Face Recognition based on a 3D Morphable Model

Volker Blanz
University of Siegen
Hölderlinstr. 3
57068 Siegen, Germany
blanz@informatik.uni-siegen.de

Abstract

This paper summarizes the main concepts of Morphable Models of 3D faces, and describes two algorithms for 3D surface reconstruction and face recognition. The first algorithm is based on an analysis-by-synthesis technique that estimates shape and pose by fully reproducing the appearance of the face in the image. The second algorithm is based on a set of feature point locations, producing high-resolution shape estimates in computation times of 0.25 seconds. A variety of different application paradigms for model-based face recognition are discussed.

1 Introduction

Looking back at the development of vision research, significant progress has been initiated by the insight that most recognition tasks, such as face identification, do not necessarily require a full reconstruction of the geometry and the optical properties of the objects that are to be recognized. Instead, many successful approaches have identified invariant features or quantities in the images that reliably characterize individual objects. For frontal views and constant lighting conditions, a number of different algorithms have achieved very impressive results (for an overview, see [14]).

With the focus of research in face recognition shifting more and more towards uncontrolled imaging conditions, it has turned out to be surprisingly difficult to find features or quantities in images of faces that remain invariant with respect to changes in pose and illumination. In the image domain, the changes induced by 3D rotations of faces and by changes in illumination are notoriously complex, despite the relatively simple nature of the underlying transformations in three-dimensional space. 3D rotation and hidden surface removal are captured by a simple matrix equation and a depth buffer operation, and illumination effects of specular non-metallic surfaces are easy to approximate by models such as the Phong illumination model [8]. More sophisticated models would be involved only if subtle effects, such as subsurface light scattering in facial tissue, were to be simulated, which does not seem to be relevant for face recognition.

An obvious invariant feature in different images of a rigid object is the 3D surface geometry with the local reflection properties of the material. The strategy pursued with 3D Morphable Models [5], therefore, has been to extract complete shape and texture estimates as invariant features, and to exploit the fact that changes in pose and illumination are much less complex in the 3D domain than in images. In other words, the approach transfers the invariance problem to a simple, intuitive, explicit and precise formalism for the pose and illumination transformations, at the price of possibly recovering more information from the image than necessary for a pure recognition task, and facing a challenging shape reconstruction problem.

For shape reconstruction, image analysis has to deal with pose and illumination changes in image space. In contrast, image synthesis may perform these transformations both in image space, for example in image based rendering, or in 3D, as it is done in standard computer graphics. Again, the Morphable Model approach uses the fact that these transformations are simpler in 3D, by performing an iterative analysis-by-synthesis with 3D transformations. The non face-specific parameters such as head pose, focal length of the camera, illumination and color contrast are modeled explicitly, and they are estimated automatically. Unlike other approaches, such as Shape-From-Shading, there is no restriction with respect to illumination models or reflectivity functions: Any model from computer graphics can be used in the synthesis iterations, and it affects only the computational complexity of the fitting algorithm.

In order to solve the ill-posed problem of reconstructing an unknown shape with unknown texture from a single image, the Morphable Model approach uses prior knowledge about the class of solutions. In the case of face reconstruction, this prior knowledge is represented by a parameterized manifold of face-like shapes embedded in the high-dimensional space of general textured surfaces of a given topology. More specifically, the Morphable Model captures the variations observed within a dataset of 3d scans of examples by converting them to a vector space representation. For surface reconstruction, the search is restricted to the linear span of these examples.

In the following three sections, we summarize the concept of Morphable Models and describe two algorithms for shape reconstruction. For further details, see [5, 6, 3]. Section 5 will then discuss the advantages and disadvantages of several different paradigms for using the Morphable Model in face recognition.

2 A Morphable Models of 3D Faces

The Morphable Model of 3D faces[13, 5, 6] is a vector space of 3D shapes and textures spanned by a set of examples. Derived from 200 textured *Cyberware* (TM) laser scans, the Morphable Model captures the variations and the common properties found within this set. Shape and texture vectors are defined such that any linear combination of examples

$$\mathbf{S} = \sum_{i=1}^{m} a_i \mathbf{S}_i, \quad \mathbf{T} = \sum_{i=1}^{m} b_i \mathbf{T}_i. \tag{1}$$

is a realistic face if $\mathbf{S}$, $\mathbf{T}$ are within a few standard deviations from their averages. In the conversion of the laser scans into shape and texture vectors $\mathbf{S}_i$, $\mathbf{T}_i$, it is essential to establish dense point-to-point correspondence of all scans with a reference face to make sure that vector dimensions in $\mathbf{S}$, $\mathbf{T}$ describe the same point, such as the tip of the nose, in all faces. Dense correspondence is computed automatically with an algorithm derived from optical flow [5].

Each vector $\mathbf{S}_i$ is the 3D shape, stored in terms of x, y, z-coordinates of all vertices $k \in \{1, \ldots, n\}$, $n = 75972$ of a 3D mesh:

$$\mathbf{S_i} = (x_1, y_1, z_1, x_2, \ldots, x_n, y_n, z_n)^T. \tag{2}$$

In the same way, we form texture vectors from the red, green, and blue values of all vertices' surface colors:

$$\mathbf{T_i} = (R_1, G_1, B_1, R_2, \ldots, R_n, G_n, B_n)^T. \tag{3}$$

Finally, we perform a Principal Component Analysis (PCA, see [7]) to estimate the probability distributions of faces around their averages $\overline{\mathbf{s}}$ and $\overline{\mathbf{t}}$, and we replace the basis vectors $\mathbf{S}_i$, $\mathbf{T}_i$ in Equation (1) by an orthogonal set of eigenvectors $\mathbf{s}_i$, $\mathbf{t}_i$:

$$\mathbf{S} = \overline{\mathbf{s}} + \sum_i \alpha_i \cdot \mathbf{s}_i, \qquad \mathbf{T} = \overline{\mathbf{t}} + \sum_i \beta_i \cdot \mathbf{t}_i. \tag{4}$$

3 Estimation of 3D Shape, Texture, Pose and Lighting

From a given set of model parameters α and β (4), we can compute a color image $\mathbf{I_{model}}(\mathbf{x}, \mathbf{y})$ by standard computer graphics procedures, including rigid transformation, perspective projection, computation of surface normals, Phong-Illumination, and rasterization. The image depends on a number of rendering parameters ρ. In our system, these are 22 variables: 3D rotation (3 angles), 3D translation (3 dimensions), focal length of the camera (1 variable), angle of directed light (2 angles), intensity of directed light (3 colors), intensity of ambient light (3 colors), color contrast (1 variable), gain in each color channel (3 variables), offset in each color channel (3 variables).

All parameters are estimated simultaneously in an analysis-by-synthesis loop. The main goal of the analysis is to find the parameters α, β, ρ that make the synthetic image $\mathbf{I_{model}}$ as similar as possible to the original image $\mathbf{I}_{input}$ in terms of pixel-wise image difference

$$E_I = \sum_x \sum_y \sum_{c \in \{r,g,b\}} (I_{c,input}(x, y) - I_{c,model}(x, y))^2. \tag{5}$$

All scene parameters are recovered automatically, starting from a frontal pose in the center of the image, and at frontal illumination. To initialize the optimization process, we use a set of feature point coordinates [6]: The manually defined 2D feature points $(q_{x,j}, q_{y,j})$ and the image positions (p_{x,k_j}, p_{y,k_j}) of the corresponding vertices k_j define a function

$$E_F = \sum_j \| \begin{pmatrix} q_{x,j} \\ q_{x,j} \end{pmatrix} - \begin{pmatrix} p_{x,k_j} \\ p_{y,k_j} \end{pmatrix} \|^2. \tag{6}$$

that is added to the image difference E_I in the first iterations.

In order to avoid overfitting effects that are well-known from regression and other statistical problems (see [7]), we add regularization terms to the cost function that penalize solutions that are far from the average in terms of shape, texture, or the rendering parameters. The full cost function is

$$E = \frac{1}{\sigma_I^2} E_I + \frac{1}{\sigma_F^2} E_F + \sum_i \frac{\alpha_i^2}{\sigma_{S,i}^2} + \sum_i \frac{\beta_i^2}{\sigma_{T,i}^2} + \sum_i \frac{(\rho_i - \overline{\rho}_i)^2}{\sigma_{R,i}^2}. \tag{7}$$

The standard deviations $\sigma_{S,i}$ and $\sigma_{T,i}$ are known from PCA of shapes and textures. $\overline{\boldsymbol{\rho}}_i$ are the standard starting values for ρ_i, and $\sigma_{R,i}$ are ad–hoc estimates of their standard deviations.

The cost function (7) can be derived from a Bayesian approach that maximizes the posterior probability of α, β and ρ, given $\mathbf{I}_{input}$ and the feature points [5, 6]. E_I is related to independent Gaussian noise with a standard deviation σ_I in the pixel values, and the regularization terms are derived from the prior probabilities. The system performs an optimal tradeoff between minimizing E_I and achieving a plausible result.

The optimization is performed with a Stochastic Newton Algorithm [6]. The fitting process takes 4.5 minutes on a 2GHz Pentium 4 processor.

Figure 1 shows a number of reconstructed test faces. Note that, even though the training set of 3D scans contained 199 Caucasian faces and only 1 Asian face, the system can still be applied to a much wider ethnic variety of

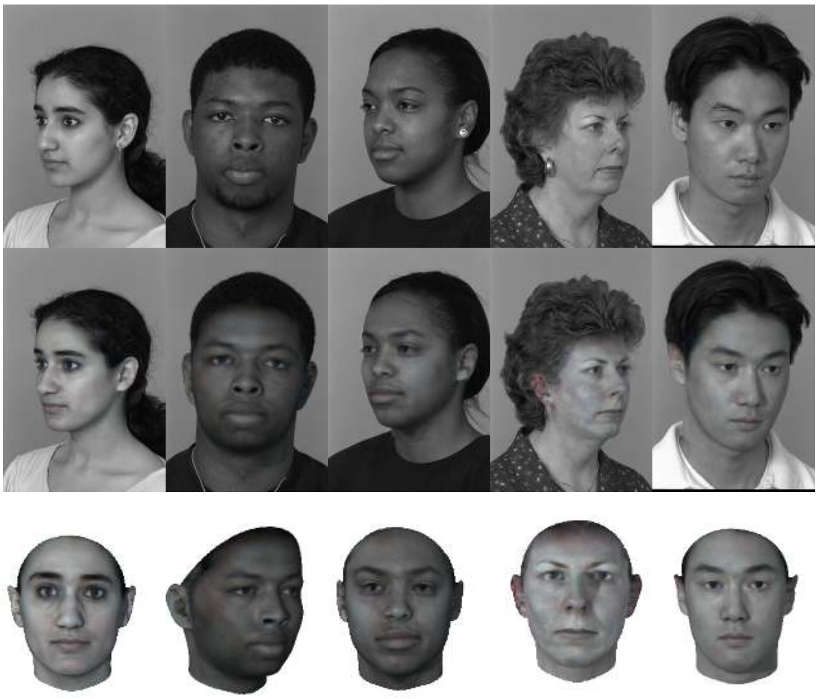

Figure 1. Reconstructions of 3D shape and texture from FERET images [11] (top row). In the second row, results are rendered into the original images with pose and illumination recovered by the algorithm. The third row shows novel views.

faces. All shapes and textures in Figure 1 are linear combinations of the 200 scanned faces.

Unlike Figure 1, we can also enhance the details of the surface texture with a method presented in [5]. This will become relevant in some of the face recognition paradigms described in Section 5. The linear combination of textures $\mathbf{T}_i$ cannot reproduce all local characteristics of the novel face, such as moles or scars. We extract the person's true texture from the image, wherever it is visible, by an illumination-corrected texture extraction algorithm [5]. At the boundary between the extracted texture and the predicted regions, we produce a smooth transition based on a reliability criterion for texture extraction that depends on the angle between the viewing direction and the surface normal. Due to facial symmetry, we reflect texture elements that are visible on one and occluded on the other side of the face.

4 Fast 3D Reconstruction from Feature Points

The most costly part of the fitting procedure, in terms of computation time, is the high-quality reconstruction of facial details, such as the shape of the nose and the lips. While the relative position of these features in the face is certainly relevant for most applications, it may sometimes be acceptable not to reconstruct their individual shape in detail. In previous work [3], we have presented an algorithm that produces a coarse reconstruction from a given set of facial feature points in 0.25 seconds (on a 1.7 GHz Pentium Xeon processor).

In this algorithm, the reconstruction is based entirely on the given 2D positions of feature points, with a cost function as in Equation (6). Unlike Chapter 3, the problem is further simplified by assuming orthographic projection, which is well justified for human faces for camera distances larger than 2 meters. Then, the 2D image positions of the vertices k_j are

$$\begin{pmatrix} p_{x,k_j} \\ p_{y,k_j} \end{pmatrix} = \mathbf{PR}\cdot\mathbf{s}\cdot\left(\begin{pmatrix} \overline{x}_{k_j} \\ \overline{y}_{k_j} \\ \overline{z}_{k_j} \end{pmatrix} + \sum_{\mathbf{i}} \alpha_{\mathbf{i}} \begin{pmatrix} x_{i,k_j} \\ y_{i,k_j} \\ z_{i,k_j} \end{pmatrix}\right) + \mathbf{t} \tag{8}$$

with an orthographic projection $\mathbf{P}$, a known, fixed rotation $\mathbf{R}$, scaling s and translation $\mathbf{t}$, an average position $(\overline{x}_{k_j}, \overline{y}_{k_j}, \overline{z}_{k_j})^T$ of the vertex k_j and principal components $x_{i,k}, y_{i,k}, z_{i,k}$. In this setting, E_F is a quadratic cost function in the shape parameters α_i, and the solution can be found directly using a Singular Value Decomposition.

In order to avoid overfitting, it is crucial in this approach to add a regularization term $\eta \cdot \sum_i \frac{\alpha_i^2}{\sigma_{S,i}^2}$ to the overall cost

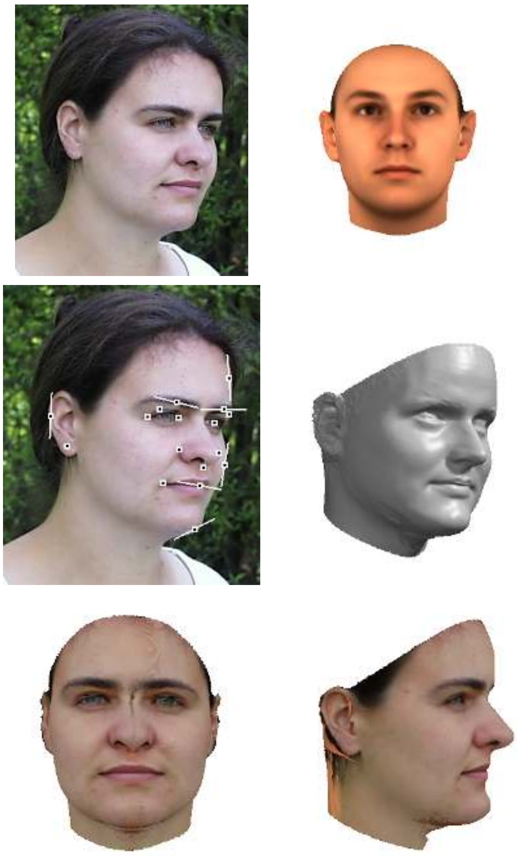

Figure 2. From an original image at unknown pose (top, left) and a frontal starting position (top, right), the algorithm estimates 3D shape and pose from 17 feature coordinates, including 7 directional constraints (second row). We used 140 principal components and 7 vectors for transformations. The third row shows the texture-mapped result. Computation time is 250ms.

function, as in Equation (7), with a regularization factor η that can be estimated from the expected noise or uncertainty in the feature point coordinates [3].

For most applications, the rotation matrix, scaling factor and translation vector are unknown. We have developed an algorithm that approximates the effect of these transformations in a linear way [3]. Residual errors due to this approximation are removed by a second pass of the same algorithm, with updated variables $\mathbf{R}$, s and $\mathbf{t}$. After the second pass, the result is stable and precise.

Finally, the color values of the input image are texture mapped on the surface. Unlike Chapter 3, it is not possible to correct for illumination, since no estimate of illumination is available in this reduced algorithm. Texture values from the right and left side of the face can be mirror reflected to obtain a fully texture face from a single side view, as shown in Figure 2.

Some facial features, such as the lips or the eyebrows, have a linear structure, so correspondence can be defined only perpendicular to this line, but not along the line. For these features, the norm $\| \begin{pmatrix} q_{x,j} \\ q_{x,j} \end{pmatrix} - \begin{pmatrix} p_{x,k_j} \\ p_{y,k_j} \end{pmatrix} \|^2$ in E_F is replaced by a squared scalar product

$$(\mathbf{n} \cdot (\begin{pmatrix} q_{x,j} \\ q_{x,j} \end{pmatrix} - \begin{pmatrix} p_{x,k_j} \\ p_{y,k_j} \end{pmatrix}))^{\mathbf{2}} \qquad (9)$$

with a vector $\mathbf{n}$ perpendicular to the feature line.

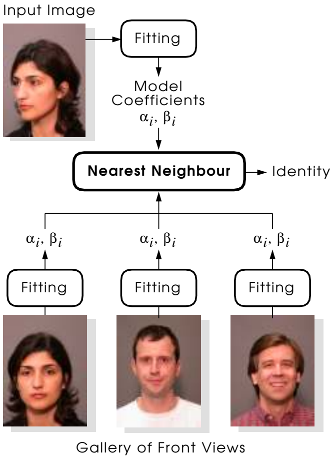

Figure 3. Coefficient-based recognition: The representation of faces in terms of model coefficients α_i, β_i for 3D shape and texture is independent of viewpoint. For recognition, all probe and gallery images are processed by the model fitting algorithm.

5 Model-based Face Recognition

In this section, we discuss a number of different approaches for face recognition across large changes in viewpoint and illumination. For enrollment, the face recognition system is provided with one *gallery* image of each individual person, and in testing, each trial is performed with a single *probe* image. In an *identification* task, the system reports the identity of the probe person, and in *verification*, it checks the claimed identity of a person. The approaches described below are easy to generalize to tasks with more than one gallery or probe image per person available.

3D Shape- and Texture-Based Approach:

The reconstructed shape and texture vectors $\mathbf{s}$ and $\mathbf{t}$ form a representation of 3D faces that can be estimated from images by the fitting algorithm described in Chapter 3 or, with less precision, with the method in Chapter 4. There are a number of options for distance measures between 3D faces to rely on for face recognition.

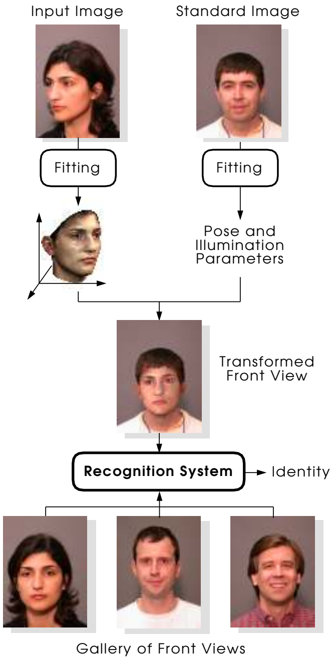

Figure 4. Viewpoint Normalization: From a probe image (top left), the Morphable Model generates a transformed front view. This is input to a view-based face recognition system for comparison with the set of frontal gallery views.

With an ideal reconstruction algorithm, shape and texture would be viewpoint- and illumination-independent However, for unknown camera parameters and an unknown 3D shape, there are a number of ambiguities in shape reconstruction, such as overall size and the ratio of the scale along the left-right and front-back axes. These would pose significant problems to 3D shape comparisons, which is why we recommend the following approach.

Coefficient-Based Approach: The Morphable Model provides a class-specific, low dimensional representation of faces in terms of model coefficients, and this representation is also adjusted to the probability distribution of faces in face-space. The linear coefficients α_i and β_i of the Morphable Model are obtained along with the 3D shape and texture by fitting the model to the input image. It is straightforward to base face recognition on a distance measure defined on these values α_i and β_i.

Due to ambiguities and residual errors in model fitting, different images of the same person produce slightly different model coefficients, so the representation is not perfectly invariant to the imaging conditions. Therefore, the ratio of variation within individuals and between individuals has to be considered by the distance measure. In [6], this is achieved by a PCA of the coefficients estimated from an arbitrary dataset of images (different from the test set used in the evaluation). Then, the model coefficients α_i and β_i are rescaled along the principal axes of within-subject variation.

Once the gallery images and the probe image are converted into model coefficients, identification is a simple nearest-neighbour search in this low-dimensional representation with the distance measure described above. This search can be performed very efficiently.

On 1940 probe images (194 individuals) of the FERET dataset [11], this approach gives an overall rate of 95.9% correct identification [6]. The images contain large variations in pose and some variation in illumination. Figure 1 shows some of the input images and the reconstructed 3D faces.

The images of the PIE database from CMU [12] cover an even larger range in pose (frontal to profile) and large variations in illumination, including harsh lighting from behind. On 4488 test images (68 individuals), the algorithm achieved 95.0% correct identification if the gallery image was a side view [6]. For verification at a 1% false alarm rate, correct verification (hit rate) is 77.5% for the CMU-PIE and 87.9% for the FERET dataset. The fitting algorithm was initialized with a set of six to eight manually defined feature points.

Viewpoint Normalization Approach: While the previous approach is applicable across large changes in viewpoint, this comes at the price of a relatively high computational cost. In contrast, most face recognition algorithms that are commercially available today are restricted to images with close-to-frontal views only, but they are more computationally efficient. In a combined approach [10, 2], we have used the Morphable Model as a preprocessing tool for generating frontal views from non-frontal images which are then input to the image-based recognition systems. This approach pays off if some of the images, for example all gallery or all probe images, are frontal views, so the fitting algorithm is only run on the remaining subset.

For generating frontal views, the Morphable Model is used to estimate 3D shape and texture of the face, and this face is rendered in a frontal pose and at a standard size and illumination. Restricted to the face, the model cannot rotate the hairstyle and the shoulders of the individual in an image. In order to obtain complete portraits that are suitable for the commercial systems, the algorithm inserts the face into an existing frontal portrait automatically (Figure 4). In other words, the hairstyle and shoulders of all preprocessed images are those of a standard person, and the inner part of the face is taken from the non-frontal input image. For details on exchanging faces in images, see [4].

In the Face Recognition Vendor Test 2002, the viewpoint normalization approach has improved recognition rates significantly for nine out of ten commercial face recognition systems tested [10]. In a comparison with coefficient-based recognition, based on the same image data and the same results of the fitting algorithm, recognition rates in the coefficient-based approach were comparable to the best results in the viewpoint-normalization approach [2]. This indicates that not much diagnostic information on identity is lost when the transformed image is rendered and subsequently analyzed in view-based recognition systems.

Synthetic Training Set Approach: Instead of generating a single standard view from each given input image, 3D face reconstruction can also help to build a large variety of different views, which are then used as a training set for learning 2D features that are invariant with respect to pose and illumination. This approach has been presented in [9]. From a small number of images of an individual, a 3D face model was reconstructed, and 7700 images per individual were rendered at different poses and illuminations. Along with the synthetic images, the rendering procedure also provides the 2D positions of features, and image regions around these features can be cropped automatically. These subimages are used for training a support vector machine classifier for each feature type and each individual. For each individual, all feature classifiers are combined to a person-specific component-based recognition system which is computationally efficient and robust at the same time.

6 Conclusions

Recent work in face recognition has demonstrated that Morphable Models of 3D faces provide a promising technique for face recognition under uncontrolled imaging conditions. The process of 3D shape reconstruction by fitting the Morphable Model to an image gives a full solution of the 3D vision problem. For face recognition, however, a full 3D reconstruction may not always be necessary. In those cases, the Morphable Model may help to improve existing image-based classifier systems by preprocessing the gallery or probe images. A comparison of the performance of coefficient-based recognition and a combined approach has demonstrated that both alternatives perform equally well, and that a combination does not imply a loss of characteristic information [2].

In addition to pose and viewpoint, facial expressions pose an interesting and relevant challenge to current face recognition systems. In the Morphable Model, facial expressions can, for example, be modeled by recording 3D scans of a face at different expressions, establishing point-to-point correspondence to the neutral reference face of the Morphable Model, and adding the facial expressions as new vectors to the dataset [1]. In this extended space, the recorded set of facial expression can be applied to any other individual face. If the extended Morphable Model is fit to an image of an expressive face, the expression can be approximately reconstructed, and by setting the coefficients of the facial expression vectors to 0, the face can be brought to a neutral expression automatically [1]. Similar to the viewpoint normalization approach, this can be used for preprocessing images and creating standard views from arbitrary images.

References

[1] V. Blanz, C. Basso, T. Poggio, and T. Vetter. Reanimating faces in images and video. In P. Brunet and D. Fellner, editors, *Computer Graphics Forum, Vol. 22, No. 3 EUROGRAPHICS 2003*, pages 641–650, Granada, Spain, 2003.

[2] V. Blanz, P. Grother, J. Phillips, and T. Vetter. Face recognition based on frontal views generated from non-frontal images. In *IEEE International Conference on Computer Vision and Pattern Recognition CVPR05*, pages 454–461, San Diego, CA, 2005. IEEE.

[3] V. Blanz, A. Mehl, T. Vetter, and H.-P. Seidel. A statistical method for robust 3d surface reconstruction from sparse data. In Y. Aloimonos and G. Taubin, editors, *2nd International Symposium on 3D Data Processing, Visualization, and Transmission, 3DPVT 2004*, pages 293–300, Thessaloniki, Greece, 2004. IEEE.

[4] V. Blanz, K. Scherbaum, T. Vetter, and H.-P. Seidel. Exchanging faces in images. In M.-P. Cani and M. Slater, editors, *Computer Graphics Forum, Vol. 23, No. 3 EUROGRAPHICS 2004*, pages 669–676, Grenoble, France, 2004.

[5] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In *Computer Graphics Proc. SIGGRAPH'99*, pages 187–194, 1999.

[6] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(9):1063– 1074, 2003.

[7] R. Duda, P. Hart, and D. Stork. *Pattern Classification*. John Wiley & Sons, New York, 2nd edition, 2001.

[8] J. Foley, A. v. Dam, S. K. Feiner, and J. F. Hughes. *Computer Graphics: Principles and Practice*. Addison-Wesley, Reading, Ma, 2. edition, 1996.

[9] J. Huang, V. Blanz, and B. Heisele. Face recognition using component-based svm classification and morphable models. In S.-W. Lee and A. Verri, editors, *Pattern Recognition with Support Vector Machines, First Int. Workshop, SVM 2002*, Niagara Falls, Canada, 2002. Springer-Verlag LNCS 2388.

[10] P. Phillips, P. Grother, R. Michaels, D. Blackburn, E. Tabassi, and M. Bone. Face recognition vendor test 2002: Evaluation report. NISTIR 6965, Nat. Inst. of Standards and Technology, 2003.

[11] P. J. Phillips, H. Wechsler, J. Huang, and P. Rauss. The feret database and evaluation procedure for face recognition algorithms. *Image and Vision Computing J*, 16(5):295–306, 1998.

[12] T. Sim, S. Baker, and M. Bsat. The CMU pose, illumination, and expression (PIE) database. In *Int. Conf. on Autom. Face and Gesture Recognition*, pages 53–58, 2002.

[13] T. Vetter and T. Poggio. Linear object classes and image synthesis from a single example image. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(7):733–742, 1997.

[14] W. Zhao, R. Chellappa, A. Rosenfeld, and P. J. Phillips. Face recognition: A literature survey. UMD CfAR Technical Report CAR-TR-948. 2000.