape, allowing shape constraints to be applied effectively during image interpretation [4]. One of the main drawbacks of the approach is, however, the need - during training - to establish a dense correspondence between shape boundaries over a reasonably large set of example images. It is important to establish the 'correct' correspondence, otherwise an inefficient parameterisation of shape can result, leading to difficulty in defining shape constraints. In practice, correspondence has often been established using manually defined 'landmarks' - a time-consuming, subjective and error-prone process.

Several previous attempts have been made to automate model building [1, 2, 6, 7, 8, 9, 10, 13] . The problem of establishing dense correspondence over a set of training boundaries can be posed as that of defining a parameterisation for each of the training shapes, leading to an implicit correspondence between equivalently parameterised points. Different arbitrary parameterisations of the training boundaries have been proposed [1, 9] , but do not address the issue of optimality. Shape 'features' (e.g. regions of high curvature) have been used to establish point correspondences, [2, 8, 13] but, although this approach corresponds with human intuition, it is still not clear that it is in any sense optimal. A third approach, and that followed in this paper, is to treat finding the correct parameterisation of the training shape boundaries as an explicit optimisation problem.

The optimisation approach has been described by several authors [3, 6, 10]. The basic idea is to find the parameterisation of the training set that yields, in some sense, the 'best' model. Kotcheff and Taylor [10] describe an approach in which the best model is defined in terms of ‘compactness’, as measured by the determinant of its covariance matrix. The parameterisation of each of a set of training shapes is represented explicitly, and a genetic algorithm search is used to optimise the model. Although this work showed promise, there were several problems: the objective function, although reasonably intuitive, could not be rigorously justified; the method was described for 2D shapes and could not easily be extended to 3D; it was sometimes difficult to make the optimisation converge; and it did not address the issue of pose transformations.

In this paper, we define a new objective function with a rigorous theoretical basis that is defined in an information theoretic framework. The key insight is that the ‘best’ model is that which describes the *entire training set* as efficiently as possible, thus we adopt a *minimum description length* criterion. We also describe a novel, continuous representation of correspondence/parameterisation that can be efficiently optimised to produce models that are significantly better than those built by hand.

2 Statistical Shape Models

A statistical shape model is built from a training set of example shapes [4]. Each shape, S_i $(i = 1 \ldots s)$, can (without loss of generality) be represented by a set of $(n/2)$ points sampled along the boundary. A variation of generalised Procrustes analysis is then performed to align each member of the training set to a reference shape by finding the rigid pose parameters that minimise the sum of squared distances between the points. By concatenating the coordinates of its sample points into a n-dimensional vector, $\mathbf{x}$, and using principal component analysis (PCA), each shape vector can be described by a linear model of the form

$$\mathbf{x}_i = \bar{\mathbf{x}} + \mathbf{P}\mathbf{b}_i \tag{1}$$

where $\bar{\mathbf{x}}$ is the mean shape vector, the columns of **P** describe a set of orthogonal modes of shape variation and **b** is a vector of shape parameters. New examples of the class of shapes can be generated by choosing values of **b** within the range found in the training set. This approach can be extended to deal with continuous boundary functions [10], but for clarity, we limit our discussion to the discrete case.

The utility of the linear model of shape shown in (1) is dependant on the set of boundary parameterisations $\{\Phi_i\}$ that are chosen. An inappropriate choice can result in the need for a large number of modes (and corresponding shape parameters) to approximate the training shapes to a given accuracy and may lead to ‘legal’ values of **b** generating ‘illegal’ shape instances. For example, consider the two models generated from a set of 17 hand outlines, shown in figure 1. Model A uses a set of parameterisations of the outlines that cause ‘natural’ landmarks such as the tips of the fingers to correspond. Model B uses one such correspondence but then uses a simple arc-length parameterisation to position the other sample points. The variance of the three most significant modes of models A and B are (1.06, 0.58, 0.30) and (2.19, 0.78, 0.54) respectively. This suggests that model A is more compact than model B. All the example shapes generated by model A using values of **b** within the range found in the training set are ‘legal’ examples of hands, whilst model B generates implausible examples as illustrated in the figure.

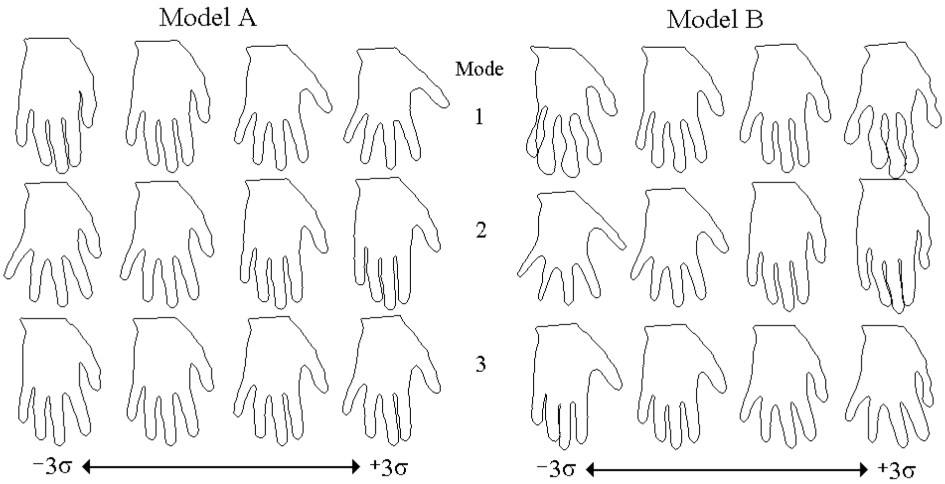

Figure 1: The first three modes of variation (±3σ) of models A and B

The set of parameterisations used for model A were obtained by marking the 'natural' landmarks manually on each training example, then using simple path length parameterisation to sample a fixed number of equally spaced points between them. This manual mark-up is a time-consuming and subjective process. In principle, the modelling approach extends to 3D, but in practice, manual landmarking becomes impractical. We aim to overcome this problem by building statistical shape models *automatically* from a set of training shapes.

3 Automatic Model Building

We wish to optimise the parameterisations $\{\Phi_i\}$ of each shape in our training set $\{S_i\}$. Since we wish to obtain a compact model with good generalisation ability, we formulate the problem as one of finding the *minimum description length* of the training shapes.

The configuration space of $\{\Phi_i\}$ is highly non-linear and has many local minima. Although stochastic optimisation techniques such as simulated annealing and genetic algorithms search for a truly *global* minima, they take many hours to converge. We overcome this problem by optimising $\{\Phi_i\}$ using a *multiresolution* approach. This allows a local optimisation method to be used at each resolution. We have used the Nelder-Mead simplex algorithm [12] to produce the results in section 4.

3.1 An Information Theoretic Objective Function

We wish to define a criterion for choosing the set of parameterisations $\{\Phi_i\}$ that are used to construct a statistical shape model from a set of training boundaries $\{S_i\}$. Our aim is to choose $\{\Phi_i\}$ so as to obtain the 'best possible' model. Ideally, we would like a model that is general (it can represent any instance of the object - not just those seen in the training set), specific (it can only represent valid instances of the object) and compact (it can represent the variation with as few parameters as possible). We therefore choose to follow the principle of Occam's razor : the simplest explanation generalises best. In our case, we need to find the simplest explanation of the training set. We formalise this by stating that we wish to find $\{\Phi_i\}$ that minimises the information required to code the whole training set, $\{\mathbf{x}_i\}$.

Suppose we have a set $\{S_i\}$ of s training shapes that are parameterised using $\{\Phi_i\}$ and sampled to give a set of n-dimensional shape vectors $\{\mathbf{x}_i\}$. Following (1) we can approximate $\{\mathbf{x}_i\}$ to an accuracy of δ in each of its elements using a linear shape model of the form

$$\mathbf{x}_i = \bar{\mathbf{x}} + \mathbf{P}\mathbf{b}_i + \mathbf{r}_i \tag{2}$$

Where $\bar{\mathbf{x}}$ is the mean of $\{\mathbf{x}_i\}$, **P** has m columns which are the m eigenvectors of the covariance matrix of $\{\mathbf{x}_i\}$ corresponding to the m largest eigenvalues λ_j, $\mathbf{b}_i$ is a vector of shape parameters, and $\mathbf{r}_i$ is a vector of residuals. The elements of $\mathbf{r}_i$ can be shown to have zero mean and a variance of $\lambda_{\mathbf{r}} = \frac{1}{n}\sum_{j=m+1}^{n} \lambda_j$ over the training set.

The total information required to code the complete training set using this encoding is given by

$$I_{Total} = I_{Model} + sI_{\mathbf{b}} + sI_{\mathbf{r}} \tag{3}$$

Where I_{Model} is the information required to code the model (the mean vector, $\bar{\mathbf{x}}$, and the eigenvectors of the covariance matrix, **P**), $I_{\mathbf{b}}$ is the average information required to code each parameter vector $\mathbf{b}_i$, and $I_{\mathbf{r}}$ the average information required to code each residual vector, $\mathbf{r}_i$.

For simplicity, we assume that the elements of the mean $\bar{\mathbf{x}}$ and the matrix **P** are uniformly distributed in the range [-1,1], and that we use k_m bits per element for the mean and k_j bits per element for the j^{th} column of $\mathbf{P}$ giving quantisation errors $\delta_m = 2^{-k_m}$ and $\delta_j = 2^{-k_j}$ respectively. Thus

$$I_{Model} = nk_m + n\sum_{j=1}^{m} k_j \tag{4}$$

The elements of $\mathbf{b}_i$ are assumed to be normally distributed over the training set with zero mean and variance λ_j. To code them to an accuracy $\delta_{\mathbf{b}}$, we require on average

$$I_{\mathbf{b}} = \sum_{j=1}^{m} [k_{\mathbf{b}} + 0.5\log(2\pi e\lambda_j)] \tag{5}$$

Where $k_b = -log(\delta_b)$. All logs are base 2.
Similarly, to code the n elements of $\mathbf{r}_i$ to an accuracy of $\delta_r = 2^{-k_r}$ we require on average

$$I_{\mathbf{r}} = n[k_{\mathbf{r}} + 0.5\log(2\pi e\lambda_{\mathbf{r}})] \tag{6}$$

Substituting (4), (5) and (6) into (3) we obtain

$$I_{Total} = nk_m + n\sum_{j=1}^{m} k_j + s\sum_{j=1}^{m} [k_{\mathbf{b}} + 0.5\log(2\pi e\lambda_j)] + sn[k_{\mathbf{r}} + 0.5\log(2\pi e\lambda_{\mathbf{r}})] \tag{7}$$

I_{Total} is a function of the quantisation parameters k_m, k_j, $k_{\mathbf{b}}$, and $k_{\mathbf{r}}$, which are related to δ, the overall approximation error. Since we wish ultimately to minimise I_{Total} with respect to $\{\Phi_i\}$ we need first to find the minimum with respect to the quantisation parameters. This can be found analytically, leading to an expression in terms of s, n, k, m and $\{\lambda_j\}$.

$$\begin{aligned} I_{Total} = & -0.5(n + nm + sm)\log(12\alpha\lambda_{\mathbf{r}}/s) + snk + 0.5(n+s)\sum_{j=1}^{m}\log(\lambda_j) \\ & +0.5ns\log(\alpha\lambda_{\mathbf{r}}) + 0.5s(n+m)\log(2\pi e) - 0.5sm\log(s) \end{aligned} \tag{8}$$

where $\alpha = (\frac{ns}{n(s-1)-m(n-s)})$.

Thus, for a fixed number of modes, m, to optimise I_{Total} we need to minimise

$$F(m) = (n+s)\sum_{j=1}^{m}\log(\lambda_j) + [n(s-1) - m(n+s)]\log(\lambda_{\mathbf{r}}) \tag{9}$$

Note that this is independent of δ. Finally, the number of modes, m, should be chosen to minimise I_{Total}. Since m must be an integer, this can be achieved using a simple, exhaustive search to find F_{min}. Note, however, that the average information required to code b_j, the j^{th} element of the shape vector $\mathbf{b}_i$, is $k_{\mathbf{b}} + 0.5\log(2\pi e\lambda_j)$. This must be greater than zero, which imposes an upper bound on m such that $\lambda_m > 12\alpha\lambda_{\mathbf{r}}/(2\pi e)$. F_{min}, the minimum of F with respect to m can be used to asses the quality of a given model.

3.2 Optimisation of Parameterisation

We select corresponding points by uniformly sampling the parameterisation, $\Phi(t)$, of each shape - see figure 2. We wish to manipulate the set of parameterisations $\{\Phi_i\}$ in a way that minimises the value of F_{min}. The method described in this section is applicable to both open and closed curves; for clarity, we will limit our discussion to the closed case.

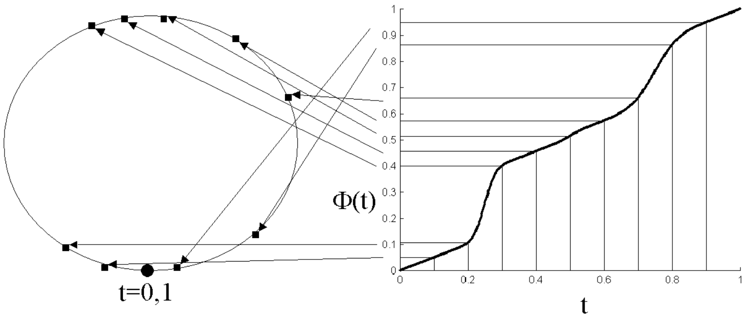

Figure 2: How a shape is sampled according to its parameterisation. The sampled points depend on the shape of the parameterisation function, Φ

A legal reparameterisation function $\Phi(t)$ is a monotonically increasing function of t, with a range $(0 \leq \Phi(t) \leq 1)$. Such a function can be represented as the cumulative distribution function of some normalised, positive definite density function $\rho(x)$, where $\Phi(t) = \int_0^t \rho(x)dx$.

We choose to represent $\rho(x)$ as a sum of Gaussian kernels:

$$\rho(x) = c\Big[1 + \sum_i \frac{A_i}{\sigma_i \sqrt(2\pi)} \exp\Big(- \frac{1}{2\sigma_i^2}(x - a_i)^2\Big)\Big] \tag{10}$$

where the coefficients A_i control the height of each kernel; σ_i specifies the width and a_i the position of the centre and c is the normalisation constant. We include the constant term to ensure that when all A_i's are zero, $\Phi(t)$ is an arc-length parameterisation.

Given this representation of $\rho(x)$, we calculate $\Phi(t)$

$$\Phi(t) = \int_0^t \rho(x)dx = c\Big[t + \sum_i \frac{A_i}{2}\mathrm{erf}\Big(\frac{t - a_i}{\sigma_i\sqrt{2}}\Big) + \sum_i \frac{A_i}{2}\mathrm{erf}\Big(\frac{a_i}{\sigma_i\sqrt{2}}\Big)\Big] \tag{11}$$

$$\begin{aligned} \text{where} \qquad c^{-1} &= 1 + \sum_i \frac{A_i}{2}\mathrm{erf}\Big(\frac{1 - a_i}{\sigma_i\sqrt{2}}\Big) + \sum_i \frac{A_i}{2}\mathrm{erf}\Big(\frac{a_i}{\sigma_i\sqrt{2}}\Big) && (12) \\ \text{and} \qquad \mathrm{erf}(x) &= \frac{2}{\sqrt{\pi}} \int_0^x e^{-r^2} dr && (13) \end{aligned}$$

We manipulate the parameterisation by varying $\{A_i\}$, the heights of the kernels; the widths $\{\sigma_i\}$ and the positions $\{a_i\}$ are fixed. If we use n_k kernels to represent the parameterisation, the configuration space becomes ($n_k s$)-dimensional. This search space is generally too large for a direct optimisation scheme to converge rapidly and reliably. We overcome this by using the following multiresolution approach:

- We begin with a single kernel of width $\sigma_1 = \frac{1}{4}$, centred at $a_1 = \frac{1}{2}$ on each shape. The height, A_1 of the kernel on each shape is initialised to zero - equivalent to an arc-length parameterisation. We employ an optimisation algorithm to find the magnitude A_1, of the kernel on each shape that minimises F.

- At each subsequent iteration k, we add an additional 2^{k-1} kernels of width $\frac{1}{4}(\frac{1}{2})^k$. The new kernels are positioned at intervals of $\frac{1}{2}^k$ between $t = 0$ and $t = 1$ so that they lie halfway between the kernels added on previous iterations. The optimisation algorithm is used to find the best height for each kernel.

- We continue recursively adding additional kernels until the parameterisation is suitably defined.

The pose of each shape affects the value of F_{min}. We therefore need to optimise the four parameters that allow a rigid transformation of each shape: translations d_x, d_y, scaling s and rotation θ. We have found, however, that adding an additional $4s$ dimensions to each iteration significantly slows the optimisation and introduces many additional false minima. Better results can be achieved by performing a procrustes analysis [5] of the reparameterised shapes inside the objective function, before calculating the value of F_{min}.

In the experiments reported below, we have assumed that a correspondence exists between the origins of each shape in the training set. If we do not have such a correspondence, (11) must be modified so that $\Phi(t) \to \big(\epsilon + \Phi(t)\big)\mathrm{mod}1$, where ϵ specifies the offset of the origin.

Although we do not give details here, this approach can be extended to reparameterising the sphere and the plane, allowing 3D statistical shape models to be constructed.

4 Results

We present qualitative and quantitative results of applying our method to several sets of outlines of 2D biomedical objects. We also investigate how our objective function behaves around the minimum and how it selects the correct number of modes to use.

4.1 Results on 2D Outlines

We tested our method on a set of 17 hand outlines, 38 left ventricles of the heart, 24 hip prostheses and 15 cross-sections of the femoral articular cartilage. The algorithm was run for four iterations, giving 16 kernels per shape. A MATLAB implementation of the algorithm takes between 20 and 120 minutes, depending on the size of training set .We compare the results to models built by equally-spacing points along the boundary and hand-built models, produced by identifying a set of 'natural' landmarks on each shape.

In figure 3, we show qualitative results by displaying the variation captured by the first three modes of each model (the first three elements of **b** varied by $\pm 2\sigma$). We also give quantitative results in table 1, tabulating the value of F_{min} and the total variance.

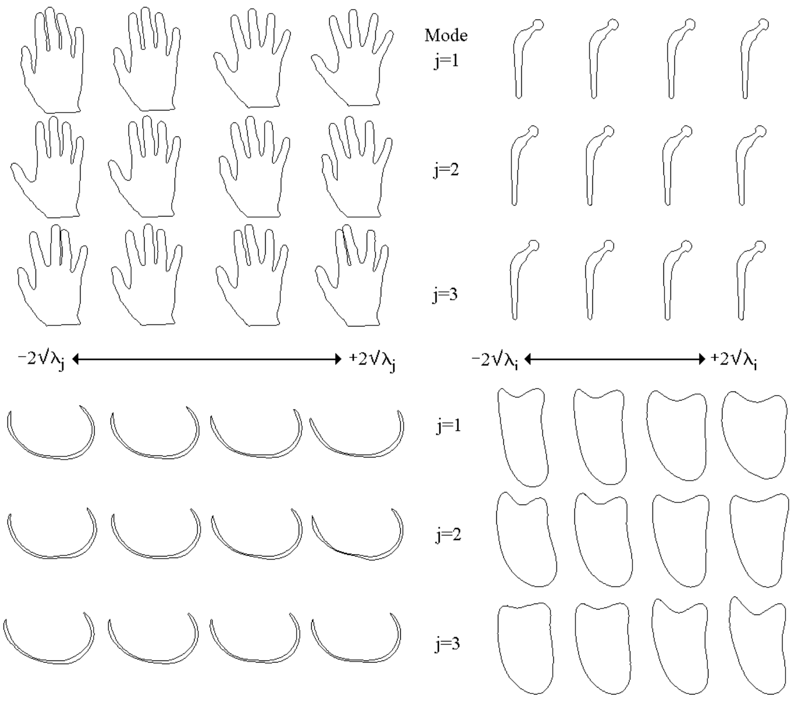

Figure 3: The first three modes $(j = 1, j = 2, j = 3)$ of variation $(\pm 2\sigma)$of the automatically generated models

Hands

	Automatic	Hand Built	Equally-spaced
V_T	6.32	9.92	17.18
F_{min}	6645	8177	9072

Hip Prostheses

	Automatic	Hand Built	Equally-spaced
V_T	6.08	13.7	13.2
F_{min}	3377	16443	11366

Knee Cartilage

	Automatic	Hand Built	Equally-spaced
V_T	8.31	10.7	11.3
F_{min}	2478	3517	3567

Heart Ventricles

	Automatic	Hand Built	Equally-spaced
V_T	4.9	14.1	14.6
F_{min}	885	1360	1470

Table 1: A quantitative comparison of each model F_{min}, the value of the objective function and V_T, the total variance.

The qualitative results in figure 3 show that the shapes generated within the allowed range of **b** are all plausible, this suggests a specific model. The quantitative results in table 1 show that our method produces models that are significantly more compact than either the models built by hand or those obtained using equally-spaced points. It is interesting to note that the model produced by equally-spacing points on the hip prosthesis is more compact than the manual model. This is because equally-spaced points suffice as there is little variation, but errors in the manual annotation adds additional noise which is captured as a statistical variation.

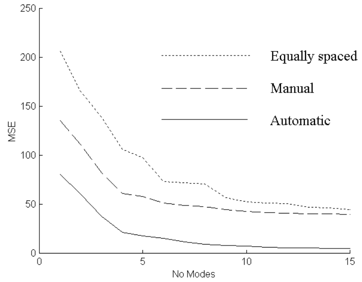

Figure 4: Leave one out tests. The plot shows the number of modes used against the mean squared approximation error

To test the generalisation ability of the models, we performed leave-one-out tests on each model described in table 1. In figure 4, we report the results on the hand outlines although the same trends appear in all datasets. As can be seen from the figure, the optimised model performs significantly better than both the manual and arc-length parameterised models for the entire range of included modes, suggesting better generalisation ability.

4.2 The Behaviour of the Objective Function

To demonstrate the behaviour of our objective function we took some corresponding points from the automatically generated hand model and added random noise to each one. Figure 5 shows a plot of F_{min} against the standard deviation of the noise. The plot shows that as the points are moved further away from their corresponding positions, the value of F_{min} increases - the desired behaviour.

4.3 Selecting the Number of Modes

We used the automatically generated heart model to show how the number of modes affects the value of the objective function. Figure 6 shows a plot of F against the number of modes used in the model. The values form an approximate quadratic with a minimum at nine modes which captures approximately 93% of the total variation.

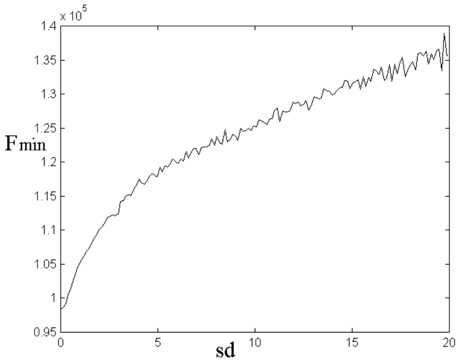

Figure 5: F_{min} increases with the s.d. of random perturbations of the ‘optimal’ correspondences

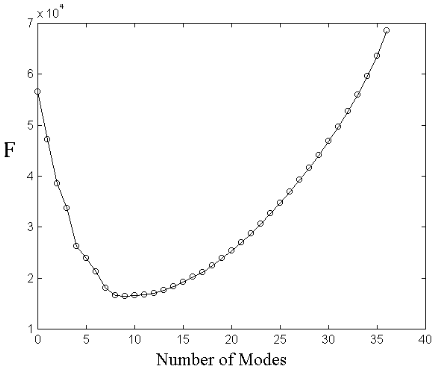

Figure 6: The values of F for a model built with different numbers of modes.

5 Conclusions

The results reported in this paper show that the automatically produced models are significantly more compact and more specific then the models built by hand - the current gold standard. The qualitative results in figure 3 show that the automatically produced models look specific. We are currently working on deriving quantitative measures of specificity.

We have derived an objective function that can be used to evaluate the quality of a statistical shape model. The expression we use has a theoretical grounding in Information Theory, is independent of quantisation error and unlike other approaches [6, 10], does not involve any arbitrary parameters. The objective function includes a $\sum \log(\lambda_i)$ term which is equivalent to the product of the λ_i's, (and thus the determinant of the covariance matrix) as used by Kotcheff and Taylor [10], but the more complete treatment here shows that other terms are also important.

Regular Principal Component Analysis can not capture non-linear variations (e.g caused by a sub-part rotating in the plane) with a single mode - this affects the generalisation ability, specificity and compactness of such linear models. The method described in this paper overcomes this by allowing points to ‘slide’ along the parameterisation to compensate for the non-linear movement - this allows the variation to be explained by a single mode.

Our ultimate aim is to build 3D statistical shape models for use in biomedical research. Although we do not give details here, the principles described in this paper extend to 3D shapes. A limitation of our current reparameterisation function is that it is not smooth at the point $(t = 0, 1)$. Although this makes no difference to 2D shapes, it will have a significant effect on 3D shapes. We intend to overcome this problem by using alternative kernel functions such as the wrapped Gaussian or the wrapped Cauchy [11].

Acknowledgements

Tim Cootes is funded under an EPSRC advanced fellowship grant. Rhodri Davies would like to thank the BBSRC and Astrazeneca pharmaceuticals [1] for their financial support.

References

[1] Baumberg, A. and D. Hogg, Learning Flexible Models from Image Sequences, in *European Conference on Computer Vision*, Stockholm, Sweden. 1994. p. 299-308.

[2] Benayoun, A., N. Ayache, and I. Cohen, Adaptive meshes and nonrigid motion computation. in *International Conference on Pattern Recognition*. 1994.

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

[4] Cootes, T., A. Hill, C. Taylor, and J. Haslam, The use of Active shape models for locating structures in medical images. *Image and Vision Computing*, 1994. 12: p. 355-366.

[5] Goodall, C., Procrustes Methods in the Statistical Analysis of Shape. *Journal of the Royal Statistical Society*, 1991. 53(2): p. 285-339.

[6] Hill, A. and C. Taylor. Automatic landmark generation for point distribution models. in *British Machine Vision Conference*. 1994. Birmingham, England: BMVA Press.

[7] Hill, A. and C.J. Taylor, A framework for automatic landmark identification using a new method of non-rigid correspondence. *IEEE PAMI*, April, 2000.

[8] Kambhamettu, C. and D.B. Goldgof, Point Correspondence Recovery in Non-rigid Motion, in *IEEE CVPR* 1992. p. 222-227.

[9] Kelemen, A., G. Szekely, and G. Gerig, Elastic model-based segmentation of 3-D neuroradiological data sets. *IEEE Medical Imaging*, 1999. 18(10): p. 828-839.

[10] Kotcheff, A.C.W. and C.J. Taylor, Automatic Construction of Eigenshape Models by Direct Optimisation. *Medical Image Analysis*, 1998. 2: p. 303-314.

[11] Mardia K., *Statistics of Directional Data*, Academic Press, 1972

[12] Press, W.H., S. A. Teukolsky, W. T. Vetterling, B. P. Flannery, *Numerical Recipes in C*, Cambridge University Press; 1993

[13] Wang, Y., B. S. Peterson, and L. H. Staib. Shape-based 3D surface correspondence using geodesics and local geometry. *IEEE CVPR 2000*, v. 2: p. 644-51.

[1] Astrazeneca pharmaceuticals, Alderely Park, Macclesfield, UK