Evaluation of implicit 3D modeling for pose invariant face recognition

Michael Hüsken[a], Michael Brauckmann[a], Stefan Gehlen[a],
Kazunori Okada[b], and Christoph von der Malsburg[c,d]

[a]Viisage Technology AG, Bochum, Germany

[b]Siemens Corporate Research, Princeton, NJ, USA

[c]University of Southern California, Los Angeles, CA, USA

[d]Ruhr-Universität Bochum, Bochum, Germany

ABSTRACT

In this paper, we describe and evaluate an approach that uses implicit models of facial features to cope with the problem of recognizing faces under varying pose. The underlying recognition process attaches a parameterized model to every enrolled image that allows the parameter controlled transformation of the stored biometric template into miscellaneous poses within a wide range. We also propose a method for accurate automatic landmark localization in conjunction with pose estimation, which is required by the latter approach. The approach is extensible to other problems in the domain of face recognition for instance facial expression. In the experimental section we present an analysis with respect to accuracy and compare the computational effort with the one of a standard approach.

Keywords: Face recognition, hierarchical graph matching, 3D head pose variation, pose estimation, pose transformation

1. INTRODUCTION

One of the main drawbacks of state-of-the-art face recognition systems is that the recognition performance is strongly influenced by environmental conditions such as pose, facial expression, and illumination. While these conditions might be well controlled in particular applications (e.g., pose and illumination are usually well controlled in access control scenarios), this is not the case in general, in particular not in uncooperative scenarios such as surveillance applications. Comparing facial images that are taken under different conditions leads usually to a loss in recognition performance. Fully invariant face recognition is possible only on the basis of a parameterized facial model that quantifies the dependency between the facial appearance or the facial features and the external conditions. Such a model is capable of counterbalancing the influence of the external parameters.

A seemingly natural approach to construct a parameterized facial model is analogous to the film industry's computer graphics methodology, employing an explicit representation of facial geometry and texture. Well known approaches are described by Murase et al.[1], Cootes et al.[2], Blanz et al.[3], and Gross et al.[4]. We are not pursuing this appearance based or *explicit 3D model* approach, as for recognition purposes in real world applications the goal of the process is rather the efficiency and quality of the comparison of *biometric features* than a *photorealistic reconstruction* of the face. *Implicit 3D models* focus, in contrast to the explicit models, on the biometric features instead of on the photorealistic reconstruction of the face. Therefore, they are more adopted to the demand of an efficient comparison of faces.

In this contribution we restrict ourselves only to the influence of pose variations; however, the concept of implicit models is a general one and can be extended towards other variations of parameters such as illumination and facial expression. A particular approach for these models is suggested by Okada et al.[5–7], together with a scheme for person independent pose estimation and pose-invariant face recognition. Experiments described in these publications underline the applicability of implicit models.

The experiments in these contributions are restricted to a large number of available data, the assumption of an accurate localization of facial landmarks (e.g., by manually setting the landmarks or by a perfect algorithm for this task), and the accurate estimation of the person's pose. All of them are usually not fulfilled in applications in real environments. We analyze this situation and perform experiments that do not rely on these preconditions. Two important issues addressed in our contribution are how the model accuracy depends on the amount of training samples and if a sufficient model training is possible even with an imperfect landmark localization and pose estimation. We demonstrate the applicability of the whole pose-invariant face recognition system without any manual assistance of a human operator as the main result. This investigation is of major interest as it demonstrates how the suggested approach of implicit 3D models might be applied in applications.

The remaining paper is organized as follows. We start with a brief description of the technical details: The structure of implicit models is introduced in section 2 and graph matching, which is used for the automatic landmark localization, is explained in the succeeding section 3. Section 4 describes the experimental setup as well as the results, in particular showing the impact of implicit models to improve recognition experiments with large variations of the pose angle. The contribution ends with a summary of the main results and a brief description of next steps.

2. IMPLICIT MODELS OF FACIAL FEATURES

A first step in every face recognition algorithm is the localization of the face in a still image or a video sequence. The last step in such a localization process is typically the localization of facial landmarks (i.e., the localization of distinct parts of the face such as the eyes, nose, mouth, border of the face, etc.). Successful approaches to face recognition usually rely on some kind of facial features, which are extracted from the localized and properly aligned images (e.g., PCA-coefficients[4,8], local features[9], and wavelet-based features[10–13]). Implicit models describe how these features depend on particular changes of certain properties of the facial image (e.g., illumination, expression, pose). Therefore, these models can overcome problems in the comparison of images with different properties, as they provide a handle to counterbalance the effect of varied image properties in the feature space.

The approach of implicit models is of general nature; it does not rely on particular properties of the facial features. In the following we will assume only, that the recognition algorithm relies on N_{nodes} localized facial landmarks and that N_{f} features, which are extracted from the image, are related to the positions of each of these landmarks. However, the ideas of implicit models can be reformulated easily if these assumptions are not met.

2.1. Structure of the Implicit Models

One might think of finding an analytical model for the relation between the facial features, the landmark positions, and pose. Gong et al.[14] analyzed this relation numerically for features resulting from a Gabor wavelet transformation and conclude that this relation is low dimensional and smooth. The properties of Gabor features are exploited by Maurer et al.[15,16], who present an analytical approach for this relation. Unfortunately, this approach is based on assumptions that might not be fulfilled in general. These examples suggest that a smooth, low dimensional mapping exists that is capable of describing how facial features change with varying pose angle. Beside these experimental hints it can be assumed that this smoothness is fulfilled in every recognition algorithm, as otherwise it is impossible to define suitable neighborhoods and distances in the feature space.

In contrast to analytical approaches for this mapping we prefer approaches that allow for an automatic model *training* (i.e., the automatic estimation of the model parameters based on training data). This property is desirable to design a general approach that can easily be reused in case of changes in the type of features or type of objects.

Okada et al. proposed linear and piecewise linear approaches for these implicit models. The simplicity of the structure allows for a fast model evaluation and a low risk of overtraining during the model estimation process. This model is the foundation of our investigations and will be introduced briefly in the following. More details can be found in the publications of Okada et al.[5,6,17].

This approach consists of two distinct mappings: The *analysis mapping*

$$\mathcal{A} : (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_{\text{nodes}}}) \rightarrow \boldsymbol{\theta} \tag{1}$$

estimates the vector of pose angles $\boldsymbol{\theta}$ depending on the vector of N_{nodes} landmark positions $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_{\text{nodes}}})$. This mapping can be parameterized to perform a person independent pose estimation with a very high accuracy. Experiments with synthetic facial data yield an average estimation error of about 2°, see Okada et al.[17]. However, one has to keep in mind that the accuracy of the estimation is influenced by the accuracy of the landmark localization, that cannot be assumed to be perfect. Small displacements of the landmarks have only a small impact to the average error of the pose estimation due to the linearity of the models.

The synthesis mapping

$$\mathcal{S}_n : \boldsymbol{\theta} \rightarrow \boldsymbol{f}_n \tag{2}$$

provides information about how the features $\boldsymbol{f}_n$ attached to the landmark with index n depend on $\boldsymbol{\theta}$. This mapping is of particular interest in the following as it allows to synthesize the features of a particular person in a particular pose. This is one conceptional difference between the analysis and the synthesis mapping: The synthesis mapping has to be established for each person separately, while the analysis mapping can be estimated for both a single person or a group of persons. In other words, the latter one can generalize towards previously unseen persons[17].

The first approach for the analysis and the synthesis mappings is a *linear model* (LM), which has the advantage of low dimensionality and therefore the chance of fast processing and low probability of overfitting. In the following we will focus only on the synthesis mapping, which will be applied in the experiments in section 4. The analysis mapping is not used in these experiments as the pose estimation is integrated into the process of landmark localization, see section 3 for further details.

The basic structure of the approach for the synthesis mapping of the facial features is

$$\boldsymbol{f}_n = \bar{\boldsymbol{f}}_n + Q_n \mathcal{K}(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}) \quad . \tag{3}$$

The mapping is, except of the non linear, trigonometric function $\mathcal{K}(.)$, linear; therefore, standard approaches for model construction can be applied with a low risk of overfitting. The mapping $\mathcal{K}(.)$ accounts for the intrinsic non linearity of the mapping functions between 3D pose angles and the features. It is given by a trigonometric transformation

$$\mathcal{K}(\boldsymbol{\theta}) = (\cos(\theta_1), \sin(\theta_1), \cos(\theta_2), \sin(\theta_2), \cos(\theta_3), \sin(\theta_3))^{\mathrm{T}} \tag{4}$$

of the three rotational degrees of freedom $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$. The parameters $\bar{\boldsymbol{f}}_n$, $\bar{\boldsymbol{\theta}}$, and Q_n are to be determined by using the training data, see section 2.2 for further details of the training algorithm. A similar approach to (3) can be applied for the synthesis of the landmark positions.

To cope with nonlinearities in the feature space, these models are extended towards *piecewise linear models* (PWLM), which are weighted superpositions of K LMs. The structure of the PWLMs is

$$\mathcal{A}^{(\text{pwlm})}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_{\text{nodes}}}) = \sum_{k=1}^{K} w_k(\boldsymbol{\theta}) \mathcal{A}_k^{(\text{lm})}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_{\text{nodes}}}) \quad , \tag{5}$$

$$\mathcal{S}^{(\text{pwlm})}(\boldsymbol{\theta}) = \sum_{k=1}^{K} w_k(\boldsymbol{\theta}) \mathcal{S}_k^{(\text{lm})}(\boldsymbol{\theta}) \quad , \tag{6}$$

where $\mathcal{A}^{(\text{lm})}$ and $\mathcal{S}^{(\text{lm})}$ denote an analysis and synthesis mapping, respectively, of a LM and $\mathcal{A}^{(\text{pwlm})}$ and $\mathcal{S}^{(\text{pwlm})}$ an analysis and synthesis mapping, respectively, of a PWLM. The LMs $\mathcal{A}_k^{(\text{lm})}$ and $\mathcal{S}_k^{(\text{lm})}$ are trained such that they are specialized to certain subdomains of the $\boldsymbol{\theta}$-space, the corresponding weight functions

$$w_k(\boldsymbol{\theta}) = \frac{\rho_k\left(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}^{(\text{lm}_k)}\right)}{\sum_{k=1}^{K} \rho_k\left(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}^{(\text{lm}_k)}\right)}, \quad \rho_k(\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{||\boldsymbol{\theta}||^2}{2\sigma_k^2}\right) \tag{7}$$

have large values in these domains and small ones elsewhere. The synthesis mapping (6) can be evaluated in a straightforward manner, as all quantities on the right hand side are assumed to be known given an independent pose estimator. In contrast to this, the analysis mapping (5) is described by an implicit equation of $\boldsymbol{\theta}$, which cannot be evaluated explicitly. Therefore, an iterative algorithm is suggested to solve the equation[6].

2.2. Training of the Implicit Models

The parameters of the models (e.g., $\bar{\boldsymbol{f}}_n$, $\bar{\boldsymbol{\theta}}$, and Q_n in (3)) are determined by using training data that describe the mapping (i.e., the facial features of one particular person for a variety of different pose angles). The parameters $\bar{\boldsymbol{f}}_n$ and $\bar{\boldsymbol{\theta}}$ are the sample means of the features and pose angles, respectively, estimated on the training data.

The matrix Q_n is a transfer matrix from the angle space to the feature space, parameterizing the relation between pose angles and facial features. Okada et al.[5,6] suggest to determine this mapping in a subspace of the feature space. The vectors of facial features are mapped on the first principle components (i.e., the principle components with the largest eigenvalues) of its distribution. In this space the mapping is determined by using singular value decomposition (SVD)[18]. In our experience the intermediate step of the principle component analysis and dimension reduction has not proven to be necessary to improve the model accuracy. Therefore, we determine Q_n directly by using a SVD. The processing time and size of the resulting model is not influenced by the choice to neglect the intermediate step of eigenvectors, as the size of the matrix Q_n depends only on the dimension of $\mathcal{K}(.)$ and f_n. Therefore, it is the same in both approaches.

Each addend in (6) is trained after a manual segmentation of the space of pose angles into K clusters in the same manner as the LMs; the application of clustering algorithms might be possible to chose the segmentation automatically. The parameterization of the weight functions (7) reflects the structure of the segmentation: $\bar{\boldsymbol{\theta}}^{(\mathrm{lm}_k)}$ and σ_k are the mean and the standard deviation of each cluster.

2.3. Pose Insensitive Face Recognition by Means of Implicit Models

One of the main issues in facial recognition is that the accuracy of the comparison of facial images strongly depends on whether or not the images show faces in the same pose. The reason for this is the dependency of the facial features of the pose. To circumvent this problem it is desirable to align the poses of the images (e.g., to reconstruct the facial features of one image in the pose of the other one.)

The combination of the pose estimation (see section 3) and the synthesis mapping provides the basis of implementing a pose insensitive face recognition. Figure 1 gives a schematic overview of the structure of such a system. The key concept is, that each person in the gallery (i.e., the database of individuals known to the

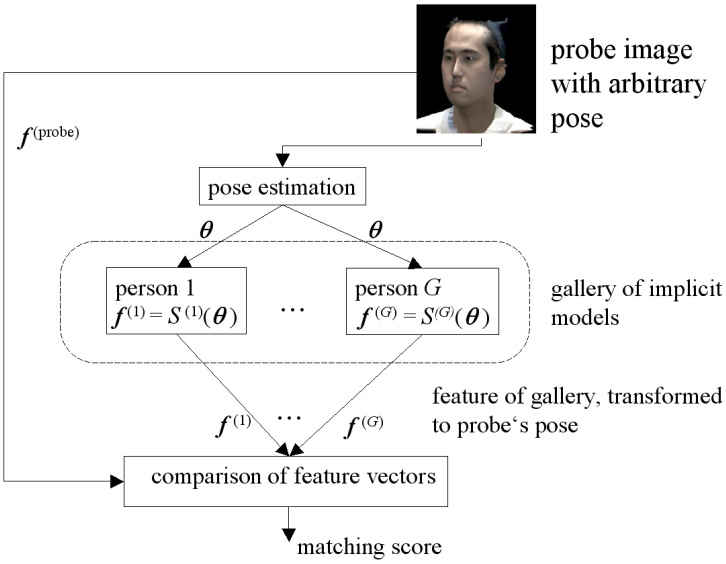


Figure 1. Basic scheme for pose-invariant face recognition: The gallery consists of implicit models instead of single images or multiple images of each person. Therefore, features of an incoming probe image can be compared with features of gallery images in the compatible pose, which can be calculated by using the synthesis mapping provided by the implicit models.

system[19]) is not represented by single or multiple facial images, but by an implicit model of its pose angle, texture, and shape, trained on multiple images of this person. Due to this representation it is possible to generate the features of the gallery entries in the pose of the probe image (i.e., the image of the unknown individual presented to the system for recognition[19]). Therefore, the comparison takes place between feature set originating from the same pose. Okada et al.[7] show the benefit of this approach in an example recognition task.

3. AUTOMATIC LANDMARK LOCALIZATION

A variety of approaches for locating faces and facial elements in images exist (e.g., Matthews et al.[20, 21], Wiskott et al.[12], Rurainsky et al.[22], Blanz et al.[3], Fröba et al.[23]). All of them have in common that some kind of static or flexible model exists that is adapted to the image. The position of the best fit is assumed to be the face's position. Only a few of them are capable of operating efficiently and automatically (i.e., without manually selecting the positions of certain landmarks, for example the eyes) on previously unseen persons and of providing an estimate of the person's pose. The hierarchical graph matching (HGM) algorithm, which is used in our experiments, overcomes all of these shortcomings.

The facial model is represented by an elastic graph that consists of information about typical shapes of faces (landmark positions) and about the facial texture. Searching for a face is equivalent to searching the position in a given image that has the highest similarity with the model faces stored in the model graph. Traditionally, graph matching is associated with Gabor features[10–12]. However, the hierarchical approach is more general and not restricted to this kind of features. Each type of feature that is used to compare faces might be used to describe the graph. Even more, the coarse to fine strategy of the hierarchy of the HGM allows to use different features in different levels of the localization process.

In case of frontal faces (faces with a pose angle less than approximately 15°) sufficient accuracy of landmark localization is achieved by using only 2D information of the face. However, for larger pose angles this model and the search space has to be extended with respect to a 3D representation.

The first step is to include 3D shape information (i.e., the position of every landmark is assumed to be in a 3-dimensional space and all landmarks are located on a grid that represents the typical shape of a face). Figure 2 depicts an example of such a grid, projected onto the image plane for three different yaw angles. This shape information represents knowledge about typical relative positions of the landmarks. This grid is subject to certain transformations: moving in the image plane, rescaling and rotation in the tilt, yaw, and roll angle. In other words, the search of the landmark's positions is constrained by the typical shape of faces. As the search is performed not only in the image plane, but also in a virtual 3D space, the extended HGM does not only provide an estimate of the position of the face, but also of the pose angles.

This extension shows a significant increase in the accuracy of landmark localization in non frontal faces (suitable up to approximately 30°) and provides a suitable estimation of the person's pose. However, for larger

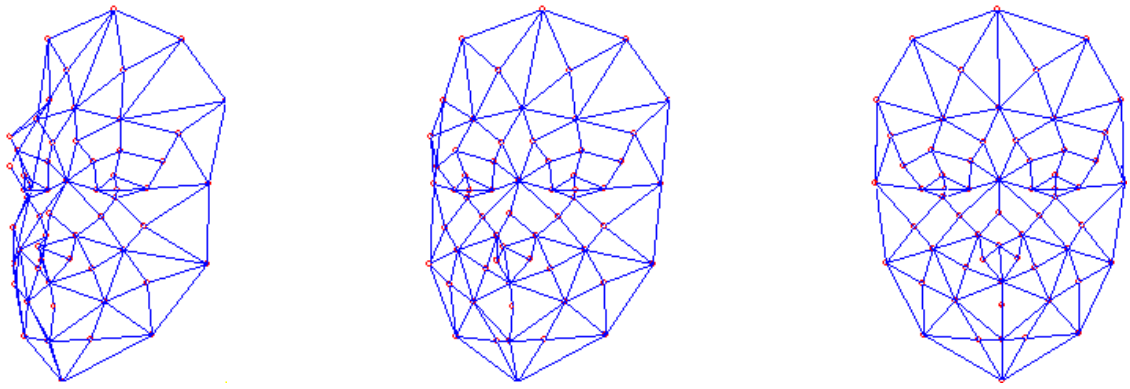


Figure 2. Example of a 3-dimensional grid containing landmarks that are used to locate frontal and non frontal faces. The nodes indicate the landmark positions and the lines edges between landmarks.

pose angles and even higher accuracies this extension with respect only to shape is not sufficient. The reason for this is as follows: The basic assumption in this model is a decomposition of shape and texture information. It is assumed that rotations in 3D influence the shape (i.e., the relative positions of the landmarks), but leave the facial features unchanged. This might be true to a certain degree (e.g., the relative position of the eyes and the nose changes rapidly while the appearance of the eyes remains more or less unchanged), but becomes more and more erroneous for larger pose angles. Therefore, we extended this shape model (without texture) with respect to an implicit model of the facial features (i.e., a model for the relation between facial features and the pose angle). The linear model (3) provides a suitable trade-off between the accuracy of the model and the evaluation time. However, to increase the accuracy a potential extension of the 3D model is to apply piecewise linear models, each of them optimized to a certain subdomain of the angle space and the integration of an explicit occlusion handling.

4. EXPERIMENTAL EVALUATION

Experiments of Okada et al.[6] have shown the benefit of implicit models to decrease the sensitivity of standard face recognition algorithms to pose variations. These experiments have been conducted based on manually located landmarks and accurate pose angles in the training data. Both kinds of information are usually not available and have to be estimated automatically during the enrollment process to instantiate the implicit models.

Our main objective here is to evaluate the concept of pose insensitive face recognition by means of implicit models under typical conditions of real applications. Therefore, we demonstrate the generation of implicit models based on the automatic enrollment of frontal and non frontal face images and the application in a face recognition scenario. Section 4.1 focuses on the enrollment process (i.e., the automatic model generation) and section 4.2 shows the application to an identification and verification scenario.

The concept of implicit models and hierarchical graph matching is of a general nature and does not rely on a particular representation of the facial features. Following Okada et al.[5–7] we have chosen Gabor features for our experimental evaluation. An introduction and a mathematical definition of these features and the measure of the similarity is given in the literature[10, 12]. However, it is straight forward to perform similar investigations with other types of features.

4.1. Automatic Generation of Implicit Models

Basically, the generation of implicit models is the estimation of the model parameters, as described in section 2.2. The necessary training data (i.e., the facial features and the pose angles) are automatically extracted from the images by means of HGM, see section 3. Therefore, the model generation is performed automatically based on a set of training images.

In the following description of the model generation we focus on the synthesis mapping (2) only, as this mapping is necessary for the representation of the gallery. The analysis mapping (1) is of minor interest here, as the automatic landmark localization, described in section 3 provides a suitable estimation of the person's pose. We restrict our work here to the linear case to keep the gallery size and the processing time in the presented application small and to show the applicability of the main principles. Furthermore, we keep the roll angle constant, only two rotational degrees of freedom are considered. However, it is straightforward to extend the experiments to the piecewise linear case.

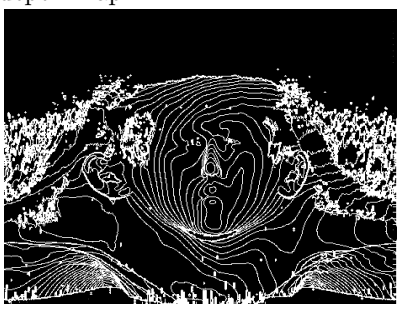
The first step is the feature extraction and pose estimation by means of HGM. This approach operates efficiently and automatically, even on non frontal faces in front of a structured background. Figure 3 depicts three examples from the PIE[24] database. Additional experiments show that this approach is valid with a satisfying degree of accuracy for pose angles up to approximately $30° - 40°$.

The second step is the choice of suitable training and test data. The recognition experiments presented in the following conducted by using images originating from 3D models rendered in arbitrary poses. The reason for this choice is to separate the effect of pose from those resulting from variations caused by changing illumination or facial expression and to have enough data available for the experiments. The *ATR-database*[26]* consists of

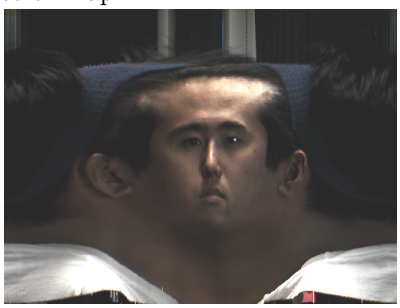
*These data have been recorded by researchers at the *Advanced Telecommunications Research (ATR) Institute International* in Kyoto, Japan.[25]

Figure 3. Three examples of landmark localization and pose estimation on the PIE-dataset.[24] The white dots indicate the estimated position of both eyes, the tip of the nose and the mouth. The top row of each image gives the estimated pose angles. Additionally, one person is rendered in the estimated pose and depicted in the lower right of each image to visualize the estimation outcome and to allow for an easier comparison between the input image from the PIE dataset and the estimation outcome.

(a) depth map



(b) color map



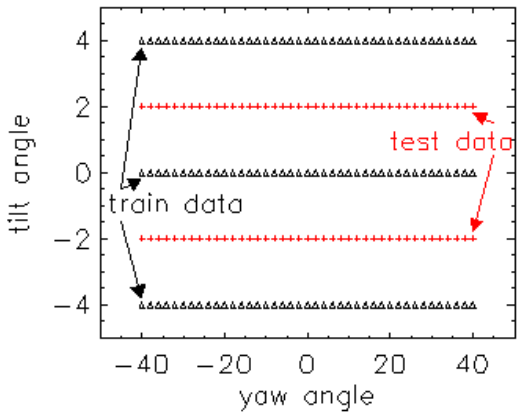
(c) rendered face views

Figure 4. ATR-data: This data set provides 3D laser scans containing texture and shape information of 116 people. These data are recorded at the Human Information Science Laboratories of ATR[25,26] by using a Cyberware laser scanner.[27] Subfigure (a) and (b) depict a typical depth and texture map of such a scan and subfigure (c) shows five rendering results of the model described by (a) and (b) in five different poses. The heads in subfigure (c) are typical training and test data for our experiments.

116 3D face models recorded by a Cyberware laser scanner[27]. A typical depth map, representing the head's shape, is shown in figure 4 (a) and the corresponding texture map is depicted in figure 4 (b). The images used to construct and test the implicit models and to conduct the recognition experiments originate from rendering the resulting 3D head models with arbitrary pose angles, see figure 4 (c) for an example.

We have chosen a scenario with strong variations in the yaw and minor variations in the tilt angle, which is a typical choice for passport scenarios. Figure 5 (a) depicts the distribution of the 205 training and test data per person we have chosen for our recognition experiments. These 205 images are separated into disjoint

(a) distribution of training and test data



(b) number of necessary training data

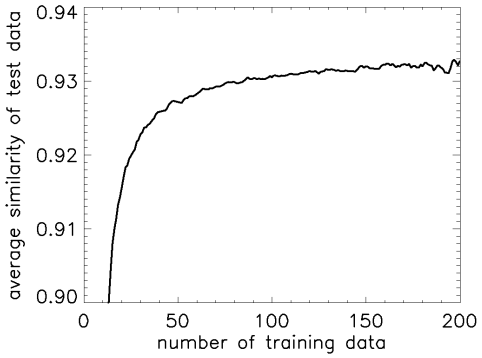


Figure 5. (a) Pose angles of the training and test images used for the automatic model generation. The yaw and tilt angles vary in steps of 2°. Each triangle and each cross represents one training or test image, respectively. (b) Result of the evaluation of the number of training data that is necessary to achieve a suitable accuracy. The accuracy is measured in terms of the average similarity between the original and reconstructed feature vector in the training set.

sets of training and test images. The amount of training data necessary to achieve accurate implicit models is determined in the following:

Implicit models have been trained with randomly chosen images out of these 205 pose variations and evaluated on the remaining images. Figure 5 (b) illustrates the dependency of the accuracy on the number of training data. The accuracy strongly increases up to 50 to 100 training images, a further increase of the number of training data has only a minor impact on the model accuracy. More than 150 images yield no further improvement. Although this observation is based on a particular distribution of data, see figure 5 (a), it coincides with the well known rule of thumb that the number of equations should be approximately 10 times larger than the number unknown model parameters to be determined. In our configuration each of the training samples implies $N_\mathrm{f} = 32$ equations at each node, the number of model parameters at each node is $N_\mathrm{f} \cdot 4 + N_\mathrm{f} + 3 = 163$. Therefore, the result can be assumed to be of more general nature.

This amount of training images seems to be acceptable in most practical applications. The data might be acquired from a short video sequence of a few seconds or from techniques capable of artificially modifying the view angle of an object and thus generating additional training data.

4.2. Application to Pose Insensitive Face Recognition

We have partitioned the 205 images per person for the recognition experiments into training and test data, depicted in figure 5 (a). The gallery is created based on only the 123 training data, the remaining 81 images per person are used as probe images to evaluate the identification and verification performance. This number of training images has proven to be sufficient to train accurate models, see section 4.1. The recognition experiment is conducted with the implicit models of 63 persons from the ATR-database, enrolled with the procedure described in section 4.1. Only the persons in the gallery are represented by implicit models, probe persons are represented only by the features extracted from a single image by using the landmark localization described in section 3. The estimation of the probe's pose is taken from its enrollment and the features of the gallery are synthesized in this pose. This scheme is visualized in figure 1.

The results of this recognition system are compared to the results of standard approaches that are typically characterized by single or multiple templates of each person in the gallery. In the single template case the gallery image of each person is given by a frontal view image, in the multiple template case by three images (frontal, half left, and half right). It is to be assumed that the recognition performance of the multiple template gallery depends on the number and view angle of the gallery images. In particular, the recognition will improve with increasing number of images. The choice of three images per person is fair in terms of processing time, as the

Table 1. Comparison between the face recognition performance for different representations in the gallery. "Rank 1", "rank 5", and "rank 10" mean the percentage of persons not on rank 1 and not better or equal to rank 5 and rank 10 in an identification scenario (i.e., the identification error). The false rejection rate (FRR) at a low false acceptance rate (FAR, given as the subscript) characterizes the performance in a verification scenario. All results are normalized to the scenario with a single template gallery; that is why the results in the first row are always 100%.

type of gallery	rank 1	rank 5	rank 10	$\text{FRR}\|_{\text{FAR}=1\%}$	$\text{FRR}\|_{\text{FAR}=0.1\%}$	$\text{FRR}\|_{\text{FAR}=0.01\%}$
single image	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
multiple images	31.9%	38.5%	46.4%	29.8%	46.5%	57.6%
implicit model	11.0%	15.4%	17.9%	28.4%	32.5%	36.7%

evaluation time of the implicit model and the multiple template gallery is the same in this case, see section 4.3 for further details.

The comparative recognition results are summarized in table 1. They show the benefit of the implicit model approach over the standard approaches with a single template or multiple templates in the gallery. The implicit models outperform the other approaches in terms of verification and identification rates. The rank statistics is the main criterion for identification systems and verification systems usually operate far away from the equal error rate (i.e., an acceptable false rejection rate at a very small false acceptance rate is usually more desirable than a low equal error rate).

To summarize, the results show the benefit of the suggested approach. Furthermore, they qualitatively coincide with recognition results based on accurate, manually localized landmarks. Hence, the automatic landmark localization can be assumed to work accurately enough to allow for an automatic model generation. The whole procedure is capable of instantiating a pose-insensitive face recognition system under real world conditions.

4.3. On Processing Speed and Template Sizes of Implicit Models

Beside the recognition performance a small template size and a low processing time are of major interest. Linear implicit models do not lead to a serious increase of these quantities compared to a single or multiple template approach, as argued in the following:

The biometric template of a single image is given by the features N_f at all N_nodes nodes, the size is $N_\text{f} \cdot N_\text{nodes}$. This size increases linearly with the number of templates in a multiple template approach. On the other hand, the size of an implicit model is mainly determined by the size of the matrices Q_n and the vectors $\bar{\boldsymbol{f}}_n$. The size of the vectors $\bar{\boldsymbol{f}}_n$ for all nodes is equal to the size of a single template and the size of each matrix Q_n is N_f times $N_\mathcal{K}$, the dimensionality of $\mathcal{K}(\boldsymbol{\theta})$ (e.g., $4 \cdot N_\text{f}$ in case of two rotational degrees of freedom). To summarize, an implicit model with 2 rotational degrees of freedom can be encoded approximately in a size of $5 \cdot N_\text{f} \cdot N_\text{nodes}$, the size of 5 templates in a multiple template approach.

As all processing times are proportional to the number of nodes, we do not consider this factor and focus only on the operations at each node. The comparison of two feature vectors is mainly determined by the normalization of two feature vectors of size N_f and its dot product (i.e., approximately 3 times the processing time of a dot product). This time increases proportional to the number of templates. The evaluation of the implicit model is mainly determined by the multiplication of Q_n with the trigonometric function $\mathcal{K}(\boldsymbol{\theta})$ ($\mathcal{K}(\boldsymbol{\theta})$ is evaluated only once for all nodes and therefore a minor contribution), which is approximately the same as $N_\mathcal{K}$ dot products of the above size. Therefore, the evaluation time of an implicit model is comparable with the comparison of one probe template with two gallery templates. Hence, a gallery of implicit models requires the same evaluation time as a multiple template approach with three templates.

Although this is just an approximation, comparable times are justified by experiments.

5. SUMMARY AND OUTLOOK

Face recognition is for a number of reasons a very important method to identify persons. However, one of the main drawbacks of this method is that the comparison of facial images is strongly influenced by external

conditions such as illumination, facial expression, and the person's pose. 3D models, which are able to predict how environmental changes influence the facial appearance (explicit models) or the facial features (implicit models) are a suitable way to circumvent this problem. We prefer implicit over explicit models as our goal is to increase the recognition rates, which rely on the facial features. Explicit modeling, which focuses on the reconstruction of the facial appearance, seems to be a time consuming way around to achieve this goal.

Our investigations are based on the implicit model approach originally proposed by Okada et al.[5–7], which models the facial features with linear or piecewise linear models. The original investigations have been restricted to a number of conditions that are usually not met in real applications, in particular the number of available training data, the accuracy of landmark localization, and the accuracy of pose estimation. Therefore, we draw particular attention to these points. The investigations yield that in the order of 100 training samples (i.e., images from different view angles of a person) are necessary to estimate an implicit model with suitable accuracy. Going beyond this amount does not yield a significant improvement. Such an amount of data is available in many applications, either based on short video sequences or based on artificial models that are initialized by far less images.

The even more interesting point is that hierarchical graph matching allows for a fully automatic estimation of these models. The whole procedure is robust enough to automatically find the positions of facial landmarks as well as the estimate the person's pose; both kind of information are necessary for the model estimation. Many investigations of approaches to solve the problem of pose invariant face recognition assume a manual localization of the face or at least some facial landmarks (e.g., the eyes), which is infeasible in most real applications. Even in the mode of an automatic model generation the implicit model approach outperforms standard approaches to face recognition that rely on a gallery with single or multiple images of a person. At the same time, the implicit models provide a very compact and fast representation of the gallery members. Storage requirements and evaluation times are comparable to a multiple template representation of 3 to 5 templates per person, respectively.

The results yield that implicit models are a very promising method to solve the problem of pose invariant face recognition in real applications. Our next step will be to integrate the implicit model approach into Viisage's face recognition algorithms and environment. Therefore, the methods will have to be combined with the proprietary feature sets as well as particular steps for pre- and post processing the data that allow for much better recognition rates than are achievable with the feature sets used in this investigation. The concept of implicit models is universal and has the potential to operate on other feature sets than the one used in this investigation and on other external parameters, for example changes in illumination and facial expression.

REFERENCES

1. H. Murase and S. K. Nayar, "Visual learning and recognition of 3-D objects from appearance," *International Journal of Computer Vision* **14**(1), pp. 5–24, 1995.
2. T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," in *Proceedings of the European Conference on Computer Vision*, H.Burkhardt and B. Neumann, eds., **2**, pp. 484–498, 1998.
3. V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25**(9), pp. 1063–1074, 2003.
4. R. Gross, I. Matthews, and S. Baker, "Appearance-based face recognition and light-fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26**(4), pp. 449 – 465, 2004.
5. K. Okada and C. von der Malsburg, "Analysis and synthesis of human faces with pose variations by a parametric piecewise linear subspace method," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, **I**, pp. 761–768, 2001.
6. K. Okada and C. von der Malsburg, "Parametric piecewise linear subspace method for processing facial images with 3D pose variations," Tech. Rep. 03-783, Computer Science Department, University of Southern California, 2003.
7. K. Okada and C. von der Malsburg, "Pose-invariant face recognition with parametric linear subspaces," in *Proceedings of the Fifth International Conference on Automatic Face and Gesture Recognition*, pp. 64–69, 2002.

8. M. Turk and A. Pentland, "Eigenfaces for recognition," *Journal of Cognitive Neuroscience* **3**(1), 1991.
9. P. S. Penev and J. J. Atick, "Local feature analysis: A general statistical theory for object recognition," *Network: Computation in Neural Systems* **7**(3), pp. 477–500, 1996.
10. M. Lades, J. C. Vorbrüggen, J. Buhmann, J. Lange, C. von der Malsburg, R. P. Würtz, and W. Konen, "Distortion invariant object recognition in the dynamic link architecture," *IEEE Transactions on Computers* **42**, pp. 300–311, 1993.
11. W. Konen and E. Schulze-Krüger, "ZN-face: A system for access control using automated face recognition," in *Proceedings of the International Workshop on Automatic Face and Gesture Recognition*, 1995.
12. L. Wiskott, J.-M. Fellous, N. Krüger, and C. von der Malsburg, "Face recognition by elastic bunch graph matching," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **19**, pp. 775–779, 1997.
13. N. Krüger and G. Peters, "Object recognition with banana wavelets," in *Proceedings of the 5th European Symposium on Artificial Neural Networks (ESANN'97)*, M. Verleysen, ed., 1997.
14. S. Gong, S. J. McKenna, and J. J. Collins, "An investigation into face pose distributions," in *Second International Conference on Automatic Face and Gesture Recognition*, pp. 265–270, 1996.
15. T. Maurer and C. von der Malsburg, "Learning feature transformations to recognize faces rotated in depth," in *International Conference on Artificial Neural Networks*, pp. 353–358, 1995.
16. T. Maurer and C. von der Malsburg, "Single-view based recognition of faces rotated in depth," in *International Workshop on Automatic Face and Gesture Recognition*, pp. 248–253, 1995.
17. K. Okada, *Analysis, Synthesis and Recognition of Human Faces with Pose Variations*. PhD thesis, University of Southern California, Los Angeles, CA, 2001.
18. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C. The Art of Scientific Computing*, Cambridge University Press, New York, 1992.
19. P. J. Phillips, P. Grother, R. J. Micheals, D. M. Blackburn, E. Tabassi, and M. Bone, "Face recognition vendor test 2002 – Evaluation report," 2003.
20. I. Matthews and S. Baker, "Active appearance models revisited," *International Journal of Computer Vision* , 2004.
21. T. Ishikawa, S. Baker, I. Matthews, and T. Kanade, "Passive driver gaze tracking with active appearance models," Tech. Rep. CMU-RI-TR-04-08, The Robotics Institute, Carnegie Mellon University, 2004.
22. J. Rurainsky and P. Eisert, "Template-based eye and mouth detection from 3D video conferencing," in *Proceedings of the International Workshop on Very Low Bitrate Video VLBV 2003*, pp. 23–31, 2003.
23. B. Fröba and A. Ernst, "Fast frontal-view face detection using a multipath decision tree," in *Proceedings of the 4th International Conference of Audio-and Video-Based Biometrie Person Authentication*, J. Kittler and M. S. Nixon, eds., *Lecture Notes in Computer Science* **2688**, pp. 911–920, Springer, 2003.
24. T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression database," *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25**(12), 2003.
25. Human Information Science Laboratories, Advanced Telecommunications Research Institute International, Kyoto, Japan. http://www.his.atr.co.jp.
26. K. Isono and S. Akamatsu, "A representation for 3D faces with better feature correspondence for image generation using PCA," Tech. Rep. HIP96-17, The Institute of Electronics, Information and Communication Engineers, 1996.
27. Cyberware, Inc., Monterey, CA, USA. http://www.cyberware.com.