

Geodesic Ground Truth Correspondence Measure for Benchmarking

A. Ericsson and J. Karlsson
Centre for Mathematical Sciences
Lund University
{anderse, johank}@maths.lth.se.

Abstract

Automatic localisation of correspondences for the construction of Statistical Shape Models from examples has been the focus of intense research during the last decade. Several algorithms are available and benchmarking is needed to rank the different algorithms. Prior work has argued that the quality of the models produced by the algorithms can be evaluated by measuring compactness, generality and specificity. In this paper severe problems of these measures are discussed and a new measure called Ground truth Correspondence Measure (gcm) is instead proposed for benchmarking. The key point with gcm is that from a database with known true correspondences, algorithms can be evaluated by measuring how close the correspondences produced by the algorithms are to the true correspondences. Minimum Description Length (MDL) with a curvature cost comes out as the winner of the automatic methods. Handmarked models turn out to be best but a semiautomatic method is shown to be nearly as good.

1 Introduction

Statistical shape modelling has turned out to be a very effective tool in image segmentation and image interpretation. A major drawback is that a dense correspondence between the shapes must be established.

In recent years there has been a lot of work on automatic construction of Shape Models. There are several different algorithms for this automatic model construction. The algorithms find optimal parameterisations of the shapes in the training set to get correspondences between the shapes.

Many have stated the correspondence problem as an optimisation problem [5, 1, 4, 7, 8, 3]. Minimum Description Length, (MDL) [2], is a paradigm that has been used in many different applications, often in connection with evaluating a model. In recent work [2, 3, 6] this paradigm is used to locate a dense correspondence between shapes.

In short the field has matured and there are many algorithms available. Benchmarking is now needed. In recent years a similar development has taken place in the field of stereo.

In order to evaluate these algorithms, different measures of the quality of the parameterisations and the resulting models can be used. The standard measures are compactness, specificity and generality [2]. It is also common to evaluate correspondences subjectively by plotting the shapes with some corresponding points marked.

The viewpoint in this paper is that a model built from true correspondences is the ideal model since it captures the true variation in the shape class. Measures like specificity, compactness and generality are clues to the quality of a model, but if ground truth correspondences are known it is possible to measure the quality of the model by measuring the closeness of the produced correspondences to the known truth. Verification using ground truth is an established method in computer vision, see for example the CAVIAR project.

There are four major contributions in this paper. (i) We show that former shape model measures have severe weaknesses and that gcm is better suited for evaluating correspondences. (ii) The gcm measure is improved by measuring geodesic distances and also a version of gcm with the Mahalanobis distance is examined. (iii) A database of datasets and matlab code for evaluating algorithms is published. (iv) Benchmarking of several state of the art algorithms is presented and MDL with curvature cost comes out as the winner of the automatic methods. (v) Contrary to former results handmade models are shown to be the best and a semi-automatic algorithm is proposed that outperforms all of the automatic algorithms.

2 Ground truth Correspondence Measure

Up until now compactness, specificity and generality [2] have been used to measure the quality of shape models. Compactness is the sum of the variances in the

shape modes. Specificity is measured by generating a large number of shapes using the model and then measuring how close they are to shapes in the training set. Generality is measured as the sum of errors when approximating a left out shape using the model built from all the other shapes. To evaluate correspondences it is common to plot the landmarks on the shapes and visually subjectively evaluate the correspondences. In our recent work and in the Experimental Validation in Section 3 it is shown that these standard measures have severe weaknesses, both in their definitions and in practical use. Instead gcm is proposed.

In order to measure the quality of the correspondences produced by an algorithm for automatic correspondence localisation, datasets with manually located landmarks have been used. These landmarks are called the ground truth. For synthetic examples these marks are exact. Let the parameterisations γ_i of the shapes \mathbf{x}_i be optimised by the algorithm that is to be evaluated. For every shape \mathbf{x}_i ($i = 1 \dots n_s$) together with its ground truth points \mathbf{g}_{ij} ($j = 1 \dots n_g$), find t_{ij} so that $\mathbf{x}_i(\gamma_i(t_{ij})) = \mathbf{g}_{ij}$. This means that the parameter values that correspond to the ground truth points on the shape are located. Now, for every shape \mathbf{x}_k ($k = 1 \dots n_s$) use the same parameter values. The idea is that the points produced should be close to the ground truth points of this shape, if the parameterisation functions represent good correspondences. That is, $\mathbf{x}_k(\gamma_k(t_{ij}))$ should be close to \mathbf{g}_{kj} . This is measured as a sum of distances over all shapes in the dataset.

$$gcm = \frac{1}{n_s(n_s - 1)n_g} \sum_{i=1}^{n_s} \sum_{j \in J} \sum_{k=1}^{n_g} \|\mathbf{x}_k(\gamma_k(t_{ij})) - \mathbf{g}_{kj}\|$$

$$J = \{1, \dots, i - 1, i + 1, \dots, n_s\}$$

Here $\|\cdot\|$ can be any norm. In our recent work the Euclidean norm was used, but in this paper the geodesic distance along the shape is used. On curves this corresponds to the arclength. The constant n_s is the number of shapes and n_g is the number of ground truth points.

Ground truth Correspondence Measure with Mahalanobis distance Due to the subjective nature of choosing ground truth points on natural shapes, an alternative ground truth correspondence measure could be used. Say that a number of people have marked ground truth points on the same data set. Means and variances can then be calculated and the norm used to calculate gcm is the Mahalanobis geodesic distance.

$$gcm = \frac{1}{n_s(n_s - 1)n_g} \sum_{i=1}^{n_s} \sum_{j \in J} \sum_{k=1}^{n_g} \frac{\|\mathbf{x}_k(\gamma_k(t_{ij})) - \bar{\mathbf{g}}_{kj}\|}{\sigma_{kj}}$$

$$J = \{1, \dots, i - 1, i + 1, \dots, n_s\},$$

where $\bar{\mathbf{g}}_{kj}$ is the mean and σ_{kj} is the standard deviation for landmark j on shape k .

3 Experimental Validation of gcm

The first experiment was to start from correct correspondences and then optimise the parameterisation so as to minimise the description length. Synthetic box bump shapes, consisting of a rectangle with a bump on different positions on the top side, have been used for this test, since we know the true correspondence here. The value of the description length (DL) and the ground truth correspondence measure (gcm) over the number of iterations is plotted in Figure 1.

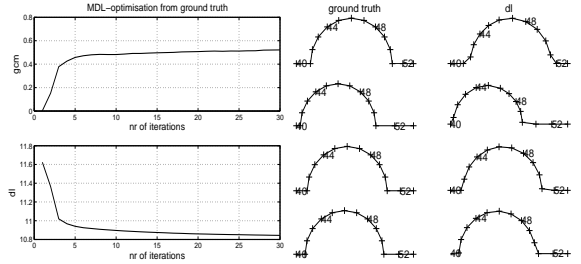


Figure 1: To the left Gcm and dl plotted over number of iterations. To the right true correspondences and optimised correspondences.

It is interesting to note here that the gcm increases as the description length decreases. The minimum, when the parameterisation is optimised with description length as goal function, does not correspond to true correspondences. In Figure 1 it can be seen that minimising the description length from true correspondences has resulted in worse correspondences. In Fig-

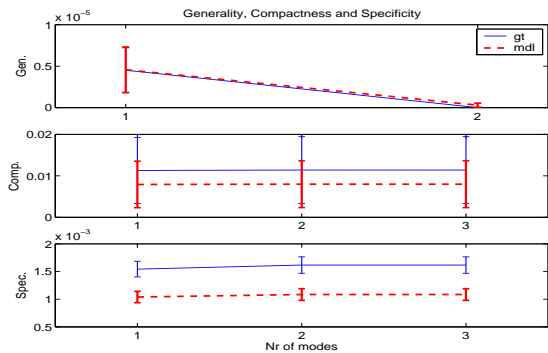


Figure 2: Generalisation, compactness and specificity of the ground truth boxbump model and the optimised box bump model.

ure 2 the compactness and specificity measures indicate that the optimised model is better, but since in this case we started from ground truth this is not correct. So these measures do not appear to measure what we want. The measure of compactness assumes that a model with less variation (more compact model) is better. With this example it turns out that this is not true

in general. A model based on true correspondences is an optimal model independent of compactness. As an extreme example it is possible to get perfect compactness (zero) by placing all landmarks in one point on all shapes.

The problem with the specificity measure is that if the training set is limited (which is often the case) it can not be assumed that all shapes of a class are close to one of the shapes in the training set.

Summing up this experiment, the conclusion is that although minimising the description length is a good method, it does not measure what we really wish to optimise and in this case it fails to recognise the true optimum. We also conclude that compactness and specificity can not, in general, decide which model is the best.

In the second experiment silhouette shapes initialised with arlength parameterisation were used. We optimise the parameterisation with respect to MDL [2] until convergence (40 iterations). Then we continue the optimisation with respect to MDL plus a curvature cost [9] until convergence (another 40 iterations).

Figure 3 shows the resulting correspondences on the part of the shapes corresponding to the eye. The plots show landmark 25 to 40. Anatomically this shows the end of the forehead and the beginning of the nose of a person looking to the left. The nose begins approximately at landmark 34 in the bottom row. The correspondences are clearly better when using curvature. Other parts of the curves are similar. The top of

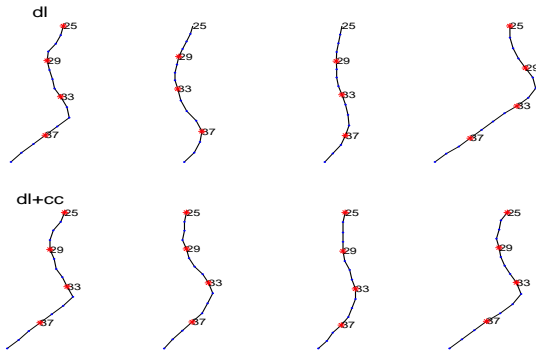


Figure 3: Corresponding landmarks on parts of silhouettes.

Figure 4 shows how gcm decreases when curvature is added. The middle plot shows how DL first decreases as it is minimised, but then when DL + curvature cost is minimised in the second part DL increases. So gcm captures an improvement in correspondences that DL misses. In Figure 4 it can be seen how the measures of generality, compactness and specificity all indicate that the model without curvature cost is better. But as seen above the correspondences improve when using curvature.

This experiment shows that gcm captures an im-

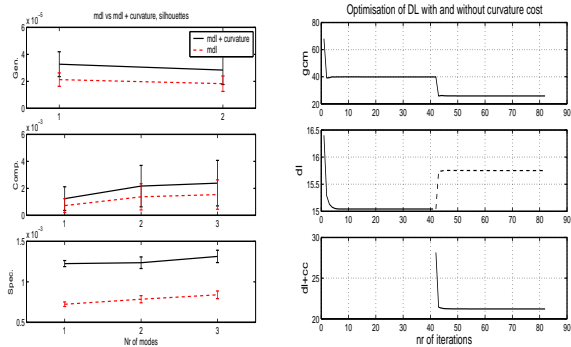


Figure 4: To the left generalisation, compactness and specificity of the DL silhouette model and the DL + curvature cost silhouette model. To the right optimisation of DL and DL + curvature cost for 4 silhouettes.

provement in correspondences that generality, compactness, specificity and DL misses.

4 Benchmarking using gcm

First a database of eight shape classes was built. The first five shape classes (sharks, birds, flightbirds, rats, forks) are curves from the database used in [8] and each of these shape classes consists of 13-23 shapes. The sixth and eight - shape classes are the silhouette- and boxbump-shapes from the previous section. The seventh shape class is a hand segmented from a video sequence. The boxbumps are synthetic with 61 ground truth points on each shape. The ground truth correspondences on the first seven data sets were (10 to 21 landmarks) manually marked. This database together with code can be downloaded from the authors web site.

The following algorithms have been benchmarked using gcm: Arlength, MDL [2], MDL with curvature (MDL+Cur) [9], MDL with parameterisation invariant scalar product (MDL+Par) [6], aias [3], aias with MDL, reparameterisation by minimising Euclidean distance (eucl), reparameterisation by minimising Euclidean and Curvature distance (eucl+cur). Also handmade models of all the datasets were built by a different person than the one marking ground truth. This was done without knowing which anatomical points were used as ground truth.

All tests were performed with 128 landmarks, 40 iterations and 7 reparameterisation control nodes for all the automatic algorithms. Table 1 shows the remaining percentage of gcm after optimising from arlength (100% means equal gcm as arlength parameterisation and 0% means perfect correspondences according to gcm).

MDL with a curvature cost added comes out as the winning algorithm. There is no algorithm that is best on all datasets and no algorithm gives as good correspondences as the correspondences marked manually.

Ground truth points have been marked for the shark

Algorithm	sharks	birds	flightbird	rats	forks	silhouettes	hands	boxbumps	mean	median
aias+mdl	25.5 ¹	80.6 ¹	62.6 ⁵	28.0 ²	22.7 ⁴	42.9 ²	21.6 ²	75.9 ³	45.0 ³	35.5 ³
MDL+Cur	26.1 ²	88.1 ⁴	58.3 ²	29.0 ⁴	22.7 ³	32.5 ¹	19.4 ¹	25.2 ¹	37.7 ¹	27.5 ¹
MDL	32.9 ⁵	90.1 ⁶	66.2 ⁶	28.6 ³	22.4 ²	46.2 ⁴	22.3 ³	30.0 ²	42.3 ²	31.4 ²
MDL+Par	29.4 ⁴	89.6 ⁵	62.4 ⁴	27.3 ¹	22.3 ¹	42.9 ³	22.3 ⁴	76.6 ⁴	46.6 ⁴	36.2 ⁴
eucl	55.9 ⁶	85.7 ²	59.5 ³	30.8 ⁵	25.2 ⁵	126.2 ⁶	34.7 ⁶	118.2 ⁵	67.0 ⁶	57.7 ⁶
eucl+cur	27.7 ³	86.6 ³	56.7 ¹	33.0 ⁶	28.8 ⁶	90.0 ⁵	22.6 ⁵	118.8 ⁶	58.0 ⁵	44.9 ⁵
semiauto	22.8	76.1	50.7	24.5	21.1	26.0	17.1	16.5	31.8	23.6
handmade	18.5	73.2	27.6	13.1	11.1	21.1	14.9	0.0	22.4	16.7

Table 1: Percentage of gcm error left after optimising from arclength. Upper index indicates rank.

dataset by 11 people. The geodesic Mahalanobis gcm was calculated for the algorithms in the same order as Table 1 and the percentage left were: 25.4, 22.2, 27.7, 25.3, 36.4, 48.4, 24.5 and 19.8. MDL+Cur comes out as the winner of the automatic methods but the handmade is best.

For the winning algorithm gcm was then used to pick optimal parameter values, such as number of landmarks and number of parameterisation nodes.

Semi-Automatic algorithm Since handmade models are best, a semi-automatic algorithm was tested. Five shapes were manually marked and then kept fixed, while the rest of the shapes in the dataset were optimised one by one using DL with curvature cost to fit the five fixed shapes. This results in an algorithm better than all the automatic algorithms, see Table 1. Seven control nodes were used for all natural datasets but for the synthetic boxbumps 15 nodes were used. Experiments with 15 nodes for the automatic algorithms results in worse correspondences for all algorithms except eucl and eucl+cur where only slightly better results were obtained.

5 Summary and Conclusions

For evaluation of correspondences located automatically there has formerly been a number of standard methods. In this paper it is shown that these methods have severe weaknesses. The new measure called Ground truth Correspondence Measure (gcm) is proposed. It is shown in two experiments on two different datasets that this measure corresponds well to subjective evaluation, whereas the standard measures do not. In this paper several state of the art algorithms were benchmarked using gcm. In Table 1 it can be seen that mdl with curvature added comes out as the winner of the automatic algorithms. There is no algorithm that is best on all datasets and no algorithm gives as good correspondences as the correspondences marked manually. The semi-automatic algorithm is better than the automatic on all datasets. In previous work it is often claimed that automatic algorithms give better models than models built by hand. These claims are often supported by measures like generality, specificity and compactness. In this paper these measures are shown to be flawed and by measuring gcm it is concluded that

models built by hand are actually very good. In some cases it may not be reasonable to manually mark the full dataset but, as seen, a semi-automatic approach, where only five shapes need to be manually marked, works very well.

References

- [1] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(24):509–522, 2002.
- [2] R. Davies. *Learning Shape: Optimal Models for Analysing Natural Variability*. PhD thesis, University of Manchester, 2002.
- [3] A. Ericsson. Automatic shape modelling and applications in medical imaging. Technical report, Centre for Mathematical Sciences, Box 118, SE-22100, Lund, Sweden, nov 2003.
- [4] A. Hill and C.J. Taylor. Automatic landmark generation for point distribution models. In *Proc. British Machine Vision Conference*, pages 429–438, 1994.
- [5] A. Hill and C.J. Taylor. A framework for automatic landmark identification using a new method of nonrigid correspondence. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22:241–251, 2000.
- [6] J. Karlsson, A. Ericsson, and K. Åström. Parameterisation invariant statistical shape models. In *Proc. International Conference on Pattern Recognition, Cambridge, UK*, 2004.
- [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] T. Sebastian, P. Klein, and B. Kimia. Constructing 2d curve atlases. In *IEEE Workshop on Mathematical Methods in Biomedical Image Analysis*, pages 70–77, 2000.
- [9] H. H. Thodberg and H. Olafsdottir. Adding curvature to minimum description length shape models. In *Proc. British Machine Vision Conference*, 2003.