Robust Autonomous Model Learning from 2D and 3D Data Sets*

Georg Langs[1,2], René Donner[2,4], Philipp Peloschek[3], and Horst Bischof[2]

[1] GALEN Group, Laboratoire de Mathématiques Appliquées aux Systèmes, Ecole Centrale de Paris, France
[2] Institute for Computer Graphics and Vision, Graz University of Technology, Austria
[3] Department of Radiology, Medical University of Vienna, Austria
[4] Pattern Recognition and Image Processing Group, Vienna University of Technology, Austria
georg.langs@ecp.fr, donner@prip.tuwien.ac.at, philipp.peloschek@meduniwien.ac.at, bischof@icg.tugraz.at

Abstract. In this paper we propose a weakly supervised learning algorithm for appearance models based on the minimum description length (MDL) principle. From a set of training images or volumes depicting examples of an anatomical structure, correspondences for a set of landmarks are established by group-wise registration. The approach does not require any annotation. In contrast to existing methods no assumptions about the topology of the data are made, and the topology can change throughout the data set. Instead of a continuous representation of the volumes or images, only sparse finite sets of interest points are used to represent the examples during optimization. This enables the algorithm to efficiently use distinctive points, and to handle texture variations robustly. In contrast to standard elasticity based deformation constraints the MDL criterion accounts for systematic deformations typical for training sets stemming from medical image data. Experimental results are reported for five different 2D and 3D data sets.

1 Introduction

Model based approaches like active shape models (ASMs) or active appearance models (AAMs) [1] capture shape and texture variation of a specific structure or object. They utilize this a priori knowledge to provide robust segmentation while allowing for repeatable identification of specific landmarks in the data. They are employed in various medical imaging tasks, like the segmentation of the diaphragm in CT data [2], vertebral morphometry in dual x-ray absorptiometry data [3], and registration in functional heart imaging [4]. The necessity for a large number of manually annotated training examples in order to obtain a sufficiently representative power of the model poses a major drawback for model based approaches, since the annotation is time consuming and the results are often sub-optimal. The problem of automatic model building or equivalently that of establishing correspondences over landmark positions in a set of images has been tackled from different directions: in [5] temporal continuity of image sequences is

* This research has been supported by the Austrian Science Fund (FWF) under the grant P17083-N04 (AAMIR). It has been partially supported from the Region Île-de-France.

N. Ayache, S. Ourselin, A. Maeder (Eds.): MICCAI 2007, Part I, LNCS 4791, pp. 968–976, 2007.

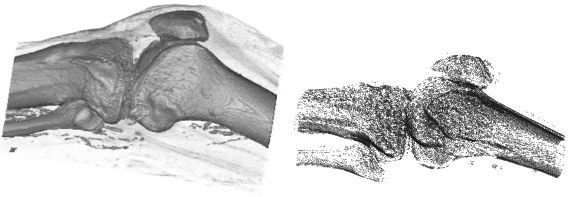

Fig. 1. Left: volume rendering of knee CT data. right: interest points on the bone structure.

used to determine correspondences. Given a set of manual continuous contour annotations in [6,7,8] landmarks are placed automatically along contours or surfaces that are mapped to a circle or a sphere using minimum description length (MDL). The reference manifold limits the approach to a topological class. Even-though these purely shape based approaches provide good landmark positions for constructing a compact shape model, in [9] the authors conclude that the lack of texture information poses a limitation hampering the capturing of *true* correspondences, like anatomical landmarks. In a line of work correspondences are established by one-to-many [10] or by group-wise registration of the entire images or volumes [11,12]. Non of these approaches can handle partially missing data. They are either dependent on a prior segmentation of objects, or deform the entire image continuously. In Fig. 1 on the left a surface rendering of the bones in knee CT data is shown. Such structures cannot be handled with a single reference manifold, and manual prior segmentation is tedious. A continuous deformation of the whole volume would not account for the compound structure, and would include large parts of soft tissue that deforms only loosely correlated with the bone surfaces. On the right the interest points, to which we restrict the calculation in this work, are depicted. Only local texture information is utilized giving sufficient information about the bone structure to perform model building.

Contribution. In this paper we propose a method to autonomously build appearance models based on group-wise registration of sparse representations of the training data. Instead of deforming dense texture maps we formulate the task as a search for correspondences between finite lists of interest points and local features in the training examples. This has several advantages: (1) the use of specific local features enables the algorithm to omit texture variations that yield no relevant information for the model, and to handle overlaps present in projective modalities, like x-ray. (2) The approach does not rely on a mapping to a reference manifold, therefore it is not constrained to an a priori topological class. (3) Occlusions and partially missing data sets are dealt with by outlier detection and robust model estimation. In contrast to purely shape based approaches local features add more specific information with regard to the correspondences of anatomical structures. These properties are relevant for complex anatomical structures that pose an obstacle to supervised model learning strategies, which would demand for an a priori definition of the topology, the structure or the connectivity constraints of the entity and a complete training set, i.e. no missing data. The approach is aimed at overcoming the necessity for the time consuming manual annotation prone to errors and variations in expert opinions.

2 Model Building

From a set of n training images or volumes $\mathbf{I}_i$, $i = 1, 2, \ldots, n$ depicting examples of a structure or an object, n sets of m_i interest points each is extracted. Initial correspondences for a random subset of k of these points are established by pairwise matching of a single reference image $\mathbf{I}_1$ to the remaining $n-1$ images. This results in correspondences for the k landmarks $\{l_1, \ldots, l_k\}$, which are encoded in a $k \times n$ matrix $\mathbf{G}$. Each column represents an image, and the entry $\mathbf{G}_{ji} \in \{1, \ldots, m_i\}$ with $j \in \{1, \ldots, k\}$ is the index of the interest point in image $\mathbf{I}_i$, at which the landmark l_j is positioned. Starting from these correspondences groupwise registration is performed by minimizing a criterion function that captures the compactness of the appearance model comprising the variation of landmark positions and local texture variation at the landmark positions in the different training images. The interest points in the images or volumes are treated as landmark candidates. Each point (i, q) with $q \in \{1, \ldots, m_i\}$ is assigned its coordinate information $\mathbf{p}(i, q)$ and local features $\mathbf{f}(i, q)$ (e.g. SIFT, steerable filters). By assigning $\mathbf{G}_{ji} = q$ the landmark l_j in image $\mathbf{I}_i$ has position $\mathbf{p}_{ij} = \mathbf{p}(i, q)$ and feature vector $\mathbf{f}_{ij} = \mathbf{f}(i, q)$. During model building the matrix $\mathbf{G}$ is modified to minimize the criterion function, resulting in *optimal* positions for each landmark in each image.

2.1 Criterion Based on Minimum Description Length

The criterion function that is minimized during model building comprises the compactness of the model that describes shape and local texture variation, and an elasticity regularization that is used during the initial phase of the optimization.

Compactness of the shape model. We use a standard linear multivariate Gaussian model for the shape representation [1]. The shape model compactness criterion is based on minimum description length, see [6] for an extensive derivation. An optimal shape model should minimize the cost L of communicating the model $\mathcal{M}$ itself and the data D (i.e. the landmark positions) encoded with the model: $L(D, \mathcal{M}) = L(\mathcal{M}) + L(D|\mathcal{M})$. Since we do not represent the entire image content but only a sparse set of landmarks and their variation a normalization term has to be introduced, that prohibits the landmarks from collapsing to a single position. The shape term is normalized by the entropy of the landmark positions in the individual examples $L_{ref} = \sum_{i=1,\ldots,N} entropy_{j=1,\ldots,k}(\mathbf{p}_{ij})$ and captures the gain of compactness achieved by the model in contrast to the complexity of the original data without exploiting its structure. The normalization is not essential to the quality of the model, but fosters the more even covering of the training images by the model landmarks. The final shape model criterion is $\mathcal{C}_S = L(\mathcal{M}_S) + L(D_S|\mathcal{M}_S) + \mathcal{R}_S - Lref$, where $L(\mathcal{M}_S)$ is the cost of communicating the shape model, $L(D_S|\mathcal{M}_S)$ is the cost of the shape data encoded with help if the model, and $\mathcal{R}_S$ is a penalty for the residual error not captured by the model.

Local texture. The image content is captured by local descriptors that extract features at the landmark candidate positions in the training images (e.g. SIFT features). For a landmark l_j the component-wise median of the individual entries in the feature vectors $\mathbf{f}_{ij}$ for $i = 1, \ldots, n$ from the landmark positions in all training images is calculated

resulting in the center $\hat{\mathbf{f}}_j$. The local appearance is then modeled by a Gaussian centered at $\hat{\mathbf{f}}_j$, and the description length is utilized as criterion for the compactness of the feature model $\mathcal{M}_T$ analogously to the shape model. Hence the criterion is $\mathcal{C}_T = L(\mathcal{M}_T) + L(D_T|\mathcal{M}_T) + \mathcal{R}_T$, where $L(\mathcal{M}_T)$ is the cost for the model, $L(D_T|\mathcal{M}_T)$ the cost of the local features encoded with the model, and $\mathcal{R}_T$ a penalty for the residual error.

Elasticity regularization. Since at the beginning of the model building the model has poor generalization behavior an elasticity cost term is introduced to regularize the deformations during the early phase of the optimization. A standard elasticity term $C_E = |\nabla \mathbf{d}(\mathbf{x})|^2$, where d is the displacement of the landmark x throughout the training set, helps avoiding a degenerate model.

Criterion function The final criterion function encompasses the compactness of models for shape and local texture information and the elasticity regularization: $\mathcal{C} = \mathcal{C}_S + \mathcal{C}_T + \alpha(t)\mathcal{C}_E$. The weight $\alpha(t)$ controls the influence of the elasticity, and is gradually decreased to 0 during optimization, to ensure a final result depending on the model costs only. During model building the criterion function is minimized by altering the matrix $\mathbf{G}$ that holds the correspondences between the landmarks and the interest points in the training images. A change of a single entry in $\mathbf{G}$ corresponds to a change of the position of a landmark in one image.

2.2 Dealing with Incomplete Data

Due to occlusions, irregular contrast or pose changes certain landmarks may not be present in all of the images. In order to still be able to build a model if parts of the data are missing the values are imputed based on previous estimates of the shape model $\mathcal{M}_S$, which is then re-estimated from the completed data vectors. The algorithm starts with estimates of the positions of the landmarks in all images i.e. no landmark is reported as missing. For a missing landmark l_j in image $\mathbf{I}_i$ the corresponding value in a matrix $\mathbf{R} \in [0,1]^{k \times n}$ is set to $\mathbf{R}_{ji} = 0$ and $\mathbf{R}_{ji} = 1$ if the landmark is present. During model building the algorithm decides at each iteration for each landmark in an image whether it is to be considered an outlier or not. The underlying idea is to perform an expectation maximization (EM) algorithm on the incomplete data set, by iteratively re-estimating mean and covariance matrix of the data [13]. Each of the shape vectors $\mathbf{x}_i$ is partitioned into $\mathbf{x}_i^a \in \mathbb{R}^{p_a}$, the vector of p_a values that are available for this particular image and $\mathbf{x}_i^m \in \mathbb{R}^{p_m}$ the vector of p_m values that are missing. Accordingly the mean is partitioned into $\bar{\mathbf{x}}_a$ and $\bar{\mathbf{x}}_m$. The relationship between available and missing records is modeled by a linear regression model $\mathbf{x}_m = \bar{\mathbf{x}}_m + (\mathbf{x}_a - \bar{\mathbf{x}}_a)\mathbf{B} + \mathbf{e}$, where $\mathbf{e}$ is the residual error. The regression matrix $\mathbf{B}$ is based on estimates of the mean $\hat{\bar{\mathbf{x}}}$ and covariance matrix $\hat{\Sigma}$ of the entire data set from the preceding model building iteration. The missing values of all shape vectors are imputed, and based on the completed data the mean $\hat{\bar{x}}$ and the covariance matrix $\hat{\Sigma}$ are re-estimated. See [14] for a concise explanation of imputation.

2.3 Optimization

Initialization For each image interest points and corresponding local features are extracted. The group-wise registration is initialized by a one-to-many registration of one

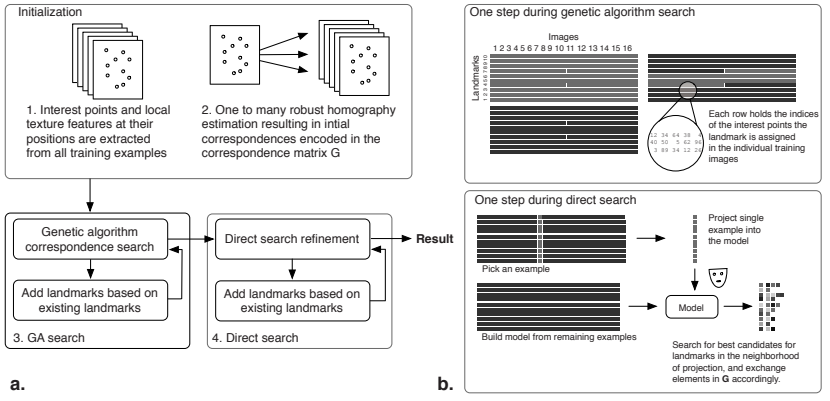

Fig. 2. a.: Scheme of the algorithm; b.: For the two ways of optimizing landmark correspondences, one step is shown: top: genetic algorithm, the correspondence matrix $\mathbf{G}$ serves as genome, two parent genomes and the resulting child are depicted. Below: direct search scheme, a model is build from all but one example, and the remaining example is changed to better fit the model.

of the images $\mathbf{I}_1$ to the remaining $n-1$ training images $\{\mathbf{I}_2, \ldots, \mathbf{I}_n\}$. In order to provide for a reliable initialization the feature vectors $\mathbf{f}_{1q}$ are matched to the feature vectors in each of the remaining images, and a transformation $\mathcal{H}_i$ (e.g. similarity transform) with a low number of parameters is estimated robustly by RANSAC. The correspondence matrix $\mathbf{G}$ is then initialized by choosing a small random set of k landmarks in $\mathbf{I}_1$ and propagating them across the images. In each image the interest points closest to the positions calculated by $\mathcal{H}_i$ are chosen as initial landmark positions, resulting in the initial correspondence matrix $\mathbf{G}_{init}$.

Optimizing the Criterion function. The algorithm is outlined in Fig. 2. After the coarse initialization of the correspondences the criterion function is minimized by updating the correspondence matrix $\mathbf{G}$. For an efficient optimization a neighborhood concept in the space of possible landmark positions has to be used. In [6] and [7] contours or surfaces of objects are mapped onto a circle or sphere. In contrast to this parameterization we employ k-D trees to efficiently search for candidates close to the current landmark position, while being independent from a parameterization reference. This enables the algorithm to adapt to complex and even changing topological configurations not defined a priori, like in three dimensional medical data where the behavior of anatomical structures needs to be modeled.

The optimization starts with a small number of landmarks (e.g. 10). After the learning process converges, additional landmarks are added to the model. This leads to an increasingly fine definition of the object. The new landmarks can either be chosen automatically by enforcing an even distribution, or they can be placed manually in a single reference frame. Their positions in the training set are estimated by interpolating the deformation field established by already existing landmark correspondences. Subsequently the entire larger set of landmarks is refined. The coarse deformations are learned by relatively few landmarks, and only after a good fit is achieved, fine local

details are modeled by a larger landmark set. Thereby a considerable speedup is achieved. We optimize the criterion function with two algorithms:

Genetic algorithm search. First the criterion is optimized by a genetic algorithm [15], with a set of correspondence matrices $\mathbf{G}$ serving as genomes of intermediate solutions (*individuals*), making a straight-forward implementation of mutations and cross-over functions possible (Fig. 2(b)).

Fine search. After the genetic algorithm converges a fine direct search is performed. It exploits two properties of the criterion to increase speed, and provide for robustness (Fig. 2(b)). During an iterative process a single example is chosen, a model is build from the remaining examples and the landmarks of the examples are projected into this model. The criterion function is calculated for interest points in the vicinity of the landmark position suggested by the model, and the landmark is moved to the position with lowest cost. This is similar to [11] but no parameterization of the landmark space is used. A search by evaluating the criterion function for small displacements of the current landmark positions is possible but results in far slower convergence. Landmark outliers are detected by comparing the current estimate for an image to the model built from the remaining images. If the cost for including the closest interest point to the landmark position estimate generated by the model is high, either due to its position or to its texture features, it is considered missing in this image. If the landmark is considered to be missing $\mathbf{R}_{ji}$ is set to 0 and its position is estimated from the model instead from an interest point candidate, as described in Sec. 2.2.

3 Experiments

Setup Evaluation results are reported for five data sets: **1.** 20 hand radiographs with a resolution of 0.34 mm / pixel, and semi-manual expert standard of reference annotations of 256 landmarks each on metacarpals and proximal phalanges. **2.** 20 randomly picked frames from a sequence of face images [16] with 576×720 pixel resolution and semi-manual ground truth annotation. The temporal coherence of the frames was not utilized. **3.** The same data as in 2. but with random occlusions covering up to 10% of the face. **4.** A synthetic data set of two 3D surfaces with an approximate diameter of 60 voxels consisting of a deformed torus and a sphere was generated using a single mode of deformation. To assess the capability of the algorithm to deal with 3D data independent of its topology, the topology of the setup changes throughout the data set. The model building was performed on dense sets of points on the surfaces. No texture information was used in this experiment, and correspondences were initialized with nearest neighbors, after the examples were centered and normalized w.r.t. to their standard deviation. **5.** 10 computed tomography data sets of the knee region (Fig. 1). For this data set no standard of reference was available and the model quality is assessed by means of the model compactness. For data sets 1.-4. the accuracy with regard to a semi-manual standard of reference annotation was assessed: the landmark correspondences define a deformation field between examples. One example was selected and the corresponding annotation was propagated to the other examples by piece wise affine interpolation according to the deformation field. The mean and median distance between propagated and standard of

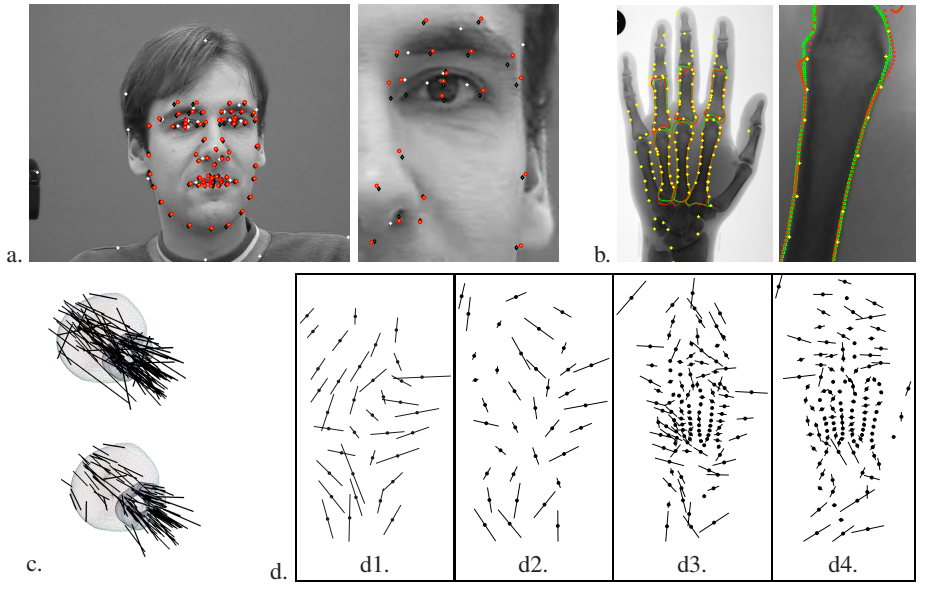

Fig. 3. a. Face data, and b. hand data: ground truth landmarks (green) and propagated landmarks, after autonomous model building (red circles), white crosses are the landmarks learned by the algorithm. c. 1^{st} shape mode for artificial 3D data, before and after optimization. d. 1^{st} and 2^{nd} shape model for hand data, d1. and d2. during early optimization, d3. and d4 after fine search. The positions of the bones correspond to b.

Table 1. Landmark deviation between propagated and standard of reference landmarks

Data	Hand	Hand cont.	3D	Face	F. occl.
Mean	5.84	2.27	2.14	10.10	13.08
Median	4.91	1.03	1.25	5.34	7.80

reference landmarks on the remaining images were recorded. It provides for a measure of how the model building captures the structure of the data.

Results In Fig. 3(a) and (b) the standard of reference (green) and the landmarks propagated according to the learned model (red) are depicted for face and hand data. White crosses show the positions of the landmarks learned automatically, used as control points for the landmark propagation. The accuracy of the resulting landmark correspondences is given in Tab. 1. The median error for the hand data is 4.91 pixels. Larger errors occur predominantly in regions where only few interest points are available because of low contrast. The error to the continuous standard of reference bone contours are reported, too, since salient features like the contours are modeled with high accuracy (mean error is 2.27 pixel). In this case a splitting of the model according to the separate bones can be expected to improve the landmark accuracy, since the variation in hand posture superposes the shape variation of individual bones [17]. For the face

data the landmark error increases if the images exhibit occlusions, however the moderate increase of the error indicates that the method can deal with incomplete data without resulting in a degenerate model. The percentage of landmarks reported missing was in the range of $8-10\%$. In Fig. 3(c) and (d) the modes of shape variation before and after optimization for sets (1) and (4) are depicted. The single mode that generates the data (4) is adequately captured by the resulting landmarks. For the hands the modes properly capture the aspect ratio change and the variation of finger positions. The results indicate that the approach is capable of generating reliable landmark correspondences which can replace semi-manual annotations, and that the resulting statistical model reflects the properties of the data. The criterion terms corresponding to shape and texture can be weighted according to the reliability of the image structure. This is subject of ongoing research. In [18] a similar evaluation was performed for registration of labeled magnetic resonance images, and the surface to surface distance after registration was in the range of 1.5 to 3.3 voxels. For the knee data (5) the compactness of the shape model increases significantly during optimization. After initialization 5 modes are necessary to represent 85% of the data variation, while after optimization 2 modes are sufficient.

4 Conclusion

In this paper we propose a method to autonomously learn appearance models. The algorithm does not need manual annotations, and establishes correspondences of landmarks by group-wise robust registration of a sparse set of interest points. No mapping onto a reference shape is used, and thereby a restriction to an a priori topological class is avoided. Instead of deforming the entire image local features are used to capture image content in an efficient way. In contrast to elasticity based registration techniques, the evolving shape model allows to deal with partially missing data in a natural way, by using the statistical properties of the training population. The results indicate that the resulting correspondences are good, that the approach produces compact models capturing the relevant information in the data, and that it has the potential to overcome the need for manual annotation. Future research will focus on the evaluation and the improvement of the method on a wider range of medical data, to model complex structures with minimal training effort and with the accuracy necessary for clinical application.

References

1. Cootes, T., Edwards, G.J., Taylor, C.J.: Active appearance models. IEEE TPAMI 23(6), 681–685 (2001)
2. Beichel, R., Gotschuli, G., Sorantin, E., Leberl, F., Sonka, M.: Diaphragm dome surface segmentation in CT data sets: A 3D active appearance model approach. In: Sonka, M., Fitzpatrick, J. (eds.) SPIE: Medical Imaging, vol. 4684, pp. 475–484 (2002)
3. Roberts, M., Cootes, T., Adams, J.: Vertebral shape: Automatic measurement with dynamically sequenced active appearance models. In: Proc. of MICCAI 2005 pp. 733–740 (2005)
4. Stegmann, M.B., Ólafsdóttir, H., Larsson, H.B.W.: Unsupervised motion-compensation of multi-slice cardiac perfusion MRI. Medical Image Analysis 9(4), 394–410 (2005)
5. Walker, K., Cootes, T., Taylor, C.: Automatically building appearance models from image sequences using salient features. IVC 20(5), 435–440 (2002)

6. Davies, R.H., Twining, C., Cootes, T.F., Waterton, J.C., Taylor, C.J.: A minimum description length approach to statistical shape modeling. IEEE TMI 21(5), 525–537 (2002)
7. Davies, R.H., Twining, C.J., Cootes, T.F., Waterton, J.C., Taylor, C.J.: 3D statistical shape models using direct optimisation of description length. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2352, pp. 3–20. Springer, Heidelberg (2002)
8. Thodberg, H.H., Olafsdottir, H.: Adding curvature to minimum description length shape models. In: Proc. of BMVC 2003, vol. 2, pp. 251–260 (2003)
9. Ericsson, A., Karlsson, J.: Benchmarking of algorithms for automatic correspondence localisation. In: Proc. of BMVC 2006. vol. 2, pp. 759–768 (2006)
10. Rueckert, D., Frangi, A., Schnabel, J.: Automatic construction of 3-D statistical deformation models of the brain using nonrigid registration. IEEE TMI 22(8), 1014–1025 (2003)
11. Cootes, T., Twining, C., Petrović, V., Taylor, C.: Groupwise construction of appearance models using piece-wise affine deformations. In: Proc. of BMVC 2005 (2005)
12. Twining, C.J., Cootes, T., Marsland, S., Petrovic, V., Schestowitz, R., Taylor, C.J.: A unified information-theoretic approach to groupwise non-rigid registration and model building. In: Proc. of Information Processing in Medical Imaging IPMI, pp. 1–14 (2005)
13. Little, R., Rubin, D.: Statistical Analysis with Missing Data. Wiley and Sons, Chichester (1987)
14. Schneider, T.: Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. Journal of Climate 14, 853–871 (2001)
15. Mitchell, M.: An Introduction to Genetic Algorithms. MIT Press, Cambridge, MA (1996)
16. (FGnet - IST-, -26434, Face and Gesture Recognition Working Group) (2000)
17. Langs, G., Peloschek, P., Donner, R., Bischof, H.: Multiple appearance models. Pattern Recognition 40(9), 2485–2495 (2007)
18. Babalola, K.O., Cootes, T.F.: Groupwise registration of richly labelled images. Proc. Medical Image Understanding and Analysis 2, 226–230 (2006)