

Annotation Propagation by MDL Based Correspondences

Georg Langs^{1,3}, Philipp Peloschek², René Donner^{1,3}, and Horst Bischof¹

¹Institute for Computer Graphics and Vision, Graz University of Technology
Inffeldg. 16 2.OG, A-8010 Graz, Austria,
langs@prip.tuwien.ac.at, donner@prip.tuwien.ac.at, bischof@icg.tu-graz.ac.at

²Department of Radiology, Medical University of Vienna,
Währinger Gürtel 18-20 A-1090 Vienna, Austria
philipp.peloschek@meduniwien.ac.at

³Pattern Recognition and Image Processing Group, Vienna University of Technology,
Favoritenstr. 9, A-1040 Vienna, Austria

Abstract *In this paper a method for the propagation of a single prototype annotation to a set of other images depicting analogous structures or objects is proposed. The correspondences between the images are established by a minimum description length guided process. A single instance of an object is annotated and the landmark positions on the remaining images are determined automatically.*

1 Introduction

The task of establishing correspondences over landmark positions in a set of images has been tackled in many fields. It is necessary for the comparison of different images with respect to spatial deformations, for atlas generation in medical applications, or for the building of statistical point distribution models (PDMs).

Among the most frequently used models are statistical shape models like active shape models (ASMs), or active appearance models (AAMs) [8] that capture shape and texture variation of a specific structure or object. During application they utilize this a priori knowledge in order to provide robust segmentation while allowing for repeatable identification of specific landmarks in the data. For PDMs ground truth annotation in a set of training examples usually is performed manually. The necessity of a large number of manually annotated training examples required for a sufficient representative power of the model poses a major drawback for model based approaches. The manual annotation is time consuming, and results are often sub-optimal.

A number of approaches that aid the establishment of group-wise correspondences have been proposed in the context of model building. Given a set of continuous contour annotations in [9] landmarks are placed automatically along these contours using the MDL principle. Thereby artifacts introduced by landmark displacements along the known contour are decreased. The approach proved widely applicable and was further refined in [14], where the local curvature is taken into account. This concept bridges the gap between mere segmentation and the repeatable identi-

fication of landmarks in images. Further advancing in this direction in [13] one-to-many non-rigid registration is used to generate a statistical deformation model (SDM) of a dense control point grid. A reference image is registered to a set of examples of an anatomical structure. The registration is based on voxel values in a set of 3D volumes. In [2] the task of model building or equivalently that of registration was concisely formulated as an image coding problem, thus allowing for an MDL based approach of simultaneous texture and pixel grid deformation model building based on a given image set. In [15] and [7] group-wise as opposed to one-to-many non-rigid registration was utilized by means of an MDL criterion to find dense correspondences in a set of images.

The method proposed in this paper registers a set of control points in a set of images in a group-wise manner. The resulting deformation field is used for the propagation of landmarks defined on one reference image. It is closest related to the latter three approaches that view the problem of registration or model building as a task of coding the information in a set of images. In contrast to those approaches we do not code the entire image, but restrict the computation to a sparse representation given by a set of interest points and local texture descriptors. This allows for reduction of computational complexity, and for taking advantage of the robust behavior of local descriptors over mere image deformation. Furthermore it affords to deal with complex image content, where non-rigid registration of the entire image is not always possible as for example for video sequences, or anatomical structures exhibiting high texture variation. In contrast to pure pair-wise point matching [4] the method takes advantage of the entire set of images. This allows for a formulation of a model based constraint that adapts to a possibly non-rigid object during optimization, instead of relying on fixed geometry based constraints.

The paper is structured as follows: in Sec.2 first an overview on the point descriptors, the shape model, and the cost function that guide the registration is given. The optimization that establishes correspondences across the im-

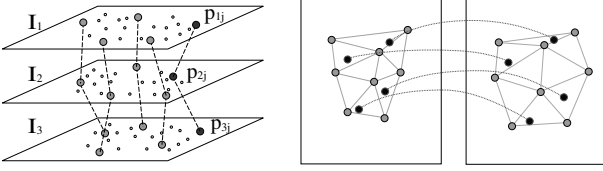


Figure 1: Scheme of the registration. 3 images with established correspondences between control points. After correspondences are established between control points (gray), landmarks (black) can be propagated with the resulting deformation field.

age set (Sec. 2.4), and the propagation of landmarks (Sec. 3) form the main part of the work. Experimental results are reported in Sec. 4. In Sec. 5 a concluding discussion is given.

2 Establishing correspondences

In the first step the method establishes correspondences of positions for a set of control points over a set of images. We use two sources of information, that tell us about the control points behavior: their spatial relation and its variation over the set of images, and the variation in the local texture extracted from the control point positions in the images. In the following these two sources will be described, and a cost function allowing for an optimization based approach will be introduced. Finally the procedure that brings point positions on all images into correspondence by utilizing that cost function will be explained.

Assume we are given a set of n images $\mathbf{I}_i, i = 1, 2, \dots, n$ each depicting an instance of a structure or an object (e.g. a set of hand radiographs) a set of interest points is extracted and treated as control point candidates. Group-wise registration is performed by minimizing the cost function that captures variation of shape and local point descriptor features \mathbf{f}_{ij} of the set of landmarks. The output of the method is a set of landmarks \mathbf{l}_j with $j = 1, 2, \dots, m$ and their positions \mathbf{p}_{ij} in all training images \mathbf{I}_i (Fig. 1). This information will be used for the propagation of a single annotation to the remaining images in the set.

2.1 Local point descriptors

The image content is captured by means of local descriptors. Any descriptor with reasonable specificity can be used. Examples are geometric blur [5, 4], shape context [3], SIFT features [11], or steerable filters [10]. A comprehensive comparison of descriptors is given in [12]. For the experiments the initial correspondences were established by the *shape context* [3]. It generates a histogram based representation of the interest point distribution in the vicinity of \mathbf{p}_{ij} and gave good results for initialization.

For the local texture description during the optimization steerable filters [10] were utilized due to their reliability and low dimensionality. For a given position \mathbf{p}_{ij} in the image they extract a feature vector \mathbf{f}_{ij} describing local frequency and directional behavior of the texture. For the steerable filters, jets comprising filters with a variance $\sigma = 4$, frequencies $\theta \in \{0.3, 0.6, 0.9\}$ and directions $\alpha \in \{0, \pi/4, \pi/2, 3\pi/4\}$ (Fig. 2) proved to give sufficiently

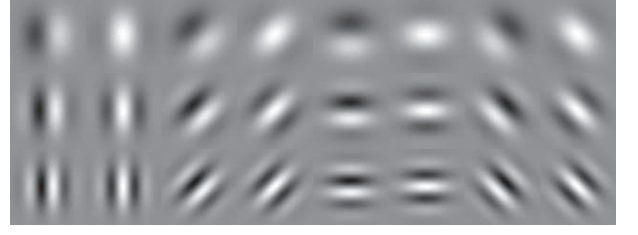


Figure 2: Filters utilized for local texture description.

reliable descriptions. By utilizing complex modulus and argument of the 12 filter responses the algorithm works with feature vectors $\mathbf{f}_{ij} \in \mathbb{R}^{24}$.

2.2 Deformation model

In order to grasp the deformation between images i.e. the displacements of a finite set of control points on the images a statistical point distribution model is employed. That is the MDL criterion is applied on a standard PCA based shape variation model as typically employed in active appearance models. Shapes are represented by a finite set of m landmarks. Each of n shapes in the training set can then be represented by a $2m$ dimensional vector \mathbf{x}_i generated by concatenation of the x and y coordinates of the points in 2 dimensional data. In order to achieve a compact representation PCA is applied on the set $\{\mathbf{x}_i, i = 1, \dots, n\}$ and thereby creates a new coordinate system that represents each of the vectors

$$\mathbf{x}_i = \bar{\mathbf{x}} + \sum_{j=1}^{n_p} a_j \mathbf{e}_j. \quad (1)$$

The modes \mathbf{e}_j are the eigenvectors of the covariance matrix sorted according to decreasing eigenvalue λ_j . $\bar{\mathbf{x}}$ is the mean shape and n_p can be chosen to fulfill a given accuracy constraint. The eigenvalues λ_j correspond to the variance of the data in the direction \mathbf{e}_j .

2.3 An MDL based cost function

The cost function reflects the compactness of a model representing shape and local texture variation at the positions of the landmarks in the set of images. The minimum description length principle states that maximizing the likelihood of a model \mathcal{M} given certain data D is equivalent to minimizing the cost of communicating the model itself and the data encoded with help of the model i.e.

$$L(D, \mathcal{M}) = L(\mathcal{M}) + L(D|\mathcal{M}). \quad (2)$$

We formulate the minimization criterion \mathcal{L} based on this principle. By minimizing the criterion one can expect to derive reasonable landmark correspondences over the set of images while the optimization ultimately leads to a compact model representing their variation:

$$\begin{aligned} \mathcal{L} = & L(\mathcal{M}_T) + L(D_T|\mathcal{M}_T) + \mathcal{R}_T \\ & + \gamma(L(\mathcal{M}_S) + L(D_S|\mathcal{M}_S) + \mathcal{R}_S) \end{aligned} \quad (3)$$

where \mathcal{M}_T is the texture model, capturing the variation of the local texture descriptor response vectors \mathbf{f}_{ij} at all

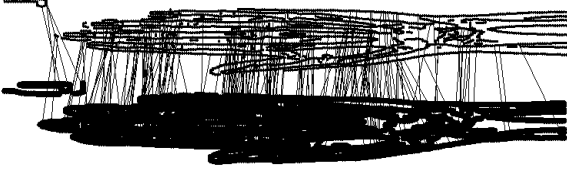


Figure 3: Correspondences between control points in two hand images.

landmarks. For a landmark \mathbf{l}_j , the feature vectors \mathbf{f}_{ij} with $i = 1, 2, \dots, n$ are modeled by a multivariate Gaussian. \mathcal{M}_S is the model of landmark shape variation. Assuming the current landmark positions \mathbf{p}_{ij} in the example images, it is modeled according to Eq. 1. $L(D_T|\mathcal{M}_T)$ and $L(D_S|\mathcal{M}_S)$ are the costs of the shape and texture data encoded with help of the two models. \mathcal{R}_T and \mathcal{R}_S are the residual errors, i.e. the data variation the model is not able to represent. γ is a scalar to balance shape and texture model costs. In Sec. 4 a discussion of the influence of γ on result accuracy in the case of ambiguous texture is given.

If we model the data with a multivariate Gaussian the modeling and encoding costs can be calculated per dimension in the eigenspace. For each dimension j of the eigenspace used to encode the data the transmission costs of the model $L(\mathcal{M}_{\mathbf{e}_j})$ are the quantized eigenvector, the variance of the data w.r.t. \mathbf{e}_j and a quantization parameter for the direction \mathbf{e}_j . $L(D|\mathcal{M}_{\mathbf{e}_j})$ is the cost of transmitting the data i.e. the quantized coefficients \hat{a}_j^i of the training set with respect to the direction \mathbf{e}_j .

The description length for the data encoded with an n_p dimensional eigenspace is the sum of the transmission costs for the data encoded using the eigenvectors $(\mathbf{e}_j)_{j=1, \dots, n_p}$ together with the cost of the residual error

$$\sum_{j=1}^{n_p} (L(\mathcal{M}_{\mathbf{e}_j}) + L(D|\mathcal{M}_{\mathbf{e}_j})) + \mathcal{R}. \quad (4)$$

An extensive derivation of the description length calculation for Gaussian models is given in [9]. Eq. 3 gives us the cost of encoding the training set with our current model. By changing landmark positions on single images and re-evaluating the criterion function we can perform iterative optimization w.r.t. Eq. 3.

Since we do not encode the entire image the additional constraint of fixed landmark positions in one example ensures that the optimization does not result in trivial solutions like a collapse of all landmarks onto the same position. However, during our experiments even with poor initialization and no additional constraint the optimization reached reasonable correspondences fairly quickly, and never converged at a global trivial solution during further refinement. In contrast to that the local distribution of landmarks did exhibit convergence towards each other if no additional constraint was posed.

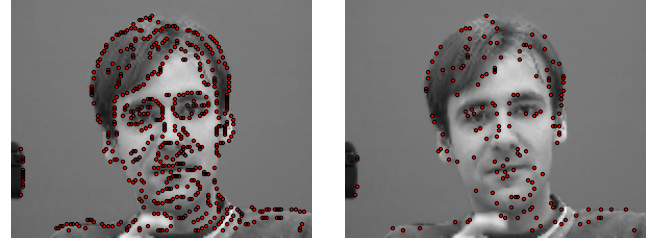


Figure 4: (a) Control point candidates and (b) control points chosen by initialization.

2.4 Registration procedure

The algorithm uses local texture descriptors of the control points and properties of the evolving point distribution model derived from the set of examples, in order to generate landmark correspondences across the data set. After an initialization relying entirely on point to point matching the group-wise registration performs optimization of the correspondences using a minimum description length criterion.

Initialization Given a set of images \mathbf{I}_i , $i = 1, 2, \dots, n$ a set of interest points is chosen as control point candidates on all images. We utilize Canny edge detector responses since they provide a sensible decrease of image points to process while retaining a density necessary for the optimization step. A subset of these points can be used for coarse initialization, additional candidates that allow more accurate positions can be added by means of a multi scale approach described later. Fig. 4(a) shows a set of control point candidates for a face. After initial pair wise establishment of correspondences with the shape context descriptor taking a single image \mathbf{I}_1 as reference and matching to the remaining $n - 1$ images only points for which correspondences in more than a minimum number of examples (5 in our experiments) could be found are taken into account. For these points \mathbf{l}_j with $j = 1, 2, \dots, m$ the best fit with respect to the shape context descriptor is found in all other images. Fig. 4(b) shows the positions of the initial control points on one image.

Optimization After initialization Eq. 3 is minimized in an iterative manner similar to [7]. In each step

1. One example \mathbf{I}_{k^*} is chosen randomly.
2. Shape and texture models are build from the remaining $n - 1$ examples. Using the control point positions \mathbf{p}_{ij} and corresponding descriptor features \mathbf{f}_{ij} where $i \in \{1, 2, \dots, n\} \setminus k^*$, \mathcal{M}_S and \mathcal{M}_T are build. For the shape model the point sets in all images \mathbf{p}_{ij} with $i \in \{1, 2, \dots, n\}$ are aligned with Procrustes analysis before PCA is calculated from the coordinate vectors of images $i \in \{1, 2, \dots, n\} \setminus k^*$ in order to obtain a shape model.
3. The best fit of the shape model to the current control point positions \mathbf{p}_{k^*j} in \mathbf{I}_{k^*} is computed.

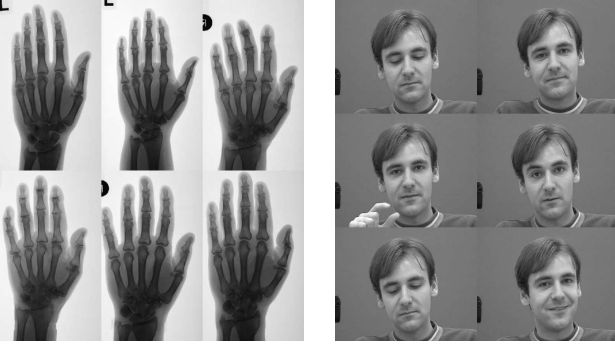


Figure 5: Data set examples for hand radiography data and face sequence [1].

4. The positions \mathbf{p}_{k^*j} are modified to \mathbf{p}'_{k^*j} out of the control point candidate pool so that the cost of encoding \mathbf{p}'_{k^*j} with M_S and M_T is minimized.
5. The procedure is repeated until convergence

The use of local sub models Instead of minimizing solely the description length of a model encompassing the entire control point set we demand the model to be locally compact as well. With each iteration a sub set of local points is chosen and the minimization step is performed only on this sub set instead of all points at the same time. In the case of shape models this includes a separate alignment for the sub set. This effects additional flexibility of the model assuming that it is harder to fulfill the linearity postulation of PCA for landmarks that are farther apart.

The objective function applied locally results in more correspondences so that all local sub sets of landmarks are part of compact sub models.

Multi scale refinement In order to speed up computation the optimization can be performed in a multi scale fashion in two ways. Optimization starts on down-sampled images where the texture descriptors capture a larger texture patch for a small number of interest points. After convergence resolution is increased and landmark positions are projected onto the new image set accordingly. The search for optimal landmark positions can then be restricted to the vicinity of the previous estimates.

Independently from the image resolution the number of control points can be increased by applying an interpolation (e.g. thin plate spline, TPS [6]) on the existing landmark positions. Thereby a dense deformation field can be generated inside of the convex hull of the landmark set and additional landmarks can be projected into the images estimating the corresponding positions in all images with help of this deformation field.

3 Propagating prototype annotations

The result of the optimization is a set of points for which corresponding positions are known in all example images. The correspondences can be used to propagate landmark positions from a prototype image to the other images. In the case

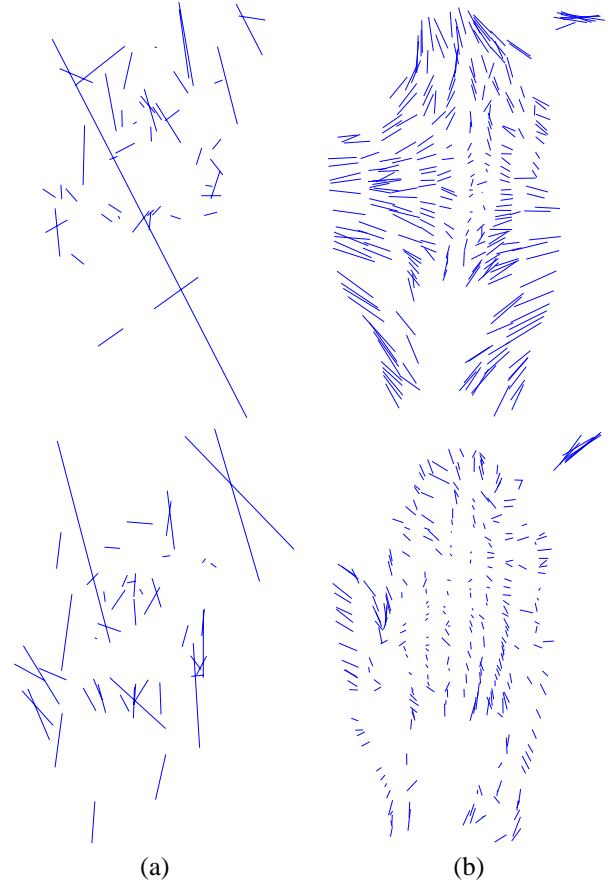


Figure 6: Bone data: (a) 1st and 2nd mode of shape variation before optimization and (b) after optimization.

of medical data an expert can annotate a single instance of the anatomical structure of interest. The landmark positions are propagated onto the remaining images by interpolation based on the established control point correspondences. For the interpolation we worked with piecewise affine and thin plate spline interpolation. In the experiment section the effect of this choice is demonstrated.

4 Experiments

Setup For evaluation of the method we performed experiments on two datasets. 40 radiographs with a spatial resolution of 0.34mm per pixel depicting the hand/wrist region and a sequence of face images [1] were used for annotation propagation and evaluation. For the face data out of 3000 available frames 40 were chosen randomly, and the succession of frames was not utilized. In both data sets background i.e. image regions that we did not expect to be included into sensible models were present. These regions were not excluded explicitly during training. For initialization control points with matches on more than 5 images were used. During optimization local sub models did comprise 20% of the entire number of control points.

The point distribution models were assessed qualitatively during the optimization. The modes of variation of the resulting models are depicted for both data sets. In order to evaluate the accuracy of the landmark propagation for

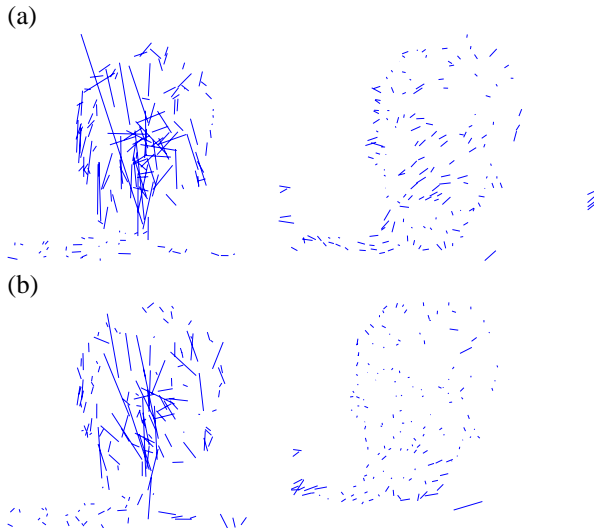


Figure 7: Face data: (a) 1st and 2nd mode of shape variation before optimization and (b) after optimization.

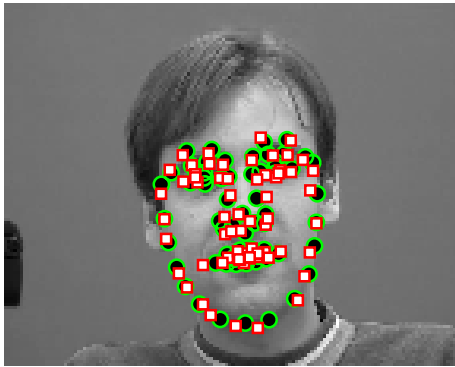


Figure 8: AAM landmark tracking (black/green circles) used as ground truth vs. propagated landmarks (white/red squares).

the hand data manual expert annotations of metacarpals and proximal phalanges were available. For the face data annotation generated by a trained AAM was provided. These annotations were not used for registration but only for the evaluation of the resulting deformation fields against (semi)manual annotations across the two image sets. One example was chosen as reference frame and its ground truth annotation was propagated to the remaining examples utilizing the deformation field defined by the automatically established control point correspondences.

Observing the evolving point distribution model In order to illustrate the effect of the cost function on the point distribution model during optimization the modes of the shape model are depicted. In Fig. 6(a) and Fig. 7(a) the modes of shape variation after point based initialization, and in Fig. 6(b) and Fig. 7(b) the modes after optimization are shown. The optimization reduces the number of modes necessary to represent 75% of the shape variation from 17 modes after initialization to 5 modes in the case of the hand data, and from 15 modes to 9 modes for the face data.

Comparison to existing annotations For the face data a set of 68 landmarks annotated with a trained AAM were available to be used as ground truth. After establishing correspondences for a set of control points by the algorithm one image was chosen as reference and the landmarks were propagated to the other images with TPS interpolation based on the control points as described above. The differences between the propagated annotations and the ground truth annotations were used to evaluate the automatically determined landmark position quality.

For 37 face images the mean/median position difference between ground truth and propagated annotation was 6.38/3.24 pixels. 3 images were excluded because of erroneous ground truth annotation due to occlusion. Fig. 8 depicts a typical frame with ground truth (black/green circles) and propagated landmarks (white/red squares).

For the hand radiographs after control point correspondences were established, the landmarks were propagated with piecewise affine and TPS interpolation. In addition to pure landmark position, errors were evaluated w.r.t. to the bone contour. With piecewise affine interpolation, this resulted in a mean/median landmark position error of 6.93/5.76 pixels (2.34/1.95 mm) and a contour difference of 2.41/1.63 pixels (0.82/0.55 mm) between ground-truth and propagated contour annotations. For TPS interpolation the landmark and contour errors are 10.14/8.10 pixels (3.43/2.74 mm) and 4.19/2.47 pixels (1.42/0.83 mm), respectively. Imperfect control points lying closely cause the rather poor performance of TPS interpolation. Piecewise affine interpolation seems to be an approach better suited to deal with local defects of the deformation field.

The autonomous model building yields better correspondence with ground truth in regions where the image contrast is high. In regions where anatomical structures are hard to discern (because of poor contrast or highly repetitive textures) as for example in the wrist region, where overlapping structures decrease the contrast significantly, the autonomous model building does perform poorer. The influence of the weight parameter γ is effective in that higher values enforce a more compact shape model and thereby improve accuracy in regions with ambiguous texture, while by degrees deteriorating performance in high contrast regions as for example the vertical bone contours. Fig. 9 shows an example of manual ground truth annotation and an annotation propagated from a different reference image as described above.

5 Conclusion

We present a method for the propagation of a single prototype annotation to a set of images depicting the same structure or instances of the same object. The algorithm is based on the MDL principle and establishes correspondences for a set of control points on all images. Thereby a deformation field can be defined. It is used for the propagation of landmarks by interpolation of this deformation field. The algorithm does not make explicit use of known properties of the structures (e.g. physical models), but an improvement of the result accuracy could be expected, if additional domain



Figure 9: Manual ground truth (yellow) vs. automatic annotation propagation on hand radiographs (red). Orange dots indicate the positions of 220 control points in the image.

specific knowledge is utilized.

The work is aimed at reducing the user interaction necessary to build statistical appearance models like active appearance models. Future work will concentrate on improved accuracy and application to image sets with higher variation in the acquisition procedure.

Acknowledgement

This research has been supported by the Austrian Science Fund (FWF) under the grant P17083-N04 (AAMIR). Part of this work has been carried out within the K-plus Competence center ADVANCED COMPUTER VISION funded under the K plus program.

References

- [1] FGnet - IST-2000-26434, Face and Gesture Recognition Working Group.
- [2] Simon Baker, Iain Matthews, and Jeff Schneider. Automatic construction of active appearance models as an image coding problem. *IEEE TPAMI*, 26(10):1380–1384, 2004.

- [3] Serge Belongie, Jitendra Malik, and Jan Puzicha. Shape matching and object recognition using shape contexts. *IEEE TPAMI*, 24(24):509–522, 2002.
- [4] A.C. Berg, T.L. Berg, and J. Malik. Shape matching and object recognition using low distortion correspondences. In *Proceedings of CVPR'05*, volume 1, 2005.
- [5] A.C. Berg and J. Malik. Geometric blur for template matching. In *Proceedings of CVPR'01*, pages 607–714, 2001.
- [6] F.L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE TPAMI*, 11(6):567–585, 1989.
- [7] T.F. Cootes, C.J. Twining, V. Petrović, and C.J. Taylor. Groupwise construction of appearance models using piece-wise affine deformations. In *Proceedings of BMVC'05*, 2005.
- [8] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. *IEEE TPAMI*, 23(6):681–685, 2001.
- [9] Rhodri H. Davies, Carole j. Twining, Tim F. Cootes, John C. Waterton, and Chris J. Taylor. A minimum description length approach to statistical shape modeling. *IEEE TMI*, 21(5):525–537, May 2002.
- [10] W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE TPAMI*, 13(9):891–906, 1991.
- [11] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.
- [12] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE TPAMI*, 27(10):1615–1630, 2004.
- [13] D. Rueckert, A.F. Frangi, and J.A. Schnabel. Automatic construction of 3-d statistical deformation models of the brain using nonrigid registration. *IEEE TMI*, 22(8):1014–1025, 2003.
- [14] Hans Hnerik Thodberg and Hildur Olafsdottir. Adding curvature to minimum description length shape models. In *Proceedings of BMVC'03*, volume 2, pages 251–260, 2003.
- [15] C.J. Twining, S. Marsland, and C.J. Taylor. A unified information theoretic approach to the correspondence problem in image registration. In *Proceedings of ICPR'04*, volume 3, pages 704–709, 2004.