

Aligning shapes by minimising the description length

Anders Ericsson and Johan Karlsson
Centre for Mathematical Sciences
Lund University, Lund, Sweden
anderse@maths.lth.se, johank@maths.lth.se

Abstract

When building shape models, it is first necessary to filter out the similarity transformations from the original configurations. This is normally done using Procrustes analysis, that is minimising the sum of squared distances between the corresponding landmarks under similarity transformations. In this article we propose to align shapes using the Minimum Description Length (MDL) criterion. We show that the Procrustes alignment with respect to rotation is not optimal.

The MDL based algorithm is compared with Procrustes on a number of data sets. It is concluded that there is improvement in generalisation when using Minimum Description Length. With a synthetic example it is shown that the Procrustes alignment can fail significantly where the proposed method does not.

The Description Length is minimised using Gauss-Newton. In order to do this the derivative of the description length with respect to rotation is derived.

1 Introduction

Statistical models of shape [5] has turned out to be a very effective tool in image segmentation and image interpretation. Such models are particularly effective in modelling objects with limited variability, such as medical organs.

The basic idea behind statistical models of shape is that from a given training set of known shapes be able to describe new formerly unseen shapes, which still are representative. The shape is traditionally described using landmarks on the shape boundary.

When building shape models, it is first customary to filter out the similarity transformations from the original configurations. The common way to align shapes before building shape models is to do a Procrustes analysis [8, 5]. It locates the unknown similarity transformations by minimising the sum of squared distances from the mean shape. Other methods exist, in Statistical Shape Analysis [10, 2, 13, 5] Bookstein and Kendall coordinates are commonly used to filter out the effect of similarity transformations.

Minimum Description Length, (MDL) [12], is a paradigm that has been used in many different applications. In recent papers [4, 3, 7, 9] this paradigm is used to locate a dense correspondence between the boundaries of shapes.

In this paper we apply the theory presented in [11] and derive the gradient of the description length [7] and propose to align shapes using the Minimum Description Length (MDL) criterion.

It turns out that when aligning shapes using MDL instead of Procrustes the translation becomes the same but the optimal rotation is different. In this paper the scale of all shapes are normalised to one.

The gradient of the description length with respect to the rotation is derived and used to optimise the description length using Gauss-Newton.

The proposed algorithm is tested on a number of datasets and the mean square error of leave one out reconstructions turns out to be lower than or the same as for Procrustes analysis. Using a synthetic example we show that the alignment using description length can get more intuitive and give much lower mean square error on leave one out reconstructions than when using Procrustes.

This paper is organised as follows. In Section 2 the necessary background on shape models, MDL and SVD is given. In Section 3, the gradient of the description length is derived by calculating the gradient of the SVD. This is used to optimise the description length using Gauss-Newton. In Section 4 results of experiments are presented and it is shown that better models are achieved using the description length to align shapes.

2 Preliminaries

2.1 Statistical Shape Models

When analysing a set of n_s similar (typically biological) shapes, it is convenient and usually effective to describe them using Statistical Shape Models. Each shape is typically the boundary of some object and is in general represented by a number of landmarks. After the shapes \mathbf{x}_i ($i = 1 \dots n_s$) have been aligned and normalised to the same size, a PCA-analysis [9] is performed. The alignment is what has been improved in this paper. The i -th shape in the training set can now be described by a linear model of the form,

$$\mathbf{x}_i = \bar{\mathbf{x}} + \mathbf{P}\mathbf{b}_i \quad , \quad (1)$$

where $\bar{\mathbf{x}}$ is the mean shape, the columns of \mathbf{P} describe a set of orthogonal modes of shape variation and \mathbf{b}_i is the vector of shape parameters for the i -th shape.

2.2 MDL

The description length (DL) is a way to evaluate a shape model. The cost in Minimum Description Length (MDL) is derived from information theory and is, in simple words, the effort that is needed to send the model and the shapes bit by bit. The MDL - principle searches iteratively for a model that can transmit the data the cheapest. The cost function makes a tradeoff between a model that is general (can represent any instance of the object), and compact (it can represent the variation with as few parameters as possible). Davies and Cootes relates these ideas to the principle of Occam's razor : the simplest explanation generalises the best.

Since the idea of using MDL for shape models was first published [4] the cost function has been refined and tuned. Here we use a refined version of the simple cost function stated in [14] and derived in [6]

$$DL = \sum_{\lambda_i \geq c} (1 + \log \frac{\lambda_i}{\lambda_c}) + \sum_{\lambda_i < \lambda_c, \lambda_i \geq \lambda_t} \frac{\lambda_i}{\lambda_c} \quad . \quad (2)$$

The scalar DL is the description length and is the cost to transmit the model according to information theory. The scalars λ_i are the eigenvalues of the covariance matrix $(\mathbf{X} - \bar{\mathbf{X}})(\mathbf{X} - \bar{\mathbf{X}})^T$, where \mathbf{X} is the matrix, which rows are the shape configurations in the training set. The constants λ_c and λ_t are calculated from Δ , which describes the precision of the landmark coordinates. Information can only be sent up to a certain degree of accuracy and Δ expresses this accuracy. The constant $\lambda_c = 2\Delta$ is the limit between what is expected to be information and what is expected to be noise. When the range of the data in a mode is smaller than Δ no information needs to be sent. This limit is set by λ_t .

There are two important properties of this cost-function. It is more intuitive than those formerly presented and the derivative is continuous.

2.3 Recapitulation of the SVD

In the rest of the paper, bold letters will be used for denoting vectors and matrices. The transpose of matrix \mathbf{M} is denoted by \mathbf{M}^T and m_{ij} refers to the (i, j) element of \mathbf{M} . The i -th non-zero element of a diagonal matrix \mathbf{D} is referred to by d_i while \mathbf{M}_i designates the i -th column of matrix \mathbf{M} . The i -th element of the vector \mathbf{x} is designated $\mathbf{x}(i)$.

A basic theorem of linear algebra states that any real or complex $M \times N$ matrix \mathbf{A} can be factored into the product of an $M \times M$ orthogonal matrix \mathbf{U} , an $M \times N$ diagonal matrix \mathbf{S} with non-negative diagonal elements (known as the singular values), and an $N \times N$ orthogonal matrix \mathbf{V} .

In other words,

$$\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^T = \sum_{i=1}^N s_i \mathbf{U}_i \mathbf{V}_i^T \quad . \quad (3)$$

The singular values are the square roots of the positive eigenvalues of the matrix $\mathbf{A}^T \mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$.

3 Optimising the DL

3.1 Computing the Jacobian of the singular values

Here we recapitulate on the theory presented in [11]. For a more mathematical investigation in this field we recommend Alan Andrew's work, especially [1].

All matrices \mathbf{A} can be factored into $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^T$, where \mathbf{S} is a diagonal matrix holding all singular values. We are interested in computing the derivatives of the singular values, $\frac{\partial s_k}{\partial a_{ij}}$ for every element a_{ij} of the $M \times N$ matrix \mathbf{A} . Taking the derivative of $\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^T$ with respect to a_{ij} gives the following equation

$$\frac{\partial \mathbf{A}}{\partial a_{ij}} = \frac{\partial \mathbf{U}}{\partial a_{ij}} \mathbf{S} \mathbf{V}^T + \mathbf{U} \frac{\partial \mathbf{S}}{\partial a_{ij}} \mathbf{V}^T + \mathbf{U} \mathbf{S} \frac{\partial \mathbf{V}^T}{\partial a_{ij}} \quad . \quad (4)$$

Clearly, $\forall (k, l) \neq (i, j), \frac{\partial a_{kl}}{\partial a_{ij}} = 0$, while $\frac{\partial a_{ij}}{\partial a_{ij}} = 1$. Since \mathbf{U} is an orthogonal matrix, we have

$$\mathbf{U}\mathbf{U}^T = \mathbf{I} \Rightarrow \frac{\partial \mathbf{U}^T}{\partial a_{ij}} \mathbf{U} + \mathbf{U}^T \frac{\partial \mathbf{U}}{\partial a_{ij}} = \omega_{\mathbf{U}}^{ijT} + \omega_{\mathbf{U}}^{ij} = \mathbf{0} \quad , \quad (5)$$

where $\omega_{\mathbf{U}}^{ij}$ is given by

$$\omega_{\mathbf{U}}^{ij} = \mathbf{U}^T \frac{\partial \mathbf{U}}{\partial a_{ij}} \quad . \quad (6)$$

From Equation (5) it is clear that $\omega_{\mathbf{U}}^{ij}$ is an antisymmetric matrix. Similarly, an anti symmetric matrix $\omega_{\mathbf{V}}^{ij}$ can be defined for \mathbf{V} as

$$\omega_{\mathbf{V}}^{ij} = \frac{\partial \mathbf{V}^T}{\partial a_{ij}} \mathbf{V} \quad . \quad (7)$$

Notice that $\omega_{\mathbf{U}}^{ij}$ and $\omega_{\mathbf{V}}^{ij}$ are specific to each differentiation $\frac{\partial}{\partial a_{ij}}$. By multiplying Equation (4) by \mathbf{U}^T and \mathbf{V} from left and right respectively, and using Equations (6) and (7), the following is obtained:

$$\mathbf{U}^T \frac{\partial \mathbf{A}}{\partial a_{ij}} \mathbf{V} = \omega_{\mathbf{U}}^{ij} \mathbf{S} + \frac{\partial \mathbf{S}}{\partial a_{ij}} + \mathbf{S} \omega_{\mathbf{V}}^{ij} \quad . \quad (8)$$

Since $\omega_{\mathbf{U}}^{ij}$ and $\omega_{\mathbf{V}}^{ij}$ are antisymmetric matrices, all their diagonal elements are equal to zero. Recalling that \mathbf{S} is a diagonal matrix, it is easy to see that the diagonal elements of $\omega_{\mathbf{U}}^{ij} \mathbf{S}$ and $\mathbf{S} \omega_{\mathbf{V}}^{ij}$ are also zero. Thus, Equation (8) yields the derivatives of the singular values as:

$$\frac{\partial s_k}{\partial a_{ij}} = u_{ik} v_{jk} \quad . \quad (9)$$

3.2 The gradient of the DL

Let $\mathbf{x}_1, \dots, \mathbf{x}_{n_s}$ be n_s shapes centred at the origin. The rotation of curve m is denoted θ_m . Differentiating (2) with respect to θ_m , we get the following expression

$$\frac{\partial DL}{\partial \theta_m} = \sum_{\lambda_k \geq \lambda_c} \frac{1}{\lambda_k} \frac{\partial \lambda_k}{\partial \theta_m} + \sum_{\lambda_k < \lambda_c, \lambda_k \geq \lambda_t} \frac{1}{\lambda_c} \frac{\partial \lambda_k}{\partial \theta_m} \quad . \quad (10)$$

We want to calculate $\frac{\partial \lambda_k}{\partial \theta_m}$. Let the m -th row of \mathbf{X} be the configuration of landmarks for shape m after moving the centre of gravity to the origin, normalising scale so that the Euclidian norm is one and rotating according to θ_m .

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 e^{i\theta_1} \\ \vdots \\ \mathbf{x}_{n_s} e^{i\theta_{n_s}} \end{bmatrix}$$

Let \mathbf{Y} be the matrix holding the deviations from the mean shape,

$$\mathbf{Y} = \mathbf{X} - \bar{\mathbf{X}} \quad ,$$

where each row in $\bar{\mathbf{X}}$ is the mean shape $\bar{\mathbf{x}}$,

$$\bar{\mathbf{x}} = \frac{1}{n_s} \sum_{j=1}^{n_s} \mathbf{x}_j e^{i\theta_j} \quad .$$

If we apply principal component analysis to \mathbf{Y} , we can describe our shapes with the linear model in equation (1). A singular value decomposition of \mathbf{Y} gives us $\mathbf{Y} = \mathbf{USV}^T$. Here \mathbf{V} corresponds to \mathbf{P} in equation (1) and the diagonal matrix $\mathbf{S}^T \mathbf{S}$ holds the eigenvalues λ_k .

Now, if y_{mj} is the j -th landmark on shape m and $\frac{\partial y_{mj}}{\partial \theta_m}$ is the derivative of the j -th landmark on shape m with respect to the rotation θ_m then

$$\frac{\partial \lambda_k}{\partial \theta_m} = \frac{\partial s_k^2}{\partial \theta_m} = 2s_k \frac{\partial s_k}{\partial \theta_m} = 2s_k \sum_{pq} \frac{\partial s_k}{\partial y_{pq}} \frac{\partial y_{pq}}{\partial \theta_m} = 2s_k \sum_{pq} u_{pk} v_{qk} \frac{\partial y_{pq}}{\partial \theta_m} \quad , \quad (11)$$

Here it is used that $\frac{\partial s_k}{\partial y_{pq}} = u_{pk} v_{qk}$, where u_{pk} and v_{qk} are elements in \mathbf{U} and \mathbf{V} , see section 3.1.

$$\frac{\partial y_{pq}}{\partial \theta_m} = \frac{\partial \mathbf{x}_p(q) e^{i\theta_p}}{\partial \theta_m} - \frac{1}{n_s} \sum_j \frac{\partial \mathbf{x}_j(q) e^{i\theta_j}}{\partial \theta_m} = \begin{cases} i(1 - \frac{1}{n_s}) \mathbf{x}_m(q) e^{i\theta_m} & p = m \\ -i \frac{1}{n_s} \mathbf{x}_m(q) e^{i\theta_m} & p \neq m \end{cases} \quad , \quad (12)$$

$$\text{since } \frac{\partial \mathbf{x}_p(q) e^{i\theta_p}}{\partial \theta_m} = \begin{cases} i \mathbf{x}_m(q) e^{i\theta_m} & p = m \\ 0 & p \neq m \end{cases} \quad .$$

If n_s is large (this is assumed in our implementation), the second term in (12) can be ignored. Then $\frac{\partial \lambda_k}{\partial \theta_m}$ can be written as

$$\frac{\partial \lambda_k}{\partial \theta_m} = 2s_k i u_{mk} \mathbf{y}_m \mathbf{V}_k \quad , \quad (13)$$

where \mathbf{V}_k is the k -th column in \mathbf{V} and \mathbf{y}_m is the m -th row in \mathbf{Y} .

3.3 Optimisation

When the gradient of an objective function is known for a specific optimisation problem, it generally pays off to use more sophisticated optimisation techniques that use the gradient. Initially the configurations are translated to the origin, since this is optimal for the description length goal function. Then the shapes are normalised so that the Euclidian norm is one for all shapes. It could be interesting to also optimise scale but the global scale must then be preserved. This means that it would be necessary to optimise under constraints. In this work only rotation is optimised using Gauss-Newton.

4 Experimental Validation

The experiments were conducted in the following way. Given a dataset the centre of gravity was moved to the origin for all shapes and scale was normalised to one according to the Euclidian norm. The initialisation for the optimisation for rotation was set to the rotation according to Procrustes. The rotation was then optimised by minimising the Description Length.

In Figure 2 the typical behaviour of the goal function can be seen. Here two rotations (x and y axis) have been optimised to align three shapes. In the left figure it can be seen that the minimum is a well defined global minimum. It seems to be several minima, but this is since the range goes from -2π to 2π in x and y. The right figure zooms in on the origin (the origin corresponds to the Procrustes solution). It can be seen that the minimum is not at the origin. When more shapes are aligned projections of the goal function looks similar to these plots.

We validate our algorithm on five real data sets, see Figure 3.

Hands 23 contours of a hand segmented out semi-automatically from a video stream. To simplify the segmentation it was filmed on a dark background. The contours were sampled in 64 landmarks using arc-length parameterisation.

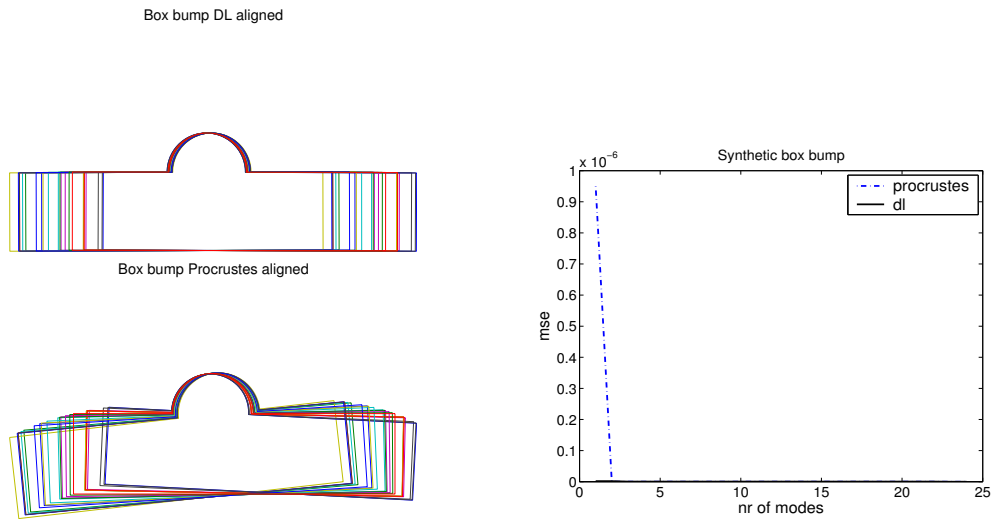


Figure 1: A synthetic example shows that Procrustes alignment can fail (lower). Note that the description length approach succeeds (upper).

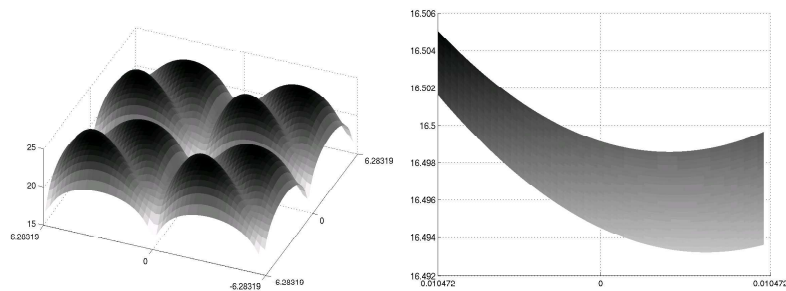


Figure 2: The description length goal function. In the left figure the range on each axis is -2π to 2π . The right figure zooms in on the origin.

Femurs 32 contours of femurs taken from X-rays in the supine projection. The contours were sampled in 64 landmarks using arc-length parameterisation.

Metacarpals 24 contours of metacarpals (a bone in the hand) deduced from standard projection radiographs of the hand in the posterior-anterior projection. The contours were sampled in 57 landmarks.

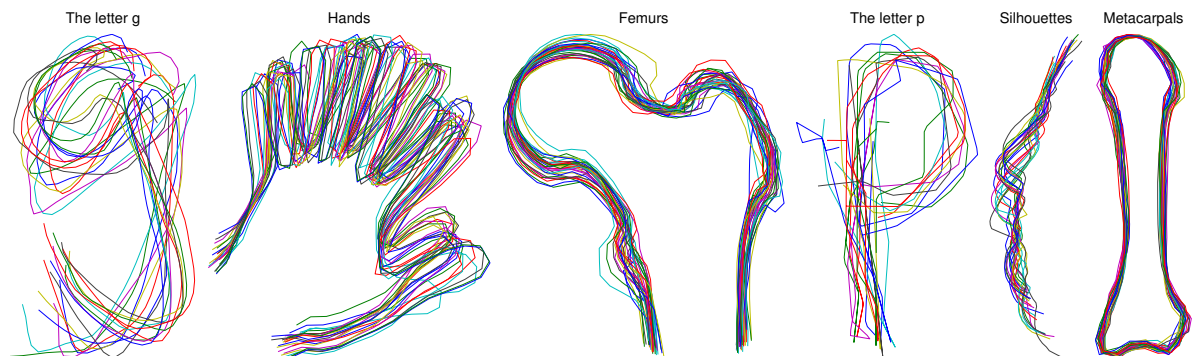


Figure 3: The description length aligned datasets.

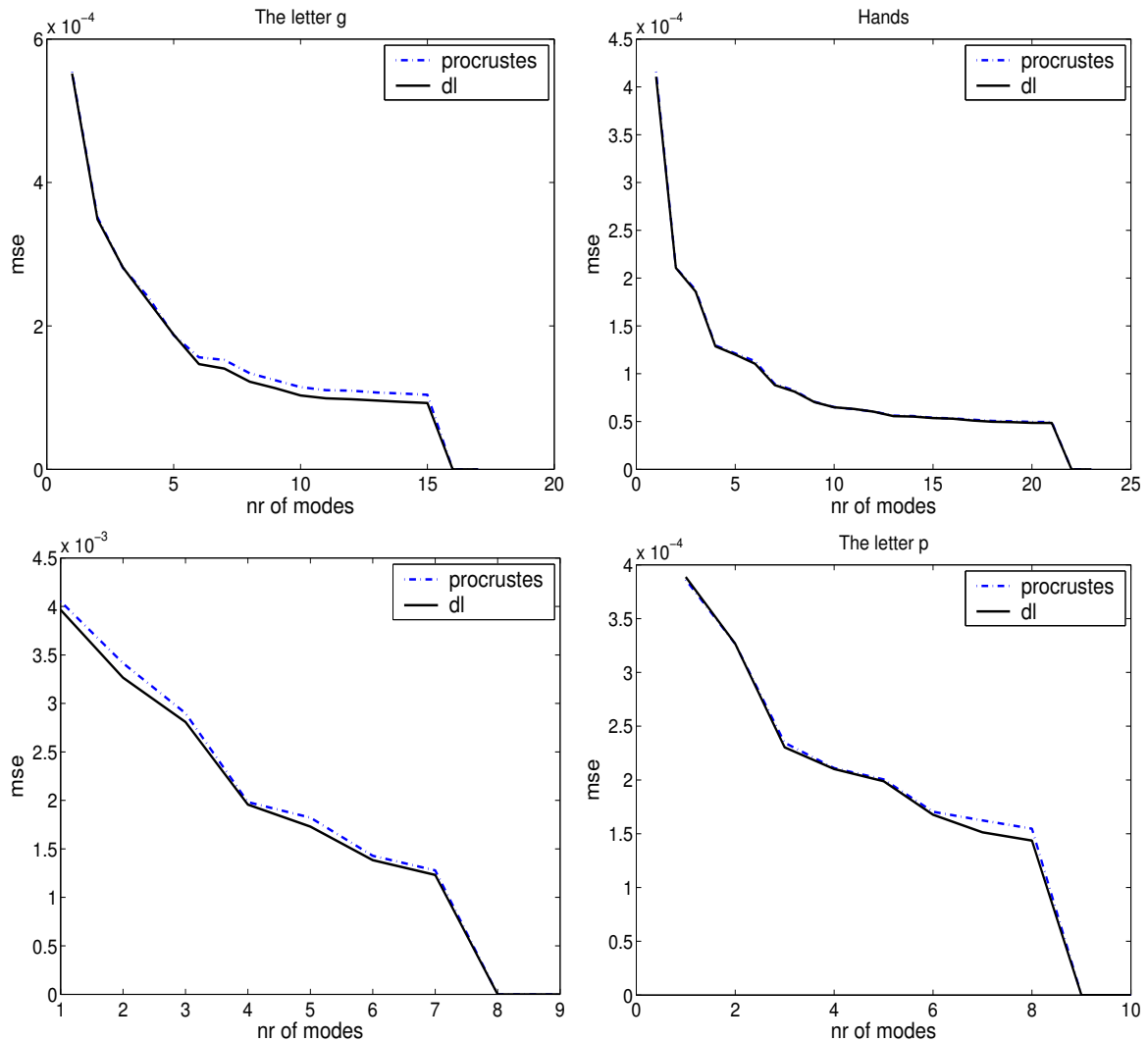


Figure 4: The mean squared error of leave one out reconstructions of the g-dataset, the hand dataset, the three dataset and the 90 p shapes.

Silhouettes The silhouette data set consists of 22 contours of silhouettes of digital camera. The silhouettes were then extracted using an edge detector. The contours were sampled in 81 landmarks.

The letter g One data set of 17 curves of the letter g. The curves of the letter g are sampled using a device for handwriting recognition. The contours were sampled in 64 landmarks using arc-length parameterisation.

The letter p One data set of 90 curves of the letter p. This letter was taken from the MIT database of Latin letters, initially collected by Rob Kassel at MIT (<ftp://lightning.lcs.mit.edu/pub/handwriting/mit.tar.Z>). The contours were sampled in 128 landmarks using arc-length parameterisation.

The quality of the models was measured as the mean square error in leave-one-out reconstructions. The model is built with all but one example and then fitted to the unseen example. This is shown in Figure 4. The plot shows the mean squared approximation error against the number of modes used. This measures the ability of the model to represent unseen shape instances of the object class.

For all examples we get models that give the same or lower error when using the description length criterion compared to Procrustes alignment. This means that the models generalise better.

In Figure 1 is an example of when the Procrustes goes visibly wrong. It is a synthetic example with 24 shapes built up by 128 landmarks. For a human it would be natural to align the boxes and let the bump be misaligned. These shapes are built up with a majority of landmarks around the bumps and therefore the Procrustes method will minimise the error between the bumps instead of the boxes. Note that this data only has one shape mode and therefore perfect alignment should give zero mean squared error on the leave one out reconstruction using just one mode. In this example the description length aligned box bump gets almost zero error on the first shape mode.

5 Summary and Conclusions

In this paper we present a new way to align shapes. The rotation is located by minimising the description length. We derive the gradient of the description length with respect to the rotation and propose to use Gauss-Newton to minimise the MDL-criterion. We have shown that the objective function is differentiable and can be written explicitly.

We have compared the proposed algorithm to Procrustes alignment and shown that better models can be achieved.

One reason why the description length alignment does not get even better results is that when there are many shapes the path to the minimum is difficult for the optimiser to follow. The derivatives gets numerically unstable close to the minimum.

Acknowledgements

Pronosco is acknowledged for providing the contours of femurs and metacarpals. We also like to thank Jesper Skjerning (IMM, DTU) for the silhouettes, Hans Henrik Thodberg (IMM, DTU) for the box bumps and Rob Kassel (MIT) for the Latin letters. And finally we thank Hans Bruun Nielsen (IMM, DTU) for the implementation of the Gauss-Newton optimiser.

References

- [1] A. Andrew, E. Chu, and P. Lancaster. Derivatives of eigenvalues and eigenvectors of matrix functions. pages 903–926, 1993.
- [2] F. L. Bookstein. Size and shape spaces for landmark data in two dimensions. *Statistical Science*, 1(2):181–242, 1986.
- [3] R.H. Davies, C.J. Twining, T.F. Cootes, J.C. Waterton, and C.J. Taylor. A minimum description length approach to statistical shape modeling. *IEEE Trans. medical imaging*, 21(5):525–537, 2002.
- [4] Rhodri H. Davies, Tim F. Cootes, John C. Waterton, and Chris J. Taylor. An efficient method for constructing optimal statistical shape models. In *Medical Image Computing and Computer-Assisted Intervention MICCAI'2001*, pages 57–65, 2001.
- [5] I. L. Dryden and K. V. Mardia. *Statistical Shape Analysis*. John Wiley & Sons, Inc., 1998.
- [6] A. Ericsson. Automatic shape modelling and applications in medical imaging. Technical report, Centre for Mathematical Sciences, Box 118, SE-22100, Lund, Sweden, nov 2003.
- [7] A. Ericsson and K. Åström. Minimizing the description length using steepest descent. In *Proc. British Machine Vision Conference, Norwich, United Kingdom*, volume 2, pages 93–102, 2003.
- [8] J.C. Gower. Generalized procrustes analysis. *Psychometrika*, 40:33–50, 1975.
- [9] J. Karlsson, A. Ericsson, and K. Åström. Parameterisation invariant statistical shape models. In *Proc. International Conference on Pattern Recognition, Cambridge, UK*, 2004.
- [10] D.G. Kendall, D. Barden, T.K Carne, and H. Le. *Shape and Shape Theory*. John Wiley & Sons, Ltd., 1999.

- [11] T. Papadopoulos and M. Lourakis. Estimating the jacobian of the singular value decomposition. In *Proc. European Conf. on Computer Vision, ECCV'00*, pages 555–559, 2000.
- [12] J. Rissanen. Modeling by shortest data description. *Automatica*, 14:465–471, 1978.
- [13] C.G. Small. *The Statistical Theory of Shape*. Springer, 1996.
- [14] H.H. Thodberg. Minimum description length shape and appearance models. In *Image Processing Medical Imaging, IPMI 2003*, 2003.