Benchmarking of algorithms for automatic correspondence localisation

Anders Ericsson and Johan Karlsson
Centre for Mathematical Sciences, Lund University, Sweden

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 focused on evaluating the quality of the models produced by the algorithms by measuring compactness, generality and specificity. In this paper severe problems with these standard measures are discussed. We propose that a ground truth correspondence measure (gcm) is used for benchmarking and in this paper benchmarking is performed on several state of the art algorithms. Minimum Description Length (MDL) with a curvature cost comes out as the winner of the automatic methods. Hand marked models turn out to be best but a semi-automatic method is shown to lie in between the best automatic method and the hand built models in performance.

1 Introduction

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 locate parameterisations of the shapes in the training set to get correspondences between the shapes.

Attempts have been made to locate correspondences on shapes using shape features, such as high curvature [10]. Many have stated the correspondence problem as an optimisation problem [1, 2, 6, 7, 8, 9, 12, 16]. Minimum Description Length, (MDL) [5], is a paradigm that has been used in many different applications, often in connection with model optimisation. In recent papers [5, 11] this paradigm is used to locate a dense correspondence between shapes.

There have been some recent interesting papers including code on the problem of matching one point set to another, [3, 18]. However these algorithms only match one shape to another, instead of working with the training set as a whole.

In short the field has matured and there are many algorithms available. So there is a need for benchmarking of these algorithms. In recent years a similar development has taken place in the field of stereo [14].

In order to evaluate these algorithms, different measures of the quality of the parameterisations and the resulting models have been used. If the model is to be used for a specific purpose, such as segmentation of the heart in scintigrams, the choice of algorithm should be made using a criterion based on the application. For a more general evaluation of shape model quality the standard measures are compactness, specificity and generality [4]. It is also common to evaluate correspondences subjectively by plotting the shapes with some corresponding points marked.

In [15] the quality of registrations of images is evaluated both by measuring the overlap with ground truth and by measuring model quality measures such as specificity and

BMVC 2006 doi:10.5244/C.20.78

generality of models constructed from the registered images. There it is claimed that ground truth correlates with generality and specificity. This is shown by doing random perturbations of ground truth and noting that all the measures increase monotonously. However, this does not show that the measures are minimal at the ground truth.

Measuring the sensitivity to perturbation of ground truth it is also claimed that specificity is more sensitive than the other two measures.

Instead of measuring sensitivity to random perturbation it would be more interesting to examine which measure is most suitable for choosing between two training sets produced by different non random strategies. This might be done by letting human experts choose which is the best of the two compared training sets.

In this paper problems with the standard general model quality measures, namely compactness, generality and specificity, are discussed. We show that especially specificity and compactness do not succeed in measuring what they attempt to measure. With practical experiments we also show that the standard measures do not correlate well with correspondence quality. Also these measures are not quantitative in the sense that they do not assign a single number to describe the quality of the shape model.

What should be considered as a shape model of high quality is highly dependent on the application. For example the model that performs best on segmentation might not be the model that performs best on classification.

The qualities that the standard measures attempt to measure are often, but certainly not always, important. However, even when they are important, as we will see, it is problematic to actually measure them. For most applications high quality of the correspondences is desirable. A shape model built from correct correspondences is a model that correctly describes the shape variation within the class, whereas, as will be shown, a simplified model can get excellent measure of specificity, generality and compactness, but relevant shape information may have been lost. We propose that a ground truth correspondence measure (gcm), measuring the quality of the correspondences at important locations, is used for measuring correspondence quality and that this is used for benchmarking.

The four major contributions in this paper are: (i) It is shown that former shape model measures have severe weaknesses. (ii) A correspondence quality measure (gcm) is proposed and it is shown that gcm together with a database of ground truth correspondences is well suited for benchmarking algorithms for correspondence localisation. (iii) A database of eight datasets consisting of a total of 28518 ground truth points for the natural datasets set by 25 people and matlab code for evaluating algorithms is published. (iv) Benchmarking of several state of the art algorithms is presented and MDL with a curvature cost comes out as the winner.

2 Measuring model quality

Generality. By generality is meant how good the model can generalise to formerly unseen shapes. The model should be able to describe all shapes of the class and not only those of the training set. This is measured by doing leave one out tests, where a model is built by using all but one of the shapes in the training set. This model tries to describe the left out shape. The error in [4] for one left out shape is the norm of the difference between the modeled shape and the true shape. Generality is measured as the mean over all left out shapes. Usually this is plotted over the number of modes used by the approximating model.

$$G(n_m) = \frac{1}{n_s} \sum_{j=1}^{n_s} ||\mathbf{x}_j - \mathbf{x}'_j(n_m)||^2 \ , \tag{1}$$

where $\mathbf{x}_j$ is the left out shape and $\mathbf{x}'_j(n_m)$ is the attempted description using the model with n_m modes. One problem with measuring generality is that the parameterisation of the left out shape is unknown. This is often solved by letting the shape be included in the correspondence localisation, but this leads to an underestimation of the error.

In [15], a different version of the generalisation measure is used. This version avoids the leave one out problem, but it does not measure generalisation ability to shapes outside the training set.

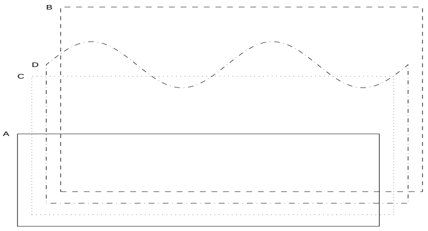

Figure 1: Problem with the specificity measure.

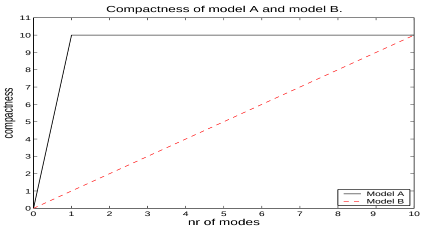

Figure 2: Problem with the compactness measure.

Specificity. A specific model can only represent shapes from the shape class for valid parameter values. This has been measured by generating a large amount (N) of shapes by picking random parameter values for the model according to the parameter space distribution. Each sample shape is then compared to the most similar shape in the training set. A quantitative measure for this [4] is:

$$S(n_m) = \frac{1}{N} \sum_{j=1}^{N} ||\mathbf{y}_j - \mathbf{y}'_j(n_m)||^2 \ , \tag{2}$$

where $\mathbf{y}'_j$ are shape examples generated by the model and $\mathbf{y}_j$ is the nearest member of the training set to $\mathbf{y}'_j$ and n_m is the number of modes used to create the samples.

To illustrate a problem, assume two models built from a training set consisting of rectangles as shape A and shape B in Figure 1. Model 1 generates samples as shape C and model 2 generates samples as shape D. These models will have equal specificity measure, since C and D have equal distance to the closest shape in the training set. However shape C belongs to the shape class, but shape D does not.

The standard specificity measure can give the same error for shapes that belong to the shape class (but are between two shapes of the training set) as it does for shapes that do not belong to the shape class.

Compactness. A compact model is a model that represents all shapes of the class with as little variance as possible in the shape variation modes and preferably with few modes. A measure of compactness is the sum of variances of the model [4],

$$C(n_m) = \sum_{j=1}^{n_m} \lambda_j \ , \tag{3}$$

where λ_j is the variance in shape mode j and $C(n_m)$ is the compactness using n_m number of modes. This measures the sum of variance of the modes. If the curve for one model is below or equal to the curve for another model for all n_m and lower for some n_m, the model represented by the lower curve is said to be more compact [4].

In Figure 2 the compactness functions of two models are plotted. The total variances of the two models are equal. For model A all variance is concentrated to the first mode. The graph for model A therefore goes up to the total variance in the first mode. For model B the variance is distributed equally over all modes. It is obvious that model A should be considered to be more compact than model B. However, using the criterion, described above, for choosing the more compact model, model B would be selected as the most compact model. Therefore we can conclude that this compactness measure is not suitable for selecting the most compact model.

Comparing quality of models In general if the curve for one model is below or equal to the curve for another model for all n_m and lower for some n_m, the model represented by the lower curve can be said to be more compact, specific or general depending on the measured quantity [4].

A problem when measuring generality, specificity and compactness is that if the curves ($G(n_m), S(n_m)$ or $C(n_m)$) for different models intersect it is not possible to choose which model is of higher quality.

3 Ground Truth Correspondence Measure

In order to measure the quality of the correspondences produced by an algorithm for automatic correspondence localisation, datasets with manually located landmarks and synthetic datasets with known corresponding points can be used. For synthetic examples these marks are exact but for manually placed marks there is of course a subjective element and also the introduction of a small error is inevitable.

Let the parameterisations γ_i of the shapes $\mathbf{x}_i$ be optimised by the algorithm that is to be evaluated. Then, for every shape $\mathbf{x}_i$ ($i = 1, \ldots, n_s$) together with its ground truth points $\mathbf{g}_{ij}$ ($j = 1, \ldots, n_g$), 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 calculated. Formally $t_{ij} = \gamma_i^{-1}(\mathbf{x}_i^{-1}(\mathbf{g}_{ij}))$. Now, for every shape $\mathbf{x}_k$ ($1 \leq k \leq n_s, k \neq i$) use the same parameter values t_{ij}. The points produced should be close to the ground truth points of this shape, if the parameterisation functions represent correspondences of high quality. That is, $\mathbf{x}_k(\gamma_k(t_{ij}))$ should be close to $\mathbf{g}_{kj}$. This is measured as the mean distance in the metric d over all shapes in the dataset.

$$gcm = \frac{1}{n_s(n_s-1)n_g} \sum_{i=1}^{n_s} \sum_{j=1}^{n_g} \sum_{k \in K} e_{ijk}$$

$$e_{ijk} = d\left(\mathbf{x}_k(\gamma_k(\gamma_i^{-1}(\mathbf{x}_i^{-1}(\mathbf{g}_{ij})))) \ , \ \mathbf{g}_{kj}\right) \quad ,$$

$$K = \{1, \ldots, i-1, i+1, \ldots, n_s\} \quad .$$

Finally we get,

$$gcm = \frac{1}{n_s(n_s-1)n_g} \sum_{i=1}^{n_s} \sum_{j=1}^{n_g} \sum_{k \in K} d\left(\mathbf{x}_k(\gamma_k(t_{ij})) \ , \ \mathbf{g}_{kj}\right) \quad ,$$

where t_{ij} is the parameterisation parameter value for the ground truth point j on shape i. The constant n_s is the number of shapes and n_g is the number of ground truth points. Any metric could be used, but in this paper $d(a,b)$ is the length of the shortest path between the point a and the point b along the shape, which has been normalised to have area one. Locally (usually also globally) the shortest path is a geodesic. Apart from the advantage of measuring the error along the shape, this also gives scale invariance. On curves the metric d corresponds to the arclength distance on curves normalised to length one.

Note that in case of intersections of the shape the path is not allowed to use the intersection as a short cut. Think of the letter α for an example of a shape with an intersection.

In [13] a similar measure is used to evaluate correspondences on registered surfaces. However, there the distance between the corresponding points is measured as the Euclidean distance. Also, the distance is measured between points on a deformed template and points on the target surface whereas here we focus on groupwise correspondence.

Due to the subjective nature of choosing ground truth points on natural shapes, statistics about the ground truth points could be taken into account. Let a number of people mark ground truth points on the same dataset. Means and variances can then be calculated and the norm used to calculate gcm can then be the Mahalanobis normalised distance,

$$gcm = \frac{1}{n_s(n_s-1)n_g} \sum_{i=1}^{n_s} \sum_{j=1}^{n_g} \sum_{k \in K} \frac{d\left(\mathbf{x}_k(\gamma_k(t_{ij}))\ ,\ \bar{\mathbf{g}}_{kj}\right)}{\sigma_{kj}} \quad ,$$

$$K = \{1, \ldots, i-1, i+1, \ldots, n_s\} \quad ,$$

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

4 Experimental Validation of gcm

The first experiment was to start from correct correspondences and then optimise the parameterisations 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, were 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 3.

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 5 it can be seen that minimising the description length from true correspondences has resulted in worse correspondences.

In Figure 4 the compactness and specificity measures indicate that the optimised model has higher quality. In this case we started from ground truth and as can be seen in Figure 5 the correspondences are worse for the optimised model. So we can conclude that compactness and specificity do not correlate with correct correspondences. The problems with compactness were already noted in [4]. In Figure 6 the specificity is evaluated visually. The upper model is built from ground truth correspondences and the lower model is built from DL-optimised correspondences. The plots show the mean shape plus three standard deviations of the first two shape modes. Since the data in this case only has one shape mode it is enough to examine the first two modes as in this plot. The model built from ground truth correspondences is clearly more specific than the DL-optimised model. The slight distortion in model 1 is due to the alignment of the shapes. The Procrustes

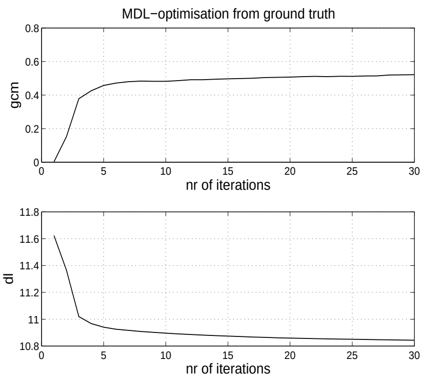

Figure 3: Gcm and DL plotted over number of iterations.

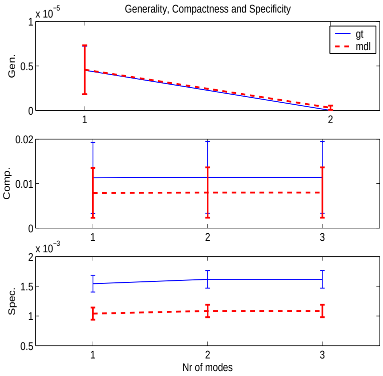

Figure 4: The standard measures of the ground truth and the dl box bump model.

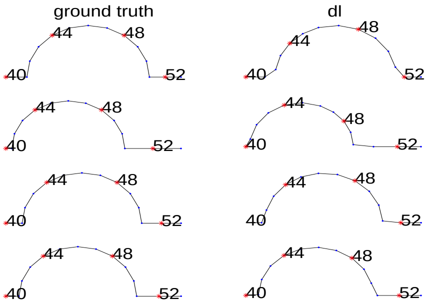

Figure 5: Ground truth and dl correspondences. The figure shows the bump part of the shapes.

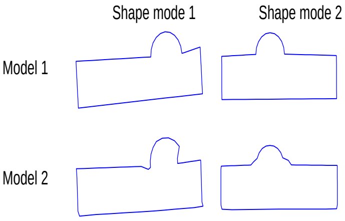

Figure 6: Mean shape +3 std of the first two shape modes for models built from ground truth correspondences (Model 1) and DL-optimised correspondences (Model 2).

alignment introduces nonlinearities in the shape modes. From Figure 6 it can be concluded that the model built from better correspondences has better specificity, contrary to the conclusion of the plot of 2 in Figure 4. So also in practice the specificity measure defined by 2 does not work.

Summing up this experiment, the conclusion is that although minimising the description length is a good method for finding approximate correspondences, in this case it fails to locate the correct correspondences and the specificity measure fails in practice. In [4] a similar experiment shows that MDL does succeed in finding optimal correspondences on the synthetic box bump data set. However, in this experiment the important Procrustes alignment step is skipped and the shapes are given perfect shape alignment.

In the second experiment silhouette shapes (22 contours of face silhouettes) initialised with arclength parameterisation were used. We optimise the parameterisation with respect to MDL [5] until convergence (40 iterations). Then we continue the optimisation with respect to MDL plus a curvature cost [17] until convergence (another 40 iterations).

Figure 7 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 9 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 8 it can be seen how the measures of generality, compactness and specificity all indicate that the model without curvature cost has higher quality. So the standard measures can not be used to measure correspondence quality.

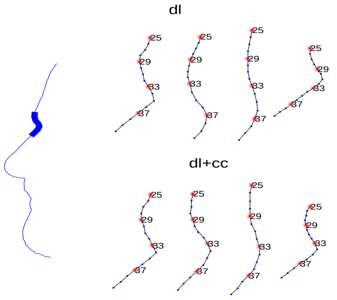

Figure 7: Corresponding landmarks on parts of silhouettes.

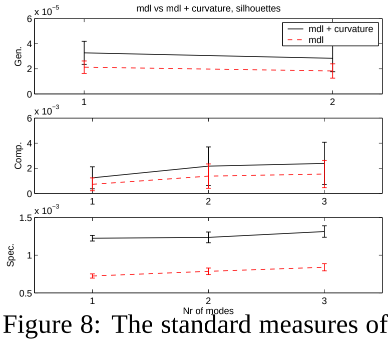

Figure 8: The standard measures of the DL and DL + Cur silhouette model.

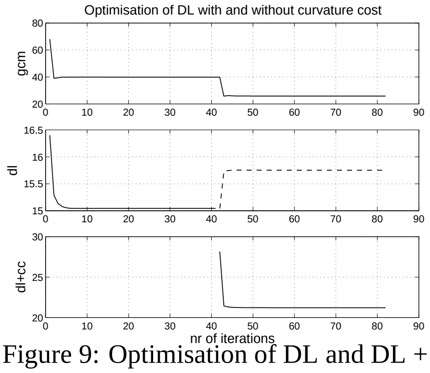

Figure 9: Optimisation of DL and DL + Cur for 4 silhouettes.

This experiment shows that gcm captures an improvement in correspondences that generality, compactness, specificity and DL all miss. This indicates that they can not be used for evaluating correspondence quality.

As an extreme example it is possible to get perfect compactness, specificity and generality (zero) by placing all landmarks in one point on all shapes.

5 Benchmarking using gcm

First a database of eight shape classes was built. The first five shape classes (sharks, birds, flight birds, rats, forks) are curves generated from images in the database used in [16] and each of these shape classes consists of 13-23 shapes with ground truth correspondences (from 10 to 21 landmarks) manually marked. The sixth and eighth shape classes are the silhouette, and boxbump shapes from the previous section. The seventh dataset consists of 23 contours of a hand.

Algorithm	sharks	birds	flightbirds	rats	forks	silhouettes	hands	mean	median	*boxbumps**
arclength (gcm)	15.55	22.88	7.12	13.37	19.11	8.93	14.00	14.42	14.00	–
MDL (%)	27.21^7	65.17^6	56.33^6	29.03^7	19.41^3	46.23^6	17.60^8	37.28^6	29.03^7	29.20^3
MDL+Cur (%)	$\mathbf{21.51^1}$	79.93^8	$\mathbf{45.02^1}$	$\mathbf{27.02^1}$	23.26^6	22.54^2	14.80^4	33.44^3	$\mathbf{23.26^1}$	$\mathbf{17.42^1}$
MDL+me (%)	29.26^9	91.67^9	61.80^{10}	30.10^8	19.83^4	47.76^8	16.81^7	42.46^8	30.10^8	50.20^5
MDL+nodecost (%)	26.13^6	66.71^7	56.34^7	28.90^6	19.21^2	47.37^7	16.24^6	37.27^5	28.90^6	30.97^4
MDL+Par (%)	24.40^5	62.40^5	47.94^4	27.80^3	$\mathbf{18.28^1}$	39.88^5	14.17^2	33.55^4	27.80^4	85.80^7
aias+mdl	21.93^3	$\mathbf{23.38^1}$	57.69^8	28.37^5	20.37^5	38.81^4	14.57^3	29.30^2	23.38^2	75.93^6
aias+mdl+cur (%)	21.61^2	23.73^2	47.73^3	27.42^2	23.92^7	26.12^3	$\mathbf{14.17^1}$	$\mathbf{26.38^1}$	23.92^3	111.00^8
eucl (%)	43.76^{10}	60.02^4	58.85^9	36.92^{10}	27.32^8	129.74^{10}	26.39^{10}	54.71^{10}	43.76^{10}	116.82^9
eucl+cur (%)	28.84^8	55.18^3	53.54^5	35.03^9	28.59^9	103.57^9	18.26^9	46.14^9	35.03^9	118.88^{10}
cur (%)	22.01^4	111.16^{10}	45.82^2	27.96^4	31.19^{10}	$\mathbf{21.12^1}$	15.84^5	39.30^7	27.96^5	17.60^2
semiauto (%)	20.31	20.40	47.68	24.58	14.32	16.67	9.27	21.89	20.31	16.52
handmade (%)	17.94	8.84	14.24	10.44	7.33	9.53	6.67	10.71	9.53	0.00

Table 1: The second row in the table shows the gcm for arclength parameterisation of the different datasets. The following rows shows the percentage of gcm error (Mahalanobis normalised) left after optimising from arclength parameterisation. The upper index indicates the rank of this algorithm on this dataset. The winning algorithm on each dataset is bold. Since the boxbumps (*) are so easy to mark, only one person has marked this dataset and the gcm without Mahalanobis has been used.

All the natural examples have been marked by 18-25 people. In total the database consist of 28518 (not including the synthetic dataset) ground truth landmarks manually set. The boxbumps are synthetic with a total of 1464 ground truth points on 24 shapes. This database together with code can be downloaded from the authors web site.

The following algorithms have been benchmarked using gcm:

arclength: All the landmarks are placed with equal arclength distance from each other.

MDL, MDL+Cur: The parameterisations are optimised so as to minimise the description length (plus a curvature cost in the second case) of the model and the dataset [5, 17].

MDL+me, MDL+nodecost, MDL+Par: These are techniques aimed at avoiding clustering of landmarks [5, 11, 17].

AIAS+MDL, AIAS+MDL+Cur: The description length with and without curvature cost is minimised for an Affine Invariant Active Shape model [7].

Eucl, Eucl+Cur: The parameterisations are optimised to minimise the Euclidean distance (plus a distance in curvature difference in the second case) between corresponding points on all the shapes.

Cur: The curvature cost used in the algorithms above is here used by itself as a cost function to be minimised.

Handmade: Also handmade models of all the datasets were built by a different person than the ones 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.

Table 1 shows the remaining percentage of gcm (with Mahalanobis) after optimising from arclength (100% means equal gcm as when using arclength parameterisation and 0% means perfect correspondences according to gcm).

AIAS+MDL+Cur is the algorithm that is best in mean. This algorithm succeeds especially well on the bird dataset. MDL+Cur has the lowest median result. Since the median is a more stable measure and since MDL+Cur is the best algorithm on three natural datasets and also performs best on the synthetic dataset, MDL+Cur is selected as the winning algorithm. There is no algorithm that is best on all datasets and no algorithm gives as good correspondences as the correspondences manually marked.

For the winning algorithm gcm was then used to pick optimal parameter values, such as number of landmarks and number of parameterisation nodes by evaluating gcm on the shark dataset. The algorithm was then run with these parameters on all datasets, which resulted in an even better 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.

6 Summary and Conclusions

For evaluation of the quality of shape models built from correspondences located automatically there have formerly been a number of standard methods. In this paper it is shown that these methods have severe weaknesses. We propose a Ground truth Correspondence Measure (gcm) for the evaluation of correspondence quality to be used in benchmarking. It is shown in experiments on different datasets that this measure corresponds well to subjective evaluation of correspondence quality, whereas the standard measures do not. It is also shown that optimising correspondences using description length initiated with correct correspondences can result in worse correspondences.

In this paper several state of the art algorithms are benchmarked using gcm. In Table 1 it can be seen that in median MDL+Cur is the best algorithm and it is also best on the synthetic dataset. 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 but the flightbird dataset.

In previous work it is often claimed that automatic algorithms give better correspondences than models built by hand. These claims are often supported by measures like generality, specificity and compactness. In this paper problems with these measures are discussed and it is shown that they do not correlate with correspondence quality. Measuring gcm, it is concluded that models carefully 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.

As a final note, of course it would be desirable to have ground truth free measures, but since the measures available all have severe problems it is our view that ground truth based measures must be used for evaluation of shape models.

7 Acknowledgments

We like to thank H. Thodberg (IMM, DTU) for the silhouettes and B. Kimia et.al. for the images of the sharks, birds, flight birds, rats and forks.

This work has been funded by the Swedish Foundation for Strategic Research (SSF) through the program 'Spatial Statistics and Image Analysis for Environment and Medicine' and the program 'Vision in Cognitive Systems' (VISCOS) and by UMAS and the Swedish Knowledge Foundation through the Industrial PhD program in Medical Bioinformatics at the Centre for Medical Innovations (CMI) at the Karolinska Institute.

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] F.L. Bookstein. Landmark methods for forms without landmarks: Morphometrics of group differences in outline shape. *Medical Image Analysis*, 3:225–243, 1999.

[3] Haili Chui and Anand Rangarajan. A new algorithm for non-rigid point matching. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume volume II, pages 44–51, 2000.

[4] R. Davies. *Learning Shape: Optimal Models for Analysing Natural Variability*. PhD thesis, University of Manchester, 2002.

[5] 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.

[6] A. Ericsson. Automatic shape modelling and applications in medical imaging. Technical report, Mathematics LTH, Centre for Mathematical Sciences, Box 118, SE-22100, Lund, Sweden, nov 2003.

[7] Anders Ericsson and Kalle Åström. An affine invariant deformable shape representation for general curves. In *Proc. 9th Int. Conf. on Computer Vision, Nice, France*, pages 1142–1149, Nice, France, 2003.

[8] A. Hill and C.J. Taylor. Automatic landmark generation for point distribution models. In *Proc. British Machine Vision Conference*, pages 429–438, 1994.

[9] A. Hill and C.J. Taylor. A framework for automatic landmark indentification using a new method of nonrigid correspondence. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22:241–251, 2000.

[10] C. Kambhamettu and D.B. Goldgof. Points correspondences recovery in non-rigid motion. In *Proc. Conf. Computer Vision and Pattern Recognition, CVPR'92*, pages 222–237, 1992.

[11] J. Karlsson, A. Ericsson, and K. Åström. Parameterisation invariant statistical shape models. In *Proc. International Conference on Pattern Recognition, Cambridge, UK*, 2004.

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

[13] Z. Mao, X. Ju, J.P. Siebert, W.P. Cockshott, and A.F. Ayoub. Constructing dense correspondences for the analysis of 3d facial morphology. *Pattern Recognition Letters*, 27(6):597–608, 15 April 2006.

[14] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1/2/3):7–42, 2002.

[15] R. Schestowitz, C. Twining, T. Cootes, V. Petrović, C. Taylor, and B. Crum. Assessing the accuracy of non-rigid registration with and without ground truth. In *Proc. IEEE International Symposium on Biomedical Imaging*, 2006.

[16] 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.

[17] H. H. Thodberg and H. Olafsdottir. Adding curvature to minimum description length shape models. In *Proc. British Machine Vision Conference*, 2003.

[18] Yefeng Zheng and David Doermann. Robust point matching for non-rigid shapes: A relaxation labeling based approach. Technical report: Lamp-tr-117, University of Maryland, College Park, 2004.