Building Shape Models from Lousy Data

Marcel Lüthi, Thomas Albrecht, and Thomas Vetter

Department of Computer Science, University of Basel, Switzerland
{marcel.luethi, thomas.albrecht, thomas.vetter}@unibas.ch

Abstract. Statistical shape models have gained widespread use in medical image analysis. In order for such models to be statistically meaningful, a large number of data sets have to be included. The number of available data sets is usually limited and often the data is corrupted by imaging artifacts or missing information. We propose a method for building a statistical shape model from such "lousy" data sets. The method works by identifying the corrupted parts of a shape as statistical outliers and excluding these parts from the model. Only the parts of a shape that were identified as outliers are discarded, while all the intact parts are included in the model. The model building is then performed using the EM algorithm for probabilistic principal component analysis, which allows for a principled way to handle missing data. Our experiments on 2D synthetic and real 3D medical data sets confirm the feasibility of the approach. We show that it yields superior models compared to approaches using robust statistics, which only downweight the influence of outliers.

1 Introduction

Statistical shape models have become a widely used tool in medical image analysis, computer vision, and computer graphics. From a technical point of view, the methods for model building are well established. The first and most challenging step is to establish correspondence among the examples. Once the shapes are in correspondence, each shape is regarded as a random observation, and standard methods from statistics can be applied. In practice, however, building statistically representative models is much more difficult. Often, acquiring a large enough data set of sufficient quality constitutes the most difficult step. This is especially true in the medical domain, where the image acquisition process is tailored to the physician's needs and to minimize harm for the patient. The data available to the researcher is therefore often noisy, incomplete, and contains artifacts.

In this paper we propose a method for building statistical shape models from data sets which can include incomplete and corrupted shapes. The main motivation for our work comes from a project involving the construction of a statistical model of the human skull from CT data. In many scans, teeth are missing completely or contain dental fillings resulting in severe metal artifacts. Others show only the region of the skull that was used to diagnose a certain pathology. As is often the case with medical data, some skulls show severe pathologies which should not be included in a model representing the normal anatomy.

To be able to build statistically representative models, we need to make sure that the corrupted parts do not distort the space of possible shapes the model can represent. Our

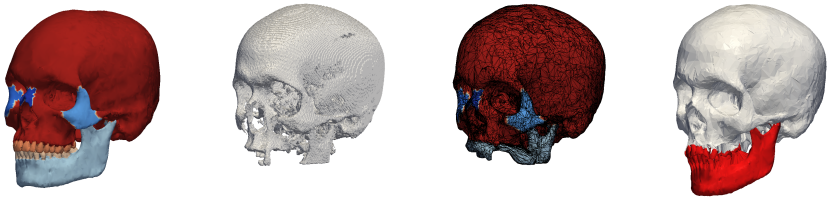
approach identifies the corrupted parts as statistical outliers, and excludes them from the model building. This is done by dividing the shapes into parts, and checking for each part individually whether it is corrupted. During model building, the best reconstruction of the corrupted parts is estimated from the remaining data sets. This is achieved in a statistically sound way using the EM algorithm for Probabilistic PCA. Performing the outlier analysis part-wise has two advantages: In statistical shape modeling, the observations are usually high-dimensional objects which can naturally be decomposed into smaller structures. Rather than throwing away all the information, we still use the parts that are intact to learn the shape variability of these sub-structures. More importantly, however, looking for outliers on individual parts makes it possible to detect small, local outliers, which would remain unrecognized if the shape was analyzed as a whole.

In our approach missing data and artifacts are just different instances of statistical outliers. There are two main approaches for dealing with outliers. *Outlier identification* can be performed to identify corrupted samples and exclude them from the data set. Methods for identifying outliers are well known in statistics [1]. Most of the traditional methods, however, consider only the case in which the number of examples is much larger than the dimensionality of the data. Such methods are not applicable to shape statistics. In recent years, outlier detection in high-dimensional data has been greatly advanced in the field of bio-informatics, where outlier-ridden data is the rule and not the exception [2, 3]. Another approach for dealing with outliers is to *robustify* the procedure, i.e. to adapt it such that outliers have less influence on the results. This can be achieved by using robust statistics [4] or by incorporating prior information [5, 6].

All steps of the workflow leading to a statistical shape model, from image denoising, segmentation, and registration to principal component analysis could benefit from being robustified [4]. In our method, however, we want corrupted parts to remain visible until the registration process has been performed. This makes it possible to detect and eliminate them completely. Therefore we do not robustify any of these pre-processing steps. Knowing that a part of a shape is an outlier allows us to choose an adequate strategy to deal with it.

2 Method

We first give a brief overview of our approach. Let a set of surfaces be given. We single out one surface as the reference shape, which we know to be complete and free of artifacts. This reference is segmented into parts as illustrated in Figure 1(a). While the method would work with arbitrary patches of reasonable size, we usually use anatomically significant parts for ease of interpretation. Before attempting any statistical analysis, we need to identify corresponding points in all the shapes. We assume that every target shape can be obtained by deforming the reference surface with a smooth vector field, which we find using a non-rigid registration algorithm. Figure 1(b) shows the result of warping the reference surface with such a vector field. We observe that both artifacts and missing data result in (locally) unnatural deformations. We aim at identifying these as statistical outliers. To do so, we rigidly align the individual parts of each shape to the corresponding part of the reference and apply an outlier identification algorithm to the locally aligned shapes. The parts that were identified as outliers are marked



(a) (b) (c) (d)

Fig. 1. Workflow of the outlier detection: 1(a) A reference surface is segmented into parts. 1(b) Some shape used for model building are incomplete or noisy. 1(c) The reference is warped to match the shape of the target. The missing parts lead to an unnatural deformation and can thus be identified as outliers. 1(d) The outlier parts are reconstructed from the remaining data.

as missing in the surface. The statistical model is built from these partial data sets using a PCA method that can handle missing data. We propose the use of the EM algorithm for probabilistic PCA (PPCA) [7, 8]. Figure 1(d) shows a reconstruction obtained by the PPCA algorithm.

In the following we provide the details of the methods we use in the individual steps of the workflow. However, the approach is general and does not depend on the particular registration or outlier identification algorithm used.

Registration To establish correspondence among the examples we use a registration algorithm based on Thirion's Demons algorithm [9]. Similar to the approach of Paragios et al. [10], we do not register the surfaces directly, but rather their distance images. After registration, each shape $\Gamma_i \subset \mathbb{R}^3$ $(i = 1, \ldots, n)$ can be represented as a warp of a reference surface Γ_1 with a vector field $\phi_i : \Gamma_1 \to \mathbb{R}^3$:

$$\Gamma_i = \{x + \phi_i(x) \,|\, x \in \Gamma_1\}. \tag{1}$$

The vector field ϕ_i can be used to transfer any discretisation of the reference Γ_1 to the shape Γ_i and thus allows us to treat the surfaces as discrete random observations (i.e. the surfaces become random vectors).

The parameter in the registration algorithm which controls the smoothness of the vector field is deliberately chosen to be small, in order to make the outliers visible and limit their influence on neighboring regions. In the case that smoother registration results are required for the final shape model, the registration can be run again after the outliers have been identified.

Procrustes Alignment The reference shape Γ_1 is partitioned into m parts, which we denote by Γ_1^j, $j = 1, \ldots, m$. Since the surfaces are in correspondence, the same partitioning is induced on all shapes. To perform outlier identification, we first align the individual parts of each shape to the corresponding part of the reference by Procrustes alignment. In this way only the shape of the part and not its position in space is considered in the outlier identification. As correspondence among the shapes has already

been established, the landmarks necessary for the Procrustes alignment only need to be labeled on the reference. These points can either be selected manually or by an automatic procedure. Let $x_k^j, k = 1, \ldots, N$ be the landmark points on the j-th part of the reference. To align the shapes, we find the rotation matrix $R \in \mathbb{R}^{3\times3}$, translation vector $t \in \mathbb{R}^3$ and scaling factor $s \in \mathbb{R}$ as:

$$(s, R, t) = \arg\min_{s,R,t} \frac{1}{N} \sum_{k=1}^{N} \|x_k^j - (sR(x_k^j + \phi_i(x_k^j) + t)\|^2. \tag{2}$$

The minimum of (2) admits a closed form solution and can be found efficiently (see Umeyama [11]).

Outlier Identification in High Dimensional Data The place in this workflow to identify and remove outliers is after the registration step, before they have a chance to corrupt the statistics, but after they have been brought into correspondence. We use the algorithm *PCOut*, proposed by Filzmoser et al. [2], which is especially designed for detecting outliers in high-dimensional spaces. As the method is quite intricate and its details are not critical for understanding our method, we only give a broad overview and refer the interested reader to the original paper [2].

The main idea of *PCOut* is to *robustly* build a PCA model and then identify those samples that do not fit well into this model. In order to build the robust PCA model, it suffices to robustly estimate the mean and covariance matrix. *PCOut* uses the robust estimators median and MAD (mean absolute deviation) to rescale the data, and performs a principal components analysis of this rescaled data. A weighting scheme using a robust kurtosis measure is used to identify the data sets that do not fit the PCA model well enough, according to a user-specified threshold. This value is referred to as the "outlier boundary".

Probabilistic Principal Component Analysis In the last step, the parts that were identified as outliers are marked as missing in the surface. There exist several methods for PCA that can deal with incomplete data [12]. One such algorithm, which is based on a sound probabilistic framework, is probabilistic PCA (PPCA) [7, 8]. Formulated in terms of the EM algorithm, PPCA can be seen as an iterative method, which simultaneously provides an estimation of the principal subspace and a reconstruction of the missing data given this subspace. It corresponds to the following generative model for an observation s:

$$s = Wx + \mu + \varepsilon. \tag{3}$$

That is, s is given as a linear mapping W of the latent variable $x \sim \mathcal{N}(0, \mathcal{I})$ plus the mean of the observation μ and some additive Gaussian noise $\varepsilon \sim \mathcal{N}(0, \sigma\mathcal{I})$. The mapping W can be found using an EM algorithm, which consists of the following steps:

$$\textbf{E-Step: } X = W^T W^{-1} W^T S \qquad \textbf{M-Step: } W^{\text{new}} = S^T X^T (X X^T)^{-1}.$$

Here, S is a matrix of all the observed data and X is the matrix of the latent variables x. Of most relevance for our work is that this EM algorithm enables us to extend the

E-Step such that missing data can be handled. To reconstruct the complete vector s from the incomplete data s^*, PPCA finds the unique pair x, s^* such that $\|Wx - s^*\|^2$ is minimized. The completed observation can be obtained explicitly by computing $s = Wx$ (i.e. s is the maximum a-posteriori reconstruction of $p(s|x)$). In each iteration of the algorithm the reconstruction is improved, as the current estimation of the subspace given by W becomes more accurate.

3 Results

We performed experiments on a synthetic data set of 2D hand contours and a 3D data set of human skull surfaces. Our implementation is solely based on freely available software packages. The registration algorithm is a variant of the Demons algorithm, as implemented in the Insight Toolkit [13]. The algorithms for outlier detection and PPCA are readily available as R packages [14, 2, 15]. The same parameter settings were used for all our experiments. To align the parts, we automatically determined 20 evenly distributed landmarks for each part. In all experiments we computed the first 10 principal components. While the individual algorithms have many parameters that could be tuned, our experiments showed that the given default values yield good results. Only the parameter *outlier boundary* for the algorithm *PCOut* critically influences the result (cf. section 2). We found a value of 0.45 to work well with all our data-sets.

For the first experiment we considered the case in which only the outlier framed in Figure 2 is present. Our algorithm successfully identifies the outlier and removes it from the analysis. The reconstruction computed by the PPCA algorithm is shown in Figure 3(c). Figure 3 clearly shows that in the presence of such outliers, standard PCA will fail. For comparison, we computed a robust PCA using the *PCAproj* algorithm as provided in the R package *pcapp* [16]. While the effect of the outlier is reduced, it still influences the model as illustrated in Figure 3(d). We performed a further experiment, now including all artifacts shown in Figure 2. Figure 4 shows the variation represented by the first two principal components. The variation in the data is captured well, without being influenced by the outliers. We observe that a cusp appears in the model. This may happen at the borders of a segment, when an outlier part is reconstructed using PPCA, but the model is not expressive enough to fit the remaining shape exactly. Table 1 shows a quantitative comparison of the different methods to a ground truth model, which is

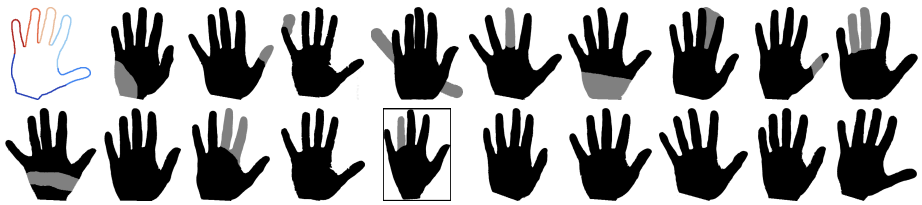
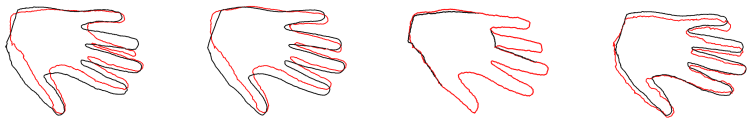


Fig. 2. The hand data set consisting of 19 hands. The hand is divided into 6 parts, as shown by the colors in the first shape. The grey area in the shape images shows manually introduced defects. The framed data-set marks the corrupted shape used in the first experiment.



(a) (b) (c) (d)

Fig. 3. Different methods for building shape models from noisy data: 3(a) The mean (black) and 2nd variation using standard PCA. 3(b) The mean and 2nd variation with our method. 3(c) The reconstruction from the PPCA algorithm (red) together with the corrupted shape. 3(d) A result from robust PCA: The outlier is still visible and leads to the thinning of the ring finger.

	mean μ	$\mu+\sigma_1$	$\mu-\sigma_1$	$\mu+\sigma_2$	$\mu-\sigma_2$	$\mu+\sigma_3$	$\mu-\sigma_3$
PCA	5.77	13.31	12.61	15.90	16.07	15.60	14.19
robust PCA (PCAproj)	5.45	7.12	8.05	6.45	7.91	8.37	8.54
outlier PPCA	1.90	6.09	4.62	6.72	6.30	5.88	5.96

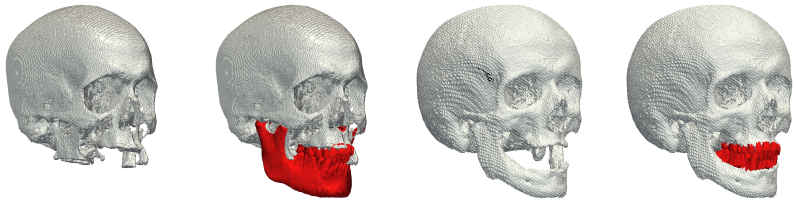
Table 1. Hausdorff distance (in mm) between the ground truth model and the models computed from data with outliers (σ_i stands for 1σ in the direction of the $i-$th principal component).

built from the data in Figure 2. We evaluated the Hausdorff distance between the mean and first three principal components of the ground truth, to the models computed with regular PCA, robust PCA, and our proposed method. Our method clearly gives the best approximation to the ground truth model.

We finally applied the algorithm to a data set of 23 human skulls. Some of the skull shapes in the data set are shown in Figure 5. As before the artifacts are detected as outliers and automatically reconstructed, as shown in the same figure. In this test, the method reaches its breaking point. As a common problem in skull data is that some or all of the teeth are missing, the reconstruction of the teeth looks slightly unnatural. This is due to the small number of examples in which the teeth are intact. However, in the final statistical model, this effect is only visible in the last few principle components. Further, as the parts are still identified as outliers, a different reconstruction strategy could be used, such as using a statistical model of the teeth. The comparison with robust PCA

mean $+2\sigma_1$ $-2\sigma_1$ $+2\sigma_2$ $-2\sigma_2$

Fig. 4. The first two principal components of the model. No artifacts are visible, despite the large number of artifacts in the data set. At segment boundaries, small discontinuities can appear (red circle), when the segment is reconstructed from limited data.



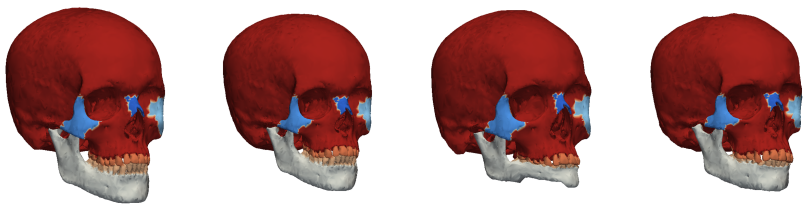
(a) (b) (c) (d)

Fig. 5. 3D Reconstruction 5(a) and 5(c) show two example surfaces from a skull data set. Their reconstruction is shown in 5(b) and 5(d) respectively.

given in Figure 6 leads again to the conclusion, that in the presence of large outliers, explicit outlier removal yields superior results than applying robust PCA.

4 Discussion

We presented an approach for building a statistical shape model in the presence of artifacts and missing data. The main idea is to divide the shapes into parts, and to perform outlier detection on each part individually. Once a part is identified as an outlier, it is removed from the data set. The remaining shape is still used to build the model. In this way, it becomes possible to build shape models from data sets in which almost every shape has some defect. Compared to robust approaches for model building, our method has the advantage that it does not only downweight the influence of an outlier, but eliminates it completely. Further, explicit identification of corrupted parts is useful, as it enables us to choose an adequate strategy to replace it. The strategy we presented here is to complete these parts implicitly during model building. Depending on the application, a different approach could be to either remove the parts completely from the analysis, or to perform a reconstruction using a dedicated shape model for this specific part. In general, some of the methods used in our workflow might not be suitable for



(a) (b) (c) (d)

Fig. 6. Models from real data. 6(a) and 6(b) show the mean and the first principal variation using our method. The outliers are clearly visible when using standard PCA (Figure 6(c)) and still influence the results of robust PCA (Figure 6(d)).

some applications. For instance, the rigid alignment removes the rotational component, and hence makes it impossible to detect rotational outliers. In the skull example, an open jaw is therefore not detected as an outlier. However, the only step that has to be changed in order to detect such cases is the local alignment. In this respect, the approach we presented here should be seen as a strategy to deal with "lousy" data sets rather than a ready-made method.

While we used anatomically significant parts to perform outlier identification, arbitrary surface patches could be used. How to choose these patches optimally is by itself an interesting problem, which will be the subject of future research.

Acknowledgements We thank Dr. Zdzislaw Krol and Dr. Stefan Zimmerer, University Hospital Basel, for their support and for providing us with the radiological data. This work was funded by the Swiss National Science Foundation in the scope of the NCCR CO-ME project 5005-66380.

References

1. Rousseeuw, P., Leroy, A., Wiley, J., InterScience, W.: Robust regression and outlier detection. Wiley New York (1987)
2. Filzmoser, P., Maronna, R., Werner, M.: Outlier identification in high dimensions. Computational Statistics and Data Analysis **52**(3) (2008) 1694–1711
3. Becker, C., Gather, U.: The largest nonidentifiable outlier: A comparison of multivariate simultaneous outlier identification rules. Computational Statistics and Data Analysis **36**(1) (2001) 119–127
4. De la Torre, F., Black, M.: Robust principal component analysis for computer vision. In: Intl. Conf. on Computer Vision (ICCV). Volume 1. (2001) 362–369
5. Cremers, D., Kohlberger, T., Schnorr, C.: Nonlinear shape statistics in mumford-shah based segmentation. Lecture Notes in Computer Science (2002) 93–108
6. Albrecht, T., Lüthi, M., Vetter, T.: A statistical deformation prior for non-rigid image and shape registration. In: IEEE Conference on Computer Vision and Pattern Recognition. (2008)
7. Tipping, M.E., Bishop, C.M.: Probabilistic principal component analysis. Journal of the Royal Statistical Society **61** (September 1999) 611–622
8. Roweis, S.: EM Algorithms for PCA and SPCA. Advances in neural information processing systems (NIPS) (1998) 626–632
9. Thirion, J.P.: Image matching as a diffusion process: an analogy with maxwell's demons. Medical Image Analysis **2**(3) (1998) 243–260
10. Paragios, N., Rousson, M., Ramesh, V.: Non-rigid registration using distance functions. Computer Vision and Image Understanding **89**(2-3) (2003) 142–165
11. Umeyama, S.: Least-squares estimation of transformation parameters between two point patterns. IEEE Trans. on Pattern Analysis and Machine Intelligence **13** (1991) 376–380
12. Little, R., Rubin, D.: Statistical analysis with missing data. Wiley (2002)
13. Ibanez, L., Schroeder, W., Ng, L., Cates, J.: The ITK Software Guide. Kitware, Inc. (2005)
14. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. (2008) ISBN 3-900051-07-0.
15. Stacklies, W., Redestig, H., Scholz, M., Walther, D., Selbig, J.: pcaMethods - a Bioconductor package providing PCA methods for incomplete data. Bioinformatics (March 2007)
16. Croux, C., Filzmoser, P., Oliveira, M.: Algorithms for Projection–Pursuit robust principal component analysis. Chemometrics and Intelligent Laboratory Systems **87**(2) (2007)