# Automatic 3D Face Reconstruction and Feature Transfer

Dissertation

zur Erlangung des Grades eines Doktors
der Ingenieurwissenschaften

vorgelegt von

Dipl.-Inf. Marcel Piotraschke

bei der Naturwissenschaftlich-Technischen Fakultät
der Universität Siegen
Siegen 2018

Gedruckt auf alterungsbeständigem holz- und säurefreiem Papier.

Betreuer und erster Gutachter
Prof. Dr. Volker Blanz
Universität Siegen
Deutschland

Zweiter Gutachter
Prof. Dr. Thomas Vetter
Universität Basel
Schweiz

Tag der mündlichen Prüfung
4. Juni 2019

# Abstract

The 3D Morphable Model (3DMM) by Blanz and Vetter [BV99] is likewise the basis and the initial motivation of this dissertation. It is based on a data set of 3D scans of example faces which are in dense point-to-point correspondence to each other. This correspondence can be used to apply a morphing between individual faces. Accordingly, it is possible to create a continuous transition between different facial shapes and textures.

The new methods which have been developed and are presented in this dissertation rely strongly on the dense point-to-point correspondence and the associated ability to morph between individual faces. Furthermore, the support of non-rigid transformations (various facial expressions) leads to a substantial advantage when comparing the 3DMM to geometry-based methods. Altogether, this enables to analyze facial attributes, including variations caused by facial expressions or aging of a specific person, as well as the transfer of these attributes between several individuals.

So far it has not been possible to reconstruct facial details like pores or facial hair when fitting a 3DMM to blurred or low-resolution input images as they exceed the level of detail captured by the 3DMM. By making use of a new hallucination approach, these missing, high spatial frequencies can be inferred from a dataset of high-resolution photos of faces. For this purpose, a search for matching candidates is performed individually for each facial region of the face in the input image. Then the additional details of all regions are combined to generate a plausible high-resolution facial texture based on a low-resolution input image.

Another new approach, which is presented in this dissertation, makes it unnecessary to manually select facial landmarks like the position of the nose, mouth, eyes, etc. when fitting the 3DMM to an input image. This new fully automated method provides the basis for analyzing large photo databases. Additionally, it creates a simplified and much more intuitive workflow, which enables even non-professionals to use the 3DMM for 3D face reconstructions. Following this, it is shown how an approach, that creates realistic lighting situations based on coarse sketches on the face in an input image, can be used to change the lighting and shading of an arbitrary face without having knowledge about the real face geometry or the underlying 3D face reconstruction method.

Finally, a novel method is introduced which extracts and combines facial information based on several images of the same person. When compared to existing multi-fitting techniques, the resulting 3D facial reconstructions are more robust and superior with respect to overall quality, especially if the processed photos contain distinct facial expressions, non-frontal poses and complex lighting.

## Zusammenfassung

Als Grundlage und zugleich als initiale Motivation dieser Dissertation dient das 3D Morphable Model (3DMM) von Blanz und Vetter [BV99]. Das 3DMM basiert auf 3D Scans von Beispielgesichtern, zwischen denen eine enge Punkt-zu-Punkt-Korrespondenz hergestellt worden ist. Diese Korrespondenz kann dazu genutzt werden ein Morphing zwischen den Gesichtsformen und -texturen unterschiedlicher Gesichter durchzuführen.

Die im Rahmen dieser Dissertation entwickelten Verfahren machen sich vor allem die Punkt-zu-Punkt-Korrespondenz und das damit verbundene Morphing zwischen unterschiedlichen Gesichtern zu nutze. Aufgrund der Unterstützung nicht rigider Transformationen (unterschiedliche Gesichtsausdrücke) durch das 3DMM ergibt sich ein entscheidender Vorteil gegenüber geometriebasierten Verfahren. So lassen sich personenübergreifende Analysen von Gesichtseigenschaften durchführen. Dazu zählen die Variationen aufgrund von Gesichtsausdrücken oder Alterung einer bestimmten Person, aber auch der Transfer dieser Attribute zwischen unterschiedlichen Personen.

Beim Fitting des 3DMM an unscharfe oder sehr niedrig aufgelöste Eingabebilder war es bislang nicht möglich feine Haar- oder Hautstrukturen wiederherzustellen, welche über den vom 3DMM repräsentierten Detailgrad hinausgehen. Mit Hilfe eines neuen Halluzinationsansatzes werden nun die fehlenden, hohen Ortsfrequenzen aus einer Datenbasis mit hochaufgelösten Fotos von Gesichtern abgeleitet. Dazu werden für jede Gesichtspartie individuelle Kandidaten gesucht, die bestmöglich zum Eingangsbild passen. Deren Details werden nahtlos zusammengefügt und dazu genutzt, aus dem ursprünglich niedrig aufgelösten Eingabebild eine passende, hochaufgelöste Gesichtstextur zu generieren.

Im Zusammenhang mit dem Fitting des 3DMM an Bilder wird in dieser Dissertation ein neu entwickelter Ansatz vorgestellt, welcher die manuelle Selektion von Landmarken (Position von Nase, Mund, Augen, etc.) im Eingabebild obsolet macht. Dieses vollautomatische Verfahren stellt die Grundlage zur Analyse großer Fotodatenbanken dar. Zugleich entsteht dadurch ein stark vereinfachter und intuitiverer Workflow, welcher es auch Laien erlaubt das 3DMM zur 3D Gesichtsrekonstruktion einzusetzen. So wird in dieser Arbeit gezeigt, wie sich auf Basis grober Skizzen realistische Beleuchtungssituationen erzeugen und auf Fotos von Gesichtern anwenden lassen, ohne Wissen über die 3D Gesichtsform oder Details des zugrundeliegenden 3D Rekonstruktionsverfahrens besitzen zu müssen.

Des Weiteren wird ein neues Verfahren vorgestellt, welches die Gesichtsinformationen einer Person aus mehreren Bildern extrahiert und kombiniert. Die resultierenden Gesichtsrekonstruktionen übertreffen dabei die visuelle Qualität existierender Multifittingansätze und sind zugleich robuster in Bezug auf extreme Gesichtsausdrücke, Posen und Beleuchtungssituationen.

## Acknowledgments

This dissertation has been written at the Media Systems Research Group at the University of Siegen and was financially supported by the German Federal Ministry of Education and Research through the INBEKI project and by the DFG Research Training Group 1564 “Imaging New Modalities”.

Special thanks go to my supervisor, Prof. Dr. Volker Blanz, who gave me the opportunity to join his team where I learned a lot about important concepts in the field of Computer Vision, Machine Learning and research in general. I also thank him for his invaluable experience and his support during all these years as well as for our fruitful discussions on current research topics and beyond. It has been an interesting time from which I keep a lot of good memories.

Furthermore, I would like to thank Prof. Dr. Thomas Vetter for accepting to be the co-advisor of my dissertation and for our discussions in Basel which helped me to take the final steps to a successful defense of this dissertation.

I also express my gratitude to Matthias Schumacher and Davoud Shahlaei, the co-authors of two of my publications, for sharing valuable ideas and for combining our individual research topics to gain exciting results.

Many thanks to my former colleagues at the Media Systems Group: Joanna Czajkowska, Mai Lan Ha, Thomas Klinkert, Sascha Nesch, Björn Schiefen and Tim Wenclawiak. And to my former colleagues at the Operating Systems and Distributed Systems Group: Prof. Dr. Roland Wismüller, Hawzhin Hozhabr Pour and Alexander Kordes. I remember countless valuable discussions and precious moments inside as well as outside of the office with all of you.

I am especially grateful to Stefanie Schmidt for all the valuable time which we have shared during all these years in Siegen and for providing strength in times of doubt.

Most of all I want to thank my mom, Ursula, for her unwavering support during my whole life. Without her encouragement this dissertation and many other achievements would not have been possible. Also, many thanks to my family and friends for all their support and for all the great moments that we have shared together.

# Contents

# 1 Introduction

This dissertation covers two main topics: On the one hand it focuses on simplifying and improving the 3D reconstruction of human faces from a single 2D photo as well as from large photo collections by developing new approaches in the field of automatic 3D shape estimation and facial texture extraction. On the other hand it introduces new techniques for transferring facial features between individual faces to alter the appearance of facial shapes and textures. These methods include synthesizing high-resolution face textures from low-resolution input images, facial aging simulation and the transfer of specific illumination situations in the context of lighting design.

## 1.1 Motivation

It is not by chance that meeting people in person and interacting with them is referred as face-to-face contact. A significant part of human communication is non-verbal, and facial expressions play a major role to guide interpersonal communication. Most of the time, the conversational partners study each others' faces closely and try to interpret even subtle expression changes. The speaker makes use of facial expressions to emphasize specific statements or to clarify if words should be taken literally, ironically or at least not too seriously. Likewise, the listener can indicate interest in a topic, signal approval or disapproval. Accordingly, when simulating human faces in rendered scenes or movies, it is extremely important to implement plausible expressions for animated characters to create convincing results.

In recent years, a lot of effort has been spent using Computer Generated Imagery (CGI) to create high-detailed facial reconstructions of various actors and to make it as hard as possible for the audience to distinguish real faces from digitally altered and rendered faces on screen. This is especially tricky when showing close-ups of an actor's face, because the human mind is well trained in recognizing facial expressions and movements, therefore often even slight inconsistencies will be noticed and compromise authenticity.

In the movie *Ant-Man* (2015) the unaltered actor Michael Douglas and in *Captain America: Civil War* (2016) the unaltered Robert Downey Jr. are each shown side by side with their digitally de-aged counterpart. Recently, in *Rogue One: A Star Wars Story* (2016) Lucasfilm and Industrial Light & Magic used CGI to digitally resurrect actor Peter Cushing who starred as the same character in the Star Wars movie of 1977 but died in 1994. A combination of motion-capture and digital re-creation of Cushing's face was used to transform the look of another actor solely by using digital technology [Itz16]. Besides high-detailed shape and texture reconstructions, a realistic lighting simulation is required to create convincing results.

A fully automated face reconstruction does not deliver the necessary details and perfection to be applicable for the above mentioned scenarios yet. Nevertheless, there are other situations where current face reconstruction techniques are extremely useful. In [BBPV03] the 3D Morphable Model (3DMM) by Blanz and Vetter [BV99] is used to animate faces in photos and even in paintings. Inspired by this idea Thies et al. [TZS+16] presented a real-time approach for the reenactment of faces in images and videos which caught the attention not only in sciences but also in public and social media around the world. Lately, Kim et al. [KGT+18] used a generative neural network to outperform prior work in quality.

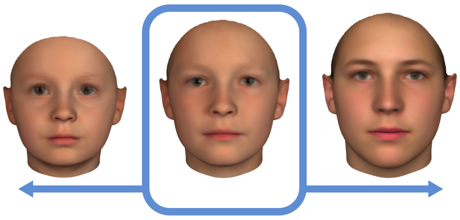
To improve face recognition for non-frontal face images, Blanz et al. [BGPV05] used the 3DMM to generate adequate frontal views. After analyzing the face recognition abilities and strategies of humans Sinha et al. [SBOR06] point out that humans show a remarkable tolerance to very low-resolutions in regard of recognizing familiar faces. In another experiment it is shown that even after drastic compressions in width faces of celebrities can still be recognized by humans. However, Jenkins et al. [JWvB11] claim that humans struggle a lot by identifying unfamiliar faces, because of the variability of the same person in individual photos caused by differences in age, facial expression, pose or illumination. The huge difference in the perception of familiar and unfamiliar faces may give a hint that humans also use some kind of model of faces which is derived from their daily experience. This would also explain why humans struggle more in recognizing faces of persons from other than their own ethnic group. Furthermore, Schumacher and Blanz [SB12] could not find any evidence that the model which humans use to deduce the profile view of a face from a frontal view is more sophisticated than a linear face model like the 3DMM. All these findings may help researchers to develop new ideas for future face recognition and reconstruction methods.

In the remainder of this dissertation the 3DMM framework is used and further developed to create and improve shape and texture reconstruction, simulate aging, neutralize facial expressions and perform sketch-based relighting.

## 1.2 Contributions

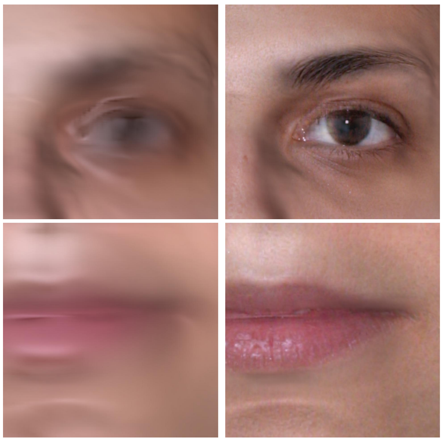
This section provides a brief overview on the contributions of this dissertation divided with respect to the individual areas of research.

### *Aging Simulation with the 3D Morphable Model of Faces*



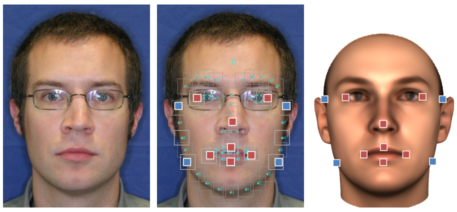
A lot of the existing face recognition software can not be applied directly to images of children. One simple reason is that the available software is just not optimized for the specifics of a child's face, but this can be corrected by using training data that includes not only faces of adults but also faces of children. A problem that is much harder to solve is caused by the aging which leads to substantial transformations even during short periods of time. Accordingly, face recognition software as well as humans struggle to recognize children after they have aged a few years. The proposed method uses a 3D Morphable Model to represent aging effects. One common practice for learning- and training-based algorithms rests upon the paradigm to find adequate training data in regard to the real world data and conditions. Therefore, 3D scans and photos of about 140 additional faces of persons that are aged between three and twenty years were captured. This method is part of a software prototype which enables, among other things, an age-independent recognition of children and was developed in the context of the BMBF project INBEKI [Kru09].

### *Hallucination of Facial Texture Details*



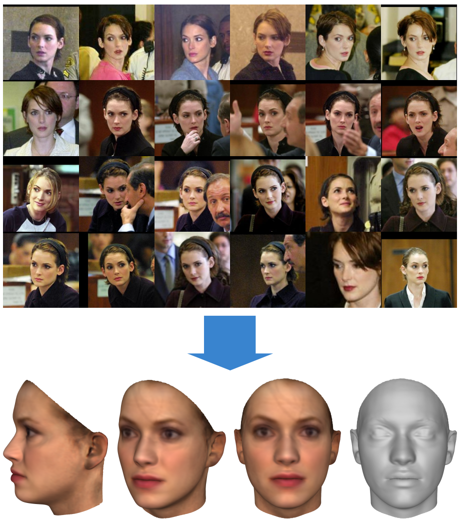
In several scenarios, i.e. when dealing with security camera footage or older photos, the quality of the input image is limited. For example, the overall resolution of an input image is low, the face that needs to be reconstructed is small in comparison to the complete captured scene or for some reason it is blurry. The proposed method uses the 3D Morphable Model to enhance the quality of images of faces during the reconstruction procedure by adding details to the facial texture that were not present in the input images. It fills in details that have been lost due to blur or low resolutions. For this purpose an additional texture enhancement algorithm is proposed which adds high-resolution details that are derived from example faces. It makes use of the Mahalanobis distance to automatically identify matching candidates and applies optical flow and warping techniques to align the transferred details to the target. This method has been published in [SPB15].

### *Automatic Initialization of the 3D Morphable Model Fitting*



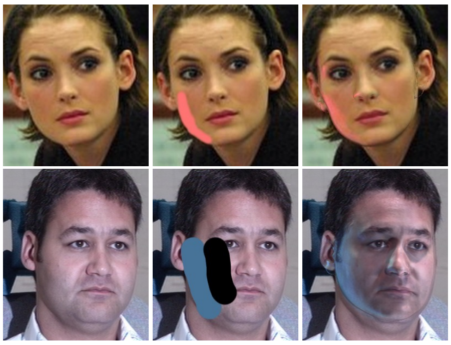
To support a fully automated initialization of the 3D Morphable Model which is no longer depending on facial landmarks which have been selected by a user, a new 3D facial reconstruction pipeline is developed. For each available input image, a face detection is applied, and if a face has been found the pose is estimated and facial landmarks are localized automatically. Next, the original input images are cropped to the region of interest and the 3D reconstruction with a 3DMM is initialized by using the position, pose and landmarks of each detected face. As inaccurate landmark locations tend to degrade face reconstruction, additionally a pairwise 3DMM multi-fitting approach is proposed. This pipeline for automatic landmark localization and 3DMM initialization is an important part of the research which has been published in [SPB16] and [PB16].

### *Automated 3D Face Reconstruction from Multiple Images*



An automatic 3D reconstruction of faces from images is challenging if the image material is difficult in terms of pose, lighting, occlusions and facial expressions, and if the initial 2D feature positions are inaccurate or unreliable, due to the usage of an automatic landmark localization method. The proposed approach reconstructs individual 3D shapes from multiple single images of one person, judges their quality and then combines the best of all results. This is done separately for different regions of the face. The core element of this algorithm is a quality measure that judges a reconstruction without having information about the true shape. In this context several different quality measures are evaluated, a method for combining individual facial shapes is developed, and a complete processing pipeline for automated 3D face reconstruction is designed. This work can be applied to arbitrary, unconstrained photo collections and has been published in [PB16].

### *Image-based Relighting of Faces*



A novel approach for realistic lighting design for faces is proposed. Starting with an arbitrary 2D image of a face, the 3DMM is used to create a 3D reconstruction of that face. By painting strokes on the original image, the user is able to add or alter the lighting effects in the input image. This method makes use of the 3DMM and a state of the art inverse lighting algorithm for faces that allows to render illumination effects such as cast shadows, specular highlights, multidirectional and colored lighting. Furthermore, methods for automatic landmark localization, 3DMM initialization and 3D reconstruction are used to automate the facial reconstruction procedure and allow to handle arbitrary images of human faces. Thus, the reconstruction as well as the lighting algorithm are invariant to pose, facial expressions and image saturation. As the lighting design approach should require only a single input image of an unknown face to produce realistic results, a quality metric is used to increase the robustness of the 3D reconstruction. This new lighting design approach has been published in [SPB16].

## 1.3 Outline

This dissertation is organized as follows:

*Chapter 2* describes the basics of the 3DMM. This includes the creation of the statistical model based on 200 3D scans of real faces and the extraction of facial features into shape and texture coefficients. Additionally, the integration of facial expressions into the model is shown.

*Chapter 3* provides details about the expansion of the original 3DMM which was limited to 3D scans of adult faces by capturing 140 additional facial scans of children. In a conclusive evaluation the benefits of the extended model are worked out, non-linear aging trajectories are extracted from the gathered facial data and age progression is simulated based on learned longitudinal data.

*Chapter 4* introduces a method to extract skin details from high-resolution scans or images of another person. Subsequently, these details are transferred and added to a low-quality facial texture from another, originally blurred input image. Additionally, an improved multi-view texture extraction and merging is presented.

*Chapter 5* presents a processing pipeline for automatic face detection, landmark localization and 3D face reconstruction. In a pose-dependent mapping useful 2D landmark positions are linked to locations on the 3D shape of a face and the 3DMM is initialized without the need of user input. To enhance the quality of the reconstructions a semi-automatic multifitting approach is discussed.

*Chapter 6* proposes a method which reconstructs individual 3D shapes from multiple single images of one person, judges their quality and then combines the best of all results. In this context several different quality metrics are evaluated, a method for combining individual facial shapes is developed, and a complete processing pipeline for automated 3D face reconstruction is designed.

*Chapter 7* focuses on an approach for realistic lighting design for faces. Based on coarse sketches of the desired lighting on the face in the input image which are drawn by a user, the algorithm creates a realistically relighted face and renders it on top of the original one.

Finally, the dissertation concludes with a closing remark in *Chapter 8* and some ideas for future work in *Chapter 9*.

# 2

# 3D Morphable Model

Most topics of this dissertation require a dense correspondence between all processed 3D faces. Therefore, this chapter describes the underlying 3D Morphable Model (3DMM) of faces introduced by Blanz and Vetter [BV99] which is derived from a dataset of 3D scans of faces. This statistical model captures the range of natural faces in terms of 3D shapes and textures and establishes a dense point-to-point correspondence of all faces with a reference face and thereby enables morphing operations between all these faces.

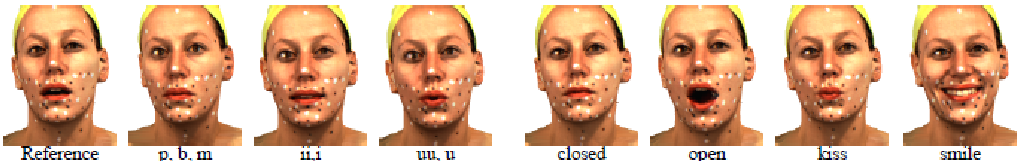
The creation of a 3DMM is presented in Section 2.1. Followed by a description of fitting the 3DMM to 2D input images in Section 2.2 and the texture extraction from such images in Section 2.3. Finally, in Section 2.4 the 3DMM fitting to 3D scans of faces is discussed.

## 2.1 Model Creation

The 3DMM is based on a vector space representation of shapes $\mathbf{S}_i = (x_1, y_1, z_1, x_2, \dots, z_n)^T \in \mathbb{R}^{3n}$ and textures $\mathbf{T}_i = (r_1, g_1, b_1, r_2, \dots, b_n)^T \in \mathbb{R}^{3n}$ with $x, y, z$ coordinates and $r, g, b$ colors of $n = 113\,753$ vertices from $i$ example faces. For simplicity, it is assumed that the number of valid texture values is equal to the number of vertices. Please note that the current version of the 3DMM is a slight modification of the model introduced in [BV99]. Besides increasing the absolute number of vertices from $n = 75\,972$ to $n = 113\,753$, especially the vertex resolution in the area of the mouth has been increased to allow for improved mouth and lip motions relating to facial expressions. With the exception of Chapter 3, where a specific 3DMM of children is created, in all experiments of this dissertation, the 3DMM of faces is constructed from 3D scans of $m = 200$ adults with a neutral facial expression (see Figure 2.1) and from $p = 35$ additional static 3D scans which capture various facial expressions of a single individual [BBPV03] (see Figure 2.2).

The 3D scans have been captured by using a Cyberware™ laser scanner. Nevertheless, a structured light scanner, a time-of-flight camera, multiview stereo or any other 3D

**Figure 2.1:** The 3DMM introduced by Blanz and Vetter [BV99] is derived from 200 3D scans.



Credit: Blanz et al. [BBPV03]

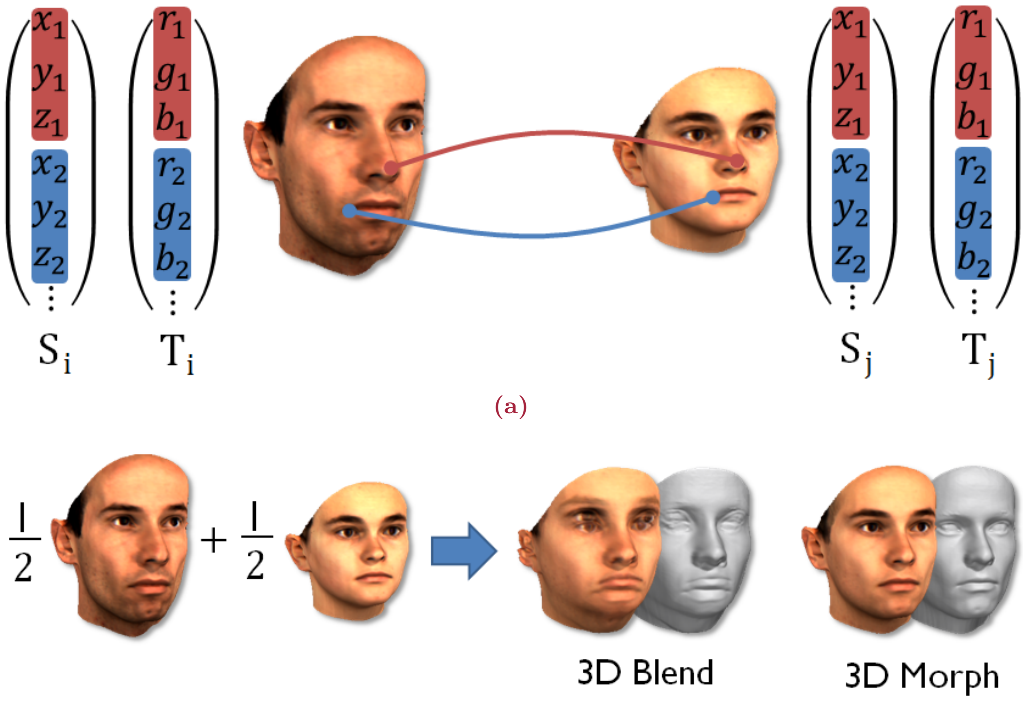
**Figure 2.2:** In a total of 35 static 3D laser scans different visemes and a gradually mouth opening are captured for the facial expression dataset [BBPV03]. Painted markers are used to achieve a more precise 3D alignment.

acquisition techniques could have been used as well. Actually, the crucial part is not capturing the 3D scans but establishing a dense point-to-point correspondence by using a modified gradient-based optical flow algorithm [BAHH92, BV99]. Afterwards, every vertex of the reference head has exactly one corresponding vertex in each of the other heads as is shown in Figure 2.3a.

This point-to-point correspondence allows 3D morphs (see Figure 2.3b) between each face of the database by applying convex combinations of each face vector using weighting factors $\mu_{S,i}$ for shape $\mathbf{S}_i$ and $\mu_{T,i}$ for texture $\mathbf{T}_i$. The result of these convex combinations is always a natural, face-like appearance in respect to shape and texture:

$$\mathbf{S} = \sum_{i=1}^{m} \mu_{S,i} \cdot \mathbf{S}_i, \qquad \mu_{S,i} \in [0,1], \qquad \sum_{i=1}^{m} \mu_{S,i} = 1 \tag{2.1}$$

$$\mathbf{T} = \sum_{i=1}^{m} \mu_{T,i} \cdot \mathbf{T}_i, \qquad \mu_{T,i} \in [0,1], \qquad \sum_{i=1}^{m} \mu_{T,i} = 1. \tag{2.2}$$


$x_1$ $y_1$ $z_1$ $x_2$ $y_2$ $z_2$
$r_1$ $g_1$ $b_1$ $r_2$ $g_2$ $b_2$
$S_i$
$T_i$
$x_1$ $y_1$ $z_1$ $x_2$ $y_2$ $z_2$
$r_1$ $g_1$ $b_1$ $r_2$ $g_2$ $b_2$
$S_j$
$T_j$
$\frac{1}{2}$ + $\frac{1}{2}$
3D Blend
3D Morph

(a)

(b)

**Figure 2.3:** The point-to-point correspondence between all faces (see Figure 2.3a) enables to generate 3D morphs which do not suffer from blending artifacts as is shown in Figure 2.3b

Instead of directly using the captured vector data, a Principal Component Analysis (PCA, see [DHS00]) is used to determine the eigenvectors (principal components) $\mathbf{s}_i$, $\mathbf{u}_i$ and $\mathbf{t}_i$ based on the individual shapes, expressions and textures, respectively. Using a PCA to transform the vector data into an orthogonal basis which is adapted to the variance of the captured data yields two major benefits: On the one hand, it enables to distinguish between relevant and irrelevant dimensions. This information can be used for data compression (see Figure 2.4). On the other hand, this basis transformation can be used to estimate the probability distribution of the shape and texture vectors which is used in the context of the Mahalanobis distance to make statements about the similarity of faces (see Sections 4.4.2 and 4.4.3) or to determine the plausibility of 3D face reconstructions (see Sections 2.2, 2.4, 3.5, 5.5.1 and 6.3.2).

The PCA is performed separately on shape, expression and texture data which means that correlations between shape, expression and texture are ignored. The advantage of forming the linear combinations for all three modalities with separate coefficients is that a wider range of faces can be created as each modality can be adjusted separately and without interfering with the other modalities. For example, arbitrary facial expressions can be added in a straightforward procedure to any existing facial shape. The latter is an extremely helpful feature when transferring facial expressions from one person to another without losing the individual look of the targeted person.

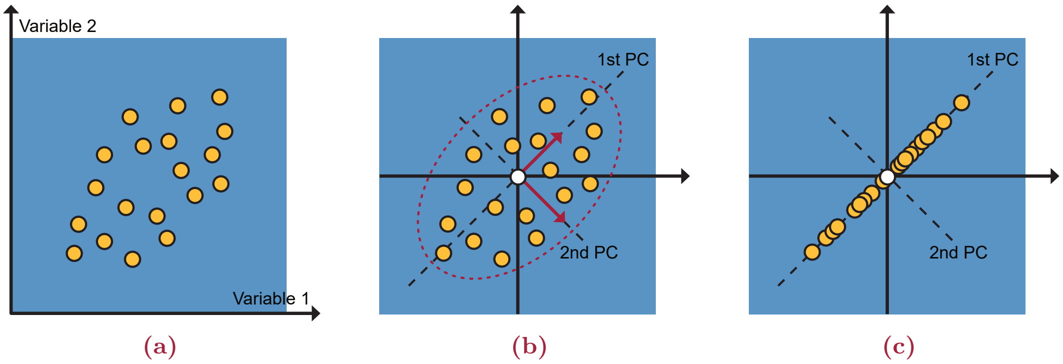


**Figure 2.4:** In Figure 2.4a a scatterplot of 2D vector points in cartesian coordinates is shown. These coordinates are used as input for the PCA. The result after centering the data and computing the first two principle components is shown in Figure 2.4b. A data reduction of the cartesian coordinates to the subspace of the first principle component to approximate the original data would result in the representation in Figure 2.4c.

In the following only the PCA of the facial shape is described in more detail but the same procedure can also be applied to the facial expression and texture: Starting with the shape vectors $\mathbf{S}_i$ of $i = 1 \ldots m$ example faces, the average shape $\overline{\mathbf{s}} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{S}_i$ is subtracted from each shape vector, so that $\mathbf{a}_i = \mathbf{S}_i - \overline{\mathbf{s}}$. Then the data matrix

$$\mathbf{A} = \begin{pmatrix} | & | & \cdots & | \\ \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_m \\ | & | & \cdots & | \end{pmatrix} \in \mathbb{R}^{3n \times m} \tag{2.3}$$

is created. It contains the centered vector data for all facial shapes and is used to compute the eigenvalues and eigenvectors of the covariance matrix

$$\mathbf{C}_S = \frac{1}{m}\mathbf{A}\mathbf{A}^T = \frac{1}{m}\sum_{i=1}^{m} \mathbf{a}_i \mathbf{a}_i^T \in \mathbb{R}^{3n \times 3n} \tag{2.4}$$

by using a Singular Value Decomposition (SVD, see [PTVF92, p. 59ff]) to factorize the data matrix $\mathbf{A} = \mathbf{U}\mathbf{W}\mathbf{V}^T$. Thus, Equation 2.4 can be written as

$$\mathbf{C}_S = \frac{1}{m}\mathbf{A}\mathbf{A}^T = \frac{1}{m}\mathbf{U}\mathbf{W}\mathbf{V}^T\,\mathbf{V}\mathbf{W}\mathbf{U}^T = \frac{1}{m}\mathbf{U}\mathbf{W}^2\mathbf{U}^T, \tag{2.5}$$

where the columns of matrix

$$\mathbf{U} = \begin{pmatrix} | & | & \cdots & | \\ \mathbf{s}_1 & \mathbf{s}_2 & \cdots & \mathbf{s}_{m-1} \\ | & | & \cdots & | \end{pmatrix} \in \mathbb{R}^{3n \times (m-1)} \tag{2.6}$$

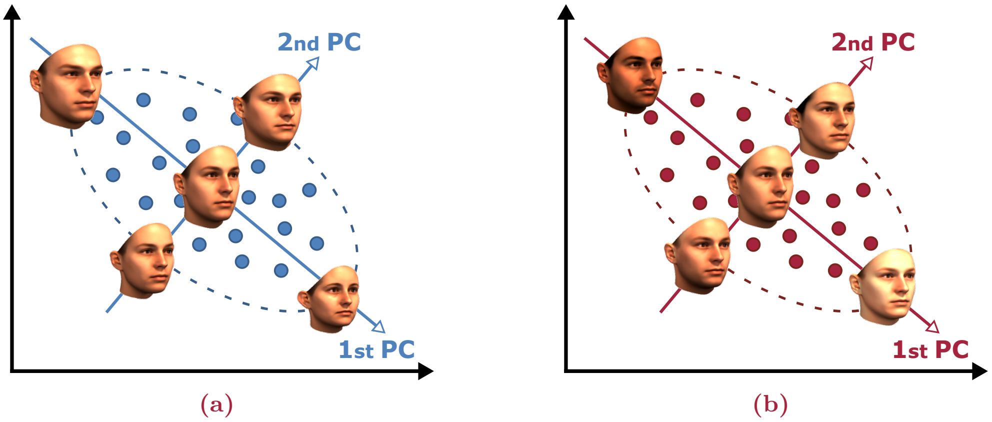


**Figure 2.5:** The shape, texture and expression are defined by individual PCAs in the 3DMM which span the face space. Above the shape (Figure 2.5a) and texture (Figure 2.5b) variations based on the the first two principle components (PC) are visualized.

are the eigenvectors $\mathbf{s}_i$ of $\mathbf{C}_S$ and the diagonal matrix $\mathbf{W} = \mathrm{diag}(w_{S,i})$ contains the eigenvalues which are the squared standard deviations $\sigma_{S,i} = \frac{1}{\sqrt{m}} w_{S,i}$ of the data along each eigenvector. As all eigenvectors are sorted by the value of their corresponding variances $\sigma^2_{S,1} \geq \sigma^2_{S,2} \geq \dots$ in descending order, the first eigenvector $\mathbf{s}_1$ represents the direction of the largest variance in the facial shape space. In this orthogonal basis the facial shape $\mathbf{S}$ is represented by

$$\mathbf{S} = \bar{\mathbf{s}} + \sum_{i=1}^{m-1} c_{S,i}\, \sigma_{S,i}\, \mathbf{s}_i = \bar{\mathbf{s}} + \sum_{i=1}^{m-1} \alpha_i\, \mathbf{s}_i = \bar{\mathbf{s}} + \mathbf{U}\boldsymbol{\alpha}, \tag{2.7}$$

where the coefficient vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_{m-1})^T \in \mathbb{R}^{m-1}$ contains all individual model coefficients $c_{S,i}$ of the shape multiplied by the standard deviation $\sigma_{S,i}$, so that $\alpha_i = c_{S,i}\, \sigma_{S,i}$. In total $m - 1$ principle components $\mathbf{s}_i$ with variances $\sigma^2_{S,i}$ are computed for the facial shape.

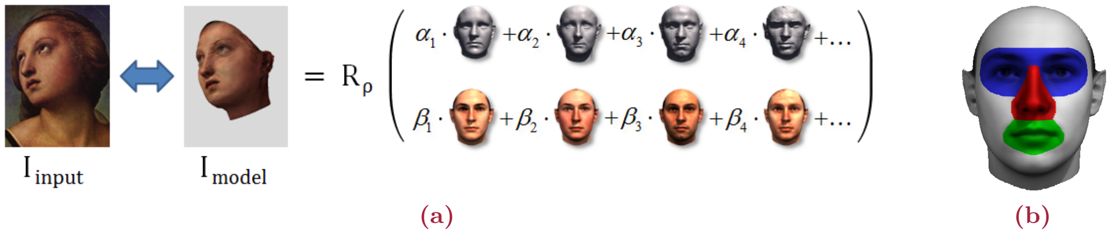
As mentioned above, the same procedure can also be applied to the facial expression and texture data. Accordingly, new faces can be approximated by the linear combinations

$$\mathbf{S} = \bar{\mathbf{s}} + \sum_{i=1}^{m-1} \alpha_i \mathbf{s}_i + \sum_{i=1}^{p-1} \gamma_i \mathbf{u}_i, \qquad \bar{\mathbf{s}} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{S}_i, \tag{2.8}$$

$$\mathbf{T} = \bar{\mathbf{t}} + \sum_{i=1}^{m-1} \beta_i \mathbf{t}_i, \qquad \bar{\mathbf{t}} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{T}_i. \tag{2.9}$$

While a separate linear combination is used for the facial texture (see Equation 2.9), the facial expressions are added to the linear combination of the facial shape in Equation 2.8.

To model individual variations in shape and texture (see Figure 2.5) the most significant


$I_{\text{input}}$
$I_{\text{model}}$
$= R_\rho \left( \alpha_1 \cdot + \alpha_2 \cdot + \alpha_3 \cdot + \alpha_4 \cdot + \ldots \quad \beta_1 \cdot + \beta_2 \cdot + \beta_3 \cdot + \beta_4 \cdot + \ldots \right)$

(a) (b)

**Figure 2.6:** Fitting the 3DMM to an input image like in Figure 2.6a is equivalent to searching for the best matching combination of shape and texture coefficients. In a subsequent refinement step the coefficients are adjusted for each facial region (see Figure 2.6b) separately to further enhance expressiveness. Please note that non-neutral facial expressions are omitted in this example.

99 eigenvectors $\mathbf{s}_i$ and $\mathbf{t}_i$ are used, plus 4 eigenvectors $\mathbf{u}_i$ for the most important degrees of freedom of facial expressions, with a focus on mouth movements. This PCA-based compression reduces the overall data consumption without abandoning essential information about the original face scans. Finally, using separate PCAs and basis vectors for the facial shape and expression enables to model 3D faces with neutral expressions even if the fitting has been performed on input images or scans with non-neutral facial expressions. This only requires to use $\gamma_i = 0$ in Equation 2.8 when the facial shape is reconstructed.

As an alternative to the described PCA-based approach, Probabilistic PCA (PPCA) [TB99] could be used as is suggested by Lüthi et al. [LAV09, LBA+12]. PPCA addresses the fact that, when creating the generative 3DMM, a finite number of example faces has been used and that the 3D scans contain some noise. As a result, PPCA might be superior when estimating the probability if a given face is valid or not. However, the PPCA-based approach is out of the scope of this dissertation.

## 2.2 Fitting to 2D Images

A 3D shape reconstruction by fitting the model to a 2D image is essentially a minimization of the image distance

$$d_{\text{image}} = \sum_{u,v} \| I_{\text{input}}(u,v) - I_{\text{model}}(u,v) \|^2 \tag{2.10}$$

in all three color channels, with respect to the linear coefficients $\alpha_i$, $\gamma_i$, $\beta_i$ and some imaging parameters $\rho_i$ that control pose, lighting and additional rendering parameters (see Figure 2.6a). Accordingly, the 2D input image $I_{\text{input}}(u,v)$ is compared to the rendered and perspectively projected 3D model to the 2D image $I_{\text{model}}(u,v)$ (for more details see [BV99]). Overfitting is avoided by a regularization term which is the Mahalanobis distance $d_{\text{maha}}$ from the starting conditions,

$$d_{\text{maha}} = \sum_i \frac{\alpha_i^2}{\sigma_{S,i}^2} + \sum_i \frac{\gamma_i^2}{\sigma_{U,i}^2} + \sum_i \frac{\beta_i^2}{\sigma_{T,i}^2} + \sum_i \frac{(\rho_i - \overline{\rho}_i)^2}{\sigma_{R,i}^2}, \tag{2.11}$$

where individual standard deviations of shape $\sigma_{S,i}$, facial expression $\sigma_{U,i}$ and texture $\sigma_{T,i}$ are estimated by PCA. As mentioned above, $\rho_i$ denotes the rendering parameters, while $\overline{\rho}_i$ are their starting values and $\sigma_{R,i}$ are the ad-hoc estimates of their standard deviations. Altogether the regularization term $d_{\text{maha}}$ penalizes solutions with low prior probabilities. Further details can be found in Blanz et al. [BBPV03].

A stochastic newton optimization algorithm minimizes the weighted sum of $d_{\text{image}}$, $d_{\text{maha}}$ and an additional term

$$d_{\text{features}} = \sum_j \| \begin{pmatrix} u_j \\ v_j \end{pmatrix} - \begin{pmatrix} P_u(x_{k_j}, y_{k_j}, z_{k_j}) \\ P_v(x_{k_j}, y_{k_j}, z_{k_j}) \end{pmatrix} \|^2 \tag{2.12}$$

which is the sum of squared distances between 2D feature positions $u_j, v_j$ and the projected positions of the corresponding vertex $k_j$, with a perspective projection $P$ [BV03]. $d_{\text{features}}$ is primarily used for initialization to pull the face model to the approximate position in the image plane during the first iterations of the fitting procedure. Its weight decreases as the fitting proceeds.

In summary, the goal is to minimize the overall error

$$E = \eta_{\text{image}} \cdot d_{\text{image}} + \eta_{\text{features}} \cdot d_{\text{features}} + \eta_{\text{maha}} \cdot d_{\text{maha}} \tag{2.13}$$

in an analysis-by-synthesis loop using the heuristically predefined weights $\eta_{\text{image}}$, $\eta_{\text{maha}}$ and $\eta_{\text{features}}$ starting with a conservative fit, where the values of $\eta_{\text{maha}}$ and $\eta_{\text{features}}$ are high. Accordingly, at the beginning the model is held close to the mean face and the selected features are highly trusted.[1] But while iteratively progressing in the analysis-by-synthesis loop the weights $\eta_{\text{maha}}$ and $\eta_{\text{features}}$ decrease.

To gain an increased expressiveness, the faces are divided into independent sub-regions like eyes, nose, mouth and a surrounding region which are optimized independently (see Figure 2.6b). The details of this Segmented Morphable Model are described in [BV99].

## 2.3 Illumination-corrected Texture Extraction

The result of the 2D model fitting algorithm in Section 2.2 is a textured 3D model of the face. Besides the shape vector **S** in Equation 2.1 an additional optimal linear combination of database vectors forms the texture vector **T** in Equation 2.2 which contains one set of RGB color values per vertex. However, for several reasons it is desirable to have true

[1] This procedure has some drawbacks in the context of imprecise or automatically selected landmark positions (see Chapter 5 and 6).

texture mapping with high-resolution textures which originate from the processed input images. First, the linear combination in Equation 2.2 has a limited number of degrees of freedom as it can only reproduce structures that are found in at least one of the database faces. Secondly, due to the fact that the resolution of the texture vectors match the resolution of the 3D shapes, only one RGB value per vertex is stored in the 3DMM. This leads to a relatively low texture resolution. Thus, fine details like eyelashes or birthmarks cannot be reproduced with the model-based approach directly. For capturing these details a high resolution texture needs to be extracted from a (high-resolution) input image. Last but not least even on blurred images, there may be individual characteristics on a low spatial frequency domain that are not in the degrees of freedom of the 3DMM, for example larger blemishes, or facial hair (see Section 4.3).

As the linear combination of texture vectors can not capture these individual details from the photo, Blanz and Vetter [BV99] propose a texture extraction procedure to map them to the model. They define $T_{\mathrm{extr}}(u, v)$ to be a 2D RGB texture map for the facial mesh with a resolution that is appropriate enough to capture all facial details from the input image. Depending on the individual requirements, mostly an overall texture resolution of $1024 \times 1024$ or $2048 \times 2048$ is used. As the 2D fitting procedure which is described in Section 2.2 already established a dense correspondence between the 3D model and the 2D input image, the resulting 2D input image position $u_k, v_k$ for each vertex $k$ which corresponds to the projection of the 3D vertex position is already known.

Now, each vertex $k$ of the model's triangular mesh has a $u_k, v_k$ coordinate in image space as well as a $u, v$ coordinate in texture space. The remaining color values for the neighbored texels in the texture map $T_{\mathrm{extr}}(u, v)$ are determined by calculating their barycentric coordinates and using these coordinates to sample the corresponding color values at position $u_k, v_k$ in image space. To reduce aliasing artifacts, a bilinear interpolation between four adjacent pixels in the input image is performed.

Up to this point, the mapping procedure preserves the illumination effects of the input image $I_{\mathrm{input}}$ and therefore prevents to render new poses and new illuminations correctly. Thus, it is important to separate the pure albedo from the influence of shading and cast shadows by using an illumination-corrected texture extraction to invert the effects of lighting (see Figure 2.7). In [BV99] Blanz and Vetter make use of the fact that the 3D shape, pose and illumination of the face are known after performing the fitting procedure because they are among the optimized parameters (see Equation 2.11). Given $I_{\mathrm{input}}$, during the sampling process the algorithm inverts the effect of color contrast, subtracts the specular reflection using the surface normals and inverts the effects of Lambertian shading. Finally the pure albedo is stored in the 2D texture map $T_{\mathrm{extr}}(u, v)$.

Subsequent rendering will then again multiply the reflectance with the Lambertian shading, add specularities and change the color contrast to obtain a realistic view in new

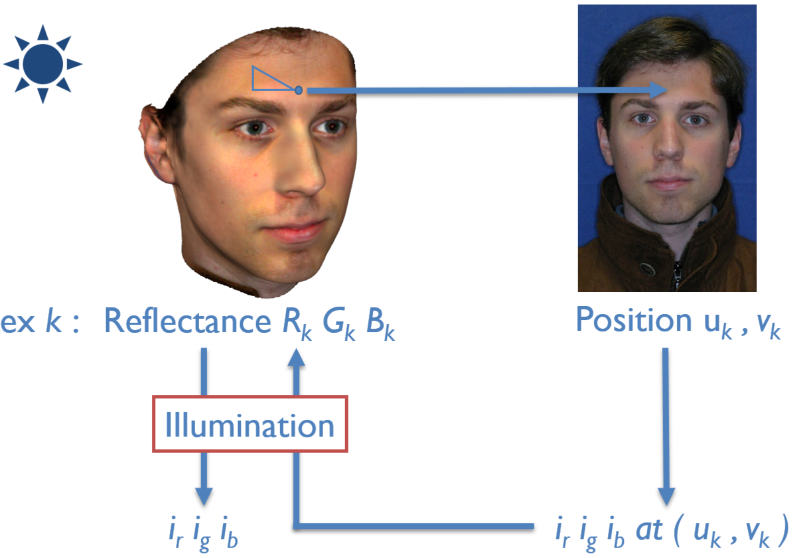

Vertex $k$ : Reflectance $R_k$ $G_k$ $B_k$
Position $u_k$ , $v_k$
Illumination
$i_r$ $i_g$ $i_b$
$i_r$ $i_g$ $i_b$ at ( $u_k$ , $v_k$ )

**Figure 2.7:** For each vertex $k$ the reflectance $(R_k, G_k, B_k)$ is estimated. By using Phong shading to simulate the lighting situation in the scene the corresponding intensity can be determined which is then compared with the real color value in the input image. Afterwards, the reflectance of each vertex $k$ is updated until the intensities of the rendering match with the input image. When extracting the texture from the image the determined reflectance values are used to invert the effects of illumination.

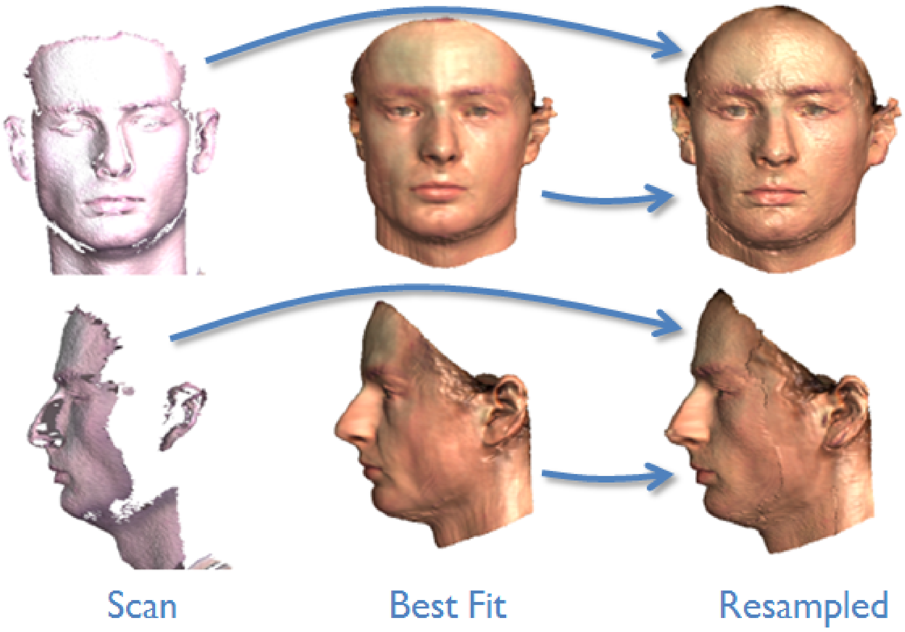
rendering conditions. Accordingly, using the estimated pose and lighting of the original photo for a subsequent rendering of the reconstructed shape $\mathbf{S}$ combined with the extracted and illumination-corrected texture $T_{\text{extr}}(u, v)$ will exactly reproduce the input image $I_{\text{input}}$.

However, texture extraction from a low resolution input image would remove the details introduced by the 3DMM. But this problem is solved by a modified procedure which is presented in Chapter 4.

## 2.4 Fitting to 3D Scans

Besides fitting the 3DMM to a 2D image, which has been discussed in Section 2.2, it can also be used directly to transform raw 3D scans into a PCA-based representation of faces (see [BSS07]). Like for the original creation of the 3DMM as described in Section 2.1 and the 2D image fitting, the 3D shape fitting algorithm needs a set of feature points for initialization.

For the initial creation of the 3DMM a modified optical flow algorithm was used to establish point-to-point correspondence between the model and the scan based on such feature points and then a PCA was applied to the cartesian coordinates of surface points. In contrast to that, the 3D shape fitting can be seen as a generalized problem of fitting to images. Assuming that the 3D scans are parametrized by image coordinates $u, v$ after



Credit: Blanz et al. [BSS07]

**Figure 2.8:** The 3D-to-3D fitting procedure establishes a correspondence between the 3DMM and the 3D shape data of the scanner. Afterwards the measured 3D point cloud data is transferred to the corresponding vertex positions of the 3DMM. For example, due to (self-) occlusion, for some areas no 3D point cloud data might be available. The missing information is filled in by the statistical model to gain a plausible result [BSS07].

applying a perspective projection, Blanz et al. [BSS07] propose to expand $I_{\text{input}}(u, v)$, so that each sample point stores not only the texture components $r, g, b$ but also the cartesian coordinates $x, y, z$. Then a 3D scan can be written as

$$\begin{aligned} I_{\text{input}}(u,v) = \quad ( \quad & r(u,v),\ g(u,v),\ b(u,v), \\ & x(u,v),\ y(u,v),\ z(u,v) \quad )^T. \end{aligned} \tag{2.14}$$

Accordingly, the optimization procedure to find the shape and texture vectors, the rigid pose transformation, camera parameters and lighting is equivalent to the one described in Section 2.2. Moreover, the determined perspective projection can also be used to fit the cartesian coordinates of the surface points to the model-based coordinates. Nevertheless, the extrinsic camera parameters as well as the intrinsic camera parameters (i.e. focal length) may differ depending on the manufacturer and type of the scanner. Therefore, these parameters need to be estimated at the beginning but stay unaltered for subsequent model fitting as described in [BSS07].

After the fitting procedure shape vector $\mathbf{S}$ and texture vector $\mathbf{T}$ are linear combinations of the faces of the 3DMM that form the optimal face representation with respect to the 3D input scan. Please note that all the details of the original 3D scan may not have been captured at this stage as only a best match to a given face within the range of the morphable model has been found, but so far no new dimensions have been added to the vector space of faces. Blanz et al. [BSS07] propose an extension of the original algorithm from [BV99] to overcome these limitations and to sample all the new facial details from the 3D scan.

Therefore, it is necessary to find the coordinates of the scanned point in camera coordinates

$$\mathbf{w_k}(u_k, v_k) = (w_{x,k}, w_{y,k}, w_{z,k})^T = \mathbf{R}_\gamma \mathbf{R}_\theta \mathbf{R}_\phi \mathbf{x}_k + \mathbf{t}_\omega \tag{2.15}$$

by predicting the image plane position $u_k, v_k$ for each vertex $k$ based on the optimized model and camera parameters, where $\mathbf{x}_k = (x_k, y_k, z_k)^T$ denotes the model-based coordinates in $\mathbf{S}$, angles $\phi$ and $\theta$ are the in-depth rotations around the vertical and horizontal axis, $\gamma$ controls the rotation around the camera axis and $\mathbf{t}_\omega$ defines the spatial shift. If the camera is located in the origin the perspective projection of the vertex $k$ to the image plane coordinates $u_k, v_k$ is defined as

$$u_k = u_0 + f\frac{w_{x,k}}{w_{z,k}}, \qquad v_k = v_0 + f\frac{w_{y,k}}{w_{z,k}}, \tag{2.16}$$

where $f$ is the focal length and $u_0, v_0$ are the coordinates of the principal point.

To obtain the corresponding vertex coordinates for $\mathbf{x}_k$ Blanz et al. [BSS07] propose to invert the rigid transformation in Equation 2.15 with the model parameters of the fitting procedure. Finally, the shape details can be sampled at point $u_k, v_k$ of the original 3D scan as is shown in Figure 2.8. Please note that the estimated value is only replaced if the sample point is not void and if the distance to the estimated position is below a given threshold. The latter avoids that outliers in the 3D scans are incorporated into the final results.

This 3D-to-3D fitting approach can be used to add new faces to the existing 3DMM and to increase its dimensionality to get beyond the existing facial space, for example, to integrate the 3D shape information of children to the 3DMM of faces as is shown in Chapter 3.

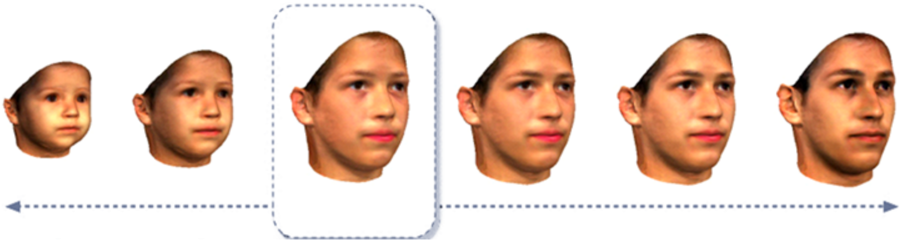
# 3

# Aging Simulation with the 3D Morphable Model of Faces

In Chapter 2 the analysis-by-synthesis approach of the 3DMM to handle varying head poses and illuminations was described from a more general point of view. By explicitly modeling the influence of these parameters, a comparison between different faces is possible, even for large varieties in pose and illuminations in the original input images. In this chapter, it is shown how the 3DMM can be used to model aging effects on faces. Again, the effects are modeled explicitly and are used in a growth simulation module which is based on learned longitudinal data. This module is part of an age-independent face recognition framework and has been developed to aid the investigation of child abuse cases.

## 3.1 Introduction

A lot of the existing face recognition software can not be applied directly to images of children. One simple reason is that the available software is just not optimized for the specifics of a child's face, but this can be corrected by using training data that includes not only faces of adults but also faces of children. A much harder to solve problem is caused by the effects of aging which lead to substantial transformations even during a relatively short period of time. Accordingly, face recognition software as well as humans struggle to recognize children after they have aged a few years. As mentioned, one common practice for learning and training based algorithms is based on the paradigm to find adequate training data in regard to the real world data and conditions. Therefore, a primary goal is to include the aging of children in a model-based system. Unfortunately, when starting the project described in this chapter, no databases were available that contain series of 3D scans of the same children over a time period of several years. To overcome this drawback the necessary longitudinal data is extracted from 2D image databases or by averaging



Credit: Blanz [Bla10]

**Figure 3.1:** The aging process of a person causes individual changes in texture (dermal characteristics) and changes to the overall shape of a face. The 3DMM takes both effects into account.

3D data of several children within certain age spans. This longitudinal data allows for an age-independent recognition of children and for the simulation of the aging process itself (see Figure 3.1). After simulating or neutralizing the facial transformations caused by aging, these synthetically aged faces can be used as input for already available face recognition systems which are otherwise sensitive to the effects of aging and would most likely fail without these preprocessing steps.

Besides improving the quality of 3D reconstructions with a neutral facial expression, another important part of this work is the processing of non-neutral facial expressions in input images of children. Although [BV99] already provided support for reconstructing facial expressions, these expressions were limited to the visemes of adults. Actually, these limitations are even more restricted because only the visemes of one single person were captured. Therefore, additional 3D scans have been recorded which capture not only the face geometry for each child with a neutral expression but also varying visemes (see Chapter 3.7). Then this additional geometry information is integrated to the PCA for facial expressions to extend the degrees of freedom for facial expressions and to improve the simulation of the facial expressions especially of children.

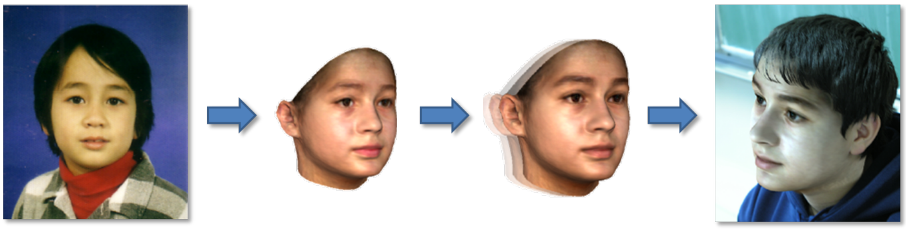
The presented work is part of a Federal Ministry of Education and Research (BMBF) project called INBEKI [Kru09] which focuses on the development of image processing methods to aid the investigation of child abuse cases. By supporting the investigation authorities in handling all relevant data and identifying the victims, this system aims to facilitate and accelerate the investigations to find the responsible criminals. The overall system was created in cooperation with several project partners including L1-Identity Solutions, rola Security Solutions, State Criminal Police Office North Rhine-Westphalia and German Research Center for Artificial Intelligence. The system has been developed between September 2009 and October 2012.

Besides implementing methods for age invariant face reconstruction and recognition, some of the major goals are generating neutral facial expressions from non-neutral faces and rendering the face in a frontal pose even if the original image captured the face in a non-frontal pose. This kind of image normalization is an important preprocessing step for the subsequent face recognition system, which is provided by another project partner. In this context, enabling the 3DMM to reconstruct children's faces more adequately made it possible to take advantage of some additional capabilities of the 3DMM that potentially enhance the results of a subsequent face recognition system: For example, after fitting the 3DMM to a 2D image that shows only the side view of a person, the face can be virtually rotated to a frontal pose. Furthermore, lighting effects and facial transformations caused by facial expressions can be normalized. Furthermore, occlusions caused by other objects in the scene can be countered.

This chapter is subdivided into the following sections: At first Section 3.2 summarizes related work. In Section 3.3 the requirements to create a 3D face model for a synthetic aging simulation are described. Subsequently, Section 3.4 discusses the procedure and the equipment that is used to acquire the necessary data. Additionally, it is shown why the data which was already available from [BV99] and [SSSB07] is not sufficient for an aging simulation that covers children and young adults between three and 18 years of age. In Section 3.5 the working steps after the acquisition of longitudinal data are described in detail. This part is crucial for aging simulation, because it shows how to extract the key face and aging information and how to compare the features of individual persons even if the input data (i.e. photos or video frames) is not limited to neutral facial expressions. Not only the 3D data of multiple scans can be combined but also the extracted textures of several, individual input images of the same person with a varying pose. This is discussed in Section 3.6. In Section 3.7 the handling of non-neutral facial expressions is described, accordingly the changes based on expressions need to be regarded in the facial space of the model. This work is based on [BV99], but increases the robustness by modifying the geometry processing especially in the area of the mouth. Unfortunately, compared to photos the acquisition of 3D scans is very time-consuming and requires special equipment. Accordingly, 3D datasets which contain longitudinal information of faces by having captured the same persons over several years are extremely rare or not existing at all. To overcome this problem Section 3.8 presents ideas to extract longitudinal information from 2D photo series and to apply the learned aging-related transformations to 3D face data. The results in the field of face reconstruction of children and aging simulation are shown in Sections 3.9 and 3.10. The coupling of the presented approach with an external face recognition module from a project partner is described in Section 3.11 in addition to a conclusion and an outlook.

## 3.2 Related Work

Although this chapter of the dissertation is based on technology introduced by Blanz and Vetter in [BV99, BV03], their work could not be applied directly to the aging simulation



Credit: Scherbaum et al. [SSSB07]

**Figure 3.2:** Starting with a 2D passport photo, a corresponding 3D model of the face is generated by the 3DMM. Afterwards, the shape and the texture of this 3D model is transformed based on a statistically determined and individually altered aging function. In a final step, it is possible to project the synthetically aged faces back into the original or into a different photo [BSVS04, SSSB07].

of children. One major drawback of the original 3DMM is that it is exclusively build from 3D scans of adults. This results in a limited performance to reconstruct the 3D shape and texture of children, especially for the desired age band (kindergarten and elementary school), as is discussed in Section 3.10.

However, Scherbaum et al. [SSSB07] have shown that the 3DMM is able to model aging effects, but again the covered age band (teenagers) is not sufficient for the desired goal. Nevertheless, some of their ideas and findings are incorporated in the work which is presented in this chapter: For example to use the learned statistical data from persons of different age to modify the original shape and texture parameters and to receive a synthetically aged 3D model of the face in the input image. This model can either be used directly as input for a face recognition module that works on 3D shapes or it can be projected back into a 2D image with any pose that may be required (see Figure 3.2). While [SSSB07] uses an Support Vector Regression, the proposed method supports both a piecewise linear interpolation and a curve-based estimation which is similar to [HBHP03].

Like the proposed work in this chapter, Shen et al. [SLSL11] focus on the growth simulation of children but they predict each facial feature individually and not in the context of the whole face.

Surveys about various age progression methods are provided by Ramanathan et al. [RCB09] and Fu et al. [FGH10]. The first approaches for modeling age progression are presented by Todd et al. [TMSP80], Burson and Schneider [BS81], and Burt and Perret [BP95] which are all limited to the image domain. The latter created an average image for two individual age groups after warping all images for 2D registration. The difference between both age groups is used to simulate age progression for novel faces. The creation of age groups is also used in the approach proposed in this chapter but it is transferred to a 3D domain now. This lifts the restrictions of the mentioned 2D approaches, including Lanitis et al. [LTC02], and the PCA-based methods of Scandrett et al. [SSG06b, SSG06a] and Tsai et al. [TLL14], which are limited to frontal or near-frontal face images.

Ramanathan and Chellappa use anthropometric measurements of a sparse set of facial features for growth prediction of adults [RC08] and for individuals which are younger than 18 years [RC06]. To improve the synthetic age progression for adults Wu et al. [WPLN99] focus on the simulation of wrinkles by using a synthetic texture, Patterson et al. [PSRB09] use an Active Appearance Model and Kumari and Dharmaratne [KD11] propose an image-based morphing technique. In this context, Ricanek and Tesafaye [RT06] published a longitudinal image database for adults, Patterson et al. [PSS14] establish a dataset for quantitatively evaluating the aging of adults, and another evaluating process is proposed by Lanitis et al. [LTS+15].

Suo et al. [SCS+12] concatenate several short-term aging patterns to create a long-term pattern and to avoid abrupt changes, while Kemelmacher-Shlizerman et al. [KSSS14] use thousands of photos to depict a prototype man and woman which is used to transfer aging effects to novel faces.

Recently, large datasets are analyzed to learn aging effects by using Dictionary Learning [STL+15], Recurrent Neural Networks [WCY+16] or Deep Restricted Boltzmann Machines [NLGB16]. To learn age transformations and preserve intrinsic subject-specific characteristics Yang et al. [YDWJ18] and Wang et al. [WTLG18] use Generative Adversarial Networks with a Convolutional Neural Network based generator.

Instead of modeling age progression, other publications focus on estimating the age of a person in an image. Lanitis et al. [LDC04] compared several different classifiers like shortest Mahalanobis distance, Neural Networks as well as a combination of these classifiers in a hierarchical method to determine to which age group a specific face belongs. They claim that the performance of machines is close to the age estimation ability of humans. Geng et al. [GZZ+06] create sequences of sparse, near frontal face images which are sorted in time order and use PCA to build up a subspace. Then, novel images are projected into this subspace, are reconstructed by using the coefficients of the known images and the minimum reconstruction error refers to the estimated age. The age estimation approach which has been proposed by Pan et al. [PHSC18] embeds the mean-variance loss and softmax loss into a Convolutional Neural Network.

Guo and Wang [GW12] analyzed the influence of facial expressions to age estimation results to develop a more robust system. To estimate the age based on a series of binary classifications Chang et al. [CCH10] use Active Appearance Models to compare the target's face to known images and Niu et al. [NZL+16] propose to use Deep Convolutional Neural Network for this task.

Finally, there are approaches for age-invariant face recognition. Like Naka et al. [NOA10] who use block matching to compensate position shifts of facial features during aging. Discriminative approaches [GM11, LPJ11, LSRJ10, OHJ12, CCH15, LGLT15, LGLT16] still contain some age-related features which interfere during the recognition process. To overcome

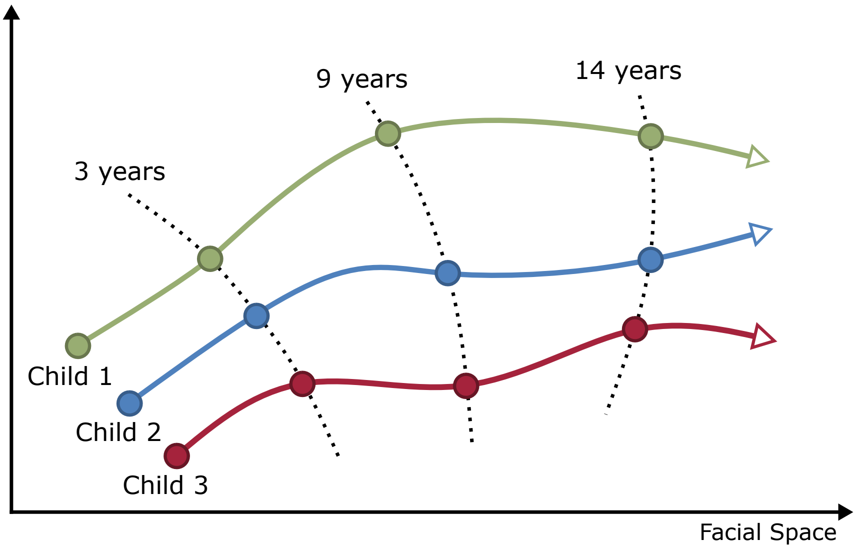


**Figure 3.3:** Besides individual facial features of a person at a specific age, the process of aging itself is individual for each child. Accordingly, a plausible growth or aging simulation is a nontrivial problem.

this drawback generative approaches like the one proposed by Park et al. [PTJ10] synthesize the target age before performing face recognition, while in [KSH12, SWT15, WLQ16] Convolutional Neural Networks are used to learn age-invariant face features.

Facial changes through age, various poses [LC05, BGPV05], lighting [ZS06] and expressions [ZLY+15] make facial recognition tasks even more difficult. Accordingly, databases like [YCS+08, CWZ+14, KSSMB16] have been introduced to support the learning of expression-based facial variations.

## 3.3 Requirements

A prerequisite for an age independent face recognition is to establish an age invariant representation of faces. Thus, in regard to the INBEKI project an age invariant representation of faces of children and teenagers is needed. One solution to achieve this goal, is the use of non-linear, individual aging trajectories, where the reconstructed face of each input image is assigned to a specific trajectory (see Figure 3.3). The distance between individual aging trajectories is used as a measure to determine the similarity of faces. This means that two faces are considered to be equal if they are located on the same aging trajectory.

As a basis for the research of possible solutions for age independent face recognition the 3DMM is used. As described in detail in Chapter 2 the 3DMM represents the shape and color of objects for a specific object class by a high-dimensional vector space of shape and texture vectors. In case of the object class for human faces a shape and a texture vector is assigned to each face. Linear combinations of these vectors generate new faces within this object class or more specifically within the facial space. This means that the linear span of a set of basis faces creates a continuum of realistic 3D face models. But this is only true if the basis faces are in dense point-to-point correspondence with each other (see Section 2.1 and 3.5), which means that the vector components of the shape and texture vectors always refer to the same surface structure, for example the tip of the nose as is shown in Figure 2.3a. Moreover, the 3DMM can be used to learn the differences between male and female faces from sample data or how the shape of the complete head or individual components (like nose or mouth) vary between different faces.

Another basis for this research is the work of Scherbaum et al. [SSSB07] who presented a system for a model based aging simulation. In addition to the 200 3D scans of adults of the original 3DMM, their extended model added 238 scans of teenagers using one 3D scan per person. Then these scans were used to generate a Morphable Model of teenager's faces as well as a non-linear model of aging trajectories by using Support Vector Regression. Using a 3D reconstruction algorithm and a method to exchange faces in images which is based on [BSVS04], they were able to place a synthetically aged or rejuvenated 3D face in the original photo or another photo. The latter allows to adapt even non-modeled features like hair style, physique or clothes of a another child which matches the target age to create a more plausible overall result (see Figure 3.2).

But their method cannot be applied to this research without additional modifications. First, the coverage and size of the dataset is not sufficient to guarantee robust results. Secondly, the desired age-set is only vastly or not at all represented in their scans: While they mostly focused on teenagers, this work needs to cover young children. Accordingly, additional and better matching 3D scans in terms of age needed to be captured for the purpose of the INBEKI project.

Although quantitative experiments in [SSSB07] showed that their non-linear age progression by using a Support Vector Regression results in less generalization error compared to a linear regression, the subtle differences were barely visible in a qualitative comparison. For that reason and because Support Vector Regression is a bit harder to control, in this work the Support Vector Regression is replaced in favor of a stepwise-linear or optionally a spline-based interpolation (see Section 3.9).

The method presented in this chapter is initialized with an arbitrary 2D photo of a child and creates a 3D reconstruction of the face based on the input photo by using the 3DMM (see Section 2.2) and by minimizing Equation 2.13. Unlike the strict specifications of biometric passport photos, the input photo may have been taken under arbitrary lighting conditions or may show arbitrary facial expressions and poses. As has been discussed in Chapter 2 theoretically, the 3DMM can adapt to all of these degrees of freedom as long as suited data was used to create the underlying face model. This is another reason why it is crucial to acquire additional 3D scans of children (see Section 3.4 and 3.5) and to learn the

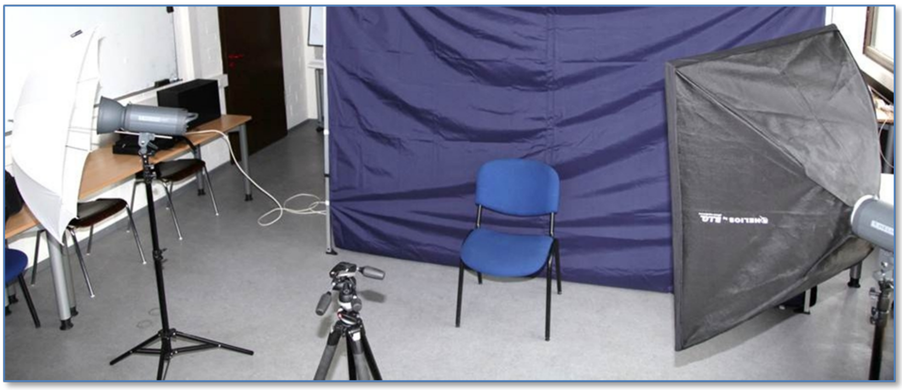


**Figure 3.4:** A DSLR camera mounted on the tripod in the foreground and two external flash units are used to capture high resolution images of a person under controlled lighting conditions. This guarantees uniform illumination of each face. Later, these photos are used to extract a high quality texture which is mapped onto the scanned 3D shape of the corresponding person.

specifics of the facial shape, texture and expressions of children (see Section 3.7).

Accordingly, the database of 3D faces was significantly increased compared to [SSSB07], especially by focusing on younger children. Finally, this ends up in covering the complete range from three year old children to 18 year old young adults, which enables age predictions exactly for the desired target group. In combination with the original 200 3D scans of adults, the possible range can even be expanded. However, it should be mentioned that the precision drops rapidly for large differences in age because of the unknown external factors that influence the aging of a person (like health, food, etc.). The improvements of this work in comparison to the original 3DMM are discussed in more detail in Section 3.10.

## 3.4 Acquisition of Photos and Scans

Over nearly a two-year period and as part of the INBEKI project [Kru09] the dataset was continuously increased by capturing additional 3D scans of children. The primary target group are children under 16 years, which are inadequately represented in the available datasets. For all newly added scans a point-to-point correspondence between each other and the existing datasets has been established by performing a 3D-to-3D fitting as is described in Section 2.4.

That way the formerly not sufficiently covered age intervals are updated and enhanced, so that especially younger children are adequately represented now. Furthermore, the age span between eight and 16 year old persons was not covered well enough to distinguish morphological variances of the faces caused by aging from random fluctuations of shape and texture within individual age groups of the dataset. This resulted in an undesirable side effect, where age spans are dominated by individual, particularly untypical faces.

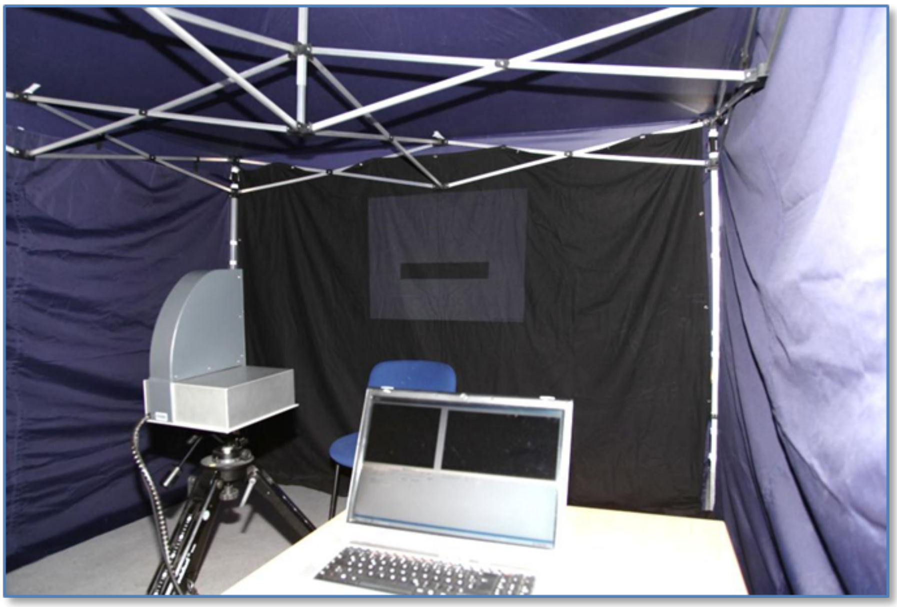


**Figure 3.5:** The structured light scanner projects a sequence of parallel lines that vary in width onto the face of a person (see Figure 3.6) and captures these 2D images. The patterns of these projected lines in the series of 2D images are used to estimate the 3D shape of the face. With this scanner it is possible to capture detailed 3D information but stray light of the environment may lead to artifacts in the scans. Therefore, the capturing process takes place in a dimmed environment inside a dark tent.

The extended dataset helps to limit these fluctuations to some degree by averaging. In a mathematical sense this is a typical overfitting problem in context of a regression task, accordingly a larger training dataset counteracts the overfitting effects.

To acquire the necessary data the children and teenagers of regional kindergartens, elementary schools and secondary schools and their parents were asked to support the INBEKI project. Their consent allowed to gather much-needed data which is used to train and improve the former face model. The scans and photos were taken on site and the equipment was designed to minimize additional effort for setup and calibration. Nonetheless, it needed to be ensured that even though the scans and photos are captured at different locations, the lighting conditions are comparable and the quality does not differ substantially.

The transportability of the equipment was a major requirement, because this made it unnecessary for the parents and their children to visit our laboratory. The idea behind this was to raise the odds for their approval by minimizing the hassle of the participants. In addition, it was guaranteed that the raw data is not published nor shared with our project partners and only used to compute a face model. But still the majority of the parents approached did not agree with taking 3D scans and photos of their children, so that it took longer than expected to gather a reasonable amount of facial data.

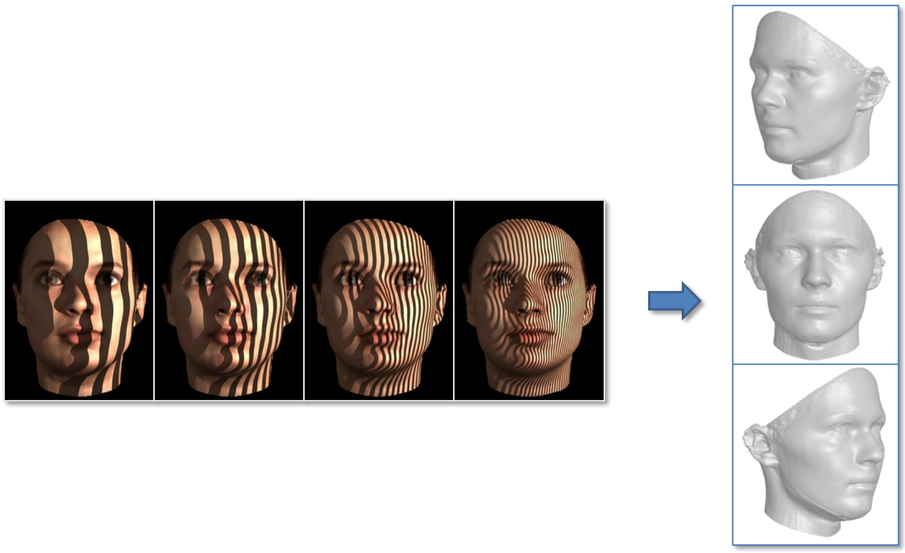


**Figure 3.6:** Based on the projected line patterns on the captured image sequence the corresponding 3D shape of the scanned object can be estimated. In each sequence the face is also captured without any line pattern which would allow to extract the facial texture directly from the photo. But due to the low resolution and the suboptimal lighting conditions the resulting quality would be clearly inferior compared to a photo taken by the described DSLR camera setup.

The equipment can be divided into two distinct groups: (1) A digital single-lens reflex (DSLR) camera and two separate flash units (see Figure 3.4) for capturing uniformly lit photos of the faces which can be used as a high-resolution facial texture map. To capture the complete face, three photos are taken (one frontal photo and two side views) of each person. (2) A 3D scanner plus a laptop that runs the appropriate software to capture the 3D geometry (see Figure 3.5). As the 3D scanner is sensitive to scattered light from the environment (see Figure 3.6), it is placed inside a tent. This setup also enables nearly constant lighting conditions for all locations.

With the presented equipment additional 3D scans of 141 persons (most of them were between three and twelve years old) and their facial expressions (visemes) were captured (see Figure 2.2). In total 813 new 3D scans have been added to the previously available face dataset of 238 persons (8-16 years). These additional scans are vital to improve the 3DMM of children and teenagers. In this context, it was important that besides the 3D scans from different views per person, also varying facial expressions had been captured. This allows not only to draw inferences about the varying appearances of different persons that sometimes also differ in age, but it also helps to predict which facial regions are affected by certain expressions and how exactly the shape of these facial regions is changed.

## 3.5 Combining Data from Multiple 3D Scans

After the acquisition of scans and photos (see Section 3.4) a series of post-processing steps are performed to combine the different views per face and to ensure point-to-point correspondence among all face data.

Although the structured light scanner captures most of the details of the face it has also several disadvantages. For example, in case of high reflective surfaces like eyes or surfaces that are not reflective at all like hair, the scanner can hardly reconstruct the shape. Instead a lot of outliers are generated which are positioned randomly in space. If these 3D positions are not removed or at least substantially reduced before fitting the 3DMM, they would interfere with the optimization procedure and would lead to visible artifacts in the resulting 3DMM face reconstructions.

Accordingly, at first the outliers need to be removed manually from each scan to ensure that mostly valid 3D information is used for the subsequent operations. For the manual removal of outliers from the point cloud data a proprietary 3D editing software which was bundled with the 3D scanner has been used, but of course software like MeshLab[2], CloudCompare[3], Autodesk Maya[4] etc. could have been used for this task as well. It can be well imagined that also an automated procedure could have been used to remove outliers at least to some extent. But then again it is unlikely that the manually procedure could have been avoided completely. In the end, having full control of the removal procedure was preferred to an overall less time consuming process.

Next, the different views which have been captured of each person (see Figure 3.7) are combined to gain a description of the face geometry which is as complete as possible. This primarily serves to fill wholes which occurred due to self-occlusion during the scanning process or because of outliers that have been removed in the previous step. Therefore, a simultaneous 3D fitting to all 3D scans of a person is performed. This procedure is based on the single scan fitting described in Section 2.4 and establishes a dense correspondence between the newly acquired data and the already available face data (see Figure 3.8). Defining a 3D scan as $I_{input}$ like in Equation 2.14, the 3D fitting is done by optimizing the overall error

$$E = \eta_{\text{image}} \cdot \sum_{view} d_{\text{image},view} + \eta_{\text{features}} \cdot \sum_{view} d_{\text{features},view} + \eta_{\text{maha}} \cdot d_{\text{maha}} \tag{3.1}$$

(see Equation 2.13) simultaneously for all available views. As has been described in Section 2.2 the overall error $E$ is a sum of the image distance $d_{image}$ (see Equation 2.10), the feature position distances $d_{features}$ (see Equation 2.12) and the Mahalanobis distance

[2] http://www.meshlab.net (Accessed on 2018-03-28)

[3] http://www.danielgm.net/cc (Accessed on 2018-03-28)

[4] https://www.autodesk.com/products/maya/overview (Accessed on 2018-03-28)

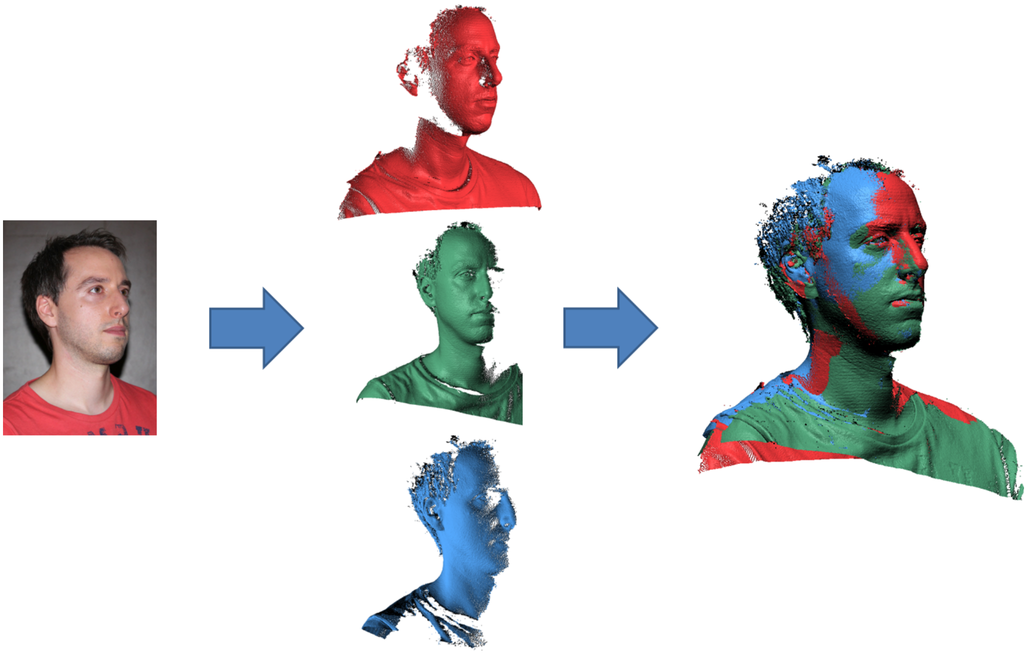


Figure 3.7: A 3D scan from a single view does not capture a face completely. Thus, the data from different views is combined to gain a 3D point cloud of the entire face.

$d_{maha}$ (see Equation 2.11). For the simultaneous fitting it is important to adapt the Mahalanobis distance to take the varying rendering parameters of each individual scan into account. This leads to the following equation, where

$$d_{maha} = \sum_i \frac{\alpha_i^2}{\sigma_{S,i}^2} + \sum_i \frac{\beta_i^2}{\sigma_{T,i}^2} + \sum_i \frac{\gamma_i^2}{\sigma_{U,i}^2} + \sum_{view} \sum_i \frac{\left(\rho_{i,view} - \overline{\rho}_{i,view}\right)^2}{\sigma_{R,i}^2}. \tag{3.2}$$

Please note that the described multifitting approach can only be performed if each scan captures the same person, so that it can be guaranteed that the shape and texture remain unchanged and that only the rendering parameters differ for each view. Furthermore, Equation 3.2 requires the same facial expression in all simultaneously fitted 3D scans[5]. But this constraint is satisfied because all sets of 3D scans with a varying pose are captured with a neutral facial expression. All 3D scans with varying facial expressions are fitted separately and by using only a single input scan, because only one 3D scan was captured per viseme.

Like in Section 2.2 the overall error $E$ is minimized by using a stochastic newton optimization algorithm, where $\eta_{\text{image}}$, $\eta_{\text{maha}}$ and $\eta_{\text{features}}$ are the predefined weights of the current processing step of the analysis-by-synthesis loop.

As a result, a high resolution triangle mesh is created based on the 3D point cloud data of the scanner and can be used for rendering and texture mapping. In addition, advanced

[5] In Section 5.5.1 a 2D multifitting approach is presented which is able to handle varying facial expressions.

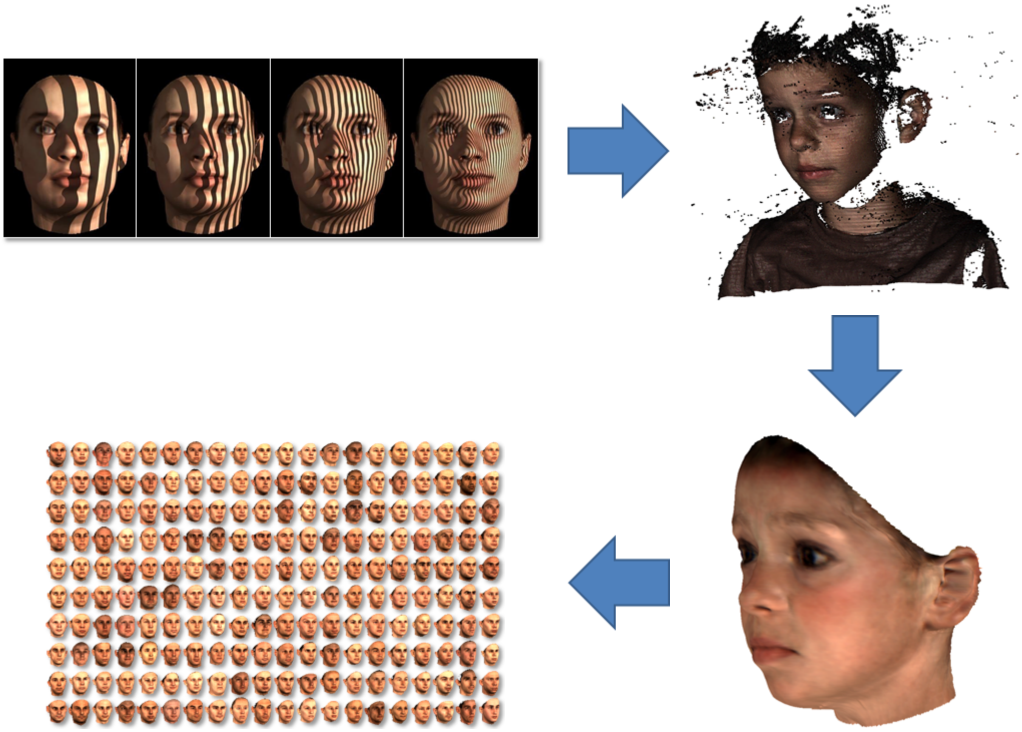


Figure 3.8: By using a structured light scanner the 3D geometry of the face is determined. Next, a triangulated mesh is computed by fitting the 3DMM to the point cloud of the scanned data. Finally, the newly acquired facial geometry is in dense point-to-point correspondence to all other faces of the 3DMM and can be incorporated to the statistical model itself.

processing steps like morphing are supported, because of the dense 3D correspondence between all faces and their vertex positions and colors.

Even more important is the fact that after the shape and texture vectors have been brought into correspondence to the existing face data, the newly gathered 3D information which is based on the additional faces can be used to extend the previously existing facial space by additional, plausible facial features (see Figure 3.8). For this purpose the 3D fitting approach by Blanz et al. [BSS07] is used: After performing the 3D fitting procedure which has been described above, the resulting shape vector $\mathbf{S}$ and the texture vector $\mathbf{T}$ are just linear combinations of the already available faces of the original adult face database and do not capture all important details of the new input data. To get beyond the currently existing facial space, after establishing a dense point-to-point correspondence, the 3D-to-3D fitting samples the new details directly from the input data to preserve the original shape and texture of the 3D scan as is shown in Figure 2.8. Further details about this procedure are provided in Section 2.4.

Accordingly, the 3D fitting of the 3DMM which has been described in Section 2.4 is used for bootstrapping the adult-based model and to include facial information of children and teenagers based on additional 3D scans. This renders the more complicated application of the optical flow that was used to initially create the 3DMM (see Section 2.1) unnecessary in this context.

## 3.6 Multiview Texture Extraction

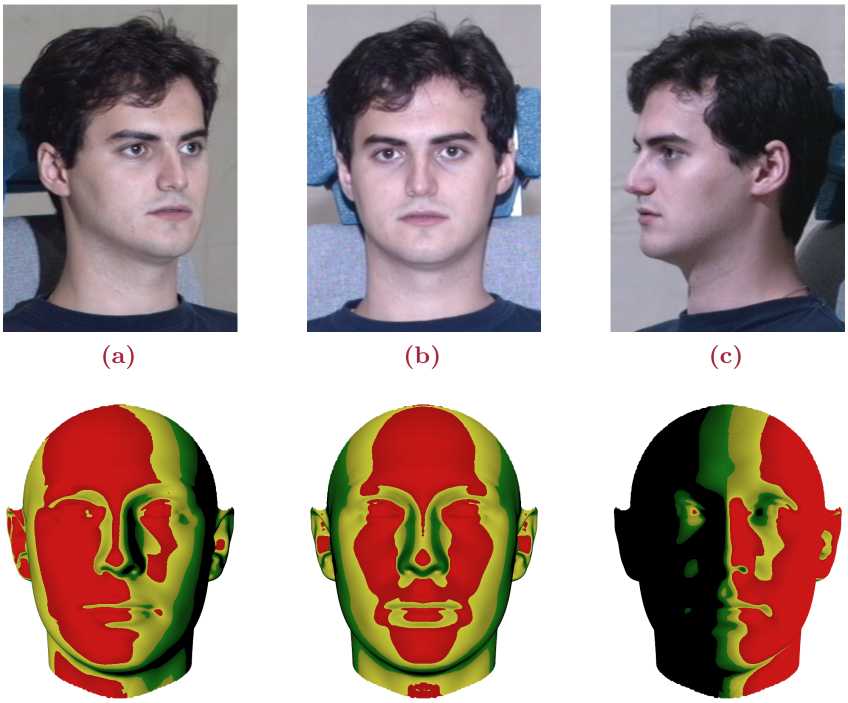
Texture extraction from a single 2D image is already included in the 3DMM presented by Blanz et al. [BV99] (see Section 2.3) and even texture extraction from multiple images in the context of a 3DMM is mentioned in [BBPV03]. Nevertheless, up to now there has been only a rudimentary implementation of this feature: If more than one input image is available to fit the 3DMM the texture is extracted sequentially in exactly the same order as the images are listed in a configuration file. Although the predefined order in which the color values are sampled is not crucial, the main problem of the current approach is that if a texel is covered by more than one facial texture, only the color values from the last defined texture are used in the final result. All other values are overwritten and therefore become irrecoverable even if they captured the facial texture more adequate than the most recently extracted texture values.

Mostly the texture with the highest available resolution, a constant lighting and no occlusions or strong shadows should be preferred. Thus, being aware of this problem, a partial solution could be to manually define the order of used images in such a way that the input image which covers the most parts of the face or which is expected to deliver the "best" overall quality is processed at last. This would ensure that the resulting facial texture contains the texels of the preferred image in any case.

Considering input images from three different views as is shown in Figure 3.9 would result in a possible sequence starting with input image 3.9c, followed by 3.9a and then 3.9b. But as can be seen in Figures 3.9d to 3.9f some facial regions are covered better by the input image in Figure 3.9c or Figure 3.9a than by the one in Figure 3.9b. Nevertheless, as most regions are covered by Figure 3.9b which is applied last – only black marked areas are not covered at all – the texture information based on Figures 3.9a and 3.9c would be overwritten nearly completely.

To solve the disadvantages of the current texture extraction implementation, a more sophisticated approach is proposed to improve the handling of multiple input images during texture extraction. Instead of using a predefined order the influence of each individual input image texture is determined by computing the angle between the normal directions of the reconstructed 3D shape and the viewing direction into the scene or photo. Accordingly, texels that are straight in the viewing direction are considered to be most accurate.

Before determining the angle the 3D shape needs to be transformed to the estimated pose that is found in the input image. This results in weight maps like the ones shown in the bottom row of Figure 3.9, where the influence of different extracted textures is



(a) (b) (c)

(d) (e) (f)

**Figure 3.9:** In Figures 3.9a to 3.9c three photos of the same person with varying poses are shown. The influence of the extracted textures for each pose based on the angle between the shape normal and the viewing direction is visualized in Figures 3.9d to 3.9f. In red regions the influence is highest, followed by yellow and green in a descending order. In areas that are marked black the extracted texture is not used at all.

visualized. Please note that for illustration purposes only four quality levels are shown: in the red area the influence of the corresponding input image is maximal, this influence decreases in the yellow area and even more in the green one. Everything which is colored black has no influence at all to the final result, because these parts are not visible in the input image based on the pose estimation of the 3DMM fitting. Unlike the artificial visualization in Figure 3.9 which only distinguishes between four levels of influence, the implementation of the combined texture extraction approach creates weight maps that contain fine nuances of different weights for each pixel from the available input images.

As mentioned above, the weight itself depends on the angle between the normal of the estimated 3D shape and the viewing direction. Accordingly, the weighting factor $\omega_{ij}$ is defined as

$$\omega_{ij} = \max\left( 0, \; \frac{\mathbf{n}_i \cdot \mathbf{v}_j}{\|\mathbf{n}_i\| \; \|\mathbf{v}_j\|} \right), \tag{3.3}$$

where $\mathbf{n}_i$ is the normal direction of the fragment $i$ and $\mathbf{v}_j$ is the estimated viewing direction into the reconstructed 3D scene of the $j$th 2D input image. To avoid negative values in

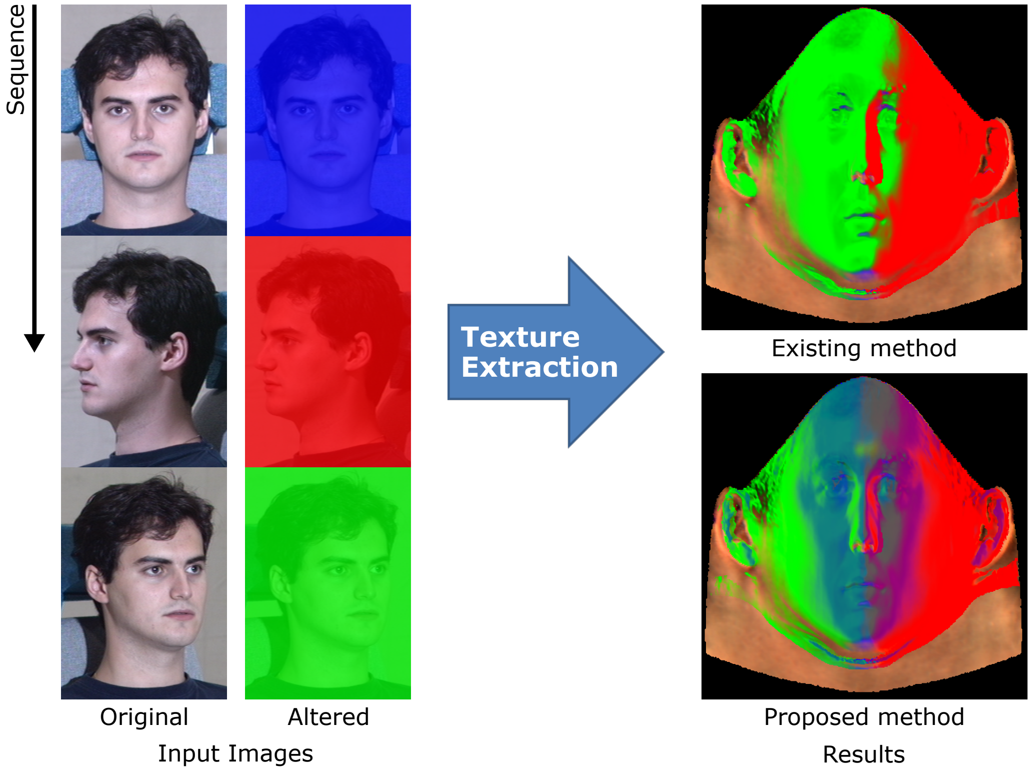


**Figure 3.10:** Processing the input images (left) from top to bottom the results of the corresponding texture extractions are shown on the right side. In case of the existing method the blue pixels from the first input image got mostly lost. Whereas the proposed method blends all pixels regardless of the predefined processing order.

cases where the angle between both vectors is greater than 90° the max function is used, so that $\omega_{ij} \in [0, 1]$. The exact normal direction of each fragment $i$ is determined through a barycentric interpolation of the corresponding vertex normal directions of the triangle mesh. Additionally, a z-buffer test is performed to check if a fragment is occluded or visible. If a fragment $i$ is not visible, the corresponding weight factor $\omega_{ij}$ is set to zero to prevent any influence of the texture extraction for this region.

Then the final color $c_i$ for fragment $i$ is defined as a weighting of the individual color values $\text{c\_tex}_{ij}$ based on the sampled textures from the $j$th input image:

$$c_i = \frac{\omega_{ij}}{\sum_{j=1}^{n} \omega_{ij}} \cdot \text{c\_tex}_{ij}. \tag{3.4}$$

The Equation 3.4 is valid if $\sum_{j=1}^{n} \omega_{ij} > 0$. Otherwise, if none of the available input images cover fragment $i$ which means that $\sum_{j=1}^{n} \omega_{ij} = 0$, instead of extracting the color value for the current fragment from the input images itself, it results from the linear combination of the 3DMM texture coefficients (see Equation 2.9). This is equivalent to the texture extraction from a single image which is described in Section 2.3. Accordingly, for the delighting of the textures the approach of Blanz and Vetter [BV99] is used. Please note that it is performed separately for each input image.

The proposed method for texture extraction and combination from multiple input images results in smooth transitions between textures that cover the same region. However, as the weighting factors are determined by computing the angle between the viewing direction and the shape normal, a wrong pose estimation may lead to errors. This can be seen by comparing the input image in Figure 3.9c with the resulting weight map in Figure 3.9f. While the right eye is not visible at all in Figure 3.9c, parts of this eye still have some influence on the final result based on Figure 3.9f, because the corresponding region contains green, yellow and even some red markings. Obviously, in this example the head rotation is estimated slightly wrong which lead to an incorrect weight map.

In Figure 3.10 a comparison between the original texture extraction method and the proposed one is shown. To visualize the influence of the individual input images to the resulting facial texture, for each image a distinct color overlay has been applied. Given that the input images on the left are processed from top to bottom the results of the corresponding texture extractions are shown on the right side. In case of the original approach it is clearly visible that the blue pixels from the first processed image have been nearly completely overwritten by the pixel colors of the two subsequent input images. This means that nearly all facial texture information of the first image got lost during the texture extraction process. Whereas in case of the proposed method the blue pixels are still present in the final result and are blended with the red and green pixels from the two other input images whenever a facial region is covered by more than just one input image. Accordingly, the order of the processed views does no longer affect the final result of the illumination-corrected facial texture extraction when using multiple views.

## 3.7 Facial Expressions

The INBEKI project demanded the support of facial expressions, because the majority of the images that needed to be reconstructed by using the 3DMM are expected to have non-neutral facial expressions. Furthermore, apart from passport photos, neutral facial expressions are extremely rare in photos, therefore a system for facial reconstruction and synthetic aging should not be limited to these kind of images. Accordingly, the acquired information about facial expressions need to be integrated into the facial space of the 3DMM for faces of children.

Before the INBEKI project only a single dataset existed that contained information about facial expressions and it covered only the visemes of a single female adult (see [BBPV03]) who is shown in Figure 2.2. In consideration of the different age spans that need to be supported, assuming constant 3D deformations for all persons and ages did not

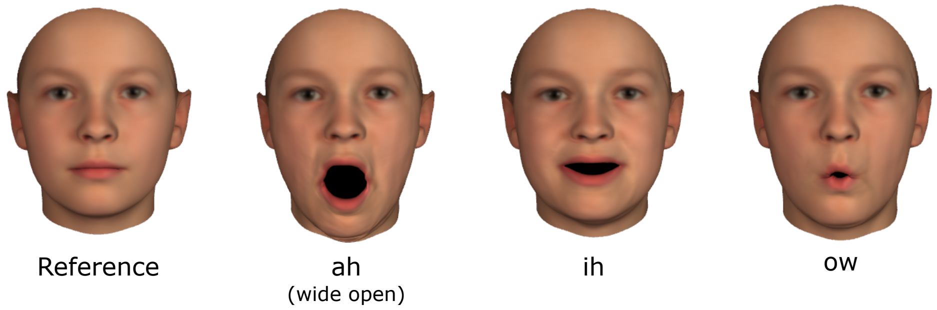
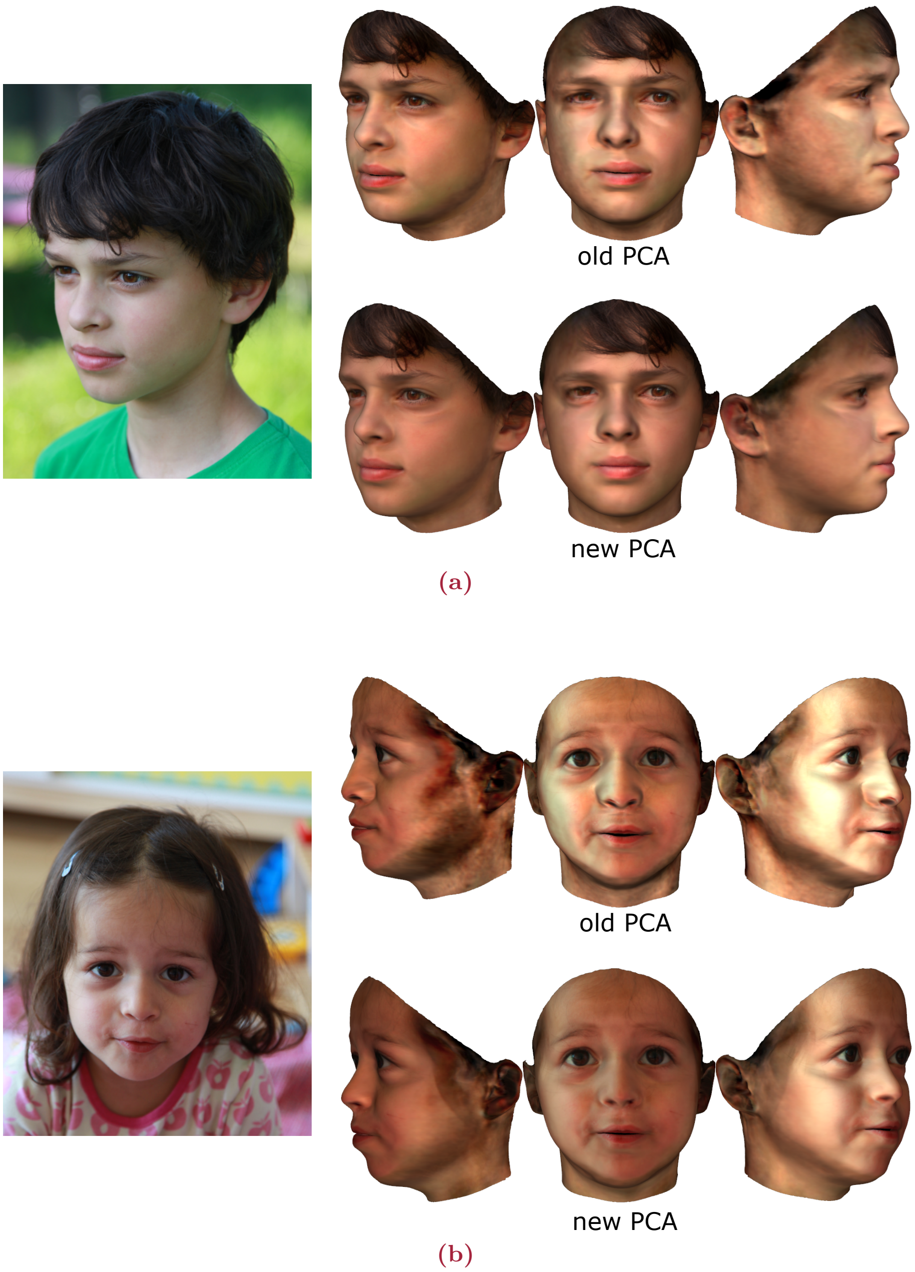


**Figure 3.11:** Besides capturing the 3D face geometry with a neutral facial expression, for each child an additional set of scans were performed to capture different visemes. After establishing a point-to-point correspondence between all scans, morphing can be used to combine individual visemes or to create an animation between different facial expressions.

seem sufficient. For example, the size and appearance of wrinkles that are caused by facial expressions differ a lot between children and adults. Therefore, additional training data for facial expressions of children are added to the existing 3DMM for the faces. As described in Section 3.4 besides capturing 3D scans of faces with a neutral facial expression also a series of visemes has been recorded for each child. But in contrast to the neutral faces that were captured from three different viewing angles, the visemes were only captured from a frontal view to keep the additional effort reasonable. Thus, the 813 captured scans contain not only 3D data of different views per person but also varying facial expressions. Establishing a correspondence between the vertices of each face and also between every facial expression (see Section 3.5), enabled to morph not only between the facial shapes of different persons but also between different facial expressions. And the ability to combine individual facial shapes to create a new one can also be transferred to the combination of several facial expressions or visemes.

Besides morphing between the captured facial expressions of one person, this technique also allows to transfer the facial expression from one person to another or to combine the same viseme of several persons to receive the mean of a specific facial expression as is shown in Figure 3.11. This can be helpful to analyze the specifics of a smile and then using these insights to transform a face with neutral expression into a smiling face or vice versa. Especially the latter is very important and helpful in regard to face recognition, because many face recognition approaches still struggle if faces with different expressions need to be compared. In the INBEKI project this problem is addressed by adding a preprocessing step that generates a neutral facial expression before a standard method for face recognition is applied. For replacing the non-neutral face in a photo with neutral variant the work of Blanz et al. [BBPV03, BSVS04] is adapted to faces of children by creating a dataset of 3D scans of children and their facial expressions as is shown in the previous sections of this chapter.



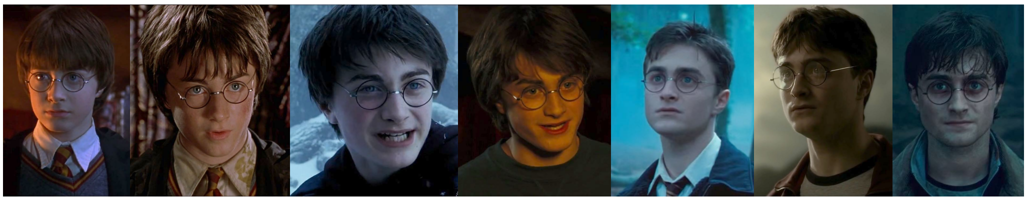
(a)

(b)

**Figure 3.12:** Two comparisons between the reconstruction results of the original 3DMM and the 3DMM with the new, extended dataset are shown. In Figure 3.12a as well as in Figure 3.12b only the underlying PCA model has been exchanged, but the same landmark locations and fitting parameters have been used for the 3D face reconstructions (right) from the 2D input images (left).

## 3.8 Extracting Longitudinal Information from Photo Series

In Section 3.4 the importance of additional 3D scans is described: After establishing a point-to-point correspondence between the newly acquired data of children and the existing face model of adults, previously missing age bands are supported now. A comparison



Credit: Images from author; The Harry Potter Saga [CCNY11]

**Figure 3.13:** The screenshots from the Harry Potter film series [CCNY11] show the actor Daniel Radcliffe at different ages. Such photo series or albums can be used to extract longitudinal information from a specific person by fitting the 3DMM to each image.

between the reconstruction results based on the old dataset and the new, extended dataset is given in Figure 3.12.

Nevertheless, the described approach is limited in regard of longitudinal data, because the available 3D scans do not capture the changes of the shape of a specific person over a long period of time but only show a snapshot of different persons at different ages. This accumulated data is indirectly used to infer longitudinal information.

During the project duration of INBEKI it was not possible to create a longitudinal dataset of 3D scans and there is no known access to a publicly available 3D dataset that corresponds to the longitudinal needs. To fill this gap, series of publicly available pictures which show the same person at different ages are analyzed. The photos of online photo albums like flickr, Google Images as well as private photo albums have been used to gain additional information about faces and their transitions over time. Other sources to collect longitudinal data based on a single person are the images of celebrities, especially if they started their career at a young age and performed in several movies over time (see Figure 3.13). Furthermore, there are some growth databases available which provide photos of the same person at different ages like VADANA [SRK11] or CACD [CCH15]. These and some additional datasets are reviewed by Ricanek et al. [RBS15]. Unfortunately, these datasets primarily contain photos of adults and the age gap between individual photos is quite huge which makes it hard to extract detailed aging trajectories from these photo collections.

However, a fruitful source are timelapse videos: photos of the same person that have been collected over a period of several years are used to create a movie about the facial transitions while growing up (see Table 3.1). Some of these videos are well planned by the parents of the child which results in a dense coverage. Accordingly, there are only a few days or weeks between each captured photo of the face. Therefore, the visible transformations are gradually which makes them likewise comprehensible and reproducible. This makes timelapse videos an optimal source to generate individual aging curves. Nevertheless, without having timelapse videos of a large number of different people, this approach provides limited feasibility for generalization. Furthermore, as these photos are used to

| from (years) | to (years) | time steps (estimate) | gender | url |
|---|---|---|---|---|
| 0 | 1 | every day | male | [Cat10] |
| 0 | 4 | every week | female | [Ste10] |
| 0 | 1 | every day | female | [Cat12] |
| 0 | 12 | every week | female | [Hof12a] |
| 0 | 9 | every week | male | [Hof12b] |
| 14 | 21 | every day | female | [Bro14] |
| 0 | 5 | every week | female | [Smy14] |
| 0 | 16 | every week | female | [Hof15] |
| 12 | 20 | every day | male | [Cor16] |
| 13 | 23 | every day | male | [Wil17] |
| 0 | 18 | every week | female | [Hof17] |

**Table 3.1:** Examples of time lapse videos that show the effects of aging. The persons themselves or their parents regularly captured a photo for several years. Then these photos have been used to create a video which shows the changes of a specific face over time.

create a video file, it is necessary to extract the individual frames before fitting the 3DMM to each individual photo.

Using the 3D reconstruction method of the extended 3DMM (see Section 3.5) which incorporated the facial data of children, enabled to extract longitudinal information about the 3D shape from these photo series to obtain a more precise model for the aging trajectories. In addition, extracting the aging trajectory of a specific person also enables to transfer the age-related features of person A to person B which works in a similar way as the transfer of facial expressions (see Section 3.7). Consequently, the individual facial features of the originally captured person are subtracted from the aging trajectory which results in a normalized description of the aging process. In a final step the facial features of another person are combined with these normalized aging features. This procedure enables to sample individual aging transformations and to transfer them arbitrarily to other persons. The details about the synthetic aging simulation and the application of aging trajectories are described in Section 3.9.

## 3.9 Synthetic Aging Simulation

In Sections 3.4 to 3.8 the acquisition of additional facial data and the necessary postprocessing steps to establish point-to-point correspondence and to combine shape and texture information of the gathered scans and photos are described. In the following, this facial information which has been extracted from persons of different age groups is used for a synthetic aging simulation on the basis of the established 3D facial space.

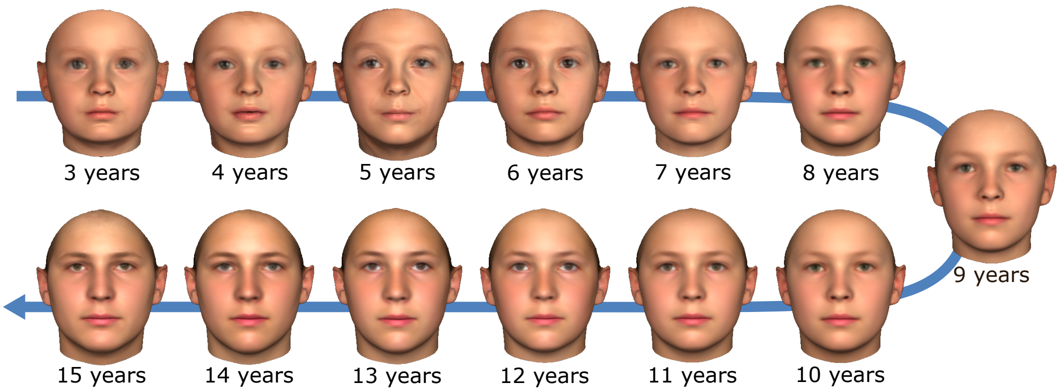


**Figure 3.14:** The corresponding mean face of each age level based on the captured faces (see Section 3.4). The number of available faces per year are shown in Table 3.2.

| Age (years) | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Number of scans | 8 | 7 | 5 | 6 | 17 | 28 | 32 | 48 | 85 | 39 | 54 | 25 | 13 |

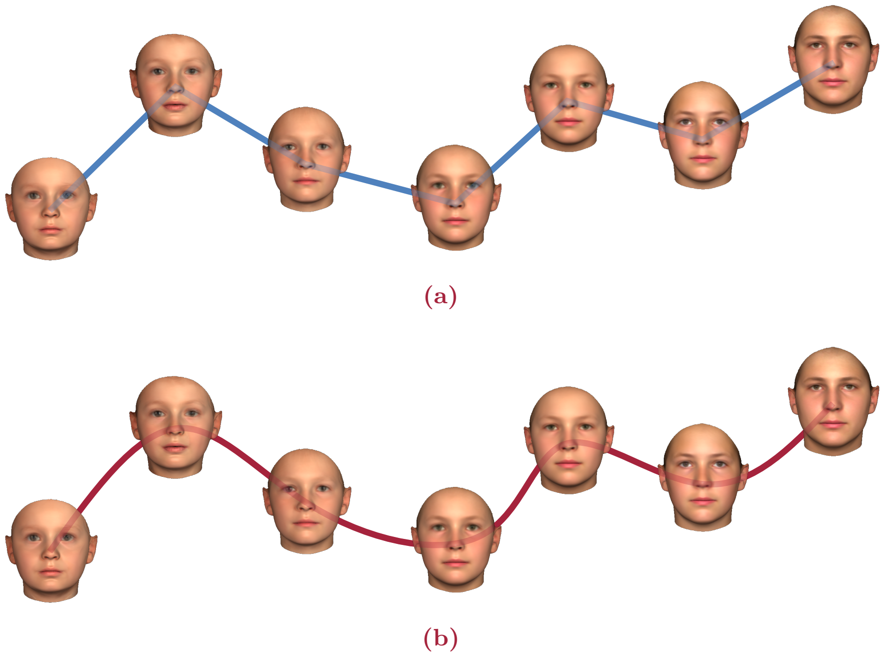
**Table 3.2:** For each full year a corresponding mean face is created (see Figure 3.14). In this table the number of used 3D scans per age interval with a neutral facial expression is shown.

In contrast to the creation of aging trajectories directly based on an individual as is discussed in Section 3.8, computing the mean of all faces which are related to a predefined age group allows to extract a more generalized aging trajectory. The resulting mean faces which are based on the available 3D data are shown in Figure 3.14. Furthermore, in Table 3.2 the numbers of available faces per age interval which are used to create each mean face are presented.

To enable smooth transitions between these mean faces a piecewise linear interpolating between neighbored mean faces is used for the morphing procedure (see Figure 3.15a). Alternatively, a spline-based approximation (see Figure 3.15b) is supported to calculate in-between aging transformations. For both approaches a series of mean faces is computed, where each mean face $\overline{\mathbf{P}}_i$ is related to a specific age range of the training data as is shown in Figure 3.14. Then

$$\overline{\mathbf{P}}_i = \frac{1}{n} \sum_{k=1}^{n} \mathbf{P}_k \tag{3.5}$$

is the corresponding mean face which is created for each age interval $i$ by computing the arithmetic mean for all $n$ faces $\mathbf{P}_k$ which share the same age. Although a lot of data has been gathered during the INBEKI project the overall amount of available face data for all persons which are between three and fifteen years old, is still not dense enough to generate meaningful mean faces for each quarter or half year, not to mention for each single month (see Table 3.2). Choosing a too short interval length would risk that individual, untypical



(a)

(b)

**Figure 3.15:** The example in Figure 3.15a illustrates the piecewise linear interpolation between mean faces of neighbored age levels (see Figure 3.14) to simulate the effects of aging. For the spline-based aging simulation in Figure 3.15b a curve is fitted through the age-dependent changes of a dataset.

faces have too much influence on the final result of the arithmetic means. Furthermore, it is not guaranteed that the data of more densely covered years is uniformly distributed over that time frame. Thus, the used interval length covers a full year.

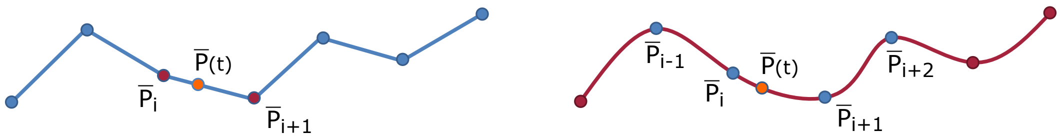
For simulating the growth of a face, the transformation which needs to be applied to the input face $\mathbf{P}_{\text{input}}$ is equal to the mean difference $\overline{\mathbf{P}}_{\text{diff}}$ between the interpolated mean value $\overline{\mathbf{P}}_{\text{target}}$ which corresponds to the target age and the appearance of the interpolated mean value $\overline{\mathbf{P}}_{\text{curr}}$ which matches the age of the input face, so that

$$\overline{\mathbf{P}}_{\text{diff}} = \overline{\mathbf{P}}_{\text{target}} - \overline{\mathbf{P}}_{\text{curr}}. \tag{3.6}$$

Then the synthetically aged face $\mathbf{P}_{\text{aged}}$ is computed by adding the mean difference $\overline{\mathbf{P}}_{\text{diff}}$ to the input face $\mathbf{P}_{\text{input}}$ by computing

$$\begin{aligned}\mathbf{P}_{\text{aged}} &= \mathbf{P}_{\text{input}} + (\overline{\mathbf{P}}_{\text{target}} - \overline{\mathbf{P}}_{\text{curr}}) \\ &= \mathbf{P}_{\text{input}} + \overline{\mathbf{P}}_{\text{diff}}.\end{aligned} \tag{3.7}$$

An equivalent interpretation of this procedure can be described as follows: all distinct features which make the input face unique compared to the mean face of the same age are preserved. These features are extracted by computing $\mathbf{P}_{\text{input}} - \overline{\mathbf{P}}_{\text{curr}}$ in Equation 3.8.



(a) (b)

**Figure 3.16:** The exact age-dependent mean appearance $\overline{\mathbf{P}}(t)$ is interpolated either by using a piecewise linear (see Figure 3.16a) or a spline-based approach as is shown in Figure 3.16b.

Then the mean appearance $\overline{\mathbf{P}}_{\text{target}}$ of the target age is added to these features to create the synthetically aged input image $\mathbf{P}_{\text{aged}}$, so that

$$\mathbf{P}_{\text{aged}} = (\mathbf{P}_{\text{input}} - \overline{\mathbf{P}}_{\text{curr}}) + \overline{\mathbf{P}}_{\text{target}}. \tag{3.8}$$

In case the current age of the input face does not hit the predefined mean faces exactly, the values of the nearest neighbors which enclose the given age need to be interpolated to determine the corresponding in-between facial appearance of $\overline{\mathbf{P}}_{\text{curr}}$. The same applies for $\overline{\mathbf{P}}_{\text{target}}$ if the target age does not exactly match with one of the given mean faces. Subsequently, the mean difference $\overline{\mathbf{P}}_{\text{diff}}$ is estimated again by using Equation 3.6. To determine the in-between mean faces the proposed framework supports both a piecewise linear and a spline-based interpolation.

For the piecewise linear interpolating the exact age-dependent mean appearance $\overline{\mathbf{P}}(t)$ is estimated by

$$\overline{\mathbf{P}}(t) = (1-t) \cdot \overline{\mathbf{P}}_i + t \cdot \overline{\mathbf{P}}_{i+1}, \tag{3.9}$$

where $t$ is the relative position between the next younger mean face $\overline{\mathbf{P}}_i$ and the next older mean face $\overline{\mathbf{P}}_{i+1}$ as is shown in Figure 3.16a. First, the correct interval and its boundaries are identified based on the current or the target age respectively. Then the relative position $t_{\text{curr}}$ for $\overline{\mathbf{P}}_{\text{curr}}$ is computed based on the current age $a_{\text{curr}}$ of the input face, the age $a_{\text{curr,L}}$ of the next younger mean face and the age $a_{\text{curr,H}}$ of the next older mean face by

$$t_{\text{curr}} = \frac{a_{\text{curr}} - a_{\text{curr,L}}}{a_{\text{curr,H}} - a_{\text{curr,L}}}. \tag{3.10}$$

Accordingly, the relative position $t_{\text{target}}$ for the appearance of $\overline{\mathbf{P}}_{\text{target}}$ is

$$t_{\text{target}} = \frac{a_{\text{target}} - a_{\text{target,L}}}{a_{\text{target,H}} - a_{\text{target,L}}}. \tag{3.11}$$

For the spline-based approach a curve is fitted through all available mean faces to describe the aging trajectory. This is implemented by using Catmull-Rom splines [HvM$^{+}$13,

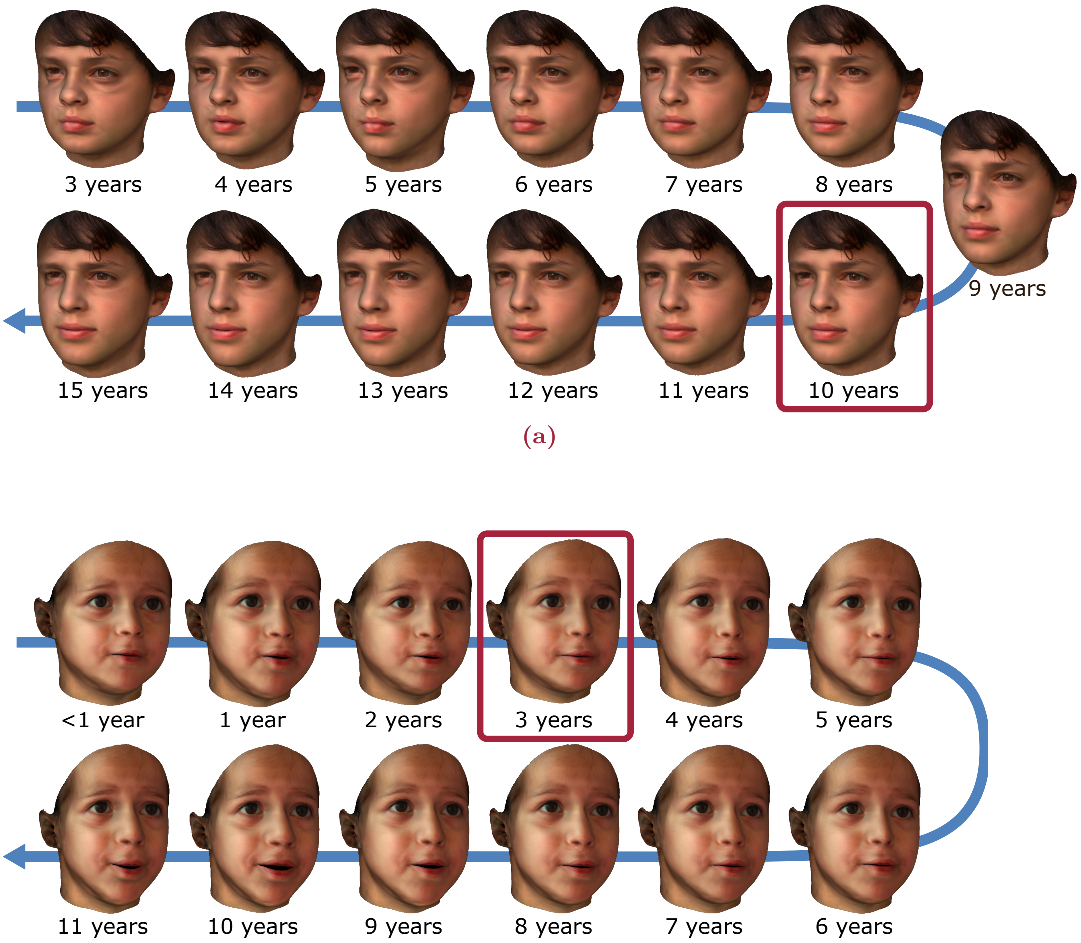
p. 600], where the age-dependent mean faces $\overline{\mathbf{P}}_i$ are used as control points. Accordingly, the dimensional space of each control point is equal to the size of the full face vector consisting of $n$ vertices with $x, y, z$ coordinates and $r, g, b$ colors (see Section 2.1). Instead of solving Equation 3.9 as has been done for the piecewise linear interpolation, for the spline-based method the age-dependent in-between facial appearance is interpolated by using

$$\overline{\mathbf{P}}(t) = \begin{bmatrix} 1 & t & t^2 & t^3 \end{bmatrix} \frac{1}{2} \begin{bmatrix} 0 & 2 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 2 & -5 & 4 & -1 \\ -1 & 3 & -3 & 1 \end{bmatrix} \begin{bmatrix} \overline{\mathbf{P}}_{i-1} \\ \overline{\mathbf{P}}_{i} \\ \overline{\mathbf{P}}_{i+1} \\ \overline{\mathbf{P}}_{i+2} \end{bmatrix} \tag{3.12}$$

which results in a C1 continuity that enables smooth transitions for the synthetic aging process. Please note that the start and end points are duplicated to ensure that the Catmull-Rom spline passes through them. In contrast to the piecewise linear interpolating the splines-based approach uses not only the next younger mean face $\overline{\mathbf{P}}_i$ and next older mean face $\overline{\mathbf{P}}_{i+1}$ but additionally their predecessor $\overline{\mathbf{P}}_{i-1}$ and successor $\overline{\mathbf{P}}_{i+2}$ respectively which is shown in Figure 3.16b. Nevertheless, the relative position $t$ between $\overline{\mathbf{P}}_i$ and $\overline{\mathbf{P}}_{i+1}$ is computed again by using Equation 3.10 for the current age interval to gain $\overline{\mathbf{P}}_{\text{curr}}$ and by using Equation 3.11 for the targeted age interval to gain $\overline{\mathbf{P}}_{\text{target}}$. Finally, Equation 3.7 is used to estimate the synthetically aged face $\mathbf{P}_{\text{aged}}$ for the original input face $\mathbf{P}_{\text{input}}$.

In Figure 3.17 two examples of the synthetic aging simulation are shown. Starting from fitting the 3DMM to a 3D scan of a child or a 3DMM reconstruction based on a 2D photo, the originally reconstructed 3D face (marked with a red box in Figures 3.17a and 3.17b) is synthetically rejuvenated or aged. As described in Section 3.5, reconstructing a face with the 3DMM framework enables a point-to-point correspondence between the newly acquired head and the existing face data of the 3DMM. By using a piecewise linear or a spline-based interpolation to model the aging trajectory through all available mean faces the original shape and texture of the scan (or photo) is transformed relative to the targeted age.

The red box in Figures 3.17a and 3.17b marks the 3D face which has been reconstructed directly from a given 2D input image of a child. Please note that the original input images and the corresponding 3D reconstructions which are based on the new, extended face dataset are shown in more detail in Figure 3.12. All other faces have been made younger or older using the proposed aging simulation approach. In case of Figure 3.17a the aging trajectory has been extracted from the newly acquired face dataset (see Section 3.4). In contrast to that the aging trajectory which is used for the aging simulation in Figure 3.17b has been extracted from a timelapse video of another female child [Hof12a] (see Table 3.1). The idea of using timelapse videos to enlarge the database or to extract longitudinal information



(a)

(b)

**Figure 3.17:** The red box marks the 3D face which has been reconstructed directly from a 2D input image of a child. All other face renderings result from the proposed growth simulation. In Figure 3.17a the aging trajectory has been extracted from the newly acquired face dataset (see Section 3.4). The corresponding mean faces are shown in Figure 3.14. Whereas in Figure 3.17b the aging trajectory has been extracted from the timelapse video of another child [Hof12a] (see Table 3.1).

has been described in Section 3.8. While the aging simulation in Figure 3.17a which is based on a series of averaged face scans produces more generalized and steady results, the synthetically aged faces based on a single timelapse video contain some inconsistencies like varying facial expressions (see Figure 3.17b) as this effect could not be neutralized completely during the 3DMM fitting to individual frames.

After rendering the transformed 3D face into an image, any standard face recognition method that works on 2D images can be used. Additionally, the lighting parameters can be modified at will and it is also possible to render the transformed 3D face on top of the face of a different person in any available image, so that the scene around the face is freely configurable and exchangeable as has been shown by Blanz et al. [BSVS04]. This can also be a helpful tool when creating missing person reports for human observers which may struggle if only the uncommon face rendering of the 3DMM is presented to them. By using the proposed method the 3D model can be extended easily by elements that are not covered explicitly by the 3DMM: for example the synthetically aged face can replace the face in the photo of another child and therefore make use of the hairstyle or clothes in the target image as has been shown by Scherbaum et al. [SSSB07] (see Figure 3.2).

## 3.10 Perception Experiments

The results of the proposed aging simulation approach are examined in a perception experiment to validate if the synthetically aged faces look plausible and if they match the targeted age. In this context another important aspect that needed clarification is the question if the acquisition of additional 3D faces of children and the associated creation of a new PCA for the 3DMM as well as the presented mathematical function in Section 3.9 for the aging simulation improved the overall results. This new PCA incorporates the original 200 faces of adults from [BV99], the 238 faces of teenagers used by [SSSB07] and the newly acquired faces of young children and teenagers during the INBEKI project which are discussed Sections 3.4 to 3.7.

The experiment is divided into three separate parts. In the first part (see Figure 3.18) two reconstructions are shown simultaneously, one is created by using the old PCA (200 adults) while for the other reconstruction the new PCA is used which is based on the extended face dataset incorporating 200 adults, 238 teenagers and 141 children. For each pair of face reconstructions the study participants are asked: "*Which reconstruction appears younger compared to the other one?*" Accordingly, the PCA that fits better to the specifics of children's faces should be identified. The hypothesis related to this question is that having additional data about the shape and look of faces of children within the newly generated PCA lead to a visually younger reconstruction result and in further consequence to an improved reconstruction result.

In the second part of the experiment (see Figure 3.19) the synthetic aging procedure is examined. Side by side two images are shown: on the one side the unaltered 3D reconstruction based on the new PCA and on the other side the result of the synthetic aging procedure that is used for rendering a face older or younger by 2-6 years. The goal is to evaluate the transformation of the face based on the aging and reunification procedure by checking if the modified face is perceived to be older or younger by a human subject. Accordingly, the following question is asked: "*The face of which of the two images seems to be younger?*"

In the third and final part of the experiment, the 3D reconstructions of the first part (see Figure 3.18) are shown once again. But this time they are supplemented with the

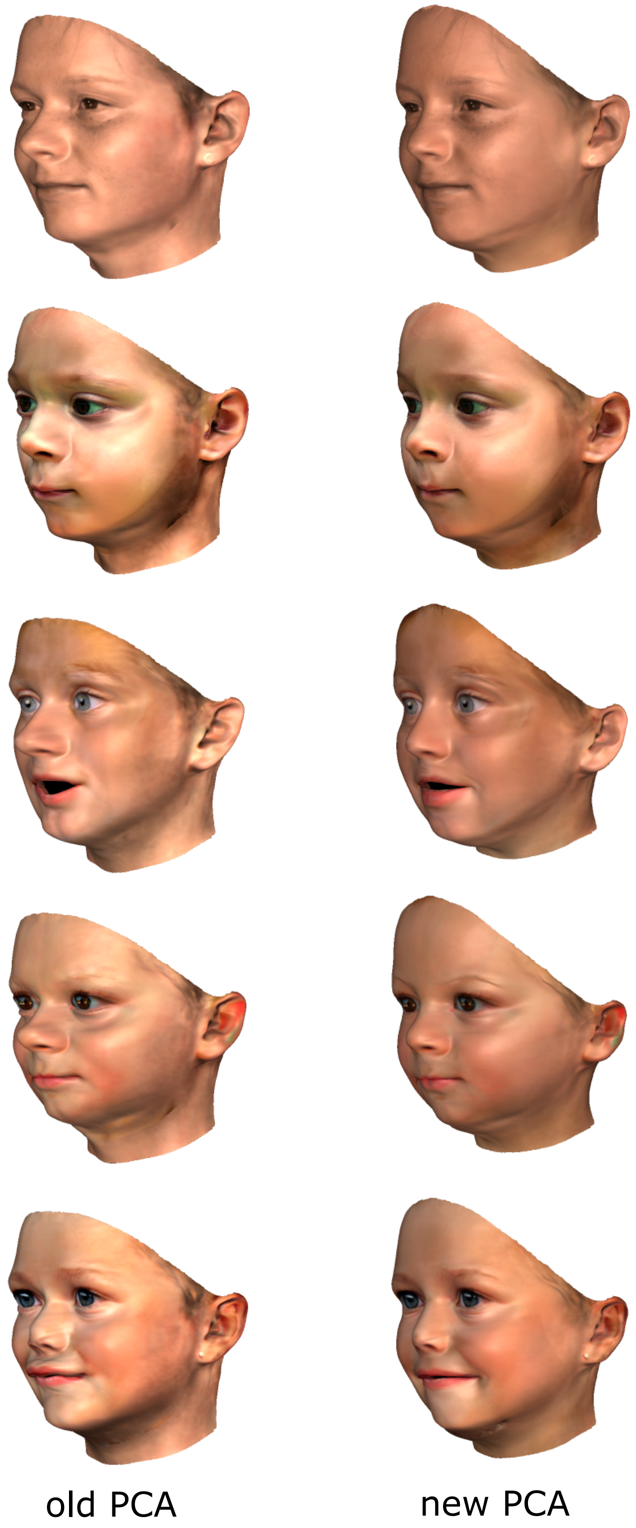


**Figure 3.18:** Comparison between the reconstruction results which are based on the PCA of the original 3DMM (left column) and the adjusted and improved PCA (right column). For both 3D reconstructions the same input images and landmark positions are used.

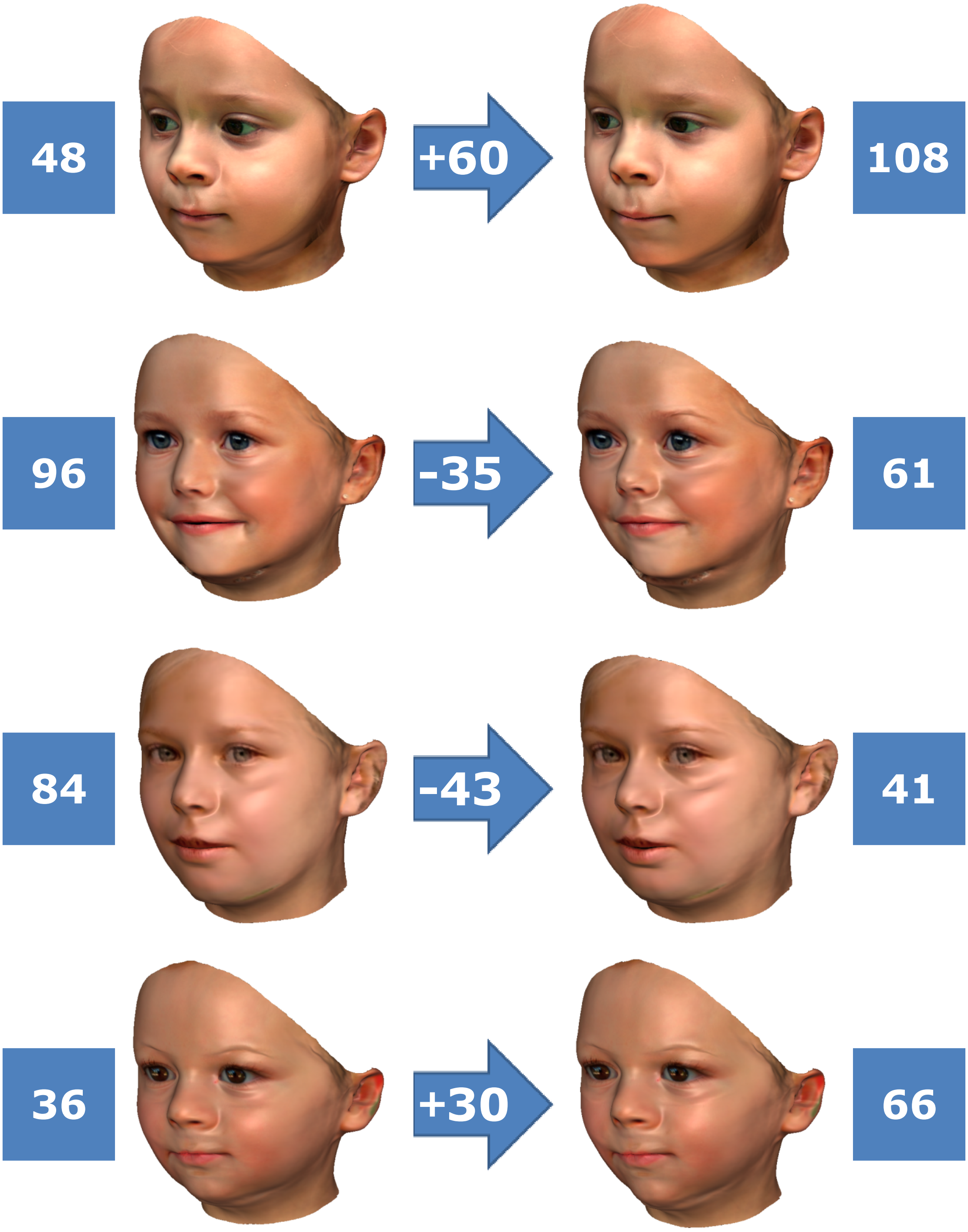


**Figure 3.19:** The left face is the original reconstruction from an input image by using the new PCA and the right face is the corresponding result after performing a synthetic rejuvenation or aging. The numbers in the blue boxes correspond to the current age in **month** of each face and the numbers in the arrows are the performed age transformations in month.

original photos which have been used for the 3D face reconstruction procedure and the related question is: "*Which reconstruction is more plausible?*" In contrast to part one of the experiment the goal is not to figure out if the new PCA is better suited to represent the faces of young children in general, instead it is more specific by evaluating if the generated reconstruction with the new PCA is indeed visually closer to the desired result in comparison to the original PCA of the 3DMM.

In every part of the experiment 51 image pairs are shown and the positioning is randomized to avoid any inferences that would help to identify which reconstruction is created by using the original or the new PCA. The experiment was run on eight participants, who were employees and students of the University of Siegen with varying scientific background.

With regard to the first part of the experiment in 79.7% of the cases the results based on the new PCA are rated to appear younger in comparison to the old PCA which uses only the facial information of adults. In the second part the synthetic aging or rejuvenation is perceived correctly in 63.0% of the cases. However, it should be mentioned that these results do not provide detailed inferences about the correlation between hit ratio and the magnitude of the changes in age. It is to be expected that the hit ratio increases with a larger difference for the aging and decreases for a smaller difference. When applying age transformations based on the acquired data it often turned out that the changes based on age transformations which are less than two years are hardly visible. Above that, these changes also depend on the original age of the reconstructed person. In sum, the results of this second part are not as clear as the ones of the first part of the experiment.

Finally, in the third part 71.1% of the reconstructions that are based on the new PCA are rated to be more plausible and better suited than the results based on the old PCA. More details regarding the individual results of the described perception experiments are provided in Tables A.1, A.2 and A.3 which can be found in the appendix of this dissertation.

## 3.11 Conclusions

The presented growth simulation system is one module of a larger image processing framework to aid the investigation of child abuse cases. It is closely interconnected with a 2D face recognition module. While the recognition task itself is performed by an additional module, the approach which is described in this chapter handles a series of preprocessing tasks on the given photos to support the face recognition module.

For example, the proposed 3DMM approach creates a 3D reconstruction from a 2D input image which enables to render arbitrary views of the face that needs to be recognized. This capability is primarily used to create a frontal face image if the face in the input image originally is shown in a non-frontal pose. Furthermore, the reconstructed 3D model is used to neutralize facial expressions to some extent, another capability of the 3DMM which has been adapted to children. As described in Section 3.7 additional 3D scans of various visemes are captured and integrated into the existing 3DMM. Unfortunately, the scanning speed of the used structured-light 3D scanner is not high enough to capture the facial motion in real-time nor in reasonable time intervals. Thus, only the final facial expression of each viseme is captured and a linear interpolation between each viseme and the neutral expression is performed to simulate the in-between facial expressions. This presumes a uniform motion, but actually facial expressions often consist of non-linear transformations of the facial shape. Furthermore, asymmetric facial transformations in regard to the left and right half of the face – like closing only one eye – are missing in the captured data. It was considered to compensate this by creating 3D reconstructions of individual frames of a video showing corresponding facial expressions, but first tests showed that the reconstruction with the existing 3DMM of adults as well as the one of children is insufficient for this task without additional training data. Thus, the proposed method can only serve as a coarse or incomplete estimation of facial expressions so far. Although there is still room for further improvements, the proposed procedures play an important role to enable 3D face reconstructions from non-neutral photos of children and to neutralize their expressions.

Another preprocessing task is the modeling of a synthetic facial image which estimates the look of a person before or in a given number of years. Subsequently, the synthetically aged or rejuvenated face is used to search a person in a given database. As has already been discussed, there are various unknown parameters which influence the growth of a person like living conditions, wealth, food or health. Thus, it is hard to give a precise prediction of the appearance of a person in several years. To compensate this imperfection, instead of only modeling one candidate, a spectrum of synthetically aged or rejuvenated facial images is created for one person as an input for the face recognition algorithm. For example, each candidate could be created by using one of the different extracted aging trajectories (see Section 3.8). To simulate gaining or losing weight during the process of aging specific series of pictures can be used to extract the corresponding trajectories or a set of principle components of the 3DMM can be altered which influence exactly these facial attributes.

To fit the 3DMM to hundreds of scans and photos it was necessary to manually select between seven and eleven facial landmarks per face which was quite time consuming. This was also necessary when additional aging trajectories were extracted from timelapse videos or if additional trajectories need to be extracted from a image collection or (timelapse) video in the future. To avoid or at least reduce this additional effort in Chapters 5, 6 and 7 different strategies for automatic landmark localization and face reconstruction are described.

The resolution and quality of the input images may vary a lot. Thus, it is important that not only the face recognition method is flexible and robust in this context but it is also beneficial if the proposed preprocessing can handle input data which is highly diverse in quality. Accordingly, Chapter 4 addresses the processing of blurry photos and the hallucination of missing facial details.

# 4 Hallucination of Facial Texture Details

The 3DMM of faces is not limited to the 3D reconstruction of faces in images (see Section 2.2) or scans (see Section 2.4). It can also be used to enhance the quality of images of faces during the reconstruction procedure by adding details to the facial texture that were not present in the original images. The 3DMM fills in details that have been lost due to blur or low resolutions. For this purpose, an additional texture enhancement algorithm is proposed in Section 4.4 that adds high-resolution details from example faces. This method has been published in [SPB15].

## 4.1 Introduction

The causes of blur and low resolution in images are highly diverse. Main reasons for blurry images or videos are that the object is out of focus, the camera or object moves while the shutter is open, the aperture setting leads to a shallow depth of field and sometimes it is just caused by a mist up or smeared lens. Likewise, there are different reasons for a low resolution of objects: Maybe the overall resolution of the camera is just low or the object in question occupies only a relatively small area of an otherwise high resolution sensor. In many scenarios it is not possible to just retake a photo, for example when analyzing the footage of a surveillance camera for law enforcement. Accordingly, various image processing algorithms have been developed to recover or enhance information that is present in images. Deconvolution methods that strive to invert the effects of blurring can be found in [KH96] as well as in [FSHR06] and [HY12]. In case of low resolution video it is possible to take advantage of the interdependence between frames [BBZ96, SS96] to improve the deconvolution. As all these methods are based on very general assumptions

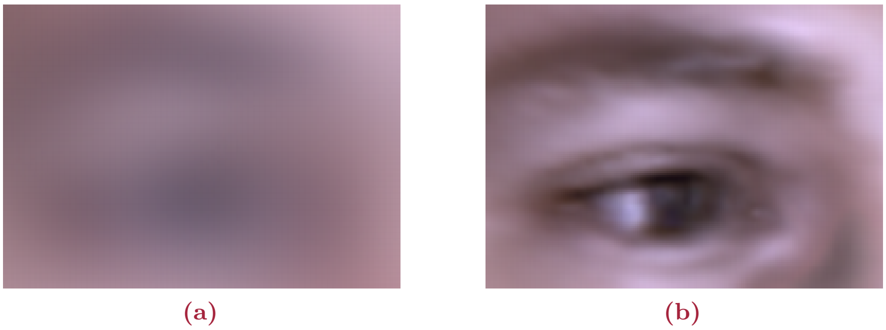


**Figure 4.1:** Figure 4.1a shows a close-up of an eye in a blurred input image, whereas Figure 4.1b shows a close-up of the deblurred reconstruction result.

about the image content, they can be applied to arbitrary images which makes them versatile but at the same time these general assumptions limit the quality of their outcome.

By having additional knowledge of the content in an image, a mathematical model can be designed to allow a model-based image enhancement. In contrast to more general approaches, model information can be used to add new information to an image, like in Figure 4.1a, where a heavily blurred eye region is shown. The deblurred result in Figure 4.1b has been created by leveraging class-specific knowledge during the restoration process which fills in eyeball, iris, eyelashes and all other details of a human eye. To achieve this result a downsampling operator has been explicitly added to the analysis-by-synthesis algorithm of the 3DMM of faces [BV99].

As the 3DMM is a statistical description of the natural shapes and textures of faces, the added image detail must not necessarily match with the original appearance of the face in the photo. Instead, it is an educated guess based on prior information about faces and remaining information in the input image. By exploiting correlations in the set of human faces the solution of the proposed method goes beyond filling in details from the average face or any other random face. Therefore, the 3DMM is fitted to the degraded image which results in a best fit after the optimization procedure. Then the facial details of this best fit are rendered into the original image to deblur the face. This enables to obtain high quality images or 3D face models even from degraded input data as is shown in Figure 4.2.

To apply the proposed method to an input image it needs to fulfill two basic prerequisites. First, the input image must contain a human face and secondly, the corresponding facial feature point coordinates must have been selected. Then the proposed framework creates a textured 3D face which corresponds to the face in the input image. Finally, this face can be rendered back into the input image and on top of the original face to create an image with a deblurred face.

This chapter is organized as follows: After introducing related work in regard to image enhancement algorithms in Section 4.2, the remainder of this chapter covers the details

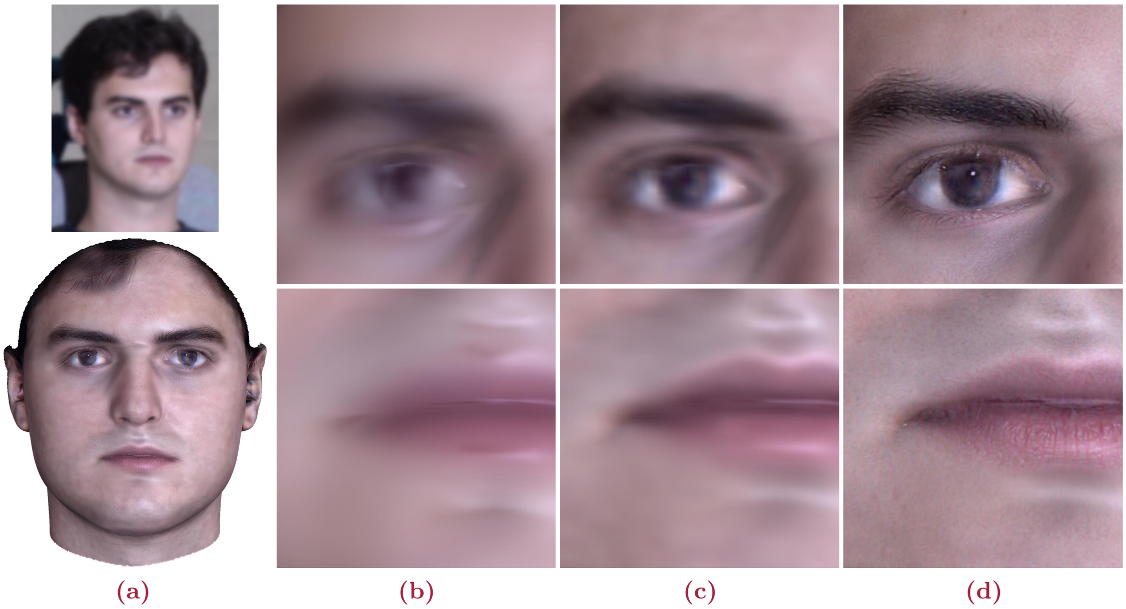


**Figure 4.2:** 4.2a shows the blurred input image *(top)* and the final result of the 3D reconstruction after applying the proposed approach *(bottom)*. The Figures 4.2b to 4.2d are zoomed in views to the eye and mouth regions of the reconstructed 3D face model. While in 4.2b only the original texture was extracted from the photo, in 4.2c the deblurring described in Section 4.3 was applied and in 4.2d the high-resolution texture transfer of Section 4.4 was added to the results of Figure 4.2c.

of the 3D model-based algorithm for face hallucination at any pose and illumination by handling blur in a 3D analysis-by-synthesis approach. Section 4.3 briefly summarizes the combination of low spatial frequency information from the input image with mid-level details of the model. Followed by a description of the transfer of high spatial frequency details across faces for face hallucination in Section 4.4. This transfer is performed on the level of eyelashes and pores. Finally a summary and outlook are given in Section 4.5.

## 4.2 Related Work

To enhance the resolution in images most approaches make assumptions about the missing details. Therefore, it is not guarantied that the enhancement is perfectly correct but it is plausible based on the available information in the image that is sometimes combined with additional model-based knowledge. Baker and Kanade [BK00] propose a hallucination method that learns a prior to incorporate missing details to faces of humans in images which is also addressed in a survey by Liang et al. [LLZC12]. In contrast to [ECT98] who propose to use Active Appearance Models (AAMs) to fill in missing details, Baker and Kanade do not account for the shape differences explicitly. They solve for the maximum a posteriori (MAP) solution using Bayes Law to estimate the unknown high resolution image for a given low resolution image. Liu et al. [LSZ01, LSF07] use a two-level approach based on global Eigenfaces and a local patch-based nonparametric Markov network to improve the image resolution. While Li et al. [LCS+08] consider the local consistency by making use of local visual primitives [CYX04], the method of Ko and Chien [KC16] is based on a fully connected feedforward neural network to enable patch-based face hallucination. Dedeoglu et al. [DKA04] learn a domain-specific prior based on spatio-temporal consistencies and image formation for face hallucination in video.

In contrast to the above mentioned methods Li and Lin [LL04] achieve pose-invariant hallucination. Therefore, they use a Support Vector Machine classifier to estimate the pose, synthesize the corresponding frontal facial image and hallucinate the details for the frontal face. By exploiting large datasets of face images, recent image matching techniques and MAP estimation Tappen and Liu [TL12] propose a method to handle even larger variations in pose. Zhu et al. [ZLLT16c] use a deep network to deal with significant pose and illumination variations.

Unlike these image-based methods, in the remainder of this chapter a 3D approach is presented which is invariant to changes in pose, size, illumination and other image parameters. Therefore, a 3DMM [BV99] is fitted to the input image using a novel fitting algorithm that is robust to the effects of blurring by explicitly simulating image blur in an analysis-by-synthesis approach (see Section 4.3). In addition to the facial details estimated by the 3DMM reconstruction which is equivalent to a MAP estimate [BV03], the proposed algorithm adds details, such as eyelashes and pores, from other faces to obtain high resolution results. While Scherbaum et al. [SRH+11] apply the makeup of one person to the face of another one and Schneider et al. [SEV18] use a parametric model to add freckles to a face, the fundamental idea of the approach which is discussed in this chapter is to transfer facial details in general.

Another 3D approach has been developed by Pan et al. [PHW08]. They propose a method for hallucinating 3D facial shapes from low resolution 3D scans by using both a Gaussian Curvature Image and a Surface Displacement Image to enhance the resolution of a 3D shape. Unfortunately, a potential texture of the object is ignored and their approach does not work on images but on 3D shapes only.

As for generative face models in 2D, such as AAMs, a major challenge in using a 3DMM for face hallucination or 3D reconstruction from low resolution images is to adapt the cost function to blurred input data. In [DBK06] AAMs are made applicable to low resolution images by using a resolution-aware formulation (RAF). It contains an explicit model of the downsampling or blurring in the cost function and as standard AAM would do, the image difference is computed in terms of pixels of the input image space and not in the shape-free texture space. But their approach is restricted to nearly frontal views.

Recently, Deep Neural Networks have been applied to face hallucination and facial

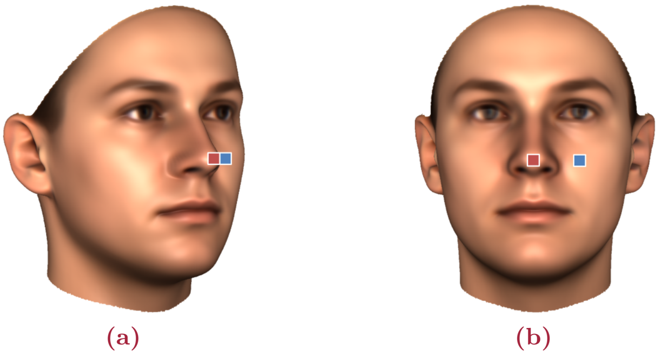


**Figure 4.3:** This figure visualizes the discontinuities between image positions and surface parametrization. While the red and the blue marked vertices of the model surface are next to each other in the half-profile view shown in Figure 4.3a, after rotating the head to a frontal view in Figure 4.3b the distance in image space between both vertices has strikingly increased.

texture inference. In [ZLLT16c] Zhu et al. use a Deep Cascaded Bi-Network for face hallucination to recover previously missing details from a low-resolution input image and Saito et al. [SWH$^{+}$16] present a method to synthesize a photo-realistic facial texture from a single image. Like the proposed approach in this chapter, Saito et al. combine the fitting of a 3D face model with a texture synthesis based on a database of high-resolution facial skin textures. But instead of using the Mahalanobis distance to find matching samples, a deep convolutional neural network is used for feature extraction.

## 4.3 Model-based Deblurring

*Section 4.3 briefly summarizes the model-based deblurring and texture enhancement techniques for low resolution input images from the joint work which has been published in [SPB15]. As these approaches are a contribution of our co-author, Matthaeus Schumacher, they are out of the scope of this dissertation. However, they are important preprocessing steps of the high resolution texture transfer which is described in Section 4.4 and therefore shall not be omitted completely at this point.*

For the analysis-by-synthesis algorithm in Chapter 2 it is assumed that the appearance of a pixel depends only on a single surface point or at least on closely neighbored surface points of the 3D model. In case of blur this assumption is no longer valid because even surface points that are located further away may influence the final pixel appearance. This effect is even stronger in case of depth discontinuities, where vertices that are far away in terms of the 3D mesh structure, may get projected next to each other in the rendered image. In Figure 4.3 the vertex that models the tip of the nose in a half-profile is rendered next the the vertex of the cheek. Accordingly, the mapping from image positions to a surface parametrization is not continuous. While this was not an issue before, after adding blur to the analysis-by-synthesis approach the image distance $d_{image}$ (see Equation 2.10) needs to be adapted adequately.

As described in Section 2.2 $I_{model}(x, y)$ is the rendered and perspectively projected 3D model, accordingly a blurred image is obtained by an image-space operator

$$I_{model}(x, y) \mapsto \varphi(I_{model}(x, y)). \tag{4.1}$$

By simulating the effect of $\varphi$ on the rendered image $I_{model}(x, y)$ the difference

$$\Delta_j = \varphi(I_{model}(x_j, y_j)) - I_{model}(x_j, y_j). \tag{4.2}$$

for each vertex $j$ to the unfiltered rendered image can be determined. While the screen positions $x_j$ and $y_j$ may vary due to adjustments of shape and pose of the 3D model during reconstruction, $\Delta_j$ is attached to a fixed vertex $j$ of the model. Based on Equation 2.10, the image distance $d_{image}$ is modified to

$$d_{image,j} = \sum_{x,y} \| I_{input}(x_j, y_j) - (I_{model}(x_j, y_j) + \Delta_j) \|^2 \tag{4.3}$$

and by substituting Equation 4.2 in 4.3

$$d_{image,j} = \sum_{x,y} \| I_{input}(x_j, y_j) - \varphi(I_{model}(x_j, y_j)) \|^2. \tag{4.4}$$

This extension to the error function allows for non-local modifications of the analysis-by-synthesis loop. In a double-stage approach the blur level $\varphi$ is estimated. First, the blur metric of the input image $I_{input}$ is calculated based on the proposed method in [MDWE04]. Secondly, to simulate different levels of blur a low-pass filter is applied subsequently on the rendered image $I_{model}$ until it matches with the blur measure in $I_{input}$.

In Section 2.3 texture extraction is used to transfer details from the input image $I_{input}$ to the 3DMM reconstruction by replacing the estimated, model-based color values of the texture vector $\mathbf{T}$ (see Equation 2.2). However, in case of blurred or low resolution input images this kind of texture extraction would remove the details introduced by the 3DMM. On the other hand individual characteristics that are not in the degrees of freedom of the 3DMM can also occur on a low spatial frequency domain. Thus, a modified procedure is presented in [SPB15] which retains the details of the 3DMM as well as individual characteristics. First, an enhanced input image

$$I_{input^+}(x, y) = I_{input}(x, y) + (I_{model}(x, y) - \varphi(I_{model}(x, y))) \tag{4.5}$$

is computed. It adds all the image details to $I_{input}(x, y)$ that otherwise would have been washed out in $\varphi(I_{model})$.

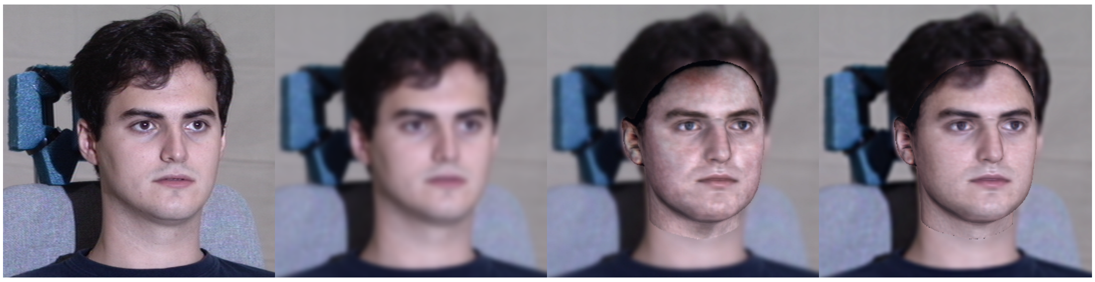


**Figure 4.4:** From left to right the following is shown: (1) The ground truth input image $I_{input}$, (2) the blurred input image, (3) the reprojected reconstruction of $I_{model}$ without texture extraction and (4) the reprojected and deblurred reconstruction $I_{input^+}$ with enhanced texture extraction as is described in Section 4.3 and [SPB15].

Afterwards, the enhanced input image $I_{input^+}$ is used for the illumination-corrected texture extraction of the 3DMM as described in Section 2.3. A result of the previously described procedure is shown in Figure 4.4. In the second column the blurred input image is shown followed by the result of the model-based deblurring and the result after the enhanced texture extraction. The corresponding ground truth is shown in the first column of Figure 4.4. More details on model-based deblurring, additional results and a comparison to the reconstruction of the original error function can be found in [SPB15].

## 4.4 High-Resolution Texture Transfer

Although the methods described in the previous section already provide substantial enhancements to blurred input images, they cannot go beyond the level of detail represented in the 3DMM itself (see Section 2). Accordingly, the fine structures of hair or skin, including pores and slight dermal irregularities, are not recovered.

To cope with this drawback, in this section an additional method is proposed which adds details above the spatial frequencies captured by the 3DMM (see Figure 4.2d). These facial details are derived from a database of high-resolution photos of faces (see Section 4.4.1) and are transferred to new individuals during a postprocessing step of the 3DMM fitting. The basic idea is to find a matching face (see Section 4.4.2) in a high-resolution photo collection for each low-resolution input image and to use it as a basis for the transfer of skin details. In Section 4.4.3 this approach is extended to enable a segment-based transfer of details. Instead of searching for an entire face that matches the one in the input image best, it looks for matching facial regions. As these regions need not necessarily belong to the same face, this modified approach allows to combine skin information of distinct faces. Optical flow is used to improve the correspondence between the extracted facial features and the targeted low-resolution texture (see Section 4.4.4). An overview of the processing

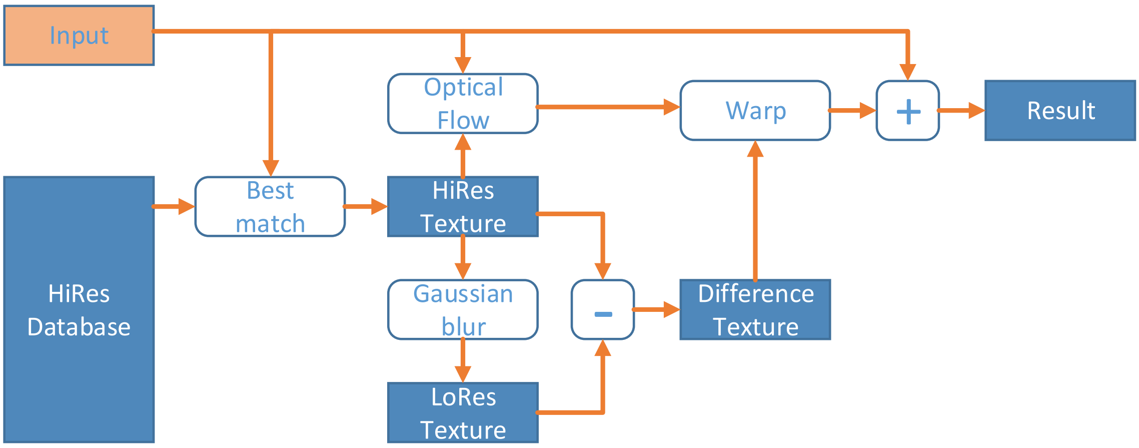


**Figure 4.5:** An overview of the high-resolution texture transfer processing pipeline. For the depicted pipeline in the diagram it is assumed that the input data and all elements of the face database have already been fitted to the 3DMM of faces.

pipeline for the high-resolution texture transfer is provided in Figure 4.5 and the results of this method are summarized in Section 4.4.5.

Currently the high-resolution database contains 221 individual faces (79 female and 142 male persons) which are extracted from the Multi-PIE face database [GMC+10] by fitting the 3DMM to each of the 2D images. The texture data of the faces in this database is used to transfer facial details, so that the formerly low-resolution texture $T_{i,L}(x, y)$ of person $i$ gets transformed into a high-resolution texture $T_{i,H}(x, y)$. In the following, the result of this process is denoted as $T_{i,L \to H}(x, y)$.

### *4.4.1 Extraction of Skin Features*

Before any facial details can be transferred, it is necessary to extract them from a pair consisting of a low-resolution texture $T_{j,L}(x, y)$ and a corresponding high-resolution texture $T_{j,H}(x, y)$ of the same person $j$. $T_{j,H}(x, y)$ is generated by fitting the 3DMM to the high-resolution input image and by applying the texture extraction method of the 3DMM which is described in Section 2.3. To reduce computational overhead caused by the fitting procedure, the 3DMM is only fitted to the high-resolution image, while low-resolution texture $T_{j,L}(x, y)$ is just the result of a Gaussian blurred high-resolution texture $T_{j,H}(x, y)$ and therefore artificially generated.

In this context it is assumed that the desired facial details are equivalent to the difference between the low and the high-resolution textures of the same person. The resulting high spatial frequencies are stored in the texture $T_{diff}(x, y)$, so that

$$T_{diff}(x, y) = T_{j,H}(x, y) - T_{j,L}(x, y), \tag{4.6}$$

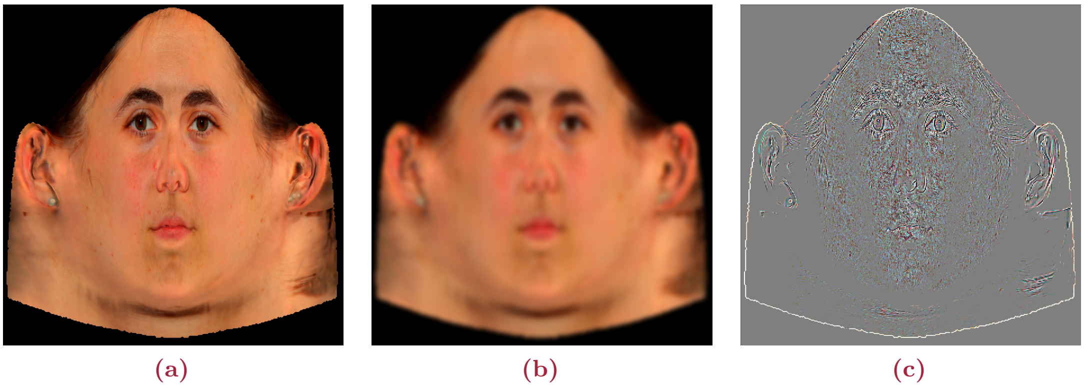


**Figure 4.6:** In Figure 4.6a the extracted high-resolution texture $T_{j,H}(u,v)$ of an input image is shown. The corresponding low resolution texture $T_{j,L}(u,v)$ is shown in Figure 4.6b and the resulting difference texture $T_{diff}(u,v)$ in Figure 4.6c.

where $T_{j,H}(x,y)$ is the extracted texture from a high-resolution image of person $j$. As described above $T_{j,L}(x,y)$ originates from a low-resolution image of the same person $j$ (see Figure 4.6). Accordingly, $T_{diff}(x,y)$ is the result after applying a Laplace operator to $T_{j,H}(x,y)$. Both images are stored in the facial database.

For simplicity, for all examples in the remainder of this section the blur level is predefined heuristically with the goal that $T_{diff}(x,y)$ contains the desired amount of high frequencies to describe skin features. Alternatively, the estimated blur level $\varphi$ of the input image (see Section 4.3) can be used to capture exactly the high frequencies in $T_{diff}(x,y)$ that are missing in the blurred input image. In this case $T_{j,L}(x,y)$ and $T_{diff}(x,y)$ need to be computed on the fly based on the blur level of the input image, instead of being precomputed and stored in the facial database. A third option would be to let the user decide the level of detail transfer interactively. Then again $T_{j,L}(x,y)$ and $T_{diff}(x,y)$ are not predefined.

A key prerequisite for a successful transfer of details is that the 3DMM guarantees a dense point-to-point correspondence between each texels $(x,y)_j$ of all fitted individuals $j$ in the database and the texels $(x,y)_i$ of the extracted texture of the current input image from an individual $i$. As long as this condition is complied, the difference texture $T_{diff}(x,y)$ can be added to the extracted (low-resolution) texture $T_{i,L}(x,y)$ of any other face to create a convincing high-resolution texture $T_{i,H}(x,y)$. Thus, the transferred texture $T_{i,L\to H}(x,y)$ is defined as

$$\begin{aligned} T_{i,L\to H}(x,y) &= T_{i,L}(x,y) + (T_{j,H}(x,y) - T_{j,L}(x,y)) \\ &= T_{i,L}(x,y) + T_{diff}(x,y). \end{aligned} \tag{4.7}$$

Besides Figure 4.6c, an additional example of a typical difference texture $T_{diff}(x,y)$ is

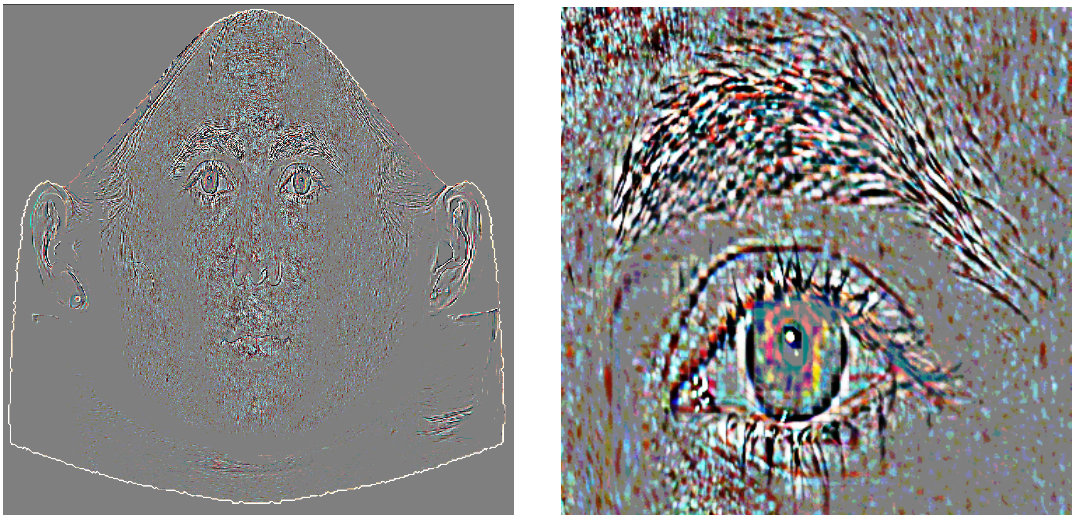


**Figure 4.7:** The schematic Figure 4.7a shows the result of $T_{diff}(u, v)$ based on Equation 4.6 for the complete facial texture map, while Figure 4.7b is a magnification of the left eye. Please note that $T_{diff}(u, v)$ contains signed values. For this image all RGB pixel values are mapped into an interval between $[0, 255]$, where a value of 128 corresponds to a value of zero in the original difference texture. Furthermore, the contrast is artificially intensified to enhance visibility.

shown in Figure 4.7. As the difference texture is a floating point texture which contains both positive and negative values, these values have been mapped into the interval $[0, 255]$ for the shown images. While in Figure 4.7a the whole cylindrical texture map is displayed, Figure 4.7b allows a more detailed look at the region of the right eye as well as its eyebrow. Not only the patterns of the eyelashes and brows are extracted, but also the high frequency differences of the dermal texture.

### *4.4.2 Search for Matching Faces*

As it is unusual that the extracted information of one face perfectly represents the missing details of any other face, in many scenarios it is necessary to locate pairs of textures that are similar to each other. Therefore, the PCA coefficients $\boldsymbol{\beta}_{in}$ which belong to the facial texture of the input image are compared with the coefficients of each sample $\boldsymbol{\beta}_{db}$ of the facial database. For this comparison the Mahalanobis distance

$$d_t(\boldsymbol{\beta}_{in}, \boldsymbol{\beta}_{db}) = \sqrt{(\boldsymbol{\beta}_{in} - \boldsymbol{\beta}_{db})^T \mathbf{C}_T^{-1} (\boldsymbol{\beta}_{in} - \boldsymbol{\beta}_{db})} \tag{4.8}$$

of the texture coefficients is calculated between vector $\boldsymbol{\beta}_{in}$ and every available vector $\boldsymbol{\beta}_{db}$, where $\mathbf{C}_T$ is the covariance matrix of the facial texture PCA of the 3DMM. Subsequently, the facial details are transferred from exactly that person, where the computed distance is minimal. Quantifying the distances in PCA space to detect the nearest neighbor instead of

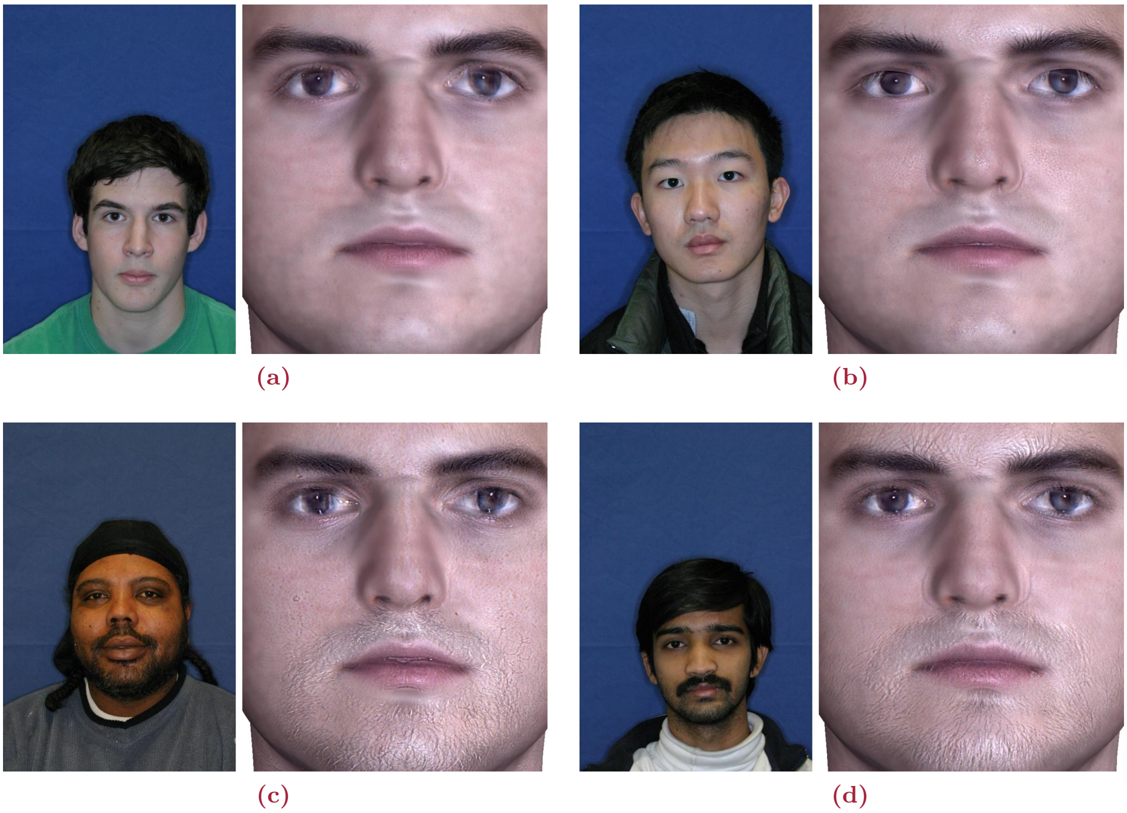


**Figure 4.8:** In the left part of Figures 4.8a to 4.8d the initial photo of four candidates from the high-resolution facial database [GMC+10] is shown. Whereas the right part is a rendering of the reconstructed 3D face of the person in the input image of Figure 4.2a using the resulting texture $T_{i,L\rightarrow H}(u,v)$ from the high-resolution texture transfer. While the Mahalanobis distance for Figures 4.8a and 4.8b is **minimal** and thus convincing facial details are added. 4.8c and 4.8d depict candidates where the distance is **maximal**. Thus, the associated difference texture $T_{diff}(u,v)$ (see Equation 4.6) does not add proper details to the 3D reconstruction: Facial hair is added where non is present in the input image and in Figure 4.8d the eyebrows do not match very well.

computing the absolute differences ensures that the most significant features are regarded for the similarity measurement.

One remaining drawback is that regions with a high influence on the overall Mahalanobis distance may lead to unwanted artifacts in the resulting detail transfer for the overall head: To prevent that similar eye, nose and mouth regions between a female and a male face lead to a transfer of beard stubble into the female face, all images in the database are labeled, so that only the details from individuals of the same gender are taken into consideration during the distance measurements and finally for the transfer of facial details. Please note that the region based approach presented in Section 4.4.3 does not presuppose a labeling of the high-resolution images in the database, because by handling each region separately the described interference problem is solved implicitly.

The best fitting candidates based on the computed Mahalanobis distance between the low-resolution input image $T_{i,L}(x, y)$ of the person in Figure 4.2 and every face in the database are shown in Figure 4.8a and 4.8b while those candidates which differ most from the input image are shown in Figure 4.8c and 4.8d. The left part of each image in Figure 4.8 shows the original photo of each candidate. The final result $T_{i,L\rightarrow H}(x, y)$ after transferring the facial details by adding the difference texture $T_{diff}(x, y)$ to the input image is depicted in the right part.

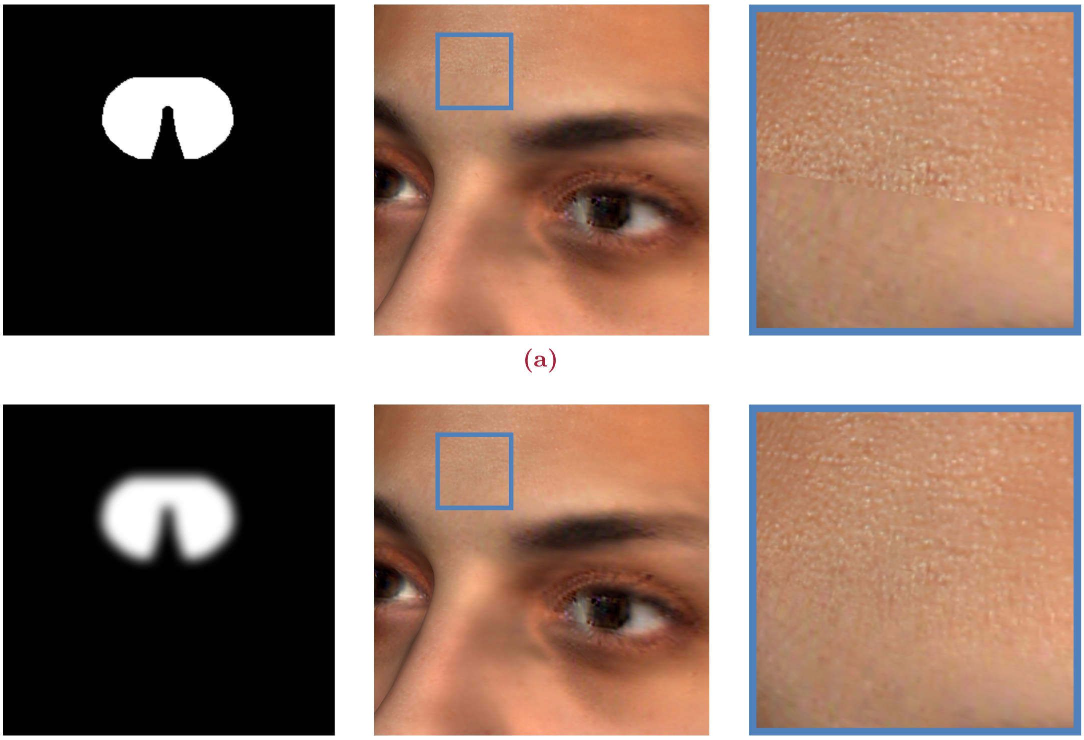
Even though the transfer of details in Figures 4.8c and 4.8d for images with a high Mahalanobis distance compared to the input image in Figure 4.2a adds unwanted facial features to the resulting textured 3D face, the strong point-to-point correspondence which is guaranteed by the 3DMM (see Section 2) prevents completely unconvincing results.

### *4.4.3 Search for Matching Facial Regions*

Sometimes it is challenging to find a face where all facial regions (eyes, nose, mouth, ears, etc.) are similar enough at once to be useful for the texture transfer. For example, there may be strains of hair on the forehead of some individuals from the image database while there is no hair on the forehead of the person in the input image. Therefore, a more sophisticated approach is to assemble the details of individual facial regions separately by looking for each corresponding subarea which matches the current candidate best. Finally, the details of several individuals are combined in the process of the high-resolution texture transfer. This is done by computing the Mahalanobis distance in Equation 4.8 separately for each region of the face and not just for the whole face. Then for each region the texture details are transferred by generating $T_{diff}(x, y)$ based on the corresponding facial region that has the minimal Mahalanobis distance $d_t$. In this context a binary texture mask is used to identify all vertices that belong to a predefined region. Only corresponding vertices which are set to 1 are taken into account when determining the region-based Mahalanobis distance while vertices with a value of 0 are neglected. An example of the binary mask of the eye segment is shown on the left side of Figure 4.9a.

To avoid visible transitions at the borders of neighbored regions as is shown in Figure 4.9, a texture blending is applied to combine the textures of different regions. Thus, blend maps are used which define the blend intensities for each facial region like is shown for the eye region on the left side of Figure 4.9b.

Searching for matching facial regions instead of looking for similar looking faces provides an additional advantage: If comparing whole faces like in Section 4.4.2 local differences were sometimes overruled by global similarities. Occasionally this leads to errors like the transfer of beard stubble into a female face. By labeling the high-resolution images based on the person's gender, this kind of artifacts could be avoided to some extent. Nevertheless, applying a region based approach to detect the best matches provides a solution to this



(a)

(b)

**Figure 4.9:** The region-based combination of extracted details can lead to visible transitions at the borders of neighbored regions as is shown in Figure 4.9a. By using a smoothed blend mask for each facial region a blending of neighbored texture values is performed which reduces these kind of artifacts as is shown in Figure 4.9b. From left to right in each row the blend mask for the eye segment (see Figure 2.6b), the rendering of the textured face and a magnification of the region of interest are shown.

and similar problems without the necessity of an additional labeling of the images in the database.

### *4.4.4 Optical Flow and Warping to Enhance Texture Quality*

Artifacts may still appear if a global similarity of two segments is co-occurring with strong local differences. This is occasionally observed in the region of the eyes or the eyebrows. An example of such an artifact is shown in the close-up of the iris in Figures 4.10a and 4.11a, where the artificially added details look like misplaced contact lenses. To handle these remaining imperfections an image warping is applied to the difference texture $T_{diff}(x, y)$ so that it fits better to the low-resolution input texture $T_{i,L}(x, y)$ before performing the transfer of facial details. Therefore, the optical flow between the input texture $T_{i,L}(x, y)$ and the best matching high-resolution texture $T_{j,H}(x, y)$ from the database needs to be estimated.

Originally, for 2D images the optical flow computes the movement $\Delta x$ and $\Delta y$ of a

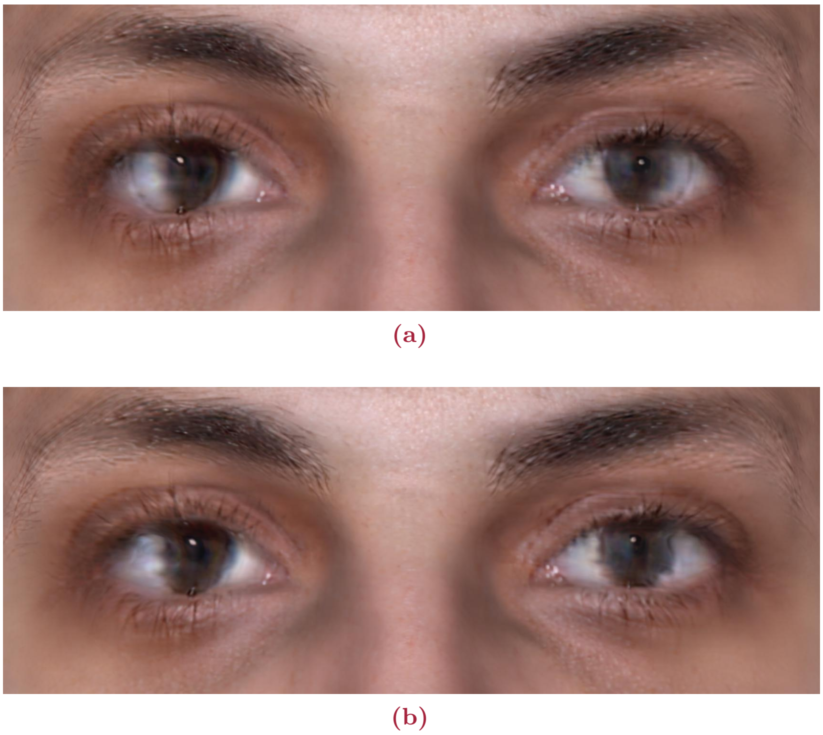


**Figure 4.10:** Figure 4.10a shows an example of occasionally appearing artifacts in the area of the iris. Whereas Figure 4.10b is the result of the application of a warped difference texture $T_{diff}(u,v)$ based on the optical flow calculations between the input texture $T_{i,L}(u,v)$ and the corresponding high-resolution texture $T_{j,H}(u,v)$ from the database.

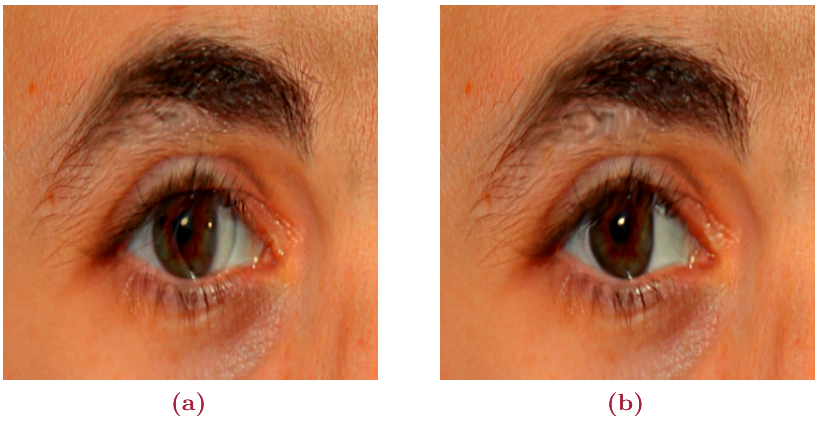


**Figure 4.11:** An additional example of the visual improvements by using optical flow and warping. Like in Figure 4.10 the artifact in the area of the iris which looks like a shifted contact lens has been corrected.

pixel with an intensity $I(x, y, t)$ between two consecutive image frames [HS81]. Assuming that the pixel intensities of corresponding points do not change and that only their position may vary, leads to the brightness constancy constraint

$$I(x, y, t) = I(x + \Delta x, y + \Delta y, t + \Delta t). \tag{4.9}$$

Using only first-order Taylor series and ignoring higher-order terms [PCF06, pp. 239–241] yields to

$$I(x + \Delta x, y + \Delta y, t + \Delta t) \approx I(x, y, t) + \frac{\partial I}{\partial x}\Delta x + \frac{\partial I}{\partial y}\Delta y + \frac{\partial I}{\partial t}\Delta t. \tag{4.10}$$

After substituting Equation 4.10 into Equation 4.9 the temporal partial derivative of $I$ is

$$\frac{\Delta I}{\Delta t} = \frac{\partial I}{\partial x}\frac{\Delta x}{\Delta t} + \frac{\partial I}{\partial y}\frac{\Delta y}{\Delta t} + \frac{\partial I}{\partial t} = 0 \tag{4.11}$$

and with $v_x = \frac{\Delta x}{\Delta t}$, $v_y = \frac{\Delta y}{\Delta t}$ being the optical flow of $I(x, y, t)$ at position $x, y$ the brightness constancy equation can be written as

$$\frac{\partial I}{\partial x}v_x + \frac{\partial I}{\partial y}v_y + \frac{\partial I}{\partial t} = 0. \tag{4.12}$$

This uniquely defines the direction of the intensity gradient $\nabla I = (\frac{\partial I}{\partial x}, \frac{\partial I}{\partial y})^T$ [Bla00, p. 23] as

$$\langle \mathbf{v}, \nabla I \rangle = -\frac{\partial I}{\partial t}, \quad \text{with} \quad \mathbf{v}(x, y) = \begin{pmatrix} v_x(x, y) \\ v_y(x, y) \end{pmatrix} \tag{4.13}$$

As the Equation 4.13 is in two unknowns which is known as the aperture problem of optical flow algorithms, additional constraints are necessary [PCF06, pp. 239–241]. Accordingly, there are different methods to estimate the optical flow between images. For example, Horn and Schunck [HS81] assume a smooth flow over the whole image, while Lucas and Kanade [LK81] as well as Bergen et al. [BAHH92] assume that the flow is constant in a local neighborhood.

Another problem is that the input texture $T_{i,L}(x, y)$ and the best matching high-resolution texture $T_{j,H}(x, y)$ are not consecutive frames but facial textures from different individuals. This means that these images most likely do not fulfill the brightness constancy constraint. Nevertheless, the optical flow between these images needs to be determined as it provides a chance to improve the correspondence between texels $T_{i,L}(x, y)$ and $T_{j,H}(x, y)$ (see Figure 4.12). This is motivated by the results in [Bla00, pp. 22–25] where it has been shown successfully that a reliable motion field for distinct image pairs using optical flow

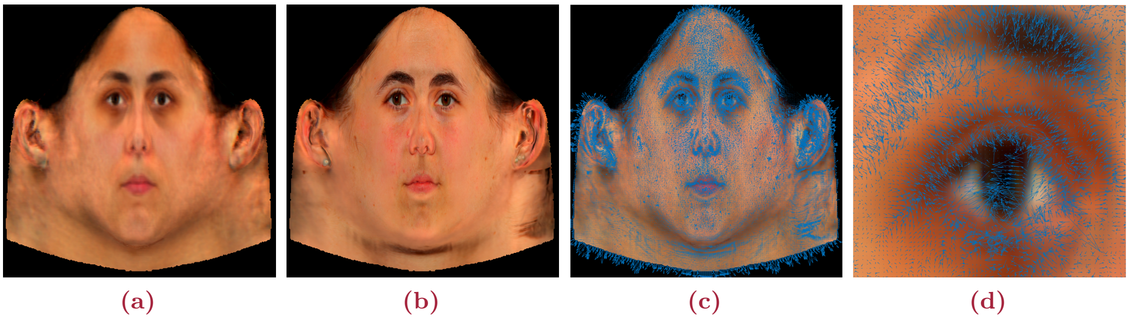


**Figure 4.12:** The optical flow estimation between the low-resolution input texture $T_{i,L}(x, y)$ (Figure 4.12a) and the high-resolution texture $T_{j,H}(x, y)$ (Figure 4.12b) is shown in Figure 4.12c and Figure 4.12d. The latter is a magnification of the right eye.

even for unsteady pixel intensities can be estimated. In [Bla00, pp. 22–25] optical flow is used to enable correspondence between two portrait images $I_1$ and $I_2$ of distinct persons by using the vector field of the translation vectors

$$\omega(x, y) = \begin{pmatrix} \Delta x(x, y) \\ \Delta y(x, y) \end{pmatrix} \tag{4.14}$$

between corresponding pixels $I_1(x, y)$ and $I_2(x + \Delta x(x, y), y + \Delta y(x, y))$. This required to change the brightness constancy constraint in Equation 4.9 to

$$I_2(x + \Delta x(x, y), y + \Delta y(x, y)) = I_1(x, y) \tag{4.15}$$

for $\omega = \mathbf{v} \cdot \Delta t$ with $I_1(x, y) = I(x, y, t)$ and $I_2(x, y) = I(x, y, t + \Delta t)$ according to [Bla00, pp. 24–27] and [Jäh97, p. 431]. As a consequence, an adjustment of the temporal derivative of $I$ is necessary, which is now approximated by

$$\partial_t I = \frac{1}{\Delta t}(I_2(x, y) - I_1(x, y)). \tag{4.16}$$

Following the proposed implementation in [Bla00, p. 25] the derivatives of $x$ and $y$ are computed by using a central differencing scheme which yields to

$$\partial_x I = \frac{1}{4}(I_1(x + 1, y) - I_1(x - 1, y) + I_2(x + 1, y) - I_2(x - 1, y)) \tag{4.17}$$

$$\partial_y I = \frac{1}{4}(I_1(x, y + 1) - I_1(x, y - 1) + I_2(x, y + 1) - I_2(x, y - 1)). \tag{4.18}$$

In the context of the given problem statement the method of Bergen et al. [BAHH92] is used to determine the optical flow between the input texture $T_{i,L}(x, y)$ and the high-resolution texture $T_{j,H}(x, y)$. It combines the algorithm of Lucas and Kanade [LK81] with a coarse-to-fine refinement strategy allowing that the correspondence is not necessarily established with the nearest pixel of equal intensity. Therefore, a low pass filter is applied to the image pair to create a low-resolution version of each image to estimate a first iteration step of the optical flow. The Gaussian pyramid [AAB+84] is used to handle varying levels of resolutions and to optimize the optical flow estimation successively.

After the flow field between $T_{i,L}(x,y)$ and $T_{j,H}(x,y)$ has been estimated a warp operation is processed to improve congruence. But instead of applying this warp operation to $T_{j,H}(x,y)$, the difference texture $T_{diff}(x,y)$ is warped based on the translation vectors $\omega(x,y)$. Accordingly, Equation 4.7 is replaced by

$$T_{i,L\to H}(x,y) = T_{i,L}(x,y) + \omega(x,y) \cdot T_{diff}(x,y). \tag{4.19}$$

By using this improved detail transfer the formerly existing artifacts in the area of the eyes disappear from the resulting facial textures. The corresponding results with the warped difference texture are shown in Figures 4.10b and 4.11b.

### *4.4.5 Results of the High-Resolution Texture Transfer*

Besides Figure 4.2 an additional result of the described transfer of texture details is shown in Figure 4.13. While strong enhancements are already achieved with the model based deblurring (see Section 4.3) as is presented in Figure 4.13c, the transfer of details improves the visual quality even more as can be seen in Figure 4.13d.

Although the transferred details for the small hairs of the eyebrow do not match perfectly to the blurred eyebrows of the input image, the presented algorithm provides a visually convincing result as can be seen at the top of Figure 4.13d. Likewise, the fine skin structure of the lips is well transferred into the formerly blurred input image.

However, it should be pointed out that it is not invariably the case that all features in the face are perfectly restored. For example, in Figure 4.10 the recovery of the iris is not as good as in Figure 4.13d, even though the corresponding Mahalanobis distances for the eye regions do not differ significantly from each other. Even after the application of the optical flow and the corresponding image warping operation (see Figure 4.11), which has been described in Section 4.4.4, the reconstruction in Figure 4.13d stays superior. Accordingly, a future improvement could be to reduce the size of the different regions and to combine all regions in a tree structure from coarse to fine. This should enable that the presented approach adapts better even to small differences between the facial textures.

## 4.5 Conclusions

In this chapter, a processing pipeline is presented which is able to reconstruct detailed 3D models of faces even from images with substantial blur. Instead of making general

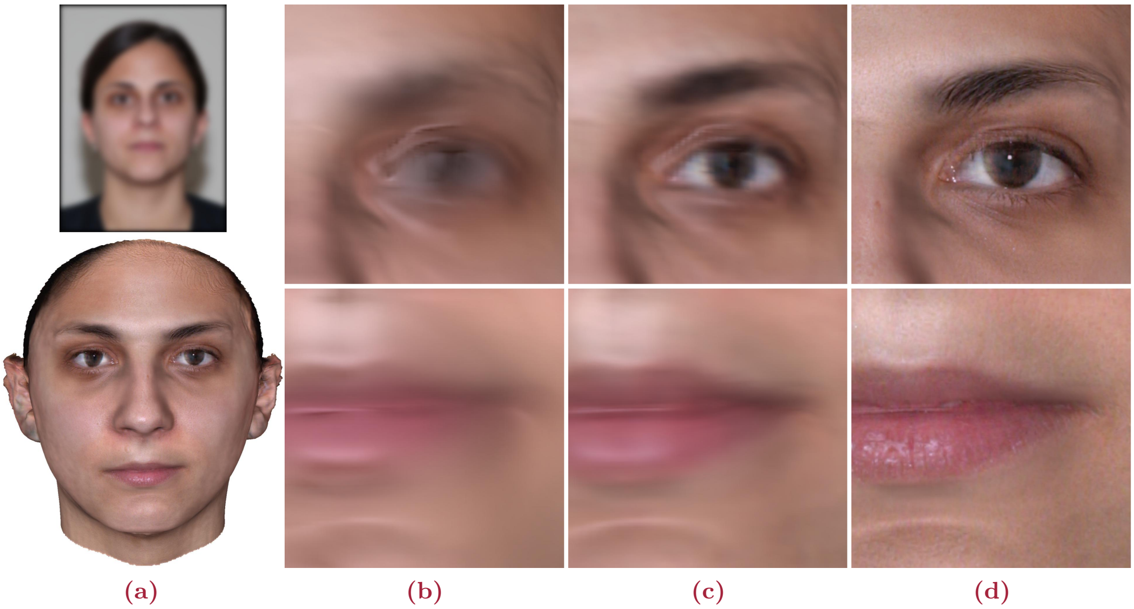


**Figure 4.13:** 4.13a shows the blurred input image *(top)* and the final result of the 3D reconstruction after applying the proposed approach *(bottom)*. The Figures 4.13b to 4.13d are zoomed in views to the eye and mouth regions of the reconstructed 3D face model. While in 4.13b only the original texture was extracted from the photo, in 4.13c the deblurring described in Section 4.3 was applied and in 4.13d the high-resolution texture transfer of Section 4.4 was added to the results of Figure 4.13c.

assumptions about the image content, it uses the explicit facial information based on the 3DMM to treat blurred faces in the input images. This makes the proposed approach very robust and general for filling in missing details to images of faces. The adapted analysis-by-synthesis algorithm allows for model-based deblurring by estimating facial details. Combined with the transfer of details on the highest level of resolution (eyelashes, pores and wrinkles), a very general tool is developed to enhance existing low quality images and to enable or at least enhance face recognition even on low quality images.

However, as has already been pointed out, this approach cannot necessarily predict the true appearance of previously missing facial details. Instead it makes an educated guess based on the correlations between features in faces which are captured by the 3DMM due to the fact that it uses global face vectors of entire faces. Using statistical inference the influence of some vector components in regard to others can be estimated. Accordingly, missing structures and details can be restored even if these are beyond the visible structures in the image.

# 5

# Automatic Initialization of the 3D Morphable Model Fitting

This chapter describes a 3D face reconstruction pipeline that automates the former manual initialization procedure of the 3DMM fitting and an implementation of a simultaneous fitting to multiple images with varying facial expressions. Starting with one or more 2D input images, for each image, a face detection is applied, and if a face has been found, the pose is estimated and facial landmarks are localized. Next, the original input images are cropped to the region of interest and the 3D reconstruction with a 3DMM is initialized by utilizing the position, pose and landmarks of each detected face. The 3D reconstruction itself and the inclusion of several images of the same person is driven by a multi-fitting approach of the 3DMM which is also able to handle facial expressions. This whole process is designed to run without the necessity of user input to select facial feature points.

The proposed pipeline for automatic landmark localization has been published in [SPB16], and insights from the multi-fitting approach lead to the improved method in Chapter 6 which has been published in [PB16].

## 5.1 Introduction

A 3D Morphable Model for faces such as the one which has been introduced by Blanz and Vetter [BV99] or the publicly available Basel Face Model [PKA+09] is a useful basis to reconstruct textured 3D shapes of faces from 2D images and to initiate further processing steps. Its key advantage is to enable a dense point-to-point correspondence between all 3D reconstructions which can be used, among other things, for morphing between faces or to transfer facial expressions. Unfortunately, to initialize the reconstruction with a 3DMM, many implementations require a manual detection of faces in the input images and furthermore the selection of feature point or landmarks in each face. If only a few images are processed, the additional effort for the user may remain negligible, but if the goal is to analyze large image databases of faces, manual landmark localization is an extremely time-consuming procedure. For example to examine effects like aging or to learn how facial expressions alter the shape of the face, it is unavoidable to use an extensive amount of input images to generate a profound dataset.

Like [BBPV03] and [BKK$^+$08] the proposed approach makes use of a 3DMM which provides dense point-to-point correspondence between all reconstructed shape vertices. In difference to [BBPV03] no manual landmark selection is required. Furthermore this method exceeds both approaches in robustness and flexibility, especially in respect to varying lighting conditions and poses, and neither [BBPV03] nor [BKK$^+$08] support the simultaneous fitting of several input images with varying facial expressions as is summarized in Chapter 2. While [KS13, SKSS14] compute 3D to 2.5D correspondence between several input images by using Collection Flow [KSS12] to reconstruct a textured 3D shape, the proposed method establishes 3D to 2D correspondence between each 3D reconstruction and its input image as well as 3D to 3D correspondence between all reconstructions. The corresponding framework fuses the facial information of several independent pictures of the same person into one final textured 3D shape by applying a reconstruction method that simultaneously handles all these images at once. One major goal of the proposed method is that it processes a huge number of images of faces and extracts as much information from these images as possible in order to broaden knowledge about faces and to incorporate it to the already existing 3DMM.

After a summary of related work in Section 5.2, it is described how the work of Zhu and Ramanan [ZR12] is utilized to automate the initialization steps for the 3D reconstruction based on the 3DMM (see Section 5.3). Followed by a discussion about some drawbacks of a naive, straightforward implementation in Section 5.4. Furthermore, the advantages of combining facial information of several images of the same person without requiring a neutral or matching facial expression through all these images are presented in Section 5.5. The results of the proposed analysis-by-synthesis approach which combines multiple images to reconstruct all collected facial attributes in a plausible way are shown in Section 5.6 and a conclusion is given in Section 5.7.

## 5.2 Related Work

In recent years, various methods in the field of automatic facial landmark localization have been published. Approaches for face detection, landmark localization and face alignment are often used as an initial step for face reconstruction or recognition methods. Please note that a detailed overview of current work in the field of 3D face reconstruction is provided in Section 6.2.

Generally, landmark localization methods can be divided into model-based and regression-based methods. The model-based approaches explicitly learn the appearance of the face shape or local templates. For example, [CET01, Liu07] use statistical models of shape variation based on complete face images, while [DSG12, AZCP13, YHZ$^+$13, TP14] use part-based models to adapt to new input images of faces. Zhou and Lin [ZBL13] argue that some landmarks can be detected more reliably (e.g., eye centers, mouth corners) than others. Therefore, they propose a graph matching to find the best candidates based on learned shape constraints, and Zhu and Ramanan [ZR12] use pre-defined Histogram of Oriented Gradients (HOG) features for each facial landmark combined with a tree structured part model as a face descriptor. To increase runtime efficiency and overall localization accuracy Hsu et al. [HCH15] propose a Regressive Tree Structure Model.

Regression-based methods estimate landmark locations by learning the relation between pixel values and facial features. By starting with a first guess of the location, the final position is refined iteratively. A series of different method is proposed to solve the regression problem. Valstar et al. [VMBP10] combine Support Vector Regression with Markov Random Fields to incorporate relationships between landmarks which narrows the search area and thereby reduces processing time. In [DGFv12, YP13, LBIC15] random forests are used for feature detection. Cao et al. [CWWS12] propose a cascaded regression method which learns the face constraints to handle non-salient landmarks directly from the training data. Thus, a parametric shape model is not needed to take the correlation between landmarks into account. To reduce outliers the cascaded regression method by Burgos-Artizzu et al. [BAPD13] explicitly detects occluded parts of the face. By learning occlusion patterns from the training data Wu and Ji [WJ15] increase the robustness of their cascade regression framework.

Besides employing boosted regressors several published regression methods for landmark localization use Deep Neural Networks [ZFC$^+$13, SWT13, FZ16, WJ16]. In this context Zhou et al. [ZFC$^+$13] propose a coarse-to-fine convolutional network which starts with an initial prediction and then locally refines the landmark positions for individual facial segments. In the cascaded network proposed by Sun et al. [SWT13] first the face is split into three overlapping regions (whole face, eyes and nose, nose and mouth), then each region is handled by an individual network and followed by local refinements which are restricted to small changes. To increase robustness Fan and Zhou [FZ16] propose to use a deeper network and added additional convolutional layers in comparison to [ZFC$^+$13, SWT13]. By combining facial action recognition and landmark localization Wu and Ji [WJ16] are able to boost the performance of both previously individual tasks. A challenge for facial landmark localization has been published by Sagonas et al. [STZP13].

In addition to the aforementioned landmark localization methods, other approaches focus on face alignment which also can be used as an initial step for facial feature localization itself as well as for 3D face model fitting. Once again these methods can be divided into model-based and regression-based approaches.

For example, Gu and Kanade [GK08] propose a model-based face alignment system which can handle facial expressions, occlusions and image noise by using a model that allows multiple candidate positions for each facial landmark. Instead of a 2D model Jourabloo and Liu [JL15] propose a 3D model for face alignment that estimates both 2D and 3D landmark locations by projecting the 3D model into the 2D input image. Asthana et al. [AZCP14] use an incremental training to reduce the computational effort to train discriminative, person-specific models from a generic model.

In contrast to that, Yan et al. [YLYL13] use a cascade regression framework with separate HOG features for the initial, rough alignment and the subsequent refinement. To increase the overall performance Zhang et al. [ZLLT14, ZLLT16a] integrate several auxiliary tasks into their method like estimating the gender and pose as well as determining if the person in the image wears glasses or smiles. By using regression trees Kazemi and Sullivan [KS14] developed a face alignment method for single images which detects landmarks in a millisecond which makes this approach suitable for real-time applications. To enable real-time 3D face alignment in videos Jeni et al. [JCK14] use a dense cascade regression which is trained on 3D face scans. Besides face alignment, they also used a deformable 3D model to reconstruct a dense 3D mesh of the face. Zhu et al. [ZLLT16b] argue that cascaded regression approaches, for example the Supervised Descent Method (SDM) by Xiong and De la Torre [XD13], do not adapt well to other than near frontal poses. Without the need of temporal prior like in video-based methods or 3D face modeling, their approach handles arbitrary poses and facial expressions very efficiently. Instead of using a single cascaded regressor, the optimization space is partitioned into multiple domains of homogeneous descent as proposed by Xiong and De la Torre [XD15]. Another SDM related approach is introduced by Tuzel et al. [TMT16]. To enhance accuracy a transformation is added before each stage in the cascade of regressions which handles global changes in pose and shape, this enables to fine-tune each regression with respect to smaller variations caused by particular facial expressions.

Like in the field of landmark localization methods Deep Neural Networks are also becoming popular for face alignment approaches. Zhang et al. [ZKSC15, ZKSC16] use deep regression networks, whereas coarse-to-fine neural network approaches are proposed by Zhang et al. [ZSKC14] and Peng et al. [PFWM16]. A Convolutional Neural Network (CNN) is used by Trigeorgis et al. [TSN+16] as well as by Kumar and Chellappa [KC18]. Furthermore, Jourabloo and Liu [JL16] and Zhu et al. [ZLL+16, ZLLL17] perform face alignment by fitting a 3D face model to an image via CNN to increase the robustness for large variations in pose. A challenge to benchmark 3D face alignment methods has been developed by Jeni et al. [JTY+16].

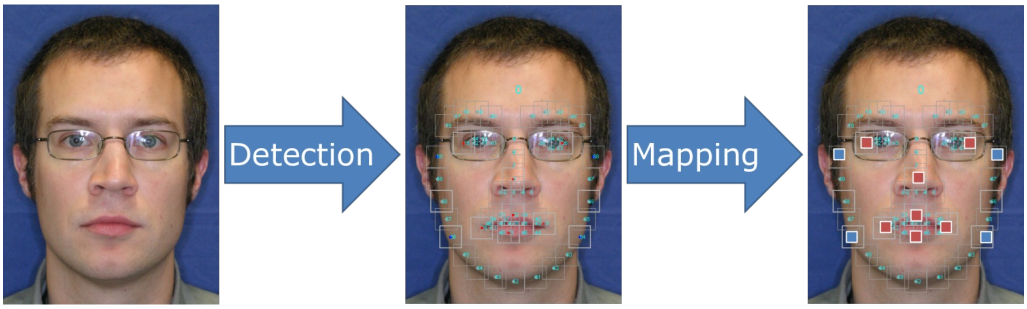


**Figure 5.1:** Landmark localization and mapping from a single image.

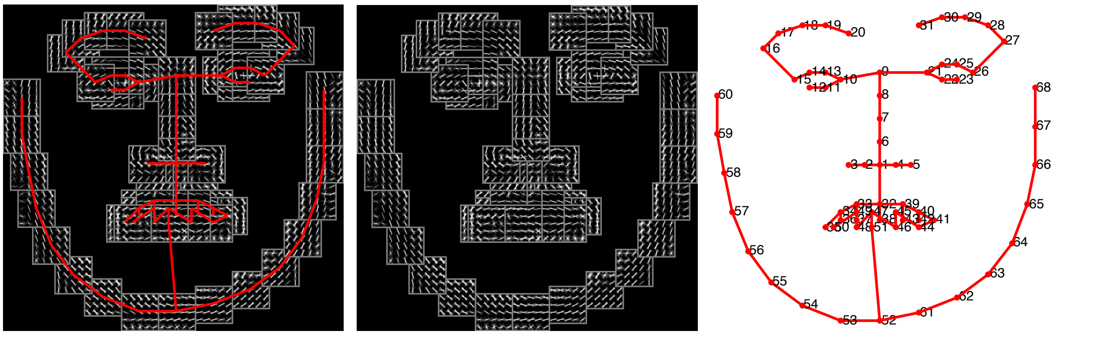
## 5.3 Automatic Initialization of the 3DMM

For an automated initialization of the 3DMM fitting algorithm, an approach by Zhu and Ramanan [ZR12] is used to identify facial landmarks automatically. Zhu and Ramanan use a tree-structured model which shows to be effective to capture the elastic deformations of faces. It can be used throughout all necessary preprocessing steps from face detection to pose estimation and landmark localization. At the same time, it is unaffected by facial expressions. Their algorithm is provided as open source with pre-trained models which localize either 39 or 68 landmarks to mark features such as eyes, nose, mouth and the contour points of the face. The exact number of landmarks depends on the pose of the processed face.

In Section 5.3.1 a brief description of Zhu and Ramanan's method is given, and Section 5.3.2 presents the adaptation of their work to allow an automatic initialization of the 3DMM framework (see Figure 5.1). Please note that in the meantime additional approaches for face detection and landmark localization have been published that outperform [ZR12] in some scenarios (see Section 5.2). Nevertheless, the availability of the complete source code is still a strong argument for using their framework as it enabled to integrate their approach seamlessly into the 3DMM framework which is introduced in Chapter 2.

### *5.3.1 Automatic Landmark Localization*

In [ZR12] a joint method is presented that starts with detecting all visible faces in an arbitrary image. After that, it locates the facial landmarks and estimates the poses by overcoming challenges like viewpoint dependent variations as well as elastic deformations caused by facial expressions. It uses a tree-structured mesh and a view-dependent topology (see Figure 5.2). The search for the locations of the facial landmarks is being repeated over multiple scales on an image pyramid. Based on their evaluation, Zhu and Ramanan report that their approach significantly outperforms Viola Jones [VJ04] in terms of face detection, and previous approaches in all tasks like pose and landmark estimation, particularly for extreme viewpoints.



(a) (b) (c)

Credit: Zhu and Ramanan [ZR12]

**Figure 5.2:** Figure 5.2a shows the full pictorial structure model for a frontal view, Figure 5.2b illustrates the local templates of the HOG features, and Figure 5.2c shows the tree-structured deformation model. More details can be found in Zhu and Ramanan [ZR12].

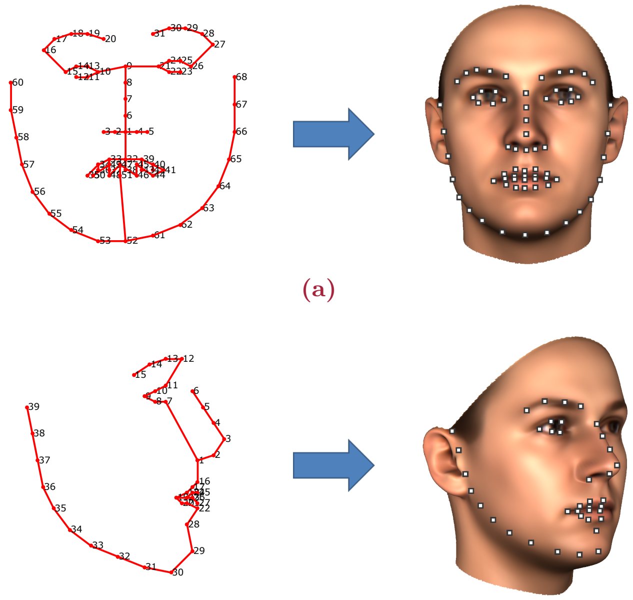
In their publicly available source code, Zhu and Ramanan provide three different pre-trained models: The first model is based on 146 template parts and works best for faces larger than 80*80 pixels. The second one is pre-trained with 99 parts and should work best for faces larger than 150*150 pixels. Finally, the third model is pre-trained with a total of 1050 template parts which gives best performance on localization compared to the two other models, especially if the resolution of the faces is larger than 150*150 pixels. Although it is significantly slower, its runtime performance is still acceptable because the current bottle neck is the 3DMM face reconstruction method of Blanz and Vetter [BV99].

The model of Zhu and Ramanan outputs the positions, sizes and poses of each detected face as well as the corresponding 2D landmark locations $L_{2\text{D}}$.

### *5.3.2 Integration into the 3DMM*

For automatic face detection, pose estimation and landmark localization high resolution input images are not necessarily required, in fact a too high resolution sometimes even impairs the results. Therefore, all images are re-sized if they exceed a maximum size of 800*800 pixels. On the other hand, the proposed processing pipeline automatically enlarges small images to a minimum of 500*500 pixels to ensure to some extent that the face shown in the image is not smaller than 150*150 pixels which is a prerequisite for the used landmark localization model (see Section 5.3.1).

Before the 3D reconstruction with the 3DMM (see Section 2.2) is applied, the results of the automatic landmark localization (see Section 5.3.1) are used to define a bounding box around each face which is used to crop the input image to the region of interest. If more than one face has been detected in an image, an individually cropped sub-image is



(a)

(b)

**Figure 5.3:** The lookup table maps each automatically detected landmark $L_{2\text{D}}$ to a vertex $L_{3\text{DMM}}$ on the 3D shape of the 3DMM. Depending on the detected pose, Zhu and Ramanan's algorithm generates two distinct landmark sets. For frontal views and small values of the pose angle $\phi$, it is the landmark set in Figure 5.3a. The set in Figure 5.3b is used for face profiles. Additionally, there is a series of different predefined lookup tables for each of the two sets and the exact pose determines which one is used (see Figure 5.4).

created for each face and the positional values of the landmark locations are updated to fit these sub-images.

Depending on the orientation of the detected face, the algorithm of Zhu and Ramanan returns one of two different landmark sets. The one shown in Figure 5.3a is used if the pose angle $\phi$ is smaller than 60°. Otherwise, if the face orientation is close to a profile view, the landmark set shown in Figure 5.3b is utilized. In addition to these two landmark sets, Zhu and Ramanan's method distinguishes viewpoints from $-90°$ to $90°$ in steps of $15°$. This more precise pose estimation is used by the proposed framework to select an optimal set of landmark points for the 3DMM initialization. Accordingly, as a function of $\phi$ the 2D landmark coordinates $L_{2\text{D}}$ are mapped to the 3D landmarks $L_{3\text{DMM}}$ of the 3DMM leading to

$$f(\phi): \quad L_{2\text{D}} \mapsto L_{3\text{DMM}}. \tag{5.1}$$

This mapping is illustrated in Figure 5.4 and is implemented by predefined lookup tables which map the pixel coordinates of detected facial features in the 2D image to 3DMM vertex indices.

Furthermore, two different types of feature points are distinguished: Fixed landmarks

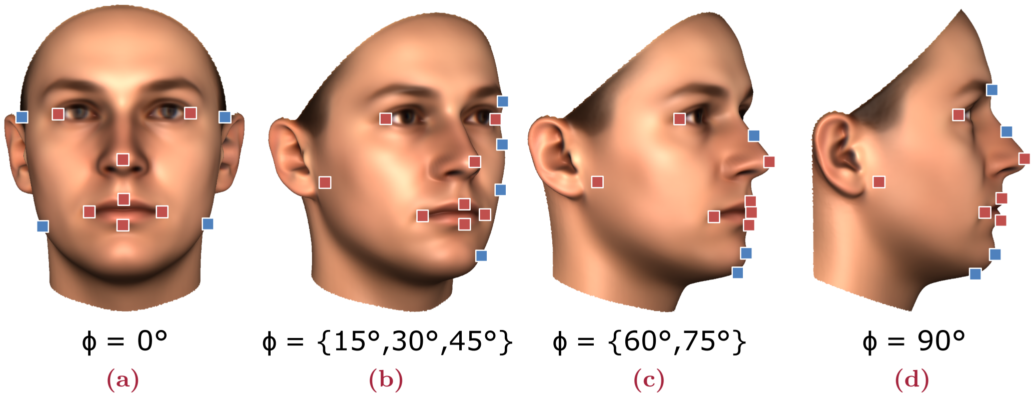


**Figure 5.4:** Depending on the pose, distinct landmarks are used to initialize the 3DMM fitting. Red squares denote fixed landmark locations and blue squares symbolize contour points. The latter are not linked to a specific vertex. Negative pose angles $\phi$ are handled in a similar way.

describe the position of the eyes, the tip of the nose and the mouth. In Figure 5.4 these landmarks are colored in red. The second kind of landmarks are contour points which are drawn as blue squares in Figure 5.4. They mark the boundary which separates the face from the background in the image. Contour landmarks are used as a starting point to match the 3D shape for an estimated pose along the complete contour of the 2D face and not only one single position during the fitting of the 3DMM. For each contour landmark, the closest vertex on the 3DMM silhouette is searched, and this mapping is updated throughout the fitting process.

Figure 5.4a to 5.4d show that not only a specific landmark set (see Figure 5.3) but also a varying subset of landmarks is used depending on the exact value of $\phi$. Even the assigned landmark type (fixed landmark versus contour point) may vary. This flexibility is a key aspect to replace manual landmark selection of a 3DMM framework with a proper automatic approach.

A comparison between Figures 5.3 and 5.4 illustrates that not all visible landmarks (white dots in Figure 5.3) are utilized. One reason is that the 3DMM should not be restricted too much during the initialization phase. Even for manually selected landmarks, it has been shown that it is not helpful to select as many landmarks as possible. Moreover, automatically detected landmarks may be false or imprecise, and the 3DMM might be able to correct misplaced landmarks during its optimization steps to some degree as long as it is not too much restricted by a large number of (imprecise) landmarks.

Another initialization step of the 3DMM fitting is to place the average head which is based on the facial data of the 3DMM at the location where the face has been detected in the input image. Besides the position of the bounding box, also the yaw and roll angles are used to pre-rotate the 3D mean face accordingly to the depicted face in the image. While the pose (yaw angle) is already estimated by the automatic landmark localization toolbox

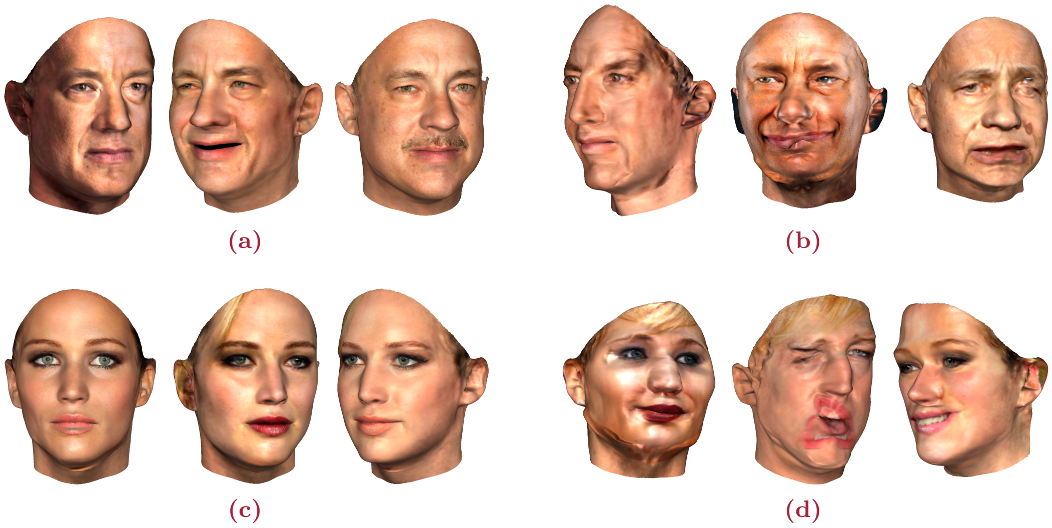


**Figure 5.5:** All images have been reconstructed using an automatic landmark localization method and an automatic 3DMM initialization. While the reconstructions based on different internet images of Tom Hanks and Jennifer Lawrence in Figure 5.5a and 5.5c are quite ok, in Figure 5.5b and 5.5d the 3D face reconstructions have failed.

of Zhu and Ramanan, the roll angle $\psi$ is approximated based on the horizontal deviation of the positions between the left eye $\mathbf{P}_{\text{left}}$ and the right eye $\mathbf{P}_{\text{right}}$ so that

$$\psi = \arccos \frac{\mathbf{P}_{\text{left}} \cdot \mathbf{P}_{\text{right}}}{\|\mathbf{P}_{\text{left}}\| \; \|\mathbf{P}_{\text{right}}\|}. \tag{5.2}$$

In case only one eye is visible, the algorithm will leave the roll angle unchanged which means that a default value for angle $\gamma = 0$ degree is assumed.

## 5.4 Degraded Reconstructions

A straightforward approach to use the automatically detected (see Section 5.3.1) and mapped (see Section 5.3.2) facial landmarks for a 3D reconstruction of arbitrary faces leads to mixed results in terms of reconstruction quality as is shown in Figure 5.5.

When using the 3DMM for face reconstructions from images the selection of landmarks is often an iterative process. First, expert knowledge about typically sensible landmark locations is used to decide which features should be selected. For example, the location of the eyes, the tip of the nose, the lobe of the ear and the corner of the mouth are mostly picked if visible, as well as some arbitrary points along the contour of the face. But depending on the pose and even the facial expression of a person additional landmark locations can be useful to improve the overall fitting quality. Accordingly, the second processing step is to fit the 3DMM to an image using the above mentioned locations and

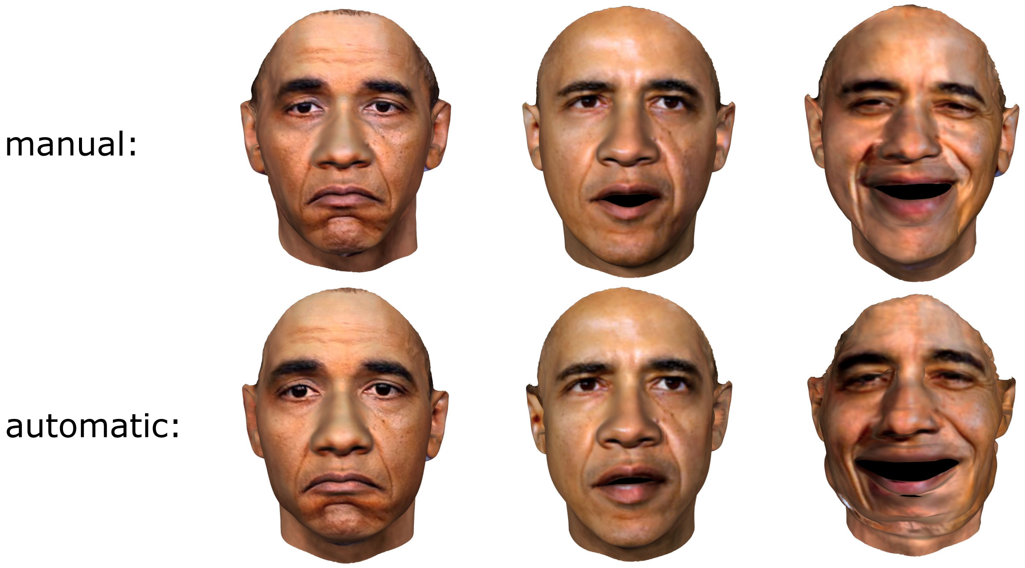


**Figure 5.6:** 3D reconstruction results based on three different images of Obama. For the reconstructions in the top row the facial landmarks were selected manually and in the bottom the automatic landmark localization method was used for the same three input images.

to evaluate the result. Depending on the outcome additional landmark locations are added or the position of already used locations is slightly modified. Sometimes these adaptation and evaluation steps are repeated several times until the quality of the final result is good enough or at least as good as possible, because for some input images it is hard or even impossible to create a sound reconstruction. This is often the case for extreme and uncommon facial expressions or for images that suffer from a low overall quality (see Chapter 4).

In Figure 5.6 a comparison between the reconstruction quality for manually (top) and automatically (bottom) selected landmarks is shown. For each column the same input image is used. All three input images have a high resolution and reasonable lighting conditions. While the quality of the reconstruction in the first two columns is more or less similar, in the last column the reconstruction using manually selected landmarks is superior.

If the selection of facial landmarks is a manual process, the above mentioned steps to improve the reconstruction quality iteratively may be time consuming but enhance the chance for a pleasing reconstruction result. But this also means that as long as the used automatic landmark localization method is not as precise as a human user, the final outcome based on manual detected landmarks is expected to be better compared to automatically detected landmarks. Furthermore, in case of automatic landmark localization and mapping methods the described iterative process is not applicable directly without the need of user interaction. Accordingly, failed reconstructions like in Figures 5.5b and 5.5d can not be fixed automatically without additional knowledge about the person's face in the image.

## 5.5 Enhancing the Appearance of 3D Reconstructions

To make use of additional knowledge about the person's face a simultaneous fitting to multiple images of the same person as is discussed in Section 5.5.1 might be helpful, but it does not necessarily lead to a higher robustness, especially if the available images contain so called 'faces in the wild'[6]. This means that the lighting conditions are completely uncontrolled, the facial expression and pose may vary a lot and parts of the faces may be occluded. Accordingly, in Section 5.5.2 an approach is described to enhance the robustness of the 3D reconstruction based on multiple images.

### *5.5.1 Simultaneous Fitting to Multiple Images*

Although the idea of performing a simultaneous fitting to multiple images with varying facial expressions $\gamma$ has already been mentioned in [BV99], it has not been implemented before the proposed work. Basically the single image fitting method which is discussed in Section 2.2 can be used to achieve a 3D reconstruction by simultaneously fitting to multiple 2D images or views of the same person. Nevertheless, some adaptations are necessary to handle multiple input images with varying facial expressions. Accordingly, the individual image distance $d_{\text{image}}$ (see Equation 2.10) of each view or input image needs to be summed up. The same applies to the feature distance $d_{\text{features}}$ (see Equation 2.12) between 2D feature positions $u_j, v_j$ and the projected positions of the corresponding vertex $k_j$ of each view. Then the corresponding cost function $E_{multi}$ is defined as

$$E_{multi} = \eta_{\text{image}} \cdot \sum_{view} d_{\text{image},view} + \eta_{\text{features}} \cdot \sum_{view} d_{\text{features},view} + \eta_{\text{maha}} \cdot d_{\text{maha}}, \tag{5.3}$$

where $\eta_{\text{image}}$, $\eta_{\text{maha}}$ and $\eta_{\text{features}}$ are the predefined weights of the current processing step of the analysis-by-synthesis loop. Furthermore, the regularization term $d_{maha}$ which is based on the Mahalanobis distance like in Equation 2.11 in Chapter 2 needs to be adapted to

$$\begin{aligned} d_{maha} = & \sum_i \frac{\alpha_i^2}{\sigma_{S,i}^2} + \sum_i \frac{\beta_i^2}{\sigma_{T,i}^2} \\ & + \sum_{view} \sum_i \frac{\gamma_{i,view}^2}{\sigma_{U,i}^2} \\ & + \sum_{view} \sum_i \frac{\left(\rho_{i,view} - \overline{\rho}_{i,view}\right)^2}{\sigma_{R,i}^2}. \end{aligned} \tag{5.4}$$

[6] A synonym for images of faces captured under arbitrary, uncontrolled conditions (lighting, facial expressions, pose, etc.). The term has been originally used in the context of the *Labeled Faces in the Wild* database [HRBLM].

Like in Section 2.2 the regularization term counteracts overfitting by regarding the standard deviation $\sigma$ during the fitting of the linear coefficients for shape $\alpha_i$, texture $\beta_i$, facial expression $\gamma_i$, as well as additional imaging parameters $\rho_i$ that control pose, lighting and rendering parameters for the corresponding principle component $i$ of each individual PCA.

As an independent facial expression for each input image is allowed, the coefficients $\alpha$ and $\beta$ for shape $\mathbf{S}$ and texture $\mathbf{T}$ of the 3D reconstruction are shared during the optimization procedure, but the coefficients $\gamma$ for the facial expression $\mathbf{U}$ are altered individually for each image. This enables to use arbitrary images of a person without being restricted to neutral or unvaried expressions.

After the 3D reconstruction has been completed it is possible to assimilate one of the expressions which have been detected in the 2D input images by applying the corresponding expression coefficients to the final shape (see Equation 2.8) or to neutralize the expression of the result by just setting all expression coefficients $\gamma_i$ to zero. Furthermore, it is also possible to transfer the expression of a completely different person to the new reconstruction by using their expression coefficients, because all reconstructions are in dense point-to-point correspondence due to the underlying 3DMM.

By testing different reconstruction scenarios Breuer [Bre10, pp. 115–122] showed that in most cases a multi-fitting approach is superior to a fit to only a single image, especially if a combination of different poses is used, for example a frontal plus a profile view of the same face. Although Breuer performed these experiments only on images which show faces with a static and neutral facial expression, her findings are also applicable to the proposed multi-fitting approach which can handle variations in facial expressions. Of course this is somehow obvious as, in contrast to a situation where only a single image is available, the combination of several input images of the same person provides additional information about the given face. But it has also been found out that using more than three input images in the multi-fitting approach does not necessarily lead to a better result. For several examples, even the contrary was the case. And as the processing time increases with each additional image, a trade-off between quality and computational effort needs to be defined. Accordingly, a two-step approach is proposed in Section 5.5.2 which uses the described multi-fitting method (see Equation 5.3 and 5.4) for image pairs and then averages all the individual, pairwise reconstructions to create the final result.

### *5.5.2 (Pairwise) Averaging of 3D Reconstructions*

Although the automatic feature detection algorithm (see Section 5.3) often works quite solid, the delivered landmarks are still a bit behind the accuracy of a manual feature point selection, especially for complex scenes caused by extreme expressions, poses or lighting conditions. Furthermore, the 3DMM (see Chapter 2) sometimes tends to deliver untypical facial shape reconstructions. To overcome these drawbacks the facial attributes of a person

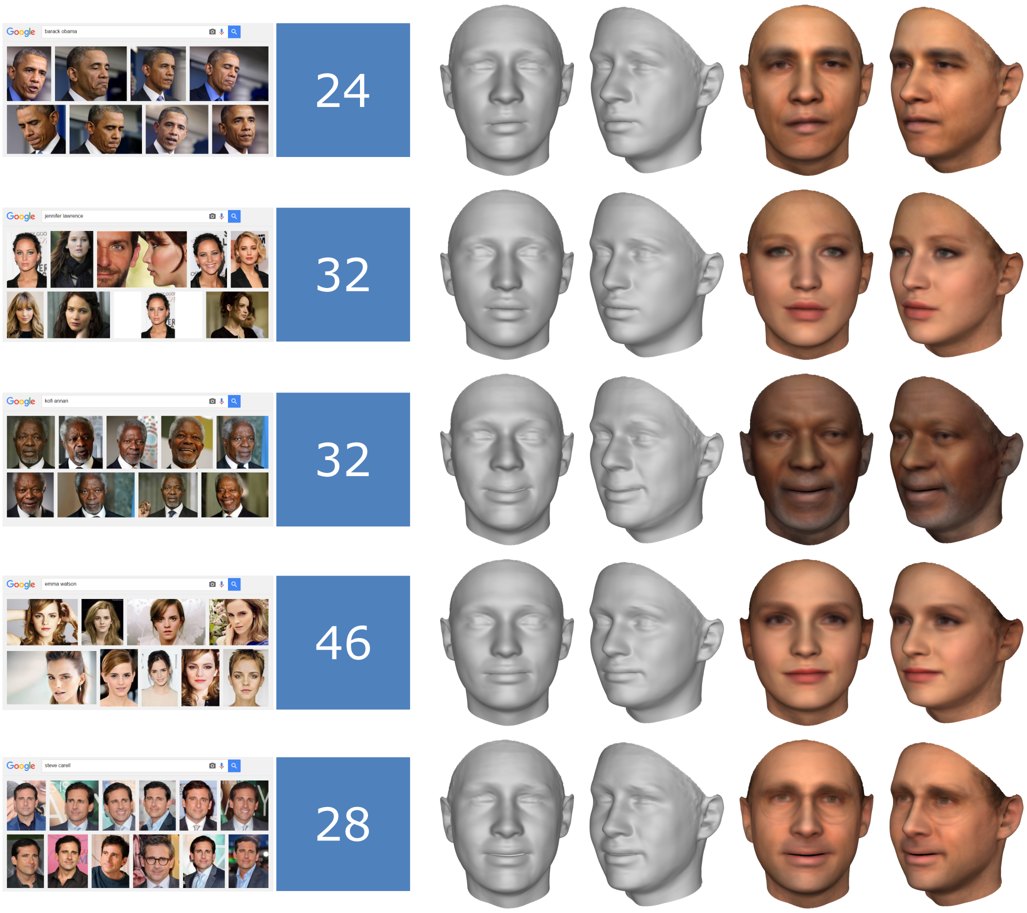


**Figure 5.7:** Averaging the results based on 3D reconstructions of all individual images from photo collections (left) results in the shown 3D facial shapes (right), regardless of whether a reconstruction from an image resulted in a 'good' or a 'bad' reconstruction. The extracted texture from each image is averaged to gain the facial color information. The number next to the photo collection is the total amount of used images including 'bad' candidates.

are gathered by reconstructing the 3D face from several images independently followed by an averaging of all individual 3D reconstructions. As the 3DMM provides a dense point-to-point correspondence between all 3D faces, the averaging can be done straight forward. This provides an enhanced 3D reconstruction which preserves individual facial features of a person while weakening fitting artifacts that appeared in some intermediate reconstructions due to a less-than-ideal optimization or misleading feature points. In Figure 5.7 five examples of the described averaging procedure are shown. Regardless of whether a reconstruction from a single image lead to a 'good' or a 'bad' reconstruction result, all individual 3D reconstructions from the 2D input images have been combined with each other to create the final, averaged 3D face reconstruction.

In nearly all cases, averaging multiple 3D reconstructions is superior to each individual reconstruction regardless of whether it is based on a single image (see Section 2.2) or if the simultaneous fitting to several images is used which is described in Section 5.5.1. Actually, the averaging of the multi-fitting results leads to the most plausible results. But as is already mentioned in Section 5.5.1 the proposed multi-fitting approach does not necessarily create better reconstruction results when increasing the number of input images that are simultaneously fitted. It is much more important that one image provides information which is not captured in the other images. Then again many different images that are processed at once may even affect the Stochastic Newton Descent algorithm [BV99] negatively from finding the local minima.

Thus, the idea is to limit the simultaneous fitting to two input images at once and additionally making use of the pose estimation from the automatic landmark localization method. In this context three different approaches are discussed; all of them avoid to pair up images from the dataset randomly.

**Option 1** lets the user manually choose one of the available images to be used as a reference image. In the following the reference image is pair-wise combined with all the other available images. A good candidate for a reference image could be a photo for which the user hopes that it counteracts misleading visual effects like glasses, extreme facial expressions or difficult lighting conditions as well as misleading feature points of other images.

**Option 2** provides an alternative way which does not require any additional user input. Here the localized facial landmarks and estimated pose are utilized to select an appropriate image candidate automatically. Thus, only those images are allowed as a reference image which have a pose that is nearly frontal and where the landmarks for the lips are arranged in a way that is similar to faces with neutral expressions (i.e. the lips are close to each other etc.). From these remaining candidates, one is randomly picked.

**Option 3** is to pair up images based on well matching poses: for example if a frontal view of a person is selected for the 3D reconstruction a profile picture of the same person would be searched in the database and vice versa. In difference to the two other options the reference image is not fixed, instead it is updated depending on the current input image.

After each input image has been paired to a reference image by one of the above presented options, for each couple a multi-fitting is performed and at the end all resulting, individual 3D reconstructions are averaged to acquire the final 3D reconstruction. Please note that Option 1 has been used to create the results in Figure 5.8, while in Figure 5.7 all single images have been used as they are to generate individual 3D reconstructions which are then averaged to create the final 3D face reconstruction per person.

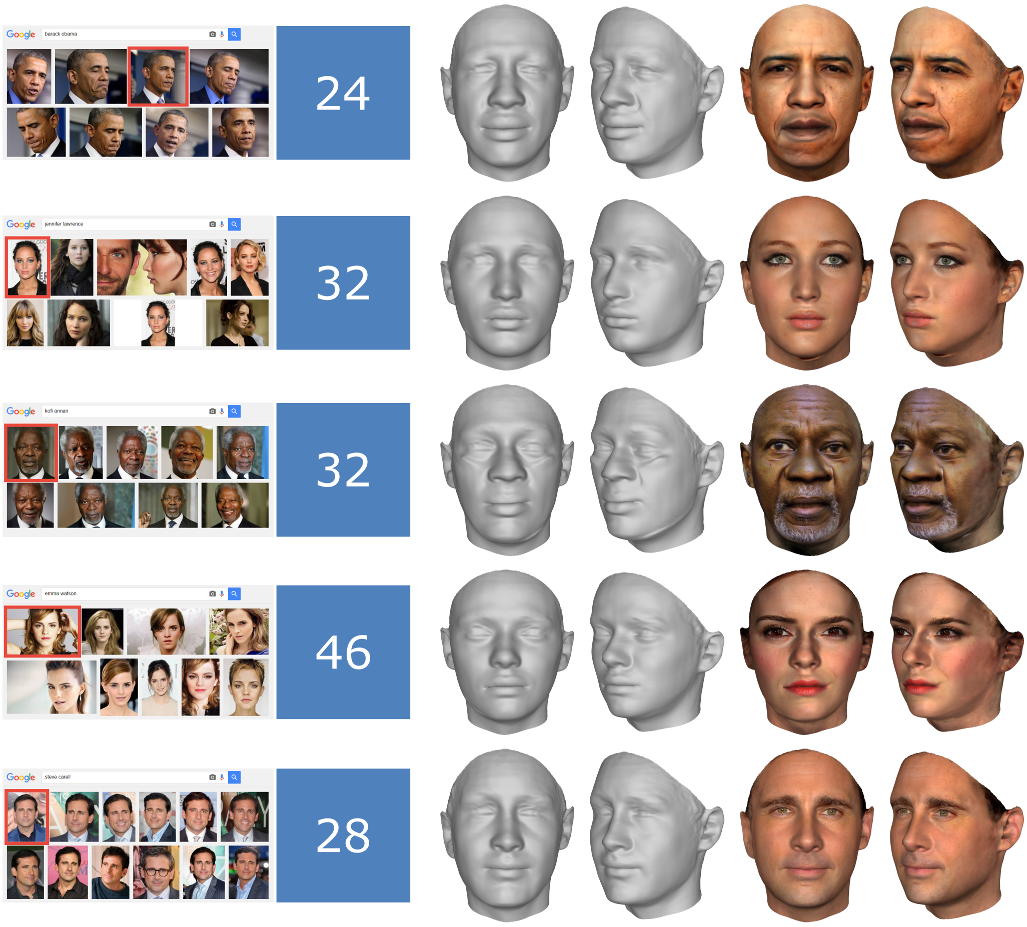


**Figure 5.8:** 3D reconstruction results for different individuals: From left to right each example contains a representative example of the photo collection with the reference image being outlined in red and the 3D reconstruction without and with texture extraction from the reference image.

## 5.6 Results

An example of the automatically generated 3D reconstruction results is shown in Figure 5.8. As described in detail in Section 5.5.2 the reconstruction starts with arbitrary images of the same person. By using the approach of Zhu and Ramanan [ZR12] the face is automatically detected in each image. Furthermore, it estimates the pose and localizes the facial landmarks. Based on pose and landmarks the original input images are cropped to the region of the face and these subimages are added to an image collection that corresponds to exactly one individual for further processing. Please note that an input image is only discarded if no face has been detected at all. If there are multiple faces in an input image only the subimage which shows the person of interest is processed. As no face recognition is implemented, this step may require some user input but only if images with multiple persons are included in the given dataset. However, this kind of user input is much less time-consuming than the manual selection of facial landmarks.

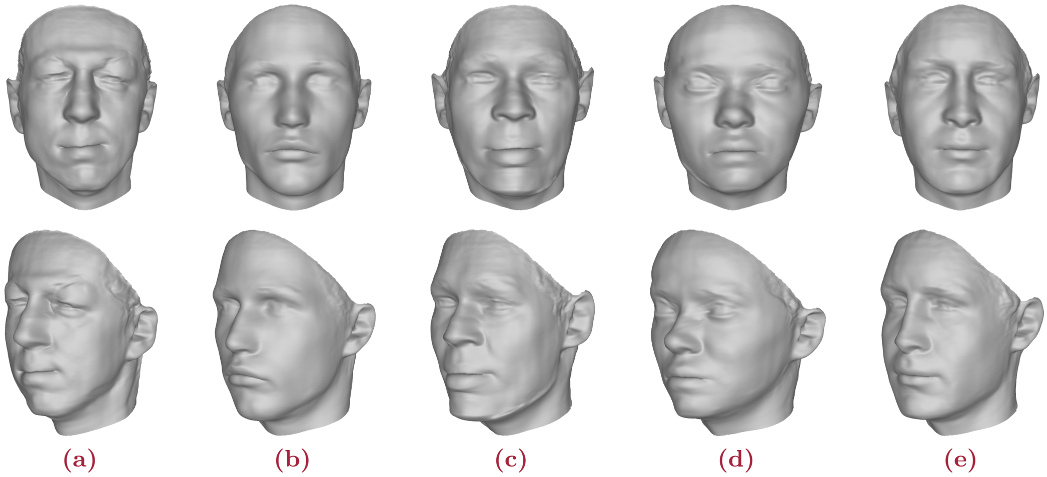


**Figure 5.9:** Single image 3D reconstructions of the reference images which are marked in Figure 5.8: 5.9a Obama, 5.9b Lawrence, 5.9c Annan, 5.9d Watson and 5.9e Carell.

Before the 3D reconstruction is started, one of the images in the collection is picked as a reference face (see Section 5.5.2). Mostly it is chosen based on the estimated pose and landmarks: the rotation along the y-axis (yaw angle) should be less than ±45 degrees and a neutral facial expression is preferred (i.e. the facial landmarks for the upper and lower lip are close to each other). The reason for this procedure is that the landmark localization method of Zhu and Ramanan is more precise for neutral faces and also the quality of the 3DMM reconstruction generally tends to be superior in case of neutral or least not too extreme facial expressions.

In the following, a simultaneous fitting for each image pair consisting of one image from the image collection and the reference image using the 3DMM is performed (see left part of Figure 5.8). In a final step, a pooling operation is performed by averaging all pairwise reconstructions to form the final facial shape as can be seen in the right part of Figure 5.8. The results are rendered once without and with applying the extracted texture of the reference image of each collection.

As described above a face with a neutral expression is more likely chosen as reference image. Additionally, only the shape and texture coefficients are used during the averaging, whereas the coefficients for the facial expressions are discarded at this point. Consequently, this results in 3D reconstructions with neutral facial expressions. But due to the 3DMM fitting a dense point-to-point correspondence is established between all reconstructions. This allows to transfer the facial expression of any processed input image of that person or even from a completely different person to the reconstructed 3D face as is shown in [BBPV03].

In comparison to the averaging in Figure 5.7 the facial reconstructions in Figure 5.8 capture the individual facial attributes a bit better as can be seen especially at the noses of Obama and Annan. Additionally, in Figure 5.9 the 3D reconstructions of each reference image are shown. These are created by using the standard single image fitting approach of the 3DMM (see Section 2.2). Especially the 3D reconstructions of Obama (Figure 5.9a) and Annan (Figure 5.9c) are inferior to the averaged results in Figure 5.7 and the pairwise approach with subsequent averaging which is shown in Figure 5.8. Accordingly, it is not the case that an outstanding well reconstructed reference image is responsible for the improved reconstruction quality in Figure 5.8.

## 5.7 Conclusions

For a long time it has been extremely time-consuming to reconstruct faces with the 3DMM as the underlying approach requires the selection of at least a few facial landmarks per input image. Although in [Bre10], [BB10] and [BKK$^+$08] ideas are presented that reduce this effort, especially in case of varying facial expressions the resulting reconstructions are often neither robust nor plausible. Unlike most already existing approaches that utilize a 3DMM, the presented method overcomes the need of manual initialization of the 3DMM framework. Furthermore, it enables to combine facial information from several different images of the same person to gain a more robust 3D reconstruction of the person's face. In contrast to geometry based approaches, it can handle arbitrary input images even if they include non-rigid deformations. Accordingly, the proposed approach generates plausible 3D reconstructions even if the input images include facial expressions.

By combining an automatic approach for landmark localization (see Section 5.3) and a simultaneous fitting approach that uses multiple images for a 3D reconstruction (see Section 5.5), a basis to analyze large image collection of faces has been established which automatically creates dense correspondence between all processed faces. For example, without the requirement of manually selected landmarks it will be easier to study effects like aging or to learn even more about facial expressions and facial textures, because such studies usually require a large number of input images.

Nevertheless, the proposed methods in Sections 5.5 still suffer from some limitations: The averaging of all individual 3D reconstructions may result in an overall more robust reconstruction than a fit of the 3DMM to a single image but all the potential fitting errors of each individual reconstruction can be neither identified nor prevented. Instead, even in the best case scenario these individual imperfections are only averaged out to some degree. The proposed method of using a promising image as a reference has some additional drawbacks. First of all, a manual selection of the reference image prevents a completely automatic reconstruction approach, even though the selection of an image may be much less time consuming than precisely selecting multiple facial landmarks in several images. Secondly, the selection may require some expert knowledge to decide which image is a good candidate and this leads to a third drawback: if the final reconstruction result is still not pleasing the reason is hard to identify. Accordingly, it would be necessary to sort out if a different reference image needs to be used, if the automatically detected landmarks are too imprecise or if the used input images themselves are the main source of the problem. To answer some these open questions a more sophisticated approach for an automatic 3D face reconstruction from multiple images is presented in Chapter 6.

# 6 Automated 3D Face Reconstruction from Multiple Images

Automated 3D reconstruction of faces from images is challenging if the image material is difficult in terms of pose, lighting, occlusions and facial expressions, and if the initial 2D feature positions are inaccurate or unreliable as is shown in Section 5.4.

Like in Chapter 5 and distinct from earlier work on 3DMM fitting [BV03], a feature detection algorithm by Zhu and Ramanan [ZR12] is used to enable an automated process. But the reduced precision of these feature locations and the suboptimal choice of features (silhouettes, ears) affect the quality of the output significantly. Thus, in this chapter a method is proposed that reconstructs individual 3D shapes from multiple single images of one person, judges their quality and then combines the best of all results. This is done separately for different regions of the face. The core element of this algorithm and the focus of this chapter is a quality metric that judges a reconstruction without information about the true shape. In the following, different quality measures are evaluated, a method for combining results is developed, and a complete processing pipeline for automated 3D face reconstruction is presented. It selects the most successful reconstructions based on a given set of images of a person and combines them to a 3D face. Accordingly, this approach is more sophisticated than the blind averaging presented in Chapter 5. It has been published in [PB16].

## 6.1 Introduction

Algorithms that reconstruct 3D faces from images by fitting a deformable face model, such as a 3D Morphable Model (see Section 2), rely on a relatively precise initial positioning of the face [BV99] or on a set of feature point coordinates [BV03]. For an automated procedure, it seems to be straight-forward to combine these algorithms with automatic face and landmark localization, such as the algorithm by Zhu and Ramanan [ZR12] or other feature detectors [BKK$^+$08]. In practice, however, this combination has turned out to be more challenging than expected, posing a number of fundamental questions. The feature point detection is a non-trivial task, especially if the image material includes complex lighting, facial expressions, wrinkles, eye glasses or facial hair. Thus, the features may be inaccurate, and some may even be outliers. Moreover, the optimal set of features for 3DMM fitting includes points that are not easy to detect, such as the facial silhouette and the ears. Those points are necessary for the 3DMM to converge to the correct pose angle, and this in turn affects the shape estimate.

Therefore, a simple combination of existing methods produces results that are substantially worse than those obtained with manually labeled features (see Section 5.4). Attempts to make 3DMM fitting more robust [BB10] are promising but still not sufficient. Instead, this work brings forward the argument that in many real-world applications more than one image of a person is available, so an automated algorithm can exploit redundant data from multiple images to gain robustness and reliability. The proposed algorithm outperforms existing methods of simultaneous 3D reconstruction from multiple images [BV99] significantly, which may be due to the fact that outliers in feature positions adversely affect the simultaneous least squares solution.

In contrast, the new algorithm calculates separate reconstructions from each input image, and then combines them to an optimal overall solution. Thus, the proposed method selects the most plausible reconstructions, operates on different regions of the face separately, and merges them into a single 3D face (see Figure 6.1).

The key component of this algorithm is a new measure for the visual quality of 3D reconstructions, based on surface normals. Automated assessment of visual quality in computer graphics and vision is a fundamental challenge. Simple image comparisons are insufficient because they are insensitive to small but important errors and artifacts. Euclidean distance in 3D overrates global shape deformations that would be irrelevant to human observers. Mahalanobis distance is also inconsistent with the quality ratings of humans. In an experimental comparison with quality ratings from human subjects, the new, normal based measure outperforms these existing criteria.

In summary, the contributions of this chapter are: First, a general measure of the quality (naturalness) of a shape reconstruction. Secondly, an algorithm for selecting and combining reconstructions of different facial regions (segments) from different input images into a single 3D face. And last but not least, an automated algorithm that produces 3D shape reconstructions from multiple images of a person, which goes beyond a simple combination of landmark localization and 3DMM fitting.

The remainder of this chapter is organized as follows: The related work in the field of 3D face reconstruction is discussed in Section 6.2. Then a series of different quality measures

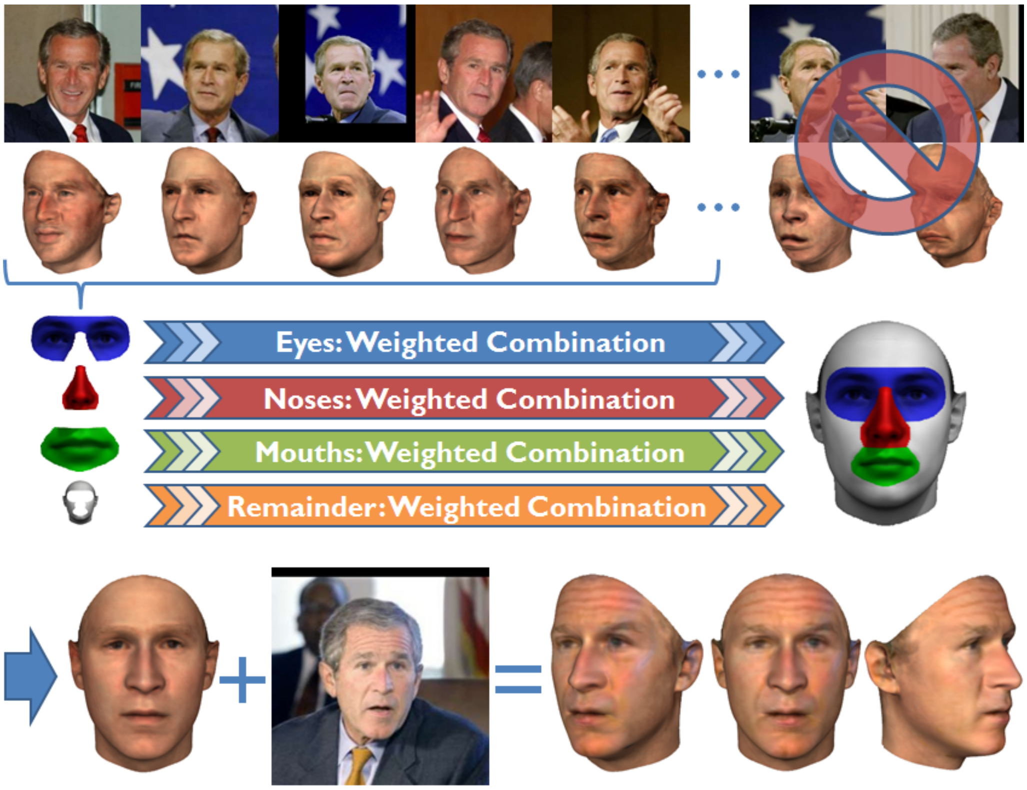

Eyes: Weighted Combination
Noses: Weighted Combination
Mouths: Weighted Combination
Remainder: Weighted Combination

**Figure 6.1:** A segment-based, weighted linear combination is used to create the final head shape. The weight decreases with the rank. Implausible segments are discarded. Note that each facial segment is handled separately. Optionally the texture can be extracted from one of the input images.

is introduced in Section 6.3 and evaluated in Section 6.4. Afterwards, the segment-based pooling (Section 6.5) and the distance driven texture extraction (Section 6.6) are presented. The corresponding results are shown in Section 6.7 and finally a conclusion is given in Section 6.8.

## 6.2 Related Work

Although several approaches which are related to high quality 3D reconstructions of faces from 2D images have been published, automated reconstruction still remains a challenging task, especially for unconstrained photos with facial expressions, difficult lighting situations or occlusions.

In the literature on face modeling several different approaches can be found. For high quality 3D reconstructions of faces which are used in computer games and movies, the state of the art techniques still require 3D scans of the person using commercial range scanners [Kon08, HP16] or multi-view camera setups [HVC08, ARL+09, BBB+10, BHPS10, GFT+11, BHB+11, KH12, AFB+13]. To achieve accurate face reconstructions the scans are performed under controlled conditions in regard of lighting, pose and facial expressions as well as known equipment. Additionally, substantial post-processing is required to combine the generated 3D data and to morph between different facial expressions and visemes to realistically animate the subject's face.

Approaches like the one presented in this chapter try to obviate the need for special equipment and work in unconstrained environments. Instead, they make use of data that can be easily produced with standard equipment or that is already available, such as photo or video data. Multi-view geometry [PKv99, HZ04, SMP07, GSC+07] is a common procedure to reconstruct 3D shapes from several single images or video frames. Although these algorithms are quite flexible in usage for different scenarios varying from the reconstruction of buildings, smaller objects and even faces, they cannot sufficiently handle varying, non-rigid transformations (e.g. facial expressions) within a series of input images. Recently, several improvements have been made in case of unconstrained 3D reconstruction from video. Video-based reconstruction approaches [GVWT13, SKSS14, SWTC14, FJA+14, IBP15, CBZB15, NFS15, TZS+16, GZC+16] take advantage of the fact that each frame is taken by the same camera and assume temporal coherence.

Other recent publications have shown promising results by aligning a 3D face to a single or multiple images as well as to video frames. The approaches by Park et al. [PHS08], Aldrian and Smith [AS10], Lee et al. [LLP+12] and Dou et al. [DWSK14] reconstruct the 3D shape from a single image. Wang et al. [WZS05] extract the silhouette from several input images to reconstruct the 3D shape, while Roth et al. [RTL15] use an image collection for photometric stereo-based normal estimation which iteratively optimizes the surface reconstruction. Most photometric stereo approaches [LK81, Woo89, YSEB99, LHK01, BJK07, HASS10, WAB+11] produce only a 2.5D face reconstruction and are therefore limited to (near) frontal images. By estimating the pose and computing the optical flow like in [KSS12, KS13, Has13, SKSS14] a high detail refinement of the 3D shape is performed, resulting in a 3D to 2D correspondence. Suwajanakorn et al. [SKSS14] even captured fine details like wrinkles and in [KSS11] Kemelmacher-Shlizerman and Seitz showed that also 'faces in the wild' can be handled properly. But these approaches lack an additional 3D to 3D correspondence. In this chapter 3D to 2D as well as 3D to 3D correspondence is addressed.

To reconstruct a 3D shape of a face from a 2D image, Blanz and Vetter [BV99] introduced the 3DMM (see Chapter 2). With the Basel Face Model [PKA+09], a 3DMM has been made available to the public and was optimized by Thies et al. [TZS+16] to allow for real-time face capture and reenactment from video. Distinct from video, temporal coherence cannot be assumed for unconstrained image collections and for single images there is no coherence at all. This makes the reconstruction process a much harder task compared to video. Patel and Smith [PS09] used statistical tools to improve the accuracy of 3DMMs, whereas Zhu et al. [ZYY+15] presented a discriminative 3DMM based on local features for accurate reconstructions. Using adaptive contour fitting Qu et al. [QMSB15] perform pose-invariant face reconstruction using a 3DMM. But a common and significant drawback of the 3DMM is its lack of robustness in case of ’faces in the wild’, especially if the facial landmarks are not perfectly detected. Although Breuer et al. [BKK+08] propose to use a Support Vector Machine for automatic 3D face reconstruction and in [BB10] an idea is presented to correct misplaced landmarks to some extent, both implementations are not robust enough to handle difficult scenarios caused by facial expressions or complex lighting conditions. The new approach in this chapter is designed to overcome the previous drawbacks of the 3DMM.

Since the publication of the proposed method in [PB16], additional approaches for 3D face reconstruction have been presented. For example, Peng et al. [PXF16] reconstruct a 3D face using B-Splines, Roth et al. [RTL16, RTL17] describe a method to reconstruct the 3D face from unconstrained photo collections and Khan et al. [KRS+16] recover the facial parts that are occluded by a virtual reality headset. Using hundreds of photos and clustering them according to their pose Liang et al. [LSKS16] reconstruct the head from internet photos and Liu et al. [LZZL16] combine face alignment and face reconstruction in a joint procedure. The approach by Bas et al. [BSBW16] uses solely hard correspondences like landmark locations and edges, whereas Schönborn et al. [SEMFV17] propose a Markov Chain Monte Carlo approach to automatically fit a 3DMM to single face images. Cao et al. [CCC+18] use a 3DMM and a sparse photometric stereo approach to reconstruct a high quality 3D face geometry from a set of five input images which have been taken in a controlled environment. A Convolutional Neural Network is trained by Richardson et al. [RSK16] and Tran et al. [THMM17]. To avoid scanning a large number of real faces the training is performed on artificial facial images which are created by a 3DMM. Inspired by Piotraschke and Blanz [PB16] in [THMM17] a ranking list is used to pool shape and texture parameters to generate training data, but instead of using the Normal Distance (see Section 6.3.5) as a quality measure, the weighting factors are generated based on the confidence measure of their landmark localization algorithm. By separating the global shape from local details Tran et al. [THM+18] achieve detailed 3D reconstructions even if parts of the face are occluded. While Booth et al. [BAP+17] and Trigeorgis et al. [TSKZ17] focused on ’faces in the wild’, Tran and Liu [TL18] created a non-linear 3DMM from a large set of unconstrained photos without collecting 3D face scans.

Additionally, there are methods which are not aiming at the reconstruction of faces directly, but provide a strong foundation for further processing by detecting faces, estimating the pose, localizing feature points or aligning face geometries [ZR12, ZBL13, PPTK13, XD13, JCK15]. More details and similar approaches are discussed in Section 5.2.

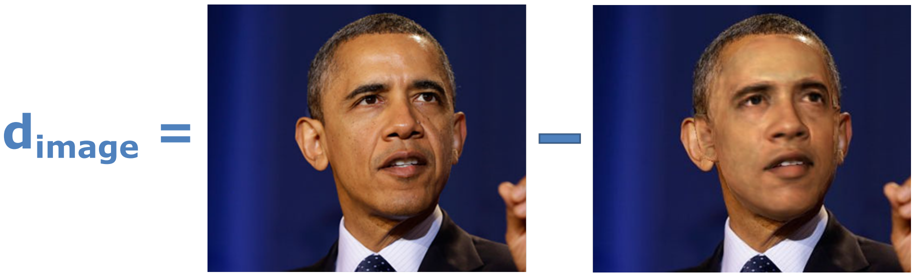


**Figure 6.2:** The image distance is computed by subtracting the input image $I_{\text{input}}(u,v)$ with a modified version $I_{\text{model}}(u,v)$ where the reconstructed face is rendered on top of the original face.

## 6.3 Quality Measures

For a meaningful quality measure, it is important to be independent of facial expression. Therefore, a 'neutralized' facial expression is generated for each reconstructed face $\mathbf{S}$ after the 3D reconstruction step and before computing each quality measure except the image distance. As the image distance compares the rendered reconstruction with the original input image, it needs to be as close as possible to the original face.

A neutral expression is gained by setting the weights $\gamma_i$ of the facial expression eigenvectors $\mathbf{u}_i$ in

$$\mathbf{S} = \overline{\mathbf{s}} + \sum_{i=1}^{m-1} \alpha_i \mathbf{s}_i + \sum_{i=1}^{p-1} \gamma_i \mathbf{u}_i$$

to zero. Like in Equation 2.8 the final 3D shape $\mathbf{S}$ is based on the average face $\overline{\mathbf{s}}$ of the 3D face model, the eigenvectors of shape $\mathbf{s}_i$ and expression $\mathbf{u}_i$ as well as their corresponding coefficients $\alpha_i$ and $\gamma_i$. Furthermore, $m$ and $p$ are the numbers of the principle components which arise from the shape and expression PCA.

### *6.3.1 Image Distance*

*Idea: Successful reconstructions are very close to the input image.*

In contrast to all others distance functions that are discussed in this section, the image distance

$$d_{\text{image}} = \sum_{u,v} \| I_{\text{input}}(u,v) - I_{\text{model}}(u,v) \|^2$$

which has been introduced in Equation 2.10 is the only distance measure that penalizes differences between the original face and the 3D reconstruction. All other distance measures only estimate the plausibility and naturalness of the reconstructed faces.

But Figure 6.2 illustrates one major drawback of this error function: it is not possible to penalize the fact that the rendered and projected face does not occlude the complete face in the input image like in case of Obama's right ear. In $I_{input} - I_{model}$, the image distance for most pixels of the right ear is zero and therefore the error is quite small. The reconstructed ear is rendered on the cheek, but due to the similar color, this has also little effect on $d_{\text{image}}$. In general, $d_{\text{image}}$ fails to capture small but relevant errors and artifacts in the reconstruction.

On the other hand, $d_{\text{image}}$ can also be high even though the faces look similar, for example this happens if the overall color tone is wrong or the face is slightly shifted. All these problems are caused by the fact that $d_{\text{image}}$ is a sum of all pixels and that many small errors might count more than a few large errors.

Please note that even though $d_{\text{image}}$ turns out to be suboptimal for rating the quality or plausibility of the 3D reconstruction, as is demonstrated in Section 6.4, it still makes sense to use $d_{\text{image}}$ in the fitting procedure because, unlike all other criteria which are discussed in the following, it measures the distance from the input face, and it is easy to compute.

### 6.3.2 Mahalanobis Distance

*Idea: Successful reconstructions are close to the average, while failed ones are further away and therefore look odd.*

The Mahalanobis distance $d_{\text{maha}}$ measures the distance of the current solution from the average face using PCA, taking into account the standard deviations observed in the training data, so that directions with a high variance are weighted lower than directions with a small variance. Thus, it is directly related to the multivariate Gaussian probability density function which is estimated by PCA. Like the image distance, the Mahalanobis distance is already integrated in the 3DMM fitting procedure (see Section 2.2) and using the Mahalanobis distance to measure the quality of a 3D face reconstruction is strongly related to the concept of the 3DMM in Chapter 2: The shape $\mathbf{S}$ of a new face can be approximated by linear combinations of principle components $\mathbf{s}_i$ of the individual shapes (see Equation 2.7) plus the average shape $\bar{\mathbf{s}}$ of all faces in the training data, so that

$$\mathbf{S} = \bar{\mathbf{s}} + \sum_i \alpha_i \mathbf{s}_i \tag{6.1}$$

For the experiments in Section 6.4, where only the quality of the reconstructed shape is rated Equation 2.11 is simplified to

$$d_{\text{maha}} = \sum_i \frac{\alpha_i^2}{\sigma_{S,i}^2}, \tag{6.2}$$

so only the distance of the neutral face shape from the average face is measured, whereas expressions, texture and rendering parameters are omitted. The motivation is that, unlike the neutral shape, the texture and expression of a successful reconstruction may be far from the average if the input image is unusual because of (facial) hair, eye glasses or smile.

### 6.3.3 Simplex Distance

*Idea: Successful reconstructions are created from convex combinations of the original 3D face scans.*

In geometry, a simplex is the corresponding volume of a convex combination (see Schoute [Sch02, p. 10] and Miller [Mil17]). While linear combinations may contain positive as well as negative weighting factors, convex combinations are restricted to positive factors. The Simplex distance which is introduced in this section penalizes the occurrence of negative weighting factors in the linear combination of 3D facial shapes while positive factors do not affect the resulting distance value at all.

In Section 2.1 it is described in detail how a new, plausible facial shape

$$\mathbf{S} = \sum_{i=1}^{m} \mu_{S,i} \mathbf{S}_i \tag{6.3}$$

is created from convex combinations of example faces $\mathbf{S}_i$ (see Equation 2.1) by using weighting factors $\mu_{S,i} \in [0,1]$ with $\sum_{i=1}^{m} \mu_{S,i} = 1$. In this context the shape vectors $\mathbf{S}_i$ serve as basis vectors of the linear object class [Bla00, p. 45] and it is assumed that all facial shapes are in dense point-to-point correspondence with each other. As long as the correspondence is correct, a convex combination of facial shapes should not lead to artifacts or implausible facial shapes.

Thus, in theory, the 3DMM approximates the 3D shape $\mathbf{S}$ of new faces by a convex combination of $m = 200$ captured 3D scans $\mathbf{S}_i$, so that only positive values for $\mu_{S,i}$ are used. But in practice, the 3DMM operates not directly on the original scans. Instead a PCA is used to reduce the dimensionality of the data and the Mahalanobis distance is used for regularization during the fitting procedure which optimizes the shape coefficients $\alpha_i$ during a linear regression. Accordingly, the final shape $\mathbf{S}$ is defined by

$$\mathbf{S} = \bar{\mathbf{s}} + \sum_{i=1}^{m-1} \alpha_i \mathbf{s}_i, \tag{6.4}$$

where $\bar{\mathbf{s}} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{S}_i$ is the mean shape of all scanned facial shapes $\mathbf{S}_i$ and $\mathbf{s}_i$ are the $m-1$ principal components for the shape (see Equation 2.7). Please note that in contrast to Equation 2.8 the facial expression is omitted in Equation 6.4, because like the Mahalanobis (see Section 6.3.2), the Euclidean (see Section 6.3.4) and the Normal distance (see Section 6.3.5) in this chapter the Simplex distance is only used for a plausibility estimation of the shape based on a neutral facial expression.

The PCA is a transformation to an orthogonal basis which is in the span of the originally scanned facial shapes and altering the shape coefficients $\alpha_i$ enables to create arbitrary facial shapes within this vector subspace. But in this PCA basis it is not obvious when the convex hull of the original face scans is exceeded. This is measured by the Simplex distance. It rests on the hypothesis that a linear combination with (many) negative values for $\mu_{S,i}$ results in an inferior overall result than a convex combination, where all weights $\mu_{S,i}$ are positive.

The concept of the Simplex distance is borrowed from the Mahalanobis distance which is used by the 3DMM to estimate face similarities. Based on the shape coefficients, a distance value is computed to determine the similarity or difference between facial shapes. But it is more restrictive than the Mahalanobis distance, because it not only penalizes absolute differences between the coefficients but also the sheer existence of negative weights $\mu_{S,i}$ for the original facial shapes.

To determine the Simplex distance of a given shape reconstruction, it is necessary to map the corresponding PCA coefficients $\alpha_i$ to the weighting factors $\mu_{S,i}$ that generate an equal shape $\mathbf{S}$. This is done by a back transformation of the shape coefficients which have been modified in an analysis-by-synthesis approach during the fitting of the 3DMM to an input image. According to Section 2.1 and Equation 2.7

$$\begin{aligned} \mathbf{S} - \bar{\mathbf{s}} &= \sum_{i=1}^{m-1} \alpha_i \, \mathbf{s}_i && (6.5)\\ &= \mathbf{U}\boldsymbol{\alpha} && (6.6)\\ &= \mathbf{U}\mathbf{W}\mathbf{V}^T \, \boldsymbol{\mu}_S && (6.7)\\ &\approx \sum_{i=1}^{99} \alpha_i \, \mathbf{s}_i && (6.8)\\ &\approx \mathbf{U}'\boldsymbol{\alpha}' && (6.9) \end{aligned}$$

where vector $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_{m-1})^T \in \mathbb{R}^{m-1}$ in Equation 6.6 contains all PCA coefficients $\alpha_i$ and $\boldsymbol{\mu}_S = (\mu_{S,1}, \mu_{S,2}, \ldots, \mu_{S,m})^T \in \mathbb{R}^m$ in Equation 6.7 is a vector of all individual weighting factors for the original 200 facial shapes (see Equation 6.3).

By default the 3DMM introduced by Blanz and Vetter [BV99] uses 99 eigenvectors to approximate the facial shape, accordingly Matrix $\mathbf{U}' = (\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_{99})$ contains only 99 of the available $m - 1$ eigenvectors $\mathbf{s}_i$ in its columns and therefore is a subset of matrix $\mathbf{U} = (\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_{m-1})$ from Equation 2.6. Thus, Equation 6.8 is an approximation of the facial shape using only the 99 most significant principal components. This can also be written as a matrix vector multiplication like in Equation 6.9, so that

$$\begin{aligned} \mathbf{U}'\boldsymbol{\alpha}' &\approx \mathbf{U}\mathbf{W}\mathbf{V}^T\boldsymbol{\mu}_S \quad | \cdot \mathbf{U}^T \\ \boldsymbol{\alpha}' &\approx \mathbf{W}\mathbf{V}^T\boldsymbol{\mu}_S. && (6.10) \end{aligned}$$

As $\mathbf{U} \neq \mathbf{U}'$, vector $\boldsymbol{\alpha}'$ is defined as $\boldsymbol{\alpha}' = (\alpha_1, \alpha_2, \ldots, \alpha_{99}, 0, \ldots, 0)^T \in \mathbb{R}^m$ which means that its size $m$ is equal to the size of $\boldsymbol{\mu}_S$ but all entries that are not defined in $\boldsymbol{\alpha}$ are set to zero. Solving Equation 6.10 for $\boldsymbol{\mu}_S$ results in

$$\boldsymbol{\mu}_S = \mathbf{V}\mathbf{W}^{-1}\boldsymbol{\alpha}' = \mathbf{V} \cdot \boldsymbol{p}, \tag{6.11}$$

where

$$p_i = \frac{1}{\sqrt{m} \cdot \sigma_i} \alpha'_i \tag{6.12}$$

and by assuming that $\alpha'_i \approx \alpha_i$, Equation 6.12 can be replaced by

$$p_i = \frac{1}{\sqrt{m} \cdot \sigma_i} \alpha_i. \tag{6.13}$$

Accordingly, the weights $\mu_{S,i}$ of each example face can be computed by substituting Equation 6.13 into Equation 6.11.

Finally, the Simplex distance $d_{\text{simplex}}$ is computed by summing up the absolute value of all negative weights $\mu_{S,i}$, accordingly

$$d_{\text{simplex}} = \sum_{i=1}^{m} \vartheta(-\mu_{S,i}), \qquad \vartheta(x) = \begin{cases} x : & x \geq 0 \\ 0 : & \text{otherwise} \end{cases}. \tag{6.14}$$

Instead of a summation of all negative values, alternative heuristics could be used to define the Simplex distance. For example, the number of all negative $\mu_{S,i}$ values or the maximal negative value of $\mu_{S,i}$ could be determined, but these alternative ideas are not pursued in the remainder of this chapter.

### 6.3.4 Euclidean Distance

*Idea: The exact vertex positions of successful reconstructions are similar to the ones of the average head.*

Another well-known and often used distance measure to compare 3D shapes is the Euclidean distance $d_{\text{eucl}}$. In the following it is used to compute the distance between the reconstructed shape vector (with neutralized expression) $\mathbf{S}$ and the average shape vector $\bar{\mathbf{s}}$ which leads to

$$d_{\text{eucl}}(\mathbf{S}, \bar{\mathbf{s}}) = \sqrt{\sum_{i=1}^{3n} (\mathbf{S}_i - \bar{\mathbf{s}}_i)^2} = ||\mathbf{S} - \bar{\mathbf{s}}||_2, \tag{6.15}$$

where $n$ is the number of vertices in the 3D face geometry.

Please note that $d_{\text{eucl}}$ is sensitive to rigid transformations of the faces. The 3DMM

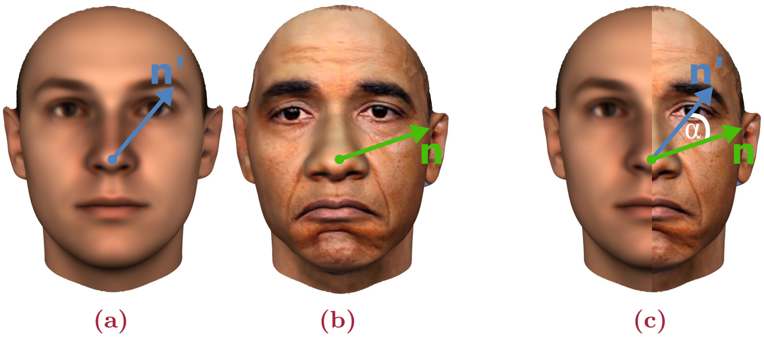


**Figure 6.3:** The Normal distance is determined by computing the angle between the normal of the average (Figure 6.3a) and the reconstructed face (Figure 6.3b) per corresponding vertex pair (see Figure 6.3c). These values are averaged per segment (see Figure 6.4a) or face to obtain a global distance value.

shape vectors are, by construction, aligned in a least-squares sense. During the 3DMM fitting, rigid transformations are applied to these externally, and captured by rendering parameters $\rho_i$ (Section 2). Still, a general drawback of $d_{\text{eucl}}$ remains with respect to simple, global transformations, e.g. anisotropic scaling, which does not affect naturalness or shape similarity, but has significant effect on $d_{\text{eucl}}$. Furthermore, Equation 6.15 tends to overrate outlier vertices in the sum of squared distances.

For the evaluation (Section 6.4), also a modified distance which is the sum of 3D vertex distances (square root on a per-vertex level)

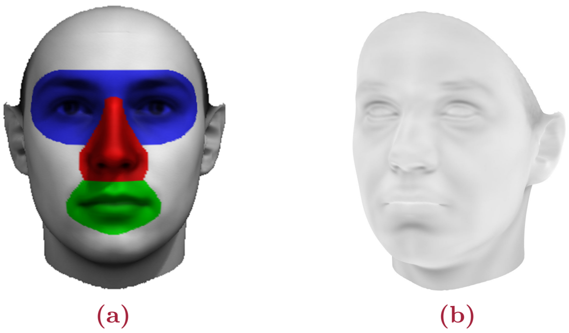
$$d_{\text{eucl2}}(\mathbf{S}, \bar{\mathbf{s}}) = \sum_{i=1}^{n} \sqrt{(x_i - \overline{x}_i)^2 + (y_i - \overline{y}_i)^2 + (z_i - \overline{z}_i)^2} \tag{6.16}$$

was considered, where $\mathbf{S} = (x_1, y_1, z_1, x_2, y_2, z_2, \dots, z_n)^T$ is the shape vector of the current reconstruction and $\bar{\mathbf{s}} = (\overline{x}_1, \overline{y}_1, \overline{z}_1, \overline{x}_2, \overline{y}_2, \overline{z}_2, \dots, \overline{z}_n)^T$ is the average shape vector according to the 3DMM vector space in Section 2.1. But as no improvement was found, when referring to the Euclidean distance $d_{\text{eucl}}$ in Section 6.4 only the definition in Equation 6.15 is used.

### *6.3.5 Normal Distance*

*Idea: The surface curvature of successful reconstructions is similar to the one of the average head.*

It has been observed that local and even global distortions of the surface are a common characteristic of failed 3D face reconstructions. This is true for most or perhaps all 3DMM algorithms (see Chapter 2) and – in a different context – even for 3D shape capture setups such as scanners or stereo and multiview techniques. For shape fitting algorithms, it is unlikely that a failed reconstruction is misaligned and still close to the average, because misalignments tend to have undesired effects on the cost functions of the fitting


(a)

(b)

**Figure 6.4:** Figure 6.4a shows the different face segments. The 3DMM based weight map is shown in Figure 6.4b.

algorithm and therefore lead away from the set of plausible faces. In the examined context, misalignments may be caused by inaccurate initial feature positions. Also, other potential reasons for failed reconstructions, such as lighting effects, occlusions or extreme facial expressions, tend to lead the algorithm far away from the average, and a very sensitive measure for this is the deviation of surface normals from the average.

It should be pointed out that regularization mechanisms, such as Equation 2.11 reduce this effect and keep the solution close to the average. Still, for practical purposes, it has been observed that (1) if the weight of the regularization is too large, it implies suboptimal results on images that would otherwise be reconstructed successfully, so there is a fundamental tradeoff between quality and robustness, and (2) the regularizer in Equation 2.11 is not a reliable measure of plausibility of faces, as will be shown in Section 6.4.

Based on the dense point-to-point correspondence between vertices $i$ of the 3DMM, the new distance measure $d_{\text{normal}}$ analyzes the difference between the surface normals $\mathbf{n}_i$ of the reconstructed face, and the normals $\mathbf{n'}_i$ of the average face:

$$d_{\text{normal}} = \sum_{i=1}^{n} \arccos \frac{\mathbf{n}_i \cdot \mathbf{n'}_i}{\|\mathbf{n}_i\| \, \|\mathbf{n'}_i\|}. \tag{6.17}$$

The idea of this Normal distance is illustrated in Figure 6.3 and is derived from cosine similarity. Note that, unlike $d_{\text{eucl}}$, $d_{\text{normal}}$ is insensitive to scaling and shifting. By segmenting the full face into distinct facial regions (eyes, mouth, nose and surrounding region, see Figure 6.4a), separate distances $d_{\text{normal}}$ can be defined which reflect the plausibility of each region separately. This idea will be used in Section 6.5.

In human faces, the normals in some vertices on the nose, the eyes or the lips vary more than others. Thus, the original 200 3D scans of the 3DMM are analyzed and different weight maps $\omega$ are created (see Figure 6.4b) which account for these local differences by scaling regions with high normal variation either up (considering them most diagnostic) or down (normalization).

| | Obama dataset | | | |
|---|---|---|---|---|
| | automatic landmarks | | manual landmarks | |
| | mean error | max error | mean error | max error |
| $d_{\text{image}}$ | 6.92 | 19 | 6.83 | 18 |
| $d_{\text{eucl}}$ | 5.67 | 16 | 8.08 | 20 |
| $d_{\text{maha}}$ | 2.25 | 8 | 3.42 | 12 |
| $d_{\text{simplex}}$ | 2.08 | 9 | 3.83 | 12 |
| $d_{\text{normal}}$ | 1.33 | 5 | 2.25 | 6 |
| $d_{\text{normalW}}$ | 1.33 | 5 | 2.17 | 6 |

**Table 6.1:** Mean and max difference of ranks for 24 reconstructions with automatically and 24 with manually selected landmarks based on the perceived quality of four naive participants (see Section 6.4.1).

In a first step, the average deviation angle $\overline{\phi}_i$ of the normal $\mathbf{n}_i$ from the average normal $\mathbf{n'}_i$ is computed in each vertex $i$ across all 200 example faces. Based on that, the weight map $\widehat{\omega}$ contains the weights $\widehat{\omega}_i = 1 - \frac{\overline{\phi}_i - \overline{\phi}_{min}}{\overline{\phi}_{max} - \overline{\phi}_{min}}$ for each vertex $i$. Accordingly, the weighted Normal distance $d_{\text{normalW}}$ is defined as

$$d_{\text{normalW}} = \sum_{i=1}^{n} \widehat{\omega}_i \ \arccos \frac{\mathbf{n}_i \cdot \mathbf{n'}_i}{\|\mathbf{n}_i\| \ \|\mathbf{n'}_i\|}, \tag{6.18}$$

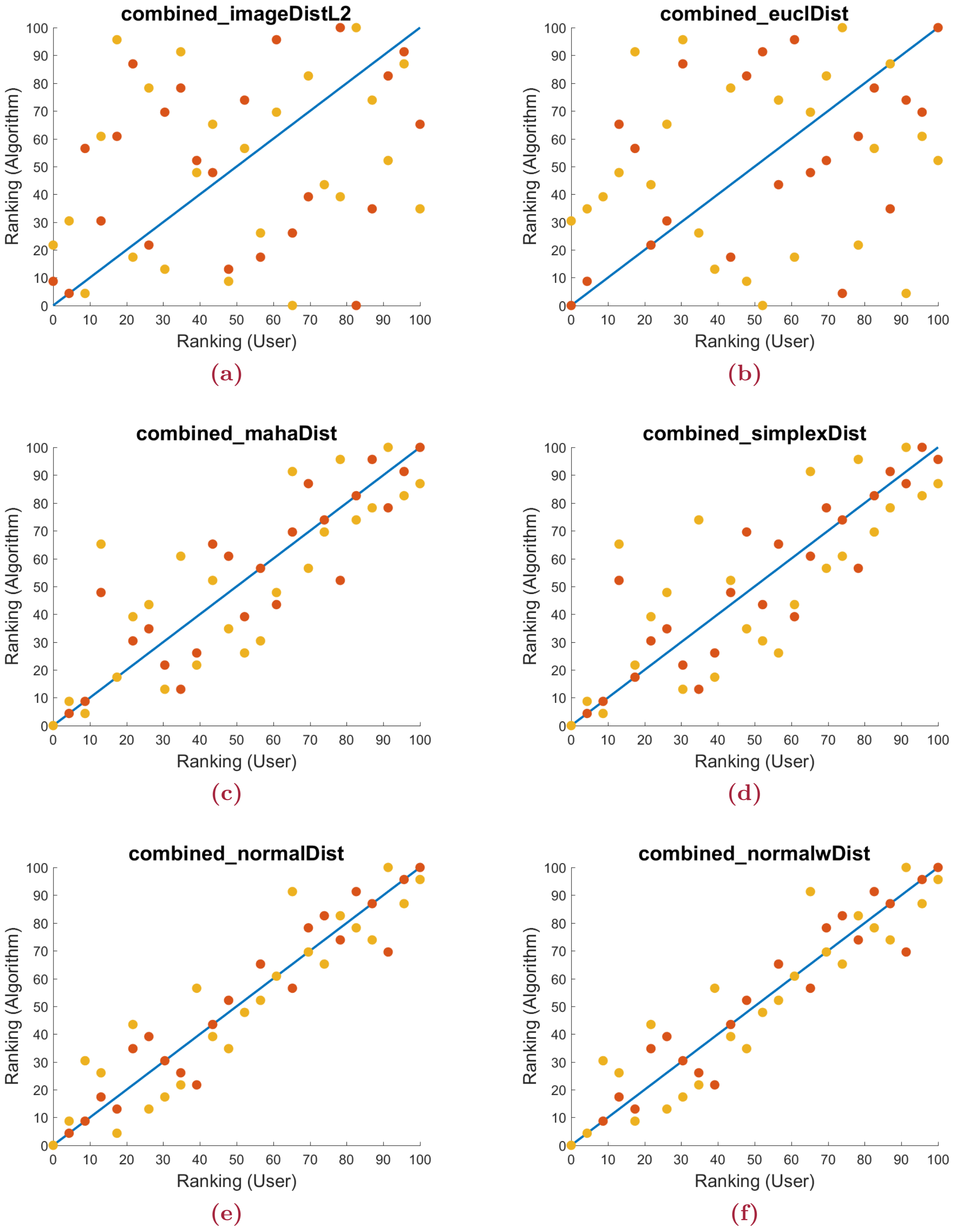
for which the experimental results are obtained that are summarized in the next section.

## 6.4 Evaluating the Distance Measures

The goal of this evaluation is to find out which quality measure is closest to the ratings that human observers would assign to different reconstructions. For humans, quality may mean how natural and plausible the 3D face looks, but also how similar it is to the person in the image. For failed reconstructions, both criteria are usually violated at the same time, so the distance measures from Section 6.3 are good candidates even though most do not measure similarity to the input face.

### *6.4.1 Evaluation 1*

The first ranking was performed on 24 3D reconstructions from pictures of Barack Obama based on automatically detected landmarks. The automatic detection of landmarks is based on the approach of Zhu and Ramanan [ZR12]. An additional set of 24 reconstructions was created by using manually selected landmarks on the same input images. Again the algorithmic distance measures introduced in Section 6.3 were used to perform a ranking. All reconstructions were created from a single image as described in Chapter 2.



(a) (b) (c) (d) (e) (f)

**Figure 6.5:** Visualization of the correlation between the average user ranking and each distant measure (100 $\widehat{=}$ very good, 0 $\widehat{=}$ very bad) for reconstructions based on automatic *(red)* and manual *(yellow)* landmark selection.

Four naive participants were asked to create a ranking in each of the two sets of 24 reconstructions, based on the perceived quality of the reconstruction. The individual user rankings were combined to define an overall ranking list, which was compared to the ranking of each distance measure. As can be seen in Table 6.1, the mean and max errors (difference of ranks assigned to each reconstruction) of Mahalanobis, Simplex and

Normal distance are much less than the ones based on Euclidean and image distance. Furthermore, based on the numbers for $d_{\mathrm{normalW}}$ (see Equation 6.18), it can be noted that the influence of the weight map is not very strong compared to the ranking based on $d_{\mathrm{normal}}$ (see Equation 6.17).

In Figure 6.5 the correlation of each distance measure is visualized: The horizontal axis describes the average user ranking, while each distance measure is mapped to the vertical axis. If a distance measure correlates perfectly with the user ranking, the dots of the scatter diagram are aligned along the diagonal. As can be seen in Figure 6.5a, for the image distance the dots are widely scattered. The same can be observed for the Euclidean distance in Figure 6.5b. Consequently, both measures are not useful to distinguish plausible from implausible reconstructions in a way that correlates to the opinion of users. For the Mahalanobis (see Figure 6.5c), the Simplex (see Figure 6.5d) and the Normal distance (see Figure 6.5e), the correlation between the user rating and the rating based on the algorithmic distance measures is clearly visible. Especially the Normal distance predicts much of the quality judgments of the participants. The correlations for all $2 \cdot 24$ reconstructions (automatic and manual) are for $d_{\mathrm{image}}$: 0.27, for $d_{\mathrm{euclidean}}$: 0.27, for $d_{\mathrm{maha}}$: 0.85, for $d_{\mathrm{simplex}}$: 0.84, for $d_{\mathrm{normal}}$: 0.94 and for $d_{\mathrm{normalW}}$: 0.94.

### 6.4.2 Evaluation 2

In a second evaluation, 3D reconstructions based on images of Obama (24 images), Lawrence (32 images), Annan (32 images), Watson (46 images) and Carell (28 images) were rated. Again two distinct sets were created, but this time only automatically selected landmarks were utilized. The first set was created by fitting to a single image, while for the second set a simultaneous fitting to two images was performed by applying the multifit approach of Blanz and Vetter [BV99]. A fixed reference image was selected and was then combined with each other image of the collection for the person. Please note that the facial landmarks differ from to the ones in Section 6.4.1. Thus, although the same input images are used for the Obama dataset, the reconstructions are different.

For each dataset, the distance measures were used to create a ranking list. Then seven naive participants were asked to rate each 3D reconstruction. Possible ratings were ’very good’, ’good’, ’acceptable’ or ’failed’. The individual ratings were averaged and then used to create a ranking. Many reconstructions obtained the same average ratings and therefore many positions in the ranking are shared. This implies higher discrepancies between the rank list derived from humans, and the rank list from distance measures than in Evaluation 1 in Section 6.4.1, where the participants were asked to create a unique ranking directly. Once again, the Normal distance matches the user rating best, as can be seen in Table 6.2 and 6.3.

As in Evaluation 1 the results of the Simplex distance, which has been introduced in

| | Obama | Lawrence | Annan | Watson | Carell |
|---|---|---|---|---|---|
| $d_{\text{image}}$ | 5.29 (13) | 7.56 (20) | 9.59 (30) | 13.15 (29) | 8.04 (18) |
| $d_{\text{eucl}}$ | 8.38 (20) | 8.38 (22) | 10.84 (23) | 13.54 (35) | 8.82 (20) |
| $d_{\text{maha}}$ | 5.79 (13) | 7.00 (18) | 5.22 (16) | 11.94 (27) | 3.46 (9) |
| $d_{\text{simplex}}$ | 5.79 (12) | 7.00 (20) | 5.41 (15) | 11.98 (27) | 3.54 (9) |
| $d_{\text{normal}}$ | 5.21 (13) | 5.44 (14) | 4.97 (12) | 11.50 (27) | 2.61 (8) |
| $d_{\text{normalW}}$ | 5.21 (13) | 5.31 (14) | 4.97 (12) | 11.41 (27) | 2.46 (8) |

**Table 6.2:** Mean and max *(in brackets)* difference of ranks for reconstructions from a single image based on the perceived quality of seven naive participants (see Section 6.4.2).

| | Obama | Lawrence | Annan | Watson | Carell |
|---|---|---|---|---|---|
| $d_{\text{image}}$ | 6.75 (14) | 8.34 (23) | 8.75 (24) | 10.80 (33) | 6.32 (19) |
| $d_{\text{eucl}}$ | 5.58 (16) | 8.47 (24) | 6.56 (19) | 10.02 (34) | 9.32 (24) |
| $d_{\text{maha}}$ | 4.33 (11) | 5.47 (14) | 5.25 (19) | 10.07 (28) | 6.82 (19) |
| $d_{\text{simplex}}$ | 4.25 (11) | 5.41 (15) | 5.38 (20) | 9.54 (30) | 6.61 (19) |
| $d_{\text{normal}}$ | 4.08 (10) | 4.41 (14) | 4.31 (15) | 7.85 (25) | 4.96 (17) |
| $d_{\text{normalW}}$ | 4.08 (10) | 4.41 (14) | 4.31 (15) | 7.80 (25) | 4.96 (17) |

**Table 6.3:** Mean and max *(in brackets)* difference of ranks for reconstructions from multiple images based on the perceived quality of seven naive participants (see Section 6.4.2).

Section 6.3.3 to serve as one possible measure for plausibility checks of 3D reconstructions of faces, are close to the ones based on the Mahalanobis distance. This might disprove the initial hypothesis that a linear combination with negative weighting factors $\mu_{S,i}$ necessarily results in an inferior overall result compared to a convex combination where all $\mu_{S,i}$ are positive. As the results of the Simplex distance and the Mahalanobis distance are not completely equal, there might be some influence but it seems to be less significant than initially suspected.

## 6.5 Weighted Linear Combinations per Segment

The automated 3D reconstruction which is proposed in this chapter compensates the reduced precision and reliability of automatically detected feature positions by using more than a single image of the face. Note that, unlike stereo and multiview algorithms, the method supports non-rigid deformations due to facial expressions, and large differences in the (unknown) imaging conditions.

The proposed strategy is to apply single image 3DMM fitting (see Chapter 2) on each of the input images of the person separately, based on landmarks detected by the algorithm by Zhu and Ramanan [ZR12]. Then selecting the $m$ best results (see Figures 6.1 and 6.6) on each segment (Figure 6.4a) using $d_{\text{normal}}$, compute weighted linear combinations of

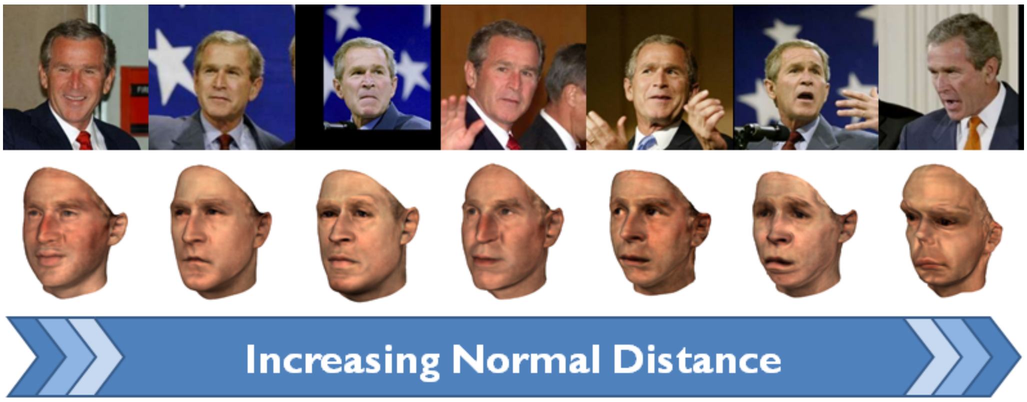


**Figure 6.6:** Plausibility rating of the single image based reconstructions using Normal distance with subsequent ordering.

these and merge them into a single 3D face. As $d_{\text{normal}}$ performed best in the evaluations in Sections 6.4.1 and 6.4.2 it seems to be the perfect metric for the following processing steps and because the performance of $d_{\text{normal}}$ and $d_{\text{normalW}}$ is equally good, the additional overhead of the weight map $\widehat{\omega}$ to estimate $d_{\text{normalW}}$ seems to be unnecessary and is not be taken into account in the remainder of this section.

The shape for each segment is determined by a weighted linear combination of corresponding segments based on the ranking list order. The weight $w_j$ decreases with the rank $j$. Thus, the combined shape for each individual segment

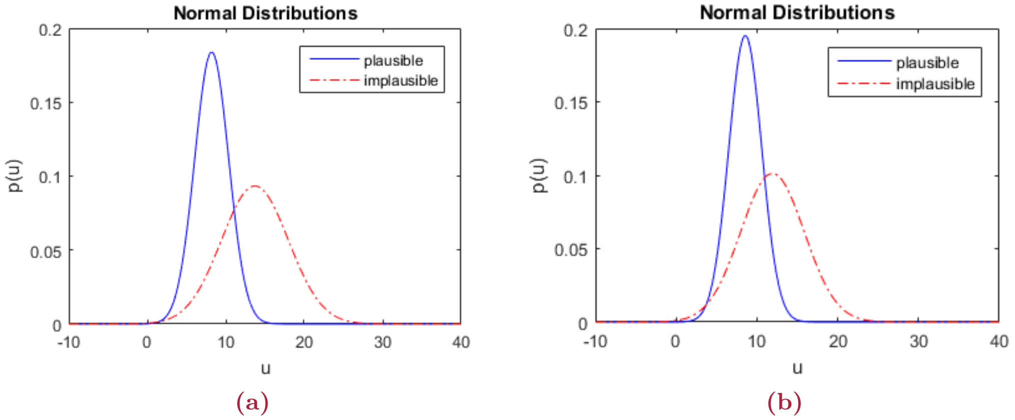
$$\mathbf{S}_{seg} = \sum_{j=0}^{m-1} w_j \, \mathbf{S}_{seg,j} \tag{6.19}$$

is determined by $m$ individual reconstructions of corresponding segments $\mathbf{S}_{seg,j}$ which are weighted by

$$w_j = \frac{1-(j \cdot \frac{1}{m})}{\sum_{k=0}^{m-1} 1-(k \cdot \frac{1}{m})} \tag{6.20}$$

with $\sum_{j=0}^{m-1} w_j = 1$. The algorithm is summarized in Figure 6.1. Note that for illustration purposes, Figures 6.1 and 6.6 refer to the shape of the entire face, and not for separate segments as in the algorithm.

An important element of this algorithm is to define a threshold quality value that determines which reconstructions are considered during the pooling based on the weighted sum. The threshold that separates plausible from implausible reconstructions is estimated based on the data from Section 6.4. In Evaluation 1, participants were also asked which faces are still plausible and which are not. In Evaluation 2, the threshold is assumed to be between ratings which are labeled “acceptable” and “failed”. For both datasets, the



(a)

(b)

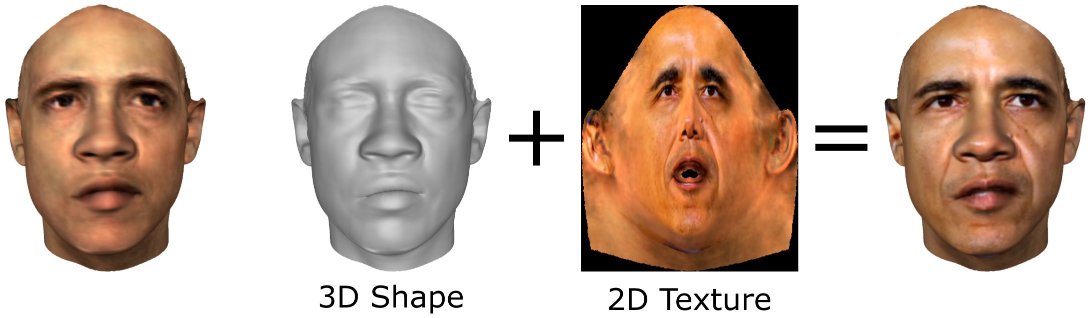
**Figure 6.7:** Gaussian distribution of $d_{normal}$ for plausible *(blue)* and implausible *(red)* 3D reconstructions. For Evaluation 1 (Figure 6.7a) the intersection is in $u = 11.08$ and for Evaluation 2 (Figure 6.7b) it is in $u = 10.95$.

Gaussian distributions $p_0(u)$ and $p_1(u)$ for plausible and implausible reconstructions are estimated using the arithmetic mean and the estimated standard deviations of $d_{\text{normal}}$ in either set.

In a maximum likelihood approach, the threshold equals the intersection point of the Gaussian distributions $p_0(u)$ and $p_1(u)$. Based on the examined data, this threshold equals $u = 11$ as is shown in Figure 6.7 and can be computed by solving $p_0(u) = p_1(u)$.

Then all segments with Normal distances larger than this threshold are discarded, as is illustrated in Figure 6.1. After the shape for each segment has been reconstructed using a weighted linear combination based on the ranking order for the remaining segments, all independent segments are combined to build the shape of the complete face using the method described in Section 6.7.2. Although a high percentage of the faces in images have a non-neutral facial expression, the separation of PCAs and basis vectors for shape and expression enables to neutralize each expression after the 3DMM reconstruction step by setting $\gamma_i = 0$ in Equation 2.8. Furthermore, the 3DMM establishes 3D correspondence between all reconstructions (see Chapter 2). Combined this allows for handling arbitrary facial expressions in each input image as well as handling varying facial expressions in the input images connected to the same person.

Finally, one of the input images, for example the one with the minimal distance $d_{\text{normal}}$, can be used for texture extraction as described in Section 2.3. Thus, the final texture is not just a linear combination of all input images, but captured from a single input image with inverse projection and lighting. In most cases this results in a higher texture resolution which captures more details and is less smoothed (see Figure 6.8). More details on the implementation of an automatic texture extraction are given in Section 6.6.



(a)

(b)

**Figure 6.8:** Comparison between the facial texture which results from the texture coefficients in Figure 6.8a and the extracted texture in Figure 6.8b from the most plausible face reconstruction (see Figure 6.9).

## 6.6 Distance Measure Driven Texture Extraction

The 2D input images can not only be used to reconstruct the 3D shape of a person but also to collect information about skin or eye color, facial hair and even fine details like wrinkles. The amount and degree of details that can be extracted from these 2D images depends on the image's quality and resolution. Accordingly, the final outcome is quite individual for each situation.

To enable an automated reconstruction of the facial shape and texture without any kind of user input, the texture extraction is based on the Normal distance like the segment-based shape reconstruction in Section 6.5. But instead of determining the Normal distances for each segment individually, in this case it is calculated for the complete head shape using the 3D reconstruction of each available input image. This allows to detect the best single image shape reconstruction. Following this, the input image $\mathbf{I}_{\text{input}}$ corresponding to the smallest Normal distance is used to extract the texture $\mathbf{T}_{\text{extr}}$ which is mapped on the final shape reconstruction $\mathbf{S}$ to improve the visual appearance as is shown in Figures 6.1 and 6.8.

For the proposed texture extraction (see Section 2.3) a facial image without an extreme facial expression, illumination, pose and without or limited occlusions delivers the best final results. Accordingly, to find a good candidate for texture extraction the quality metric needs to take care of these factors. Although the Normal distance does not measure the texture quality directly, it implicitly regards the above listed criteria for a useful texture candidate. For example, if the face in the input image is exposed to extreme facial expression, illumination, pose or occlusions the automatic landmark localization algorithm is not as accurate as it would be without these effects. Consequently, the single image based reconstruction will tend to deliver a inferior result for the 3D shape which means that the ranking based on the Normal distance also will be worse.

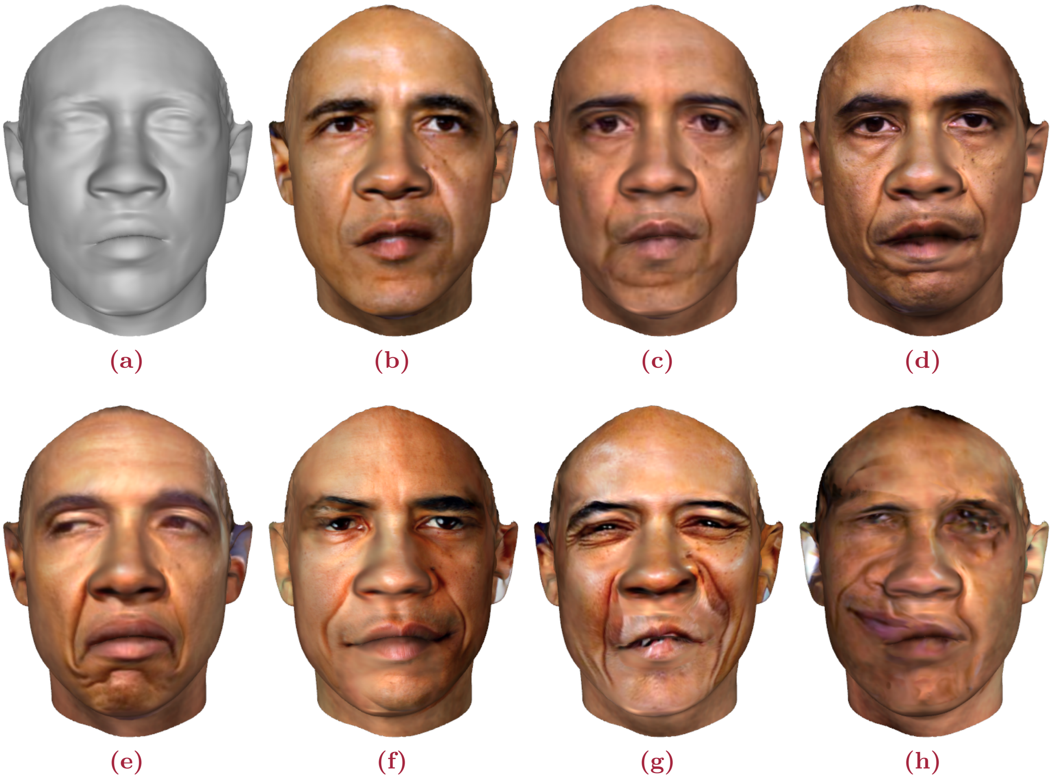


**Figure 6.9:** Figure 6.9a shows the plain 3D shape which is used for the texture mapping of all individual textures. The textures in Figures 6.9b–6.9d originate from the three most plausible reconstructions, while 6.9e shows the texture of the lowest ranked, but still plausible, reconstruction. In Figures 6.9f –6.9h texture mapping is done using the textures of the most implausible rated reconstructions.

In Figure 6.9 a set of extracted textures is mapped on the most plausible facial shape of Obama (see Section 6.7.3 and Figure 6.14). Therefore, the 3DMM has been fitted to 24 individual input images and each reconstruction has been rated by using the Normal distance. As all reconstructions – including their textures – are in dense correspondence, interchanging the textures requires no additional effort. The first row of Figure 6.9 starts with the plain 3D shape followed by the rendering of this shape with the extracted textures from the three most plausible reconstructions. At the beginning of the second row the texture of the lowest ranked, but still plausible, reconstruction is shown and mapped on the shape shown in Figure 6.9a, followed by the textures of the three lowest ranked and accordingly most implausible reconstructions of all input images of Obama.

Especially in Figures 6.9g and 6.9h it is clearly visible that the facial features of the shape no longer match the ones in the textures.

## 6.7 Results

This section presents the automatic 3D facial reconstruction results using the Normal distance as a quality measure. Therefore, in Section 6.7.1 3D reconstructions of constrained photo collections are generated to compare the proposed approach with the multifitting method of Blanz and Vetter [BV99]. Furthermore, the reconstruction results for unconstrained photo collections are shown by using images from the Labeled Faces in the Wild [HRBLM] database as input for the proposed method. In Section 6.7.2 the segment-based quality measure and merging is presented for one specific example and the smoothing effects in some of the reconstructions are discussed in Section 6.7.3.

### *6.7.1 Automatic 3D Reconstructions*

The proposed approach is compared with an existing method for simultaneous 3D reconstruction from multiple images [BV99]. For this purpose sets of 8 to 15 images are used that show the same face from different angles. A subset of these images is shown in the second column of Figures 6.10 and 6.11.

The first column in each of these figures shows the results of the existing approach, whereas the results of the proposed approach are presented in Column 3 with a uniform color and in Column 4 with the combined texture colors. Therefore, the textures of each individual segment have been linearly combined in exactly the same way as has been described for the shape in Section 6.5.

In respect of shape estimation the new approach outperforms the existing method if a fully automated approach is demanded and if the landmark locations may not fit the input image perfectly. As [BB10] and [BKK$^+$08] share the weakness of [BV99] that in a multifit scenario even few outliers reduce the quality of the reconstruction result, at this point a side-by-side comparison is skipped.

In addition, it should be mentioned that the proposed method is not directly rivaling with the concepts of [BB10]: replacing the single image fitting based on [BV99] with the one in [BB10] which automatically corrects misplaced landmark locations to some degree is also supported, but for 'faces in the wild' these corrections are not sufficient. Accordingly, no noteworthy quality improvements are noticeable. It only increased computation time.

While the input images in Figures 6.10 and 6.11 are taken under controlled lighting conditions and lack facial expressions, in an additional experiment the new approach is tested with images from the Labeled Faces in the Wild [HRBLM] database. Here pose, expression and lighting differ in each image and the overall image resolution is only 250*250px. The corresponding results are shown in Figure 6.12.

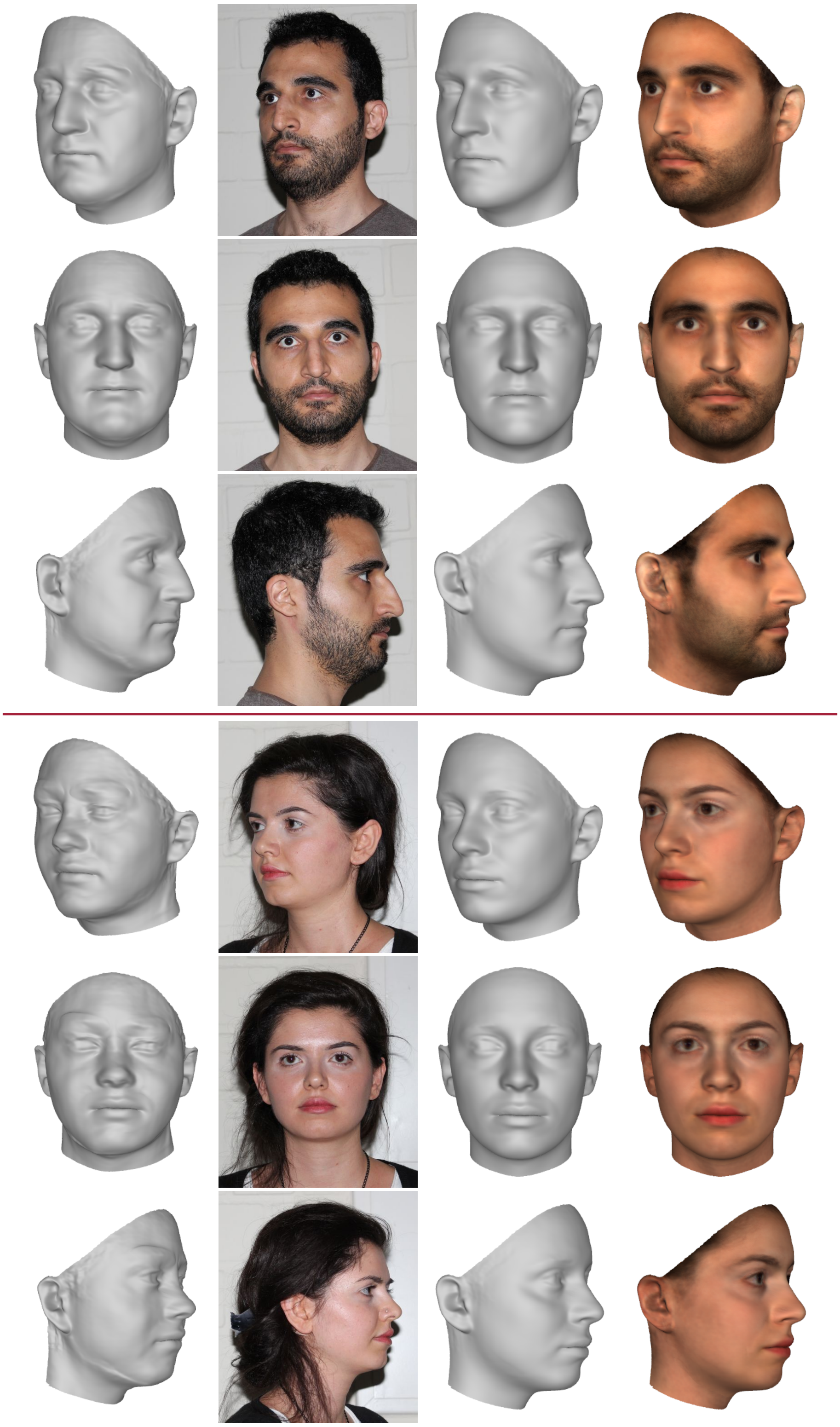


**Figure 6.10:** Two face reconstructions from multiple images using the existing 3DMM approach (Col. 1), the proposed method with Normal distance (Col. 3) plus combined color (Col. 4). A subset of the input images is shown in Col. 2. For the reconstructions 8 (top) and 11 (bottom) images are used, respectively.

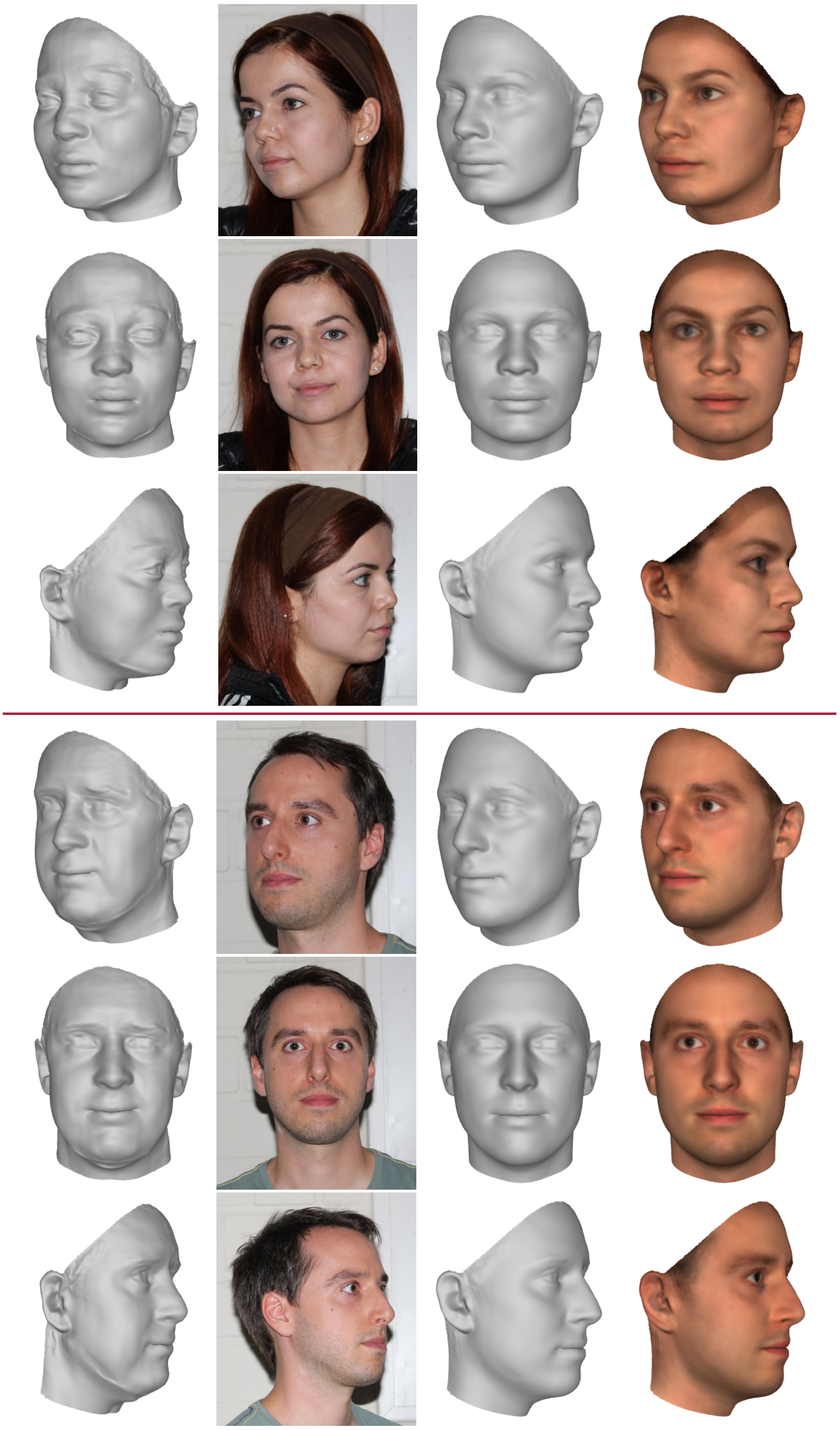


**Figure 6.11:** Two face reconstructions from multiple images using the existing 3DMM approach (Col. 1), the proposed method with Normal distance (Col. 3) plus combined color (Col. 4). A subset of the input images is shown in Col. 2. For the reconstructions 10 (top) and 15 (bottom) images are used, respectively.

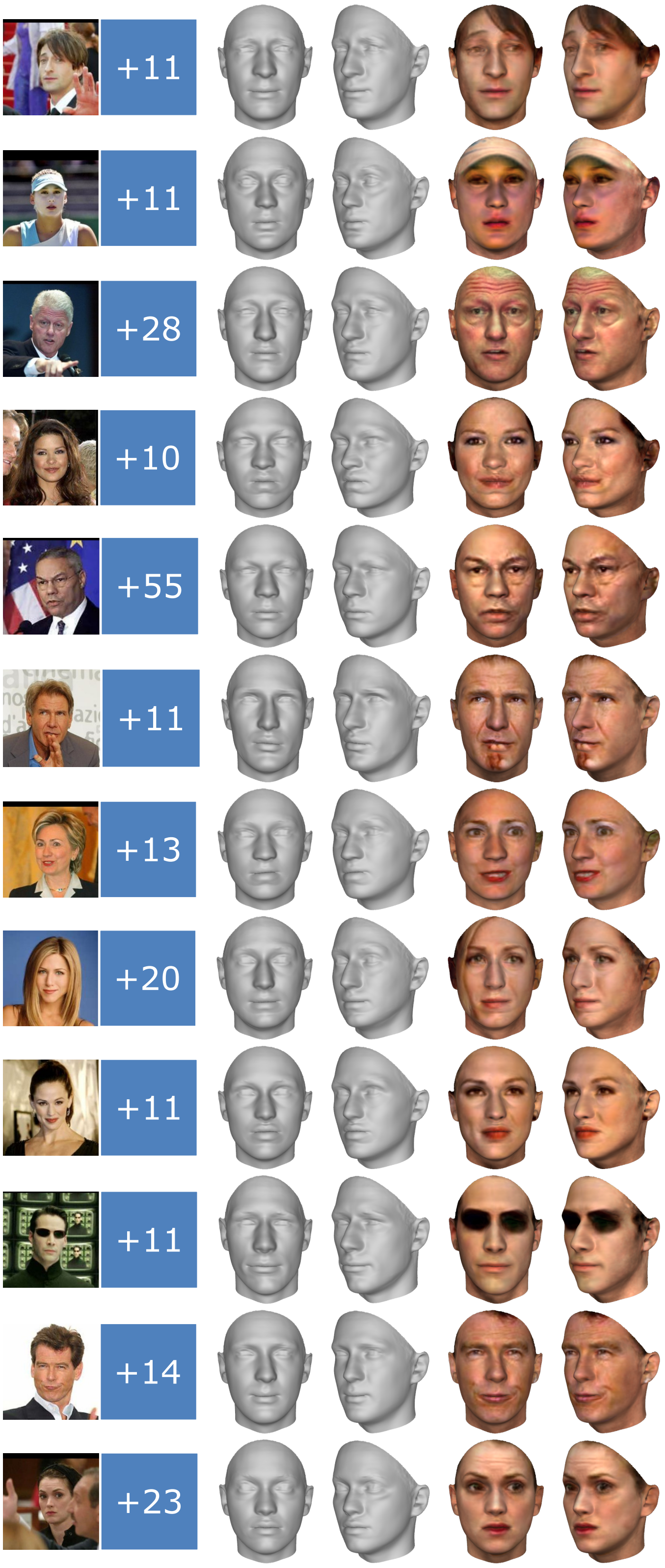


**Figure 6.12:** 3D face reconstructions for image sets from the LFW [HRBLM] database. From left to right the columns contain an example image from the image set plus the number of additional images, two views of the reconstructed shape and two views with extracted textures.

| LFW dataset: Harrison_Ford | | | | | |
|---|---|---|---|---|---|
| rank | id_eyes | id_mouth | id_nose | id_remainder | id_all |
| 1 | 2 | 2 | 2 | 2 | 2 |
| 2 | 10 | 8 | 8 | 8 | 8 |
| 3 | 12 | 12 | 12 | 10 | 10 |
| 4 | 8 | 10 | 11 | 6 | 6 |
| 5 | 11 | 11 | 10 | 12 | 12 |
| 6 | 7 | 6 | 6 | 9 | 7 |
| 7 | 6 | 7 | 1 | 7 | 9 |
| 8 | 9 | 9 | 4 | 1 | 1 |
| 9 | 4 | 1 | 9 | 4 | 11 |
| 10 | 1 | 4 | 7 | 11 | 4 |
| 11 | 5 | 5 | 5 | 5 | 5 |
| 12 | 3 | 3 | 3 | 3 | 3 |

**Table 6.4:** Rankings of image ID's based on segments (Columns 2-5) or the entire face (Column 6) for the result in Figure 6.12, 6th row. Images with numbers in *black* color are used, those with *red* numbers are discarded. Figure 6.13 visualizes a subset of these segments.

From left to right the columns of Figure 6.12 contain one image per dataset and person plus the number of used images and two views of the reconstructed shape. Starting with a uniform coloring and then with the extracted texture from the input image on the left.

To retain a fully automated process, the input image which is used for texture extraction corresponds to the most plausible single image reconstruction (see Figure 6.6) as is described in Section 6.6. As has also been discussed in Section 6.6, any other input image can be used for texture extraction as well, because the 3DMM enables a 2D to 3D correspondence between the image and each reconstruction as well as a 3D to 3D correspondence between all reconstructions (see Chapter 2).

### *6.7.2 Details on Merging Segments*

The following examples provide additional details on the ranking and the subsequent merging of plausible facial segments. To create the shape of the complete face, all independent facial segments are combined by using the method which has been published by Blanz and Vetter [BV99]. As has already been described, the proposed method distinguishes between four different facial regions (see Figure 6.4a): eyes, mouth, nose and the remainder. In Table 6.4 the individual rankings for one of the image sets from the LFW dataset [HRBLM] is shown.

Although some reconstructions fail in all segments (image ID's 1, 3, 4, 5), often some parts of the face are reconstructed better than others. For that reason, a segment-based merging is used. As is shown in Table 6.4, based on the Normal distance a distinct ranking

| Photo collection: Figure 6.11 (top) | | | | | |
|---|---|---|---|---|---|
| rank | id_eyes | id_mouth | id_nose | id_remainder | id_all |
| 1 | 346 | 346 | 346 | 354 | 354 |
| 2 | 354 | 354 | 354 | 356 | 356 |
| 3 | 355 | 355 | 355 | 351 | 346 |
| 4 | 344 | 344 | 344 | 346 | 355 |
| 5 | 351 | 356 | 356 | 355 | 351 |
| 6 | 356 | 351 | 352 | 344 | 344 |
| 7 | 345 | 352 | 345 | 352 | 352 |
| 8 | 352 | 345 | 351 | 345 | 345 |
| 9 | 353 | 353 | 343 | 353 | 353 |
| 10 | 343 | 343 | 353 | 343 | 343 |

**Table 6.5:** Rankings of image ID's based on segments (Columns 2-5) or the entire face (Column 6) for the person at the top of Figure 6.11. Images with numbers in *black* color have been used, those with *red* numbers have been discarded.

list is used for each segment to create the final result in Figure 6.12, 6th row. Besides differences in sorting, some reconstructions are used for one segment but are discarded for another (image ID's 6, 7, 9, 11). This is also valid for the other reconstructions in Figure 6.12. For the sake of completeness, Column 6 in Table 6.4 shows the ranking if the Normal distance had not been computed separately for each segment but for the whole face.

Another example of the merged and discarded segments for unconstrained photo collection is shown in Table A.4 where the Normal distance has been computed for the input images of Obama shown in the first row of Figure 5.8. The resulting shape reconstruction is shown in Figure 6.14 along with additional reconstruction results that take a varying number of plausible segments into account.

For less challenging photo collections like in Figures 6.10 and 6.11, considerably fewer reconstructions are discarded, but differences in sorting per segment remain. The details of the exact order of the segments are shown in Tables 6.5, A.5, A.6 and A.7. As these images lack of facial expressions and are captured under stable lighting conditions and without moving the camera, implausibilities mostly arise from the changes in pose.

It can also be noticed, that the ranking of the *remainder* segment often matches with the ranking result which would result if applying the Normal distance to the complete face and not to each region separately. Actually, this is not very surprising, because this region contains the largest amount of available head vertices. But this fact also emphasizes that additional expressiveness is gained from the other, smaller segments. These details would have been lost without a segment-based approach.

In Figure 6.13 the three most and least plausible facial segments according to Table 6.4

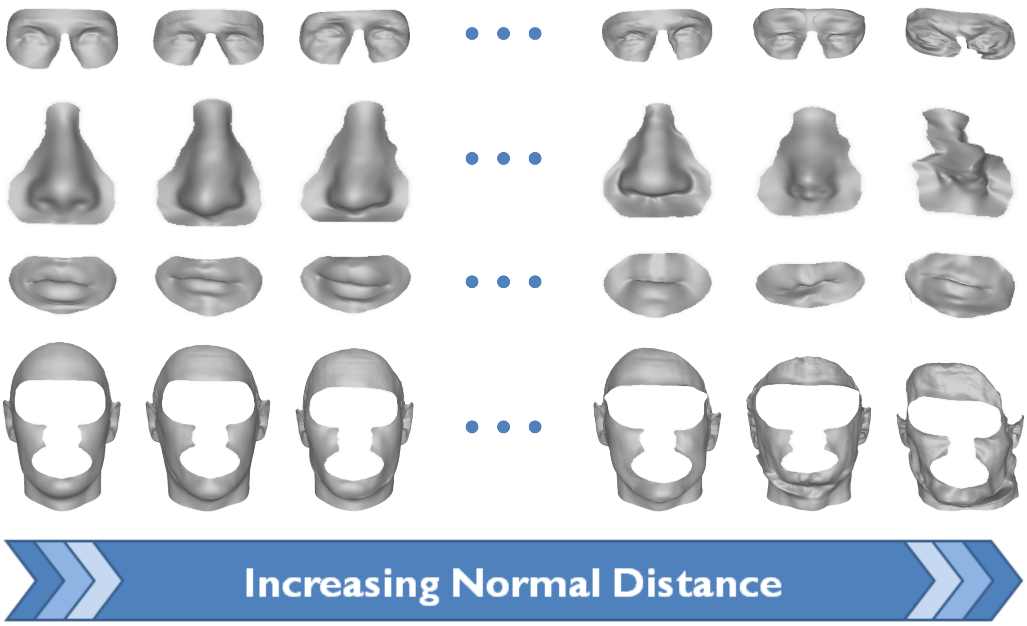


**Figure 6.13:** This figure shows renderings of the three most and least plausible facial segments using Normal distance. The reconstructions are based on the photo collection in Row 6 of Figure 6.12.

are shown to provide additional qualitative results for the proposed segment-based approach which uses the Normal distance as a quality measure. Merging all plausible segments using the weighted linear combination in Section 6.5 and the distance measure driven texture extraction (see Section 6.6) results in the facial reconstruction shown in Row 6 of Figure 6.12.

### *6.7.3 Smoothing Effects*

The smoothing effect in the resulting reconstructions especially in Figure 6.12 is a tradeoff for robustness under real world conditions. Two independent sources have been identified that lead to this kind of smoothing effect: First, the used 3DMM from [BV99] does not capture all shape details which leads to a smoother reconstruction based on each single input image. Especially, in the case of the low-resolution images from the LFW dataset [HRBLM]. Another reason is an averaging effect when merging plausible candidates (per segment) that vary in specific shape estimations, because being able to identify plausible candidates of complete faces or facial regions does not necessarily mean that they look alike. While Section 5.5.2 took advantage of the smoothing effects because they reduced individual artifacts without requiring additional clues about the processed data, the same effects lead to disadvantages here, because expressiveness gets smoothed out by combining too many plausible reconstructions with varying shapes.

But while [BV99] and [BB10] failed on the LFW dataset with automatically detected landmarks, the proposed approach at least delivers a robust basis for further improvements

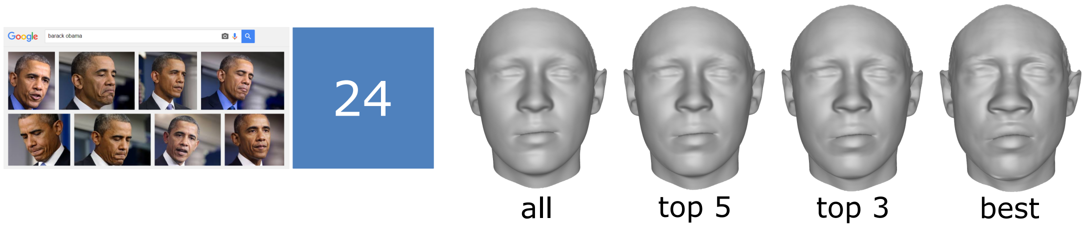


**Figure 6.14:** Varying expressiveness in regard of the number of plausible candidates per segment for 24 images of Obama (see Figure 5.7 and 5.8). From left to right the (weighted) combinations using (1) all plausible, (2) the top five, (3) the top three and (4) only the best candidates per segment are shown.

in reconstructing facial details like using shading cues [PS12] or applying a photometric stereo approach [RTL16] on the more reliable views.

Another way to counteract smoothing effects is to reduce the number of plausible reconstructions that are combined with each other. For example, by choosing only the $n$ most plausible ones per facial segment and not all of them like in Section 6.5. In Figure 6.14 an example of the relationship between the number of plausible reconstructions and the resulting expressiveness is shown. Therefore, a facial reconstruction which has been created exactly according to the approach proposed in Section 6.5 is compared with three reconstructions that use only a small subset of the available, plausible reconstructions for each facial segment: More precisely the top five, top three and finally only the most plausible segment is used. Please note that in the shown case decreasing the number of candidates indeed leads to a higher expressiveness but this also increases the chance for artifacts in the final result. While the reconstruction of the nose gets improved from left to right, the shape of the mouth does not get more plausible in regard to the groundtruth in the input images. In the end it is a tradeoff between expressiveness and robustness. Thus, in each situation the preferred strategy strongly correlates with the underlying objective.

## 6.8 Conclusions

In this chapter an algorithm has been proposed that reconstructs a 3D face from a set of arbitrary images of a person. The core idea is to perform separate reconstructions on each image and combine the best of all reconstructions into the final shape. An important element of this work is to evaluate different quality measures for 3D reconstructions. Combined with a feature point detector, an automated algorithm for 3D reconstruction is obtained that accounts for errors in the feature coordinates. The proposed method is modular, scalable and flexible, and it overcomes some of the problems that have restricted 3DMMs so far.

On a more fundamental level, it is the combination of results (multiple images, multiple segments) which makes this algorithm robust, and it is an alternative strategy to combining all input data into a single optimization problem.

It is a non-trivial result that multiple suboptimal 3D faces can be combined into a single, much more appealing one, and that this result is not just the average face. Another non-trivial result is that the reconstruction quality can be assessed without knowing the ground truth shape.

Besides the presented solution for an automatic 3D reconstruction of faces from multiple 2D input images by using the Normal distance, there are two aspects which might be interesting topics for further research: (1) An additional and more substantial analysis of the performance of the Normal distance to find out why it was superior when evaluating the different distance measures in Section 6.4. In this context one hypothesis is that the Normal distance dealt best with the not perfect correspondence computation between the 3D example faces of the 3DMM (see Blanz [Bla00, p. 96]). (2) A closer investigation of the Simplex distance and a more thorough comparison to the Mahalanobis distance. In this context it might also be interesting to replace the Mahalanobis distance in the optimization steps of the 3DMM fitting by the Simplex distance and compare the results of both approaches.

# 7

# Image-based Relighting of Faces

In this chapter an approach for realistic lighting design for faces is proposed. Starting with an arbitrary 2D image of a face, the 3DMM is used to create a 3D reconstruction of that face (see Section 2.2), and by painting strokes on the original image the user is able to add or alter the lighting effects in the input image. Besides the 3DMM framework, the proposed method makes use of a state of the art inverse lighting algorithm for faces by Shahlaei and Blanz [SB15] that enables to render illumination effects such as cast shadows, specular highlights, multidirectional and colored lighting. Furthermore, the methods for automatic landmark localization, 3DMM initialization and 3D reconstruction introduced in Chapters 5 and 6 are used to automate the facial reconstruction procedure and enable to handle arbitrary images of faces. Thus, the reconstruction as well as the lighting algorithm are invariant to pose, facial expressions and image saturation. As the lighting design approach requires only a single input image of an unknown face to produce realistic results, the Normal distance which has been introduced in Chapter 6 as a quality measure to determine plausible candidates in photo collections is adapted to increase the robustness of 3D reconstructions based on a single image. This new lighting design approach has been published in [SPB16].

## 7.1 Introduction

Professional photo shootings are often well planned and prepared, this includes not only the object or model that is to be captured in a well defined scene but also the light position, intensity and color, as lighting is a key component – especially in the field of portrait photography. Besides special and expensive equipment a lot of expertise is required; not only during the photo shooting, but also during the post-processing steps. Retouching unwanted imperfections of texture or shape is as common as varying the lighting situation at some local spots or for larger regions. As light and shadows always follow the underlying shape, this kind of post-processing again requires extensive expertise. The necessary prerequisites are not met in perfection by semi-professional photographers and definitely not by the general public who just want to take a nice photo in a completely unconstrained and unplanned situation. Nevertheless, many people are interested in gaining the desired effect and atmosphere in the final photo.

The proposed method primarily aims to solve arbitrary cases of portrait relighting. Although it can also be used as an additional tool for professional portraits, in the current state it is not designed to replace all post-processing steps in the field of lighting design completely. The major advantage of this approach is its simplicity: the 3D reconstruction can be done automatically and therefore completely transparent for the user. The only remaining, manual input is to draw coarse sketches of the desired lighting on the face in the input image. Then the algorithm creates a realistically relighted face and renders it on top of the original one. This procedure is not limited to photos but can also be applied to paintings of faces as is shown in Section 7.6.

For this post-hoc relighting procedure, several challenges need to be faced: First, the 3D information of the face needs to be extracted from a single 2D input image including the surface normals to be able to determine the reflected light as well as cast shadows. This requires an alignment of the 3D model to the face in the image which is done by using the facial landmark localization approach of Zhu and Ramanan [ZR12] to detect the face and its landmarks, and to initialize the fitting algorithm of the 3DMM (see Section 5.3). To gain more robust results the Normal distance driven approach for photo collections in Chapter 6 is adapted to improve the 3D shape reconstruction from a single input image. Secondly, the albedo or the intrinsic texture of the face needs to be extracted by using the inverse lighting approach by Shahlaei and Blanz [SB15] which delivers an improved facial texture compared to the original 3DMM by Blanz and Vetter [BV99]. It can handle harsh illumination conditions and supports soft cast shadows, colorful lighting, multiple light sources, specular highlights and Fresnel effects. Finally, the coarse sketch of the user needs to be interpreted to create a new and suitable illumination environment.

The remainder of this chapter begins with a presentation of related work in Section 7.2 and a brief description of the lighting model of Shahlaei and Blanz in Section 7.3. In Section 7.4 the sketch-based lighting design pipeline and in Section 7.5 the adapted fitting procedure are introduced. After a presentation of the results in Section 7.6, a conclusion is given in Section 7.7.

## 7.2 Related Work

This section concentrates on related work in the field of lighting design, as methods for facial landmark localization are already discussed in Section 5.2 and related methods for

3D face reconstruction are presented in Section 6.2.

Many image-based relighting approaches like [BK98, DHT+00, ADW04, PTMD07, RDL+15] require multiple images to estimate the lighting and use light stages to create varying lighting conditions. In this context, Anrys et al. [ADW04] and Peers et al. [PTMD07] promise high quality relighting, but they require high quality light stage data of the corresponding scene to acquire the object's appearance. Debevec et al. [DHT+00] use a high dynamic range reflectance field of a human face which is captured by a light stage that consists of a two-axis rotation system and a directional light source. Lighting reproduction approaches such as [DKN+95, NSD95, GH00, DWT+02, ML04, FBS05, RDL+15] use the captured illumination information, for example by using an environment map, and then apply it to a different scene or object. To improve the lighting of objects that are in the foreground of a video, Wang et al. [WDC+08] balance the intensity and color by incorporating infrared illumination. A survey of image-based relighting is given by Choudhury et al. [CCH07] and a survey of varying lighting design methods has been published by Kerr and Pellacini [KP09]. In difference to all these methods the proposed approach needs only a single input image and a coarse sketch of the desired lighting design.

While sketch-based lighting design has already been used for synthetic scenes by [SDS+93, OMSI07, PBMF07] or for near-diffuse real objects by Anrys and Dutré [ADW04], the approximation of surface normals of objects that are visible only in a single image is a challenging task. Therefore, Okabe et al. [OZM+06] rely on a pen-based interface which enables the user to draw an approximate normal map on the input image, in contrast to that Henz and Oliviera [HO15] use a user-provided guess for the shading to refine the shading of the target. Distinct from the other sketch-based approaches the proposed method does not require the user to have any knowledge about the shape of an object or being able to draw an image of a shaded face or object. Accordingly, no additional experience is required and the sketching process is fast and simple.

In contrast to [PF92, PRJ97, CSF99, SL01, PTG02, ADW04, AP08, HO15] the proposed method focuses on the relighting of faces. Like in [BV99, AS13, LZL14, SB15] a morphable model for faces is used to estimate the shape normals and to address the inverse rendering problem in a single input image, but in case of the 3DMM fitting the proposed method does not require manually selected landmarks. Still it can handle varying facial structures, arbitrary poses and complex, regionally varying reflectance behavior. Accordingly, realistic illumination effects such as the desired glossy behavior in grazing angles, colorful lighting and cast shadows can be replicated.

## 7.3 Realistic Inverse Lighting

*Section 7.3 sums up relevant aspects of the lighting model introduced by Shahlaei and Blanz [SB15] and by Conde et al. [HSBL15] which are used in Section 7.4 to create a lighting design pipeline. As I did not contribute to these approaches, they are out of the scope of this dissertation.*

The lighting model proposed by Shahlaei and Blanz [SB15] requires only a single input image to simulate a variety of lighting effects. This is a crucial difference compared to other image-based relighting methods, where a complete series of input images is used to capture all possible lighting effects. The simulation requires the 3D facial shape and the corresponding surface normals as an input to estimate the general shading effects as well as cast shadows. That is why the 3DMM framework of Blanz and Vetter [BV99] is used to create a 3D face from a single 2D image. Although the 3DMM is also capable to estimate the light in a scene (see Section 2.2), it is limited to a single light source and the Phong reflection model [Pho75]. Consequently, the estimated lighting parameters as well as the extracted texture which are both based on the 3DMM are discarded and they are replaced by a more sophisticated lighting model and a realistic reflectance function which supports additional lights and enables realistic lighting effects.

### *7.3.1 Light estimation*

As described in [SB15] the facial shape which is estimated by the 3DMM is used as a light probe. Therefore, 100 images are rendered in a virtual light stage using the same face alignment and pose as the face in the input image $I$ which is obtained from the 3DMM fitting. In each of these light stage images $C_i$ only a single light source is active. Based on the superposition principle which states that light is additive, the full lighting of the face is calculated by using non-negative coefficients of a weighted summation of $m = 100$ images of that face, leading to

$$I = \boldsymbol{C}\mathbf{x} = \sum_i \vec{x}_i \cdot C_i. \tag{7.1}$$

A non-negative iterative Newton method is used to find best matching coefficients $\vec{x}_i$ which represent the RGB color values of the i-th light source for the input image $I$ by solving the cost function

$$\underset{\mathbf{x}}{\operatorname{argmin}} \; || \; \boldsymbol{C}\mathbf{x} - I \; ||_2^2, \tag{7.2}$$

where the columns of the matrix $\boldsymbol{C}$ contain the light stage images $C_i$.

To support the complex characteristics of human skin the reflectance parameters are based on the measurements of Weyrich et al. [WMP$^+$06] and the diffuse and specular functions proposed by Jensen and Buhler [JB02] are used. This enables region-dependent reflectance attributes, Fresnel effects and subsurface scattering.

Finally, color correction is used to reproduce the tone and contrast of the input image. RGB offset values $\vec{o}$ and a $3 \times 3$ color transformation matrix $\mathbf{K}$ are provided by the 3DMM fitting method proposed by Blanz and Vetter [BV99]. Accordingly, the following mapping

$$(r', g', b')^T = \vec{o} + \mathbf{K}\ (r, g, b)^T \tag{7.3}$$

leads to the corrected RGB color values $(r', g', b')^T$ of the described image synthesis. This color mapping is applied after computing the weighted sum in Equation 7.1 by using the coefficients $\vec{x}_i$ and the light stage images $C_i$ based on the cost function in Equation 7.2 as proposed by Shahlaei and Blanz [SB15].

### *7.3.2 Intrinsic Texture Extraction*

Section 2.3 already outlined the limitations of using a linear combination of texture vectors to model the facial texture, as it can not capture small details like birthmarks, the structure of eyelashes or fine wrinkles. Accordingly, a texture extraction from the original image is proposed and to gain the intrinsic texture the pure albedo needs to be separated from the influence of shading and cast shadows. In contrast to the de-illumination in Section 2.3 which is limited to a single light source and the Phong reflection model, now the virtual light stage approach (see Section 7.3.1) is used to estimate the illumination and to revert its effects to the extracted texture $T_{\text{extr}}$ and to gain a more accurate albedo $M(p)$ for each pixel $p$.

Therefore, the effects of the color correction (Equation 7.3) are reverted for each pixel $p$ in the input image $I(p)$. This is done by using the color offset $\vec{o}$ and the color transformation matrix $\mathbf{K}$ to gain color corrected texture values. After that the diffuse and the specular effects which have been estimated in Section 7.3.1 are removed for each pixel to gain the albedo $M(p)$.

As the 3DMM fit establishes a dense correspondence between the face model and the input image, the texel values for the extracted texture $T_{\text{extr}}$ can be sampled directly from the input image $I(p)$ or to be more precise from the albedo $M(p)$. For vertices at grazing angles and for vertices hidden in the input image (almost) no information about the texture is available directly, in these cases the estimated values based on the linear combination of the 3DMM are used.

## 7.4 Paint-based Lighting Design Pipeline

*The paint-based lighting design is a joint work and has been published in [SPB16]. In this context, the combination of a stroke-based solution with an inverse lighting approach*

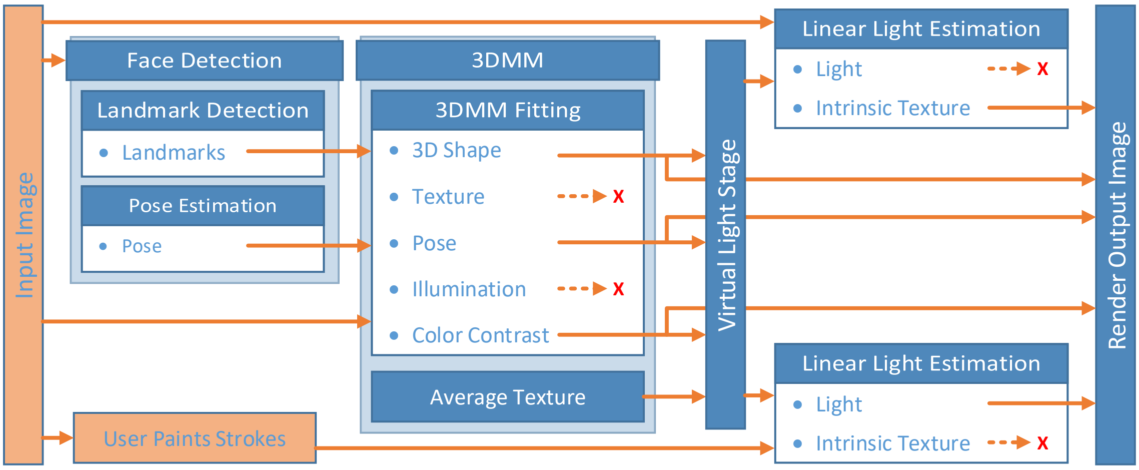


**Figure 7.1:** Overview of the lighting design pipeline: First the face and its landmarks are detected in the input image to enable 3DMM-based fitting. While parameters like 3D shape, pose and color contrast are used in the virtual light stage, the estimated texture and illumination of the 3DMM are discarded, because the virtual light stage and the linear light estimation module (top) deliver more precise estimations (see Section 7.3). In a parallel process the strokes drawn by a user on top of the input images are used for an additional light estimation (bottom). Finally, the shape and intrinsic texture are rendered combined with the desired lighting derived from the strokes and the estimated color contrast and pose from the original image.

*to enable synthetic relighting is a contribution of our co-author, Davoud Shahlaei. Its details are only briefly explained and are out of the scope of this dissertation. Instead the focus of this dissertation lies on the seamless integration of an automatic face detection and landmark localization algorithm (see Section 5.3), plus a new and adaptive 3D face reconstruction approach (see Section 7.5) which makes use of the Normal distance (see Section 6.3.5). The conjunction of these approaches enables a user-friendly and intuitive relighting procedure.*

According to Kerr and Pellacini [KP09] most users perform poorly with paint-based methods because they tend to sketch rather than accurately paint the desired images. Therefore, the proposed paint-based approach is designed to create physically correct and plausible lighting, including shadows and highlights, from very rudimentary sketches as can be seen in Figure 7.2a and in Section 7.6.

For the user-defined lighting design pipeline, the inverse lighting algorithm [SB15] which is outlined in Section 7.3 is used to derive the desired lighting from coarse sketches on the input image. The complete pipeline that enables to sketch a lighting situation in the context of lighting design is shown in Figure 7.1.

Like the inverse lighting in Section 7.3 it starts with an input image that shows a photo or painting of a face. Then the automatic face detection and landmark localization method which is described in detail in Section 5.3 is used to initialize the 3DMM fitting for the face

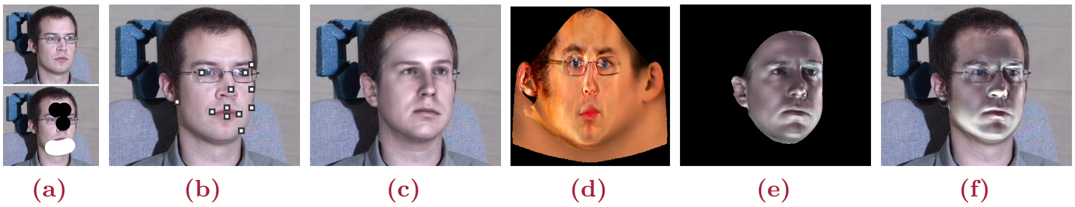


**Figure 7.2:** The process of lighting design: Based on the input image 7.2a (top) from Multi-PIE [GMC+10], the algorithm locates landmarks 7.2b and reconstructs a 3D model 7.2c using 3DMM fitting with a simplified lighting model. The 3DMM texture is replaced by an improved intrinsic texture estimation 7.2d. Lighting estimation on the painted-on image 7.2a (bottom) and a relighting of the 3D face using an empirical BRDF model simulate the new appearance 7.2e, which is composited into the original input image 7.2f.

in the image. After the 3D face reconstruction with the 3DMM of Blanz and Vetter [BV99] (see Section 2.2) has been performed, the 3D shape, the estimated pose and color contrast of the face in the input image and the average texture of the 3DMM are used to create a virtual light stage which is equivalent to the one used for the inverse lighting computation in Section 7.3. But instead of using the virtual light stage to create the intrinsic texture and the light setup in a single linear light estimation as has been proposed for the inverse relighting, for the lighting design this processing step is split into two separate parts: In the first linear light estimation the intrinsic texture is computed based on the unaltered input image like in Section 7.3 but the estimated lighting is discarded at this point. Instead, a second linear light estimation analyzes the colored patches that have been drawn by a user on top of the face in the original image. These painted sketches are used to add new cast shadows or highlights to the face by adjusting the light setup adequately. This time only the attributes for the light setup are used while the intrinsic texture is discarded.

For rendering the face with the new lighting, the 3D shape, pose, color contrast and the intrinsic texture from the original input image are combined with the light setup based on the manual sketches. Finally, the rendered face is composited into the input image to replace the original face with a relighted substitute. While it is important that the silhouette of the face remains sharp, there should be smooth transitions to parts of the human body that are not modeled explicitly like hair or lower parts of the neck and shoulders. This is achieved by using a blend map which labels most parts of the 3D head as completely opaque but gradually increases transparency at the edges of the head mesh to enable alpha blending.

In Figure 7.2 the intermediate results of the previously described processing steps are shown by starting with a typical input image. These results include automatic landmark localization, 3DMM fitting, texture extraction, estimation of designed lighting and compositing.

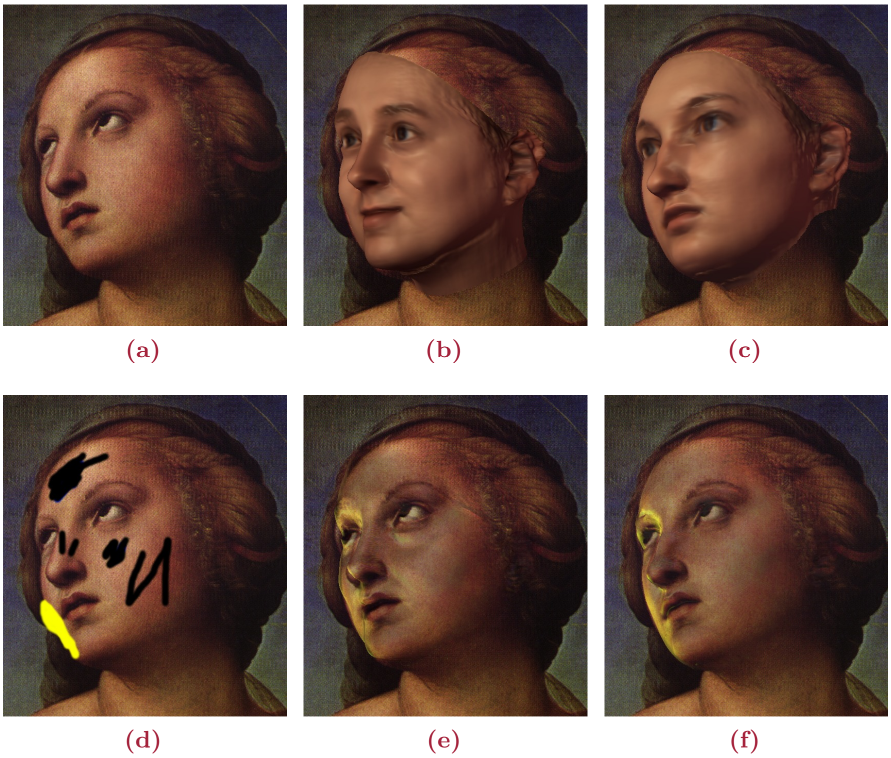


**Figure 7.3:** In difficult cases the automatic landmark localization may be suboptimal as is shown in Figure 7.3b. A manual selection of landmarks may improve the pose and 3D shape estimation (see Figure 7.3c) and consequently the relighting result (see Figure 7.3f) compared to the automated result in Figure 7.3e.

## 7.5 3DMM Fitting Revisited

In Section 5.4 it is reported that using the automatically extracted landmark locations to fit a 3D morphable model to a single input image of a face is prone to produce degraded 3D face reconstructions. As the 3DMM fitting is vulnerable to false input, sometimes performing even small corrections to the landmark locations is enough to achieve the desired results. Thus, Shahlaei et al. [SPB16] propose to use manual selected landmarks as a fallback strategy if the automatic reconstruction has failed like it is shown in Figure 7.3. While the 3D shape reconstruction in the given example is not perfect but still plausible, the adaptation to the pose in the input image has failed. This results in an implausible relighting. By using carefully considered and manually chosen landmark locations the 3DMM may adapt better to the pose as can be seen by comparing the result based on manually selected landmarks in Figure 7.3f with the relighting result in Figure 7.3e which is based on automatically selected landmarks.

In Chapter 6 a quality metric driven approach has been proposed to improve the 3D reconstruction results from multiple, unconstrained 2D input images [PB16]. But in contrast to the face reconstruction from photo collections, the presented scenario for image-based relighting of faces typically uses only a single input image instead of a complete series of photos. Accordingly, a quality metric which identifies implausible reconstructions might be helpful to suggest a manual selection of facial landmarks for failed cases, but it can not be used directly to solve the actual problem.

However, with some additional modifications, the idea of the quality metric driven approach in Chapter 6 can be adapted to the current scenario to improve the quality of 3D shape reconstructions from a single input image. Therefore, it incorporates the Normal distance (see Section 6.3.5) as an additional regularization term into the fitting procedure of the 3DMM. In the following the original 3D fitting approach (Section 7.5.1) is compared to the proposed modifications (Section 7.5.2).

### *7.5.1 Non-Adaptive Processing Steps*

By default the 3DMM fitting of Blanz and Vetter [BV99] follows a number of predefined processing steps as is schematically outlined in Table 7.1. For example, for each processing step the number of iterations in stochastic gradient descent (Column 2), the number of principle components of shape (Column 3) and texture (Column 4) that are taken into account, and the regularization intensity (Column 5) are specified as is discussed in Chapter 2.

When taking a closer look at these values, it can be seen that the fitting starts with a rough adaptation to the shape in the input image using only the five most significant principle components of the 3DMM for the shape. During the course of the fitting additional shape principle components $\mathbf{s}_i$ are taken into account and likewise the magnitude of the regularization decreases allowing the model to move further away from the mean head in terms of Mahalanobis distance (see Sections 2.2 and 6.3.2). Additionally, not every model parameter is adapted at the same time. This is shown in the last column of Table 7.1. First, only rigid transformations are performed, by using the (automatically) selected facial landmark positions the head model is translated to the position of the head in the image and adapts to its pose and size. Later the light estimation is done, followed by further adaptation steps that jointly optimize for all parameters. While the procedure in Table 7.1 is only used for fitting the complete face, subsequently similar predefined steps are used for a segment-based refinement to enhance the expressiveness of individual facial regions (see Figure 6.4a).

### *7.5.2 Adaptive Processing Steps*

In contrast to the non-adaptive processing steps in Section 7.5.1 which have been used for the 3DMM fitting so far, the pseudocode in Algorithm 7.1 shows a way to integrate the plausibility estimation based on the Normal distance directly into the 3DMM fitting

| Step | Iterations | No. PCs $\mathbf{s}_i$ | No. PCs $\mathbf{t}_i$ | Sigma | Type |
|---:|---:|---:|---:|---:|---|
| 1 | 800 | 5 | 1 | 800 | rigid |
| 2 | 800 | 14 | 1 | 200 | rigid |
| 3 | 800 | 14 | 1 | 200 | rigid |
| 4 | 800 | 14 | 1 | 100 | rigid |
| 5 | 400 | 24 | 1 | 100 | rigid |
| 6 | 400 | 24 | 1 | 100 | rigid |
| 7 | 400 | 24 | 1 | 50 | rigid |
| 8 | 2000 | 24 | 1 | 50 | rigid |
| 9 | 2000 | 24 | 1 | 50 | rigid |
| 10 | 1500 | 24 | 1 | 20 | light |
| 11 | 1500 | 24 | 1 | 20 | light |
| 12 | 1000 | 24 | 20 | 20 | all |
| 13 | 1000 | 24 | 20 | 20 | all |
| 14 | 1000 | 34 | 30 | 15 | all |
| 15 | 1000 | 44 | 40 | 10 | all |
| 16 | 1000 | 74 | 70 | 10 | all |
| 17 | 1000 | 104 | 99 | 5 | all |
| 18 | 2000 | 104 | 99 | 2 | all |

**Table 7.1:** The 3DMM fitting parameters are predefined. For example, this includes the number of iterations per step (Col. 2), the number of used shape (Col. 3) and texture (Col. 4) principle components, and the regularization intensity (Col. 5). A flag states the type of the fitting operations (Col. 6): only *rigid* transformations, only *light* estimations or adapting *all* together including non-rigid shape and texture adjustments.

procedure. However, many of the outlined steps in Algorithm 7.1 are congruent to the 3DMM fitting described in Section 2.2. First, the manually or automatically selected landmarks and the predefined parameters (see Table 7.1) are loaded. Then the rigid model parameters get optimized based on the selected landmark locations to generate a rough fit between the 2D image and the 3D proxy geometry of the 3DMM which corresponds to lines 1–6 in Algorithm 7.1.

Distinct from the original fitting approach, the proposed algorithm creates a copy of the current shape, rigid, texture and lighting coefficients (Algorithm 7.1, line 8) to be able to undo all changes (Algorithm 7.1, line 12) of the subsequent optimization procedure which uses stochastic gradient descent in the current processing step. After the gradient descent has been performed, the resulting coefficients are used to create an intermediate 3D face geometry. This 3D geometry is used for the plausibility estimation (Algorithm 7.1, lines 10–11) based on the Normal distance $d_{\text{normal}}$ as is explained in detail

**Algorithm 7.1:** Adaptive 3D face reconstruction

```
Data   : Input image $I$
Result : 3D face geometry
load estimated facial landmarks locations for image $I$
load predefined fitting parameters
foreach Step do
    if Type == rigid then
        match 3DMM proxy geometry to image $I$ using landmarks
        optimize only rigid coefficients
    else
        store current shape, rigid, texture and lighting coefficients
        optimize coefficients using stochastic gradient descent
        compute Normal distance $d_{\text{normal}}$
        if $d_{normal} >= threshold$ then
            restore previous coefficients
        end
    end
end
create final 3D face geometry using all coefficients
```

in Section 6.3.5. Choosing the Normal distance for this task is motivated by the results of the evaluation in Chapter 6, where it performed best in distinguishing plausible from implausible reconstructions compared to the other evaluated distance measures.

Like in Chapter 6 a predefined threshold value is used to determine if an intermediate 3D reconstruction result is still plausible. If that is the case the altered face coefficients are passed to the subsequent processing step. But if the reconstruction is rated to be implausible during the current stochastic gradient descent, the altered coefficients are replaced by the ones that have been saved, before starting the next optimization procedure. This undo procedure allows to use the unaltered coefficients again in the subsequent step. As each step uses different, predefined fitting parameters and the stochastic gradient descent uses a randomly selected subset of all available vertices for the optimization, the basic idea behind the proposed method is that undoing an incorrect optimization step provides the chance to recover from failure. To not take any chances and to completely avoid an infinite loop the current step is not repeated, even though this misses the opportunity that a repetition of the current processing step with a different subset of vertices might result in a plausible intermediate result. Instead the coefficients are reset to their last state and passed to the next predefined processing step.

For simplicity the region-based optimization steps are excluded in Algorithm 7.1. Like for the complete face, there are also predefined steps for the segment-based fitting which are executed after the fitting to the full head is completed. The plausibility estimation based on the Normal distance $d_{\text{normal}}$ is performed for each individual segment the same way as it is described for the complete face.

After all processing steps are performed the final face geometry is created by using all optimized coefficients and by merging all facial segments together.

### *7.5.3 Comparison between Non-Adaptive and Adaptive Fitting*

Each row of Figure 7.4 contains the results of a series of 3D face reconstructions based on the same input image (Column 7.4a). In Columns 7.4b to 7.4d the corresponding facial reconstructions based on automatically detected landmarks (see Sections 5.3.1 and 5.3.2) and in Columns 7.4e to 7.4g the results based on manually selected landmarks are shown. This enables to compare not only the influence of the adaptive fitting approach on reconstructions based on automatically detected facial landmarks, but also its effect on manually detected landmark locations.

Furthermore, two different magnitudes for the regularization by Normal distance are used. While the reconstructions in Columns 7.4b and 7.4e have been created using the non-adaptive fitting, in Columns 7.4c and 7.4f the adaptive fitting with a threshold value (see Figure 6.7) for the Normal distance of $u = 11$ is used, whereas a value of $u = 6$ has been applied to create the results in Columns 7.4d and 7.4g. To be more specific, a threshold of $u = 11$ means that the average difference between the normal directions of each vertex of the reconstructed geometry should not differ more than 11° from the normal directions of the average face.

As is discussed in Chapter 6 the quality metric can be used to check the shape plausibility for the whole head as well as for individual facial regions. In the presented examples both capabilities are used. First, the Normal distance is used to examine the plausibility of the whole face and later during refinement of individual regions the same threshold value is used again (see Section 7.5.2). Based on the findings in Section 6.5 a threshold of $u = 11$ is proposed to separate plausible from implausible reconstructions (see Figure 6.7), in the current scenario this exact threshold value is used again along with an even stricter value of $u = 6$ for comparison.

By taking a closer look at the results in Figure 7.4 it can be seen that incorrectly selected landmarks lead to degraded facial reconstructions (see Section 5.4). This is particularly noticeable in the fourth row of Column 7.4b but actually (local) artifacts are visible in all reconstructions of Column 7.4b. By using the proposed adaptive fitting approach and a threshold of $u = 11$ (see Column 7.4c) several artifacts can be reduced. With an even more aggressive threshold (see Column 7.4d) most of the previously seen artifacts disappear. But it should be noted, that this comes at the expense of expressiveness in the final result, because the stronger the regularization the more the final shape is forced to stay close to the average shape of the model.

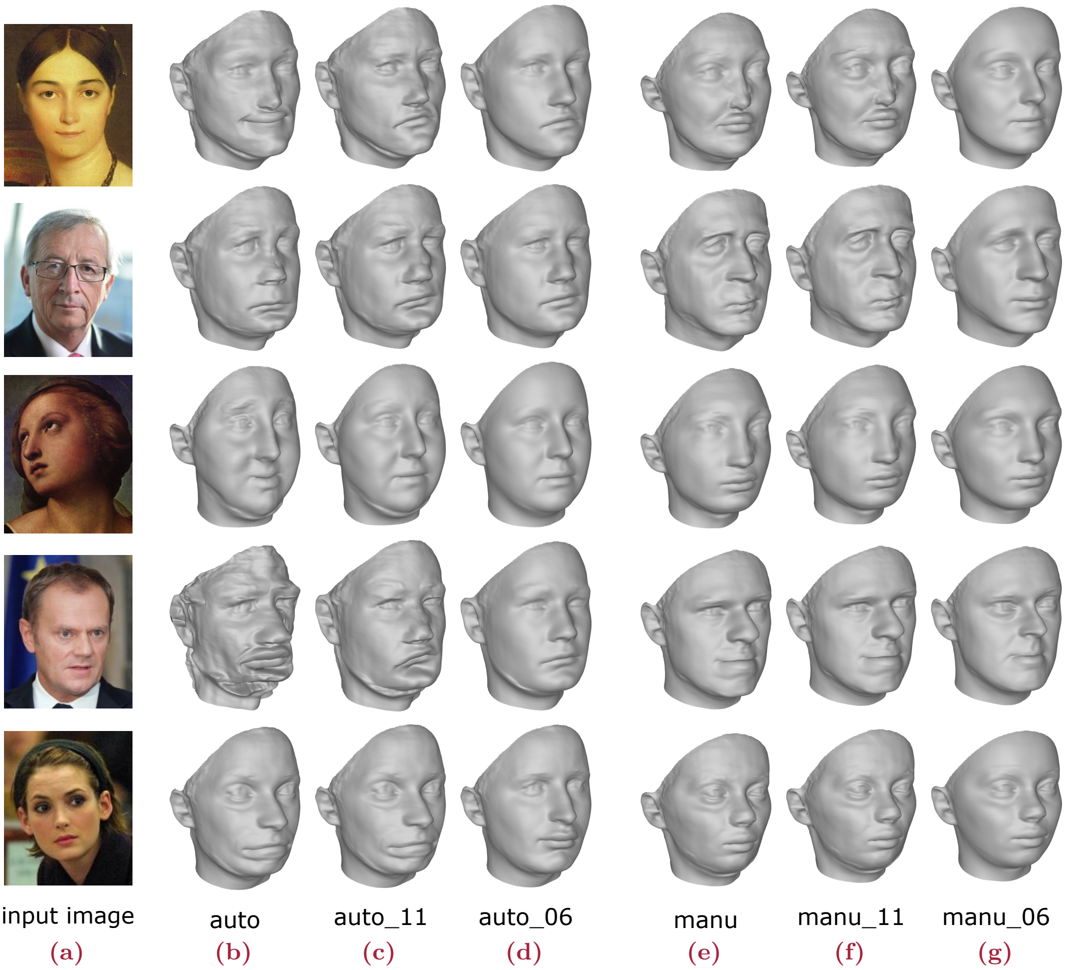


**Figure 7.4:** Column 7.4a contains the input images. In Columns 7.4b – 7.4d the corresponding facial reconstructions based on automatically detected landmarks (see Sections 5.3.1 and 5.3.2) and in Columns 7.4e – 7.4g the results based on manually selected landmarks are shown. While the reconstructions in Columns 7.4b and 7.4e have been created using the non-adaptive fitting, in Columns 7.4c and 7.4f adaptive fitting with a threshold value (see Figure 6.7) of $u = 11$ and a value of $u = 6$ in Columns 7.4d and 7.4g is used.

Even though the adaptive fitting approach was originally designed to reduce the artifacts that are introduced because of inaccurate landmark locations due to the automatic landmark localization it can also be used in combination with manually selected landmarks as is shown in Columns 7.4e to 7.4g. Like the self-adapting feature layer approach by Breuer and Blanz [BB10] which tries to correct inaccurately selected landmark locations to same degree, the proposed adaptive fitting can be seen as an alternative or additional strategy to deal with this kind of imperfections.

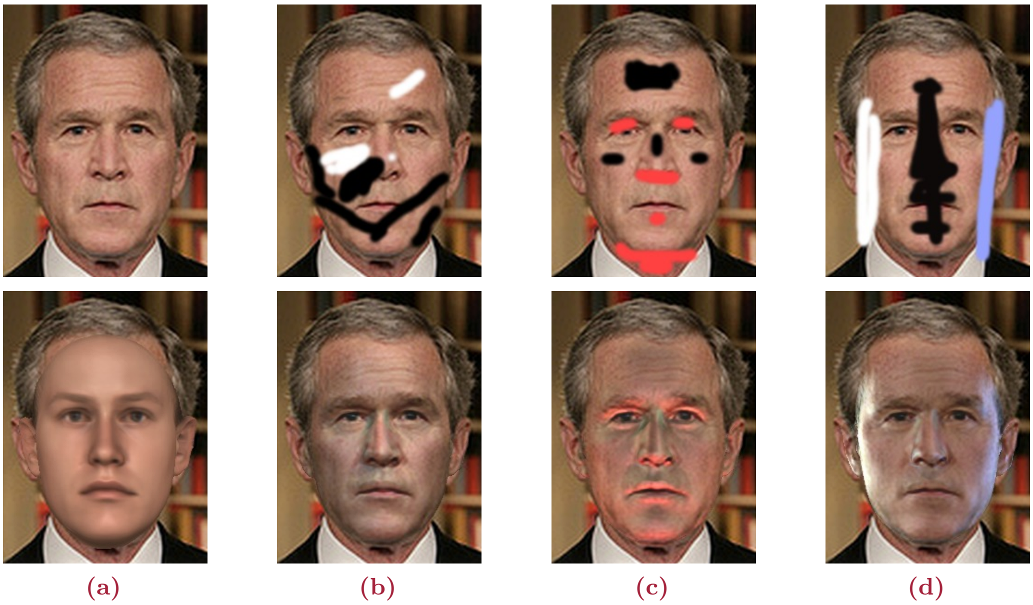


**Figure 7.5:** Figure 7.5a shows the original image (top) and the estimated 3D face with average human face texture (bottom). Figures 7.5b, 7.5c and 7.5d contain the automatically generated results (bottom) based on the proposed algorithm for the user-defined lightings (top).

## 7.6 Results

The proposed lighting design method has been applied to photographs of the Labeled Faces in the Wild (LFW) dataset [HRBLM] which has also been used for the 3D reconstructions from unconstrained image collections in Chapter 6, to images of the Multi-PIE dataset [GMC$^+$10], and to some paintings and portraits from public domain. The Figures 7.3, 7.5, 7.6 and 7.7 show the results for different subjects in situations with varying lighting designs.

Based on the paint strokes of the user the algorithm can be used to emphasize silhouettes or structures, or to add depth to the image. The proposed method supports mild (Figure 7.7c) and intense modifications (Figure 7.5d). Rim-lights are added in Figure 7.5d and 7.7h (right side of the face), or are removed like in Figures 7.7f to 7.7h (left side of the face). In Figure 7.7b the positions of main and fill light are swapped. Furthermore, additional lights with an arbitrary color and light direction can be added to a scene, changing the original look of a portrait completely, like in Figure 7.5c by adding light from below. An example of the estimated face shape that is used for the lighting computation is illustrated in Figure 7.5a (bottom).

The results in Figures 7.5 and 7.6 are created by using the automatic landmark localization approach introduced in Section 5.3. Combined with the adaptive 3DMM fitting based on Normal distance the chances to produce plausible reconstructions are increased

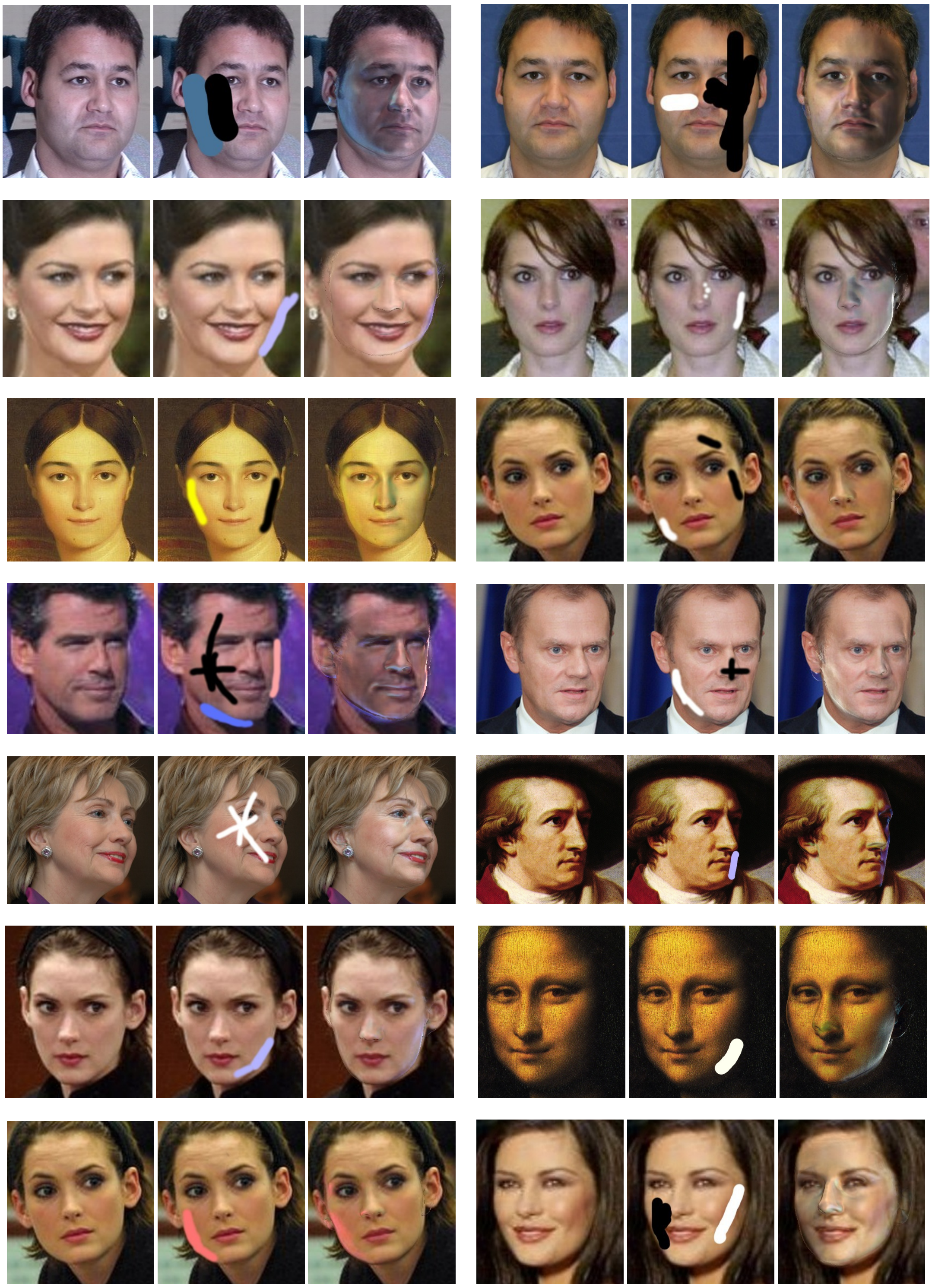


**Figure 7.6:** Each row contains two sets of three images showing the original image, the painted-on image and the automatically generated result.

without losing too much of the expressiveness of the face in the input image. However, even with the adaptive fitting from Section 7.5, there are cases where 3DMM fitting based on manually located landmark remains superior, especially as it enables the user to iteratively correct the location of landmarks or to use varying landmark sets until the final result is pleasing. Although this is more time consuming, the additional effort may pay of in some cases. For example, if the image contains uncommon poses, strong expressions or occlusions. If the automatically detected facial landmarks are too far away from the correct solution even a stronger regularization can not fix the problem, because one of the first processing steps of the 3DMM is to use these landmarks to initially translate the face model to the location of the face in the image, to adapt its size and to estimate its pose. Accordingly, it needs to trust these landmark position to some extent.

Results based on the manual selected landmarks are shown in Figure 7.7. While the original input image can be seen in the middle, the painted strokes are at the top and the result is at the bottom of each column. In Figures 7.7f to 7.7h the rim-light in the original image is removed by using different methods to demonstrate the tolerance and flexibility of the algorithm. For example, in Figure 7.7f the lighting of both sides has been balanced through the additional strokes, while in Figure 7.7g the original highlights are completely overpainted.

Inferring the facial appearance from a single input image without any additional knowledge about the lighting situation remains a challenging problem. This can be seen in some results, e.g. Figure 7.7e, where the lack of detail in the face geometry and the absence of a measured reflectance lead to visible errors, i.e. wrong highlights on the left side and unrealistically looking cast shadows under the eyes. These kind of artifacts appear often at the silhouettes, the nose, mouth or eyes if the 3D shape is not perfectly aligned with the input image. Depending on the respective lighting conditions the extent of these errors varies even for the same input image as is illustrated in Figures 7.7f to 7.7h.

## 7.7 Conclusions

The proposed paint-based method for post-hoc lighting design can be applied to arbitrary portraits. It adapts to the pose and facial expression of the face in the input image, estimates the original lighting conditions and replaces or augments them based on paint strokes drawn by the user. With a minimum effort comprehensible light modifications are applied to the original face by using a virtual light stage that changes the attributes of each single light source. This kind of automatic adaptation provides a direct interface for lighting design which enables a creative exploration of different illumination scenarios. It can be used to emphasize silhouettes and facial structures or to make images more appealing by creating a different, more suitable illumination.

As has already been discussed in Chapters 5 and 6, the 3D reconstruction of human faces from arbitrary images is a challenging task. Furthermore, in the current scenario only

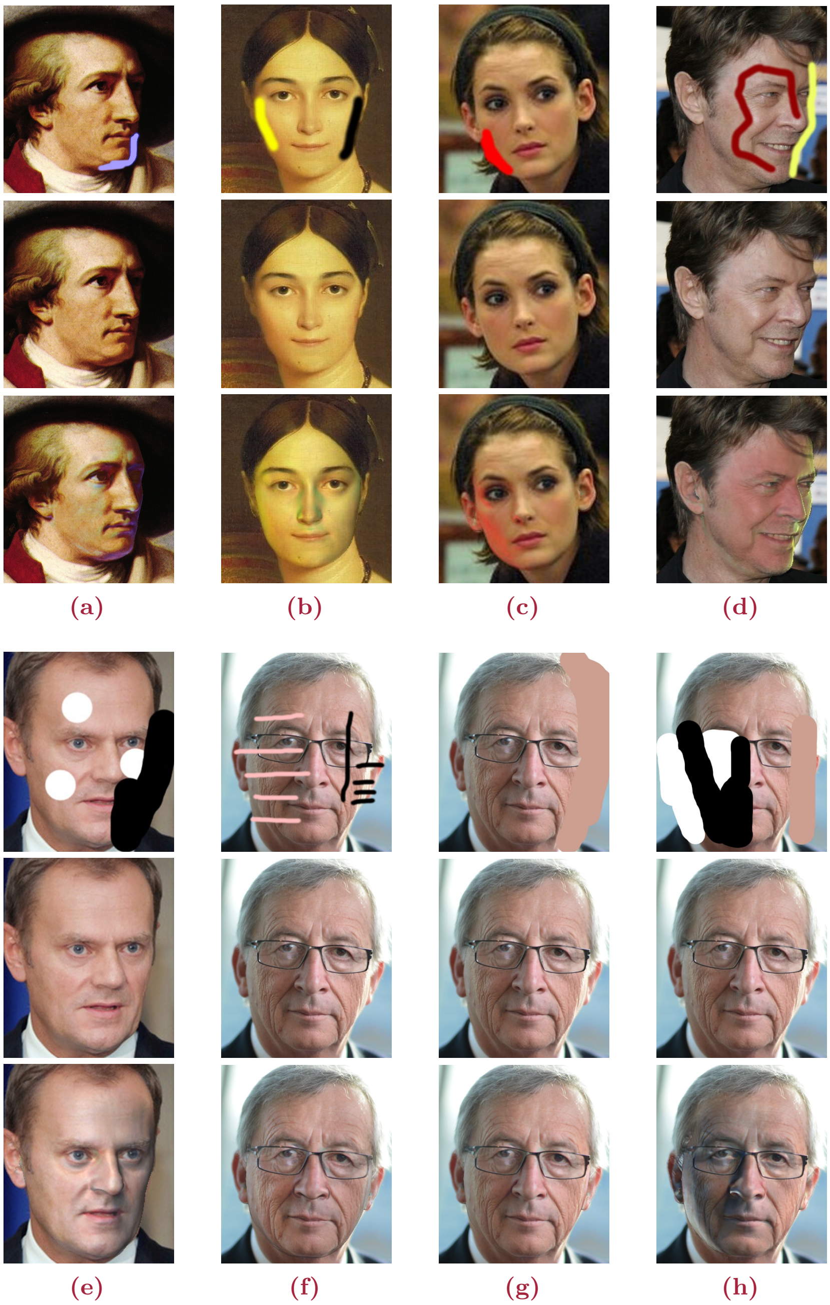


**Figure 7.7:** Eight examples are shown, each structured as painted-on image (top), original image (middle) and result image (bottom). The input image in Figure 7.7d is from the German Wikipedia page for David Bowie, the others are from the LFW [HRBLM] dataset. The landmark localization has been done manually by the user.

a single input image is available which increases the chance of implausible reconstructions. Nevertheless, in Section 7.5.3 it has been shown that the integration of a quality metric for plausibility estimation and using an adaptive fitting procedure is not only useful in combination with photo collection but is also a quite effective tool to drive the 3D reconstruction of a single input image.

Besides a good 3D reconstruction of the facial shape, the adaptation to the pose in the input image is an extremely important aspect to create high quality relighting results. Accordingly, the landmark localization and pose estimation approach by Zhu and Ramanan [ZR12] is not applicable for all kind of input images even if it is combined with the proposed adaptive fitting procedure. Therefore, manual landmark selection might still be necessary for hard cases.

Although the proposed method cannot yet cope with the quality of manual image relighting done by a professional artist, it might be a helpful tool for ambitious photographers to explore and study the effects of lighting on the appearance of human faces.

# 8 Conclusions

In this dissertation, several different methods are presented to enhance the shape, texture and lighting during the 3D face reconstruction from a single image as well as a photo collection. To some extent, all proposed approaches make use of the original 3DMM introduced by Blanz and Vetter [BV99] (see Chapter 2). Besides the possibility of reconstructing 3D faces from 2D images, which is driven by prior knowledge of faces, a key advantage of the 3DMM is the dense point-to-point correspondence. Creating correspondence between the vertices of each reconstruction and also between the 3D shape and the 2D input image allows straight forward morphing operations between all reconstructions, texture extraction from the input image, and the exchange of textures and facial expressions from arbitrary reconstructions.

In Chapter 3, the 3DMM is used to create a new model for children and teenagers and to combine it with the original one which is limited to faces of adults. The original model is extended by fitting the 3DMM to the 3D point cloud data created by a structured light scanner and to high-resolution portrait photos of these children. During these fitting steps the exact 3D geometry is extracted from the scans and a matching high-resolution texture is created from the photos. As the age of each person in the newly acquired scans and photos is known, the point-to-point correspondence is also used to generate age-related geometry and texture information to simulate the aging of a person without requiring extensive longitudinal 3D data of each individual person. Instead longitudinal information of several individuals is fused to create a model for aging simulation. Additionally, longitudinal 2D data of different persons has been analyzed and can be transferred to the 3D shape and texture of any other person. As already mentioned in Chapter 3, it is not claimed that this kind of simulation perfectly predicts the facial changes related to aging as there are several unknowns like nutrition, health, living conditions, etc., but it can be used as an approximate estimate. Thus, a method is introduced which extracts age dependent data and learns effects of aging from a collection of unconstrained scans and photos, and applies these learned features to the facial geometries of other people.

By identifying regional similarities in faces of individuals and by using hallucination techniques, it has been shown in Chapter 4 that missing details in blurred input images can be effectively augmented. Therefore, fine structures like wrinkles, facial hair, pores and slight dermal irregularities are extracted from another face or facial region which is similar apart from the additional details. Then exactly these missing features are transferred into the previously blurred facial texture to improve its appearance. Again the dense point-to-point correspondence provided by the 3DMM is a crucial prerequisite for this method.

Subsequent chapters take advantage of the point-to-point correspondence by combining facial shape information of several unconstrained 3D face reconstructions of the same person to gain a more plausible and advanced overall result. While in Chapter 5 the combining process is done more or less blindly by averaging all (pair-wise) reconstructions, in Chapter 6 different quality measures to drive the combination of facial features based on plausibility criteria have been evaluated. In this context, it turned out that the Normal distance matches the human perception best in distinguishing plausible from implausible 3D face reconstructions. The proposed approach rates the plausibility of each facial region independently and uses a weighted-linear combination technique to find a reasonable compromise between expressiveness and robustness. In Chapter 7 the quality metric is directly integrated into the fitting and optimization procedure of the 3DMM to reduce implausible 3D face reconstructions when only a single input image is available. This enables to stepwise revoke changes of the geometry during the fitting if the plausibility criterion is violated and consequently improves the image-based relighting of faces. The proposed method can be used for shape adaptations with respect to the whole head as well as for local modifications during the subsequent segment-based fitting.

With the basic concept of the 3DMM being present all along in this dissertation, improvements and extensions are made in every major area including the reconstruction of the shape, texture and lighting. And most notably, it has been shown that the 3DMM does not necessarily need to be initialized by manually selected landmark locations, even for 'faces in the wild'. Accordingly, now the whole fitting process is much more convenient for the user and the required expert knowledge has been reduced to a minimum.

# 9 Future Work

After a sound basis for automatic face reconstruction and facial texture enhancement has been established during the course of this dissertation, these achievements can be used for further improvements in future work.

In recent publications [ZC01, DB04, PF05, SH06, YCSS14] it has been shown that shape-from-shading techniques are quite effective in recovering fine local surface details. On the other hand, Patel and Smith [PS12] argue that the estimation of the global shape using shape-from-shading is inferior to a morphable model and suggest a combination of both methods. Accordingly, the proposed face reconstruction method in Chapter 6 could make use of shading cues from the more reliable views, which have been identified based on the Normal distance, to enhance the overall expressiveness of the resulting face reconstructions. For 'faces in the wild' and similar unconstrained photo collections an additional clustering which classifies different facial expressions and poses might be necessary before using the 2D image data for shape estimation.

Furthermore, the texture extraction in Chapter 6 could be extended to handle multiple images simultaneously which would enable to automatically combine textures from multiple facial poses. The proposed method in Section 6.6 which has been published in [PB16] extracts the texture only from the image that corresponds to the face reconstruction with the minimal overall Normal distance. Thus, if the image with the minimal Normal distance shows the left side of a face, texture details of the right side are not visible at all or are just mirrored. By creating the final facial texture from multiple images, it would be possible to extract texture information from facial parts that are occluded in several images as long as they are visible in at least one photo. A method to combine textures of different poses has already been introduced in Section 3.6, where it is used in an environment with controlled lighting conditions. For unconstrained image collections with varying lighting situations Poisson blending [PGB03] might be required to create seamless transitions between neighbored textures which are extracted from different photos. Additionally, the work proposed by Shahlaei and Blanz [SB15] or Conde et al. [HSBL15] could be used to neutralize lighting effects before extracting the facial textures. As has been discussed in Section 7.3 these methods are superior to the original lighting estimation of the 3DMM by Blanz and Vetter [BV99].

A segment-based texture extraction could also be useful to deal with occlusions of facial parts which are caused by (sun) glasses, beards, hats, other objects or persons in the image. In [SPB15] Schumacher et al. already proposed an occlusion handling for the 3DMM, but the occluded regions needed to be marked manually. To retain the automatic face reconstruction process a detector for occluded facial areas like the one presented by Ghiasi and Fowlkes [GF15] is required.

Sometimes photo collections contain only low-resolution images of a person like the LFW dataset [HRBLM]. This also applies to faces from the footage of security cameras. In the latter case reasons can be an overall low resolution of the camera or that, distinct from portraits, the face occupies only a small part in each frame. It is shown in Section 4.4 that detail transfer for these kind of low-resolution images is possible and Chapter 6 points out that an automatic 3D reconstruction can be done as well. A combination of the proposed transfer of texture details with the automatic face localization and reconstruction method might lead to a plausible 3D reconstruction which contains a higher degree of texture details than every available input image.

Finally, it might be possible to replace the manually generated, pose-dependent mapping table for facial landmarks in Section 5.3.2 with an automatically learned one which is more flexible and adapts better to changes in pose. Of course, a significant reason of the superior reconstruction quality when using manually detected landmarks is the higher precision of human users to identify and mark the landmark locations compared to current landmark localization algorithms. However, another important aspect is the possibility for trial and error during the manual reconstruction procedure. It allows users to iteratively improve the result of the 3D reconstruction by trying different combinations of landmark locations for individual faces, poses and expressions. For this purpose a classifier could be trained by analyzing all already processed photos and their manually located landmark positions in respect to pose, facial expression and landmark type, e.g. fixed or contour points (see Figure 5.4). As for most images in the Multi-PIE dataset [GMC+10] the landmarks have already been selected manually and all images are labeled in respect to facial pose and expression, so that suitable ground truth data is already available. In this context, the Normal distance (see Section 6.3.5) might be helpful to drive the learning approach as it provides feedback on the quality of all resulting reconstructions.

# A
# Appendix

The appendix of this dissertation contains additional tables regarding the aging simulation experiments in Section 3.10 and additional details on the ranking and merging of specific facial segments in Section 6.7.2. As these tables take up excessive amounts of space, they have been moved to the end of this dissertation so that they do not negatively affect the reading flow. Nevertheless they might provide interesting details and therefore are not omitted completely.

## A.1 Additional Tables: Experiments on Aging Simulation

In Section 3.10 three perception experiments regarding aging simulation have been discussed. These have been used to validate if the synthetically aged faces look plausible and if they match the targeted age. In the attached Tables A.1, A.2 and A.3 the results for each of the eight participants with respect to the 51 image pairs that have been shown during the experiment are summarized.

| Experiment 1 | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| photo id | $p_1$ | $p_2$ | $p_3$ | $p_4$ | $p_5$ | $p_6$ | $p_7$ | $p_8$ | average |
| 001 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 2/8 |
| 002 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 003 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 5/8 |
| 004 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 005 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 006 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 007 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 6/8 |
| 008 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 009 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 7/8 |
| 010 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 4/8 |
| 011 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 012 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 7/8 |
| 013 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 7/8 |
| 014 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 015 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 6/8 |
| 016 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 017 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 018 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 6/8 |
| 019 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 6/8 |
| 020 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 6/8 |
| 021 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 5/8 |
| 022 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 6/8 |
| 023 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 3/8 |
| 024 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 025 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 7/8 |
| 026 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 028 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 3/8 |
| 029 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 6/8 |
| 031 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 033 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 4/8 |
| 034 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 3/8 |
| 035 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 036 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 7/8 |
| 037 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 7/8 |
| 038 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 039 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 2/8 |
| 040 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 041 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 3/8 |
| 042 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 4/8 |
| 043 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 5/8 |
| 044 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 045 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 6/8 |
| 046 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 7/8 |
| 047 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 6/8 |
| 049 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 050 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 6/8 |
| 051 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 7/8 |
| 052 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 053 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 7/8 |
| 054 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 055 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 7/8 |
| SUM | 40 | 41 | 42 | 43 | 44 | 44 | 37 | 34 | 40.6 |
| % | 78.4 | 80.4 | 82.4 | 84.3 | 86.3 | 86.3 | 72.5 | 66.7 | **79.7** |

**Table A.1:** The results of the **1st** perception experiment on the proposed aging simulation approach. Matches are labeled with the number '1' and mismatches with '0'. The last column contains the average result of all eight participants ($p_i$) and the last row shows the percentage of the matches.

| | | | | Experiment 2 | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| photo id | $p_1$ | $p_2$ | $p_3$ | $p_4$ | $p_5$ | $p_6$ | $p_7$ | $p_8$ | average |
| 001 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0/8 |
| 002 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 7/8 |
| 003 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 5/8 |
| 004 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 005 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 5/8 |
| 006 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 6/8 |
| 007 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 6/8 |
| 008 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 5/8 |
| 009 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 5/8 |
| 010 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 6/8 |
| 011 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 5/8 |
| 012 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 013 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 3/8 |
| 014 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1/8 |
| 015 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 5/8 |
| 016 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 017 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 6/8 |
| 018 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 019 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 4/8 |
| 020 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 021 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 2/8 |
| 022 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1/8 |
| 023 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 5/8 |
| 024 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 6/8 |
| 025 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0/8 |
| 026 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 5/8 |
| 028 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 029 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 7/8 |
| 031 | 1 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 5/8 |
| 033 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 6/8 |
| 034 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 4/8 |
| 035 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 4/8 |
| 036 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 6/8 |
| 037 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 7/8 |
| 038 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1/8 |
| 039 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 5/8 |
| 040 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 7/8 |
| 041 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 5/8 |
| 042 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 4/8 |
| 043 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 7/8 |
| 044 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 045 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 4/8 |
| 046 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 6/8 |
| 047 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 5/8 |
| 049 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 4/8 |
| 050 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1/8 |
| 051 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 2/8 |
| 052 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 7/8 |
| 053 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 5/8 |
| 054 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 6/8 |
| 055 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 5/8 |
| SUM | 26 | 35 | 39 | 32 | 33 | 29 | 29 | 34 | 32.1 |
| % | 51.0 | 68.6 | 76.5 | 62.7 | 64.7 | 56.9 | 56.9 | 66.7 | **63.0** |

**Table A.2:** The results of the **2nd** perception experiment on the proposed aging simulation approach. Matches are labeled with the number '1' and mismatches with '0'. The last column contains the average result of all eight participants ($p_i$) and the last row shows the percentage of the matches.

| | | | | Experiment 3 | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| photo id | $p_1$ | $p_2$ | $p_3$ | $p_4$ | $p_5$ | $p_6$ | $p_7$ | $p_8$ | average |
| 001 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 2/8 |
| 002 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 5/8 |
| 003 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 6/8 |
| 004 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 5/8 |
| 005 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 006 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 4/8 |
| 007 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 008 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 5/8 |
| 009 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 7/8 |
| 010 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 2/8 |
| 011 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 012 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 013 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 014 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 6/8 |
| 015 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 016 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 017 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 018 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 7/8 |
| 019 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 7/8 |
| 020 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 4/8 |
| 021 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 5/8 |
| 022 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 4/8 |
| 023 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1/8 |
| 024 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 025 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 5/8 |
| 026 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 5/8 |
| 028 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 4/8 |
| 029 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 6/8 |
| 031 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 033 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 5/8 |
| 034 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0/8 |
| 035 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 6/8 |
| 036 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 5/8 |
| 037 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 7/8 |
| 038 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 6/8 |
| 039 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 6/8 |
| 040 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 041 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 2/8 |
| 042 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 5/8 |
| 043 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 3/8 |
| 044 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 6/8 |
| 045 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 3/8 |
| 046 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 5/8 |
| 047 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 3/8 |
| 049 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 7/8 |
| 050 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 6/8 |
| 051 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 052 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 7/8 |
| 053 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 7/8 |
| 054 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 8/8 |
| 055 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 7/8 |
| SUM | 36 | 45 | 41 | 37 | 30 | 25 | 35 | 41 | 36.3 |
| % | 70.6 | 88.2 | 80.4 | 72.6 | 58.8 | 49.0 | 68.6 | 80.4 | **71.1** |

**Table A.3:** The results of the **3rd** perception experiment on the proposed aging simulation approach. Matches are labeled with with the number '1' and mismatches with '0'. The last column contains the average result of all eight participants ($p_i$) and the last row shows the percentage of the matches.

## A.2 Additional Tables: Details on Merging Segments

In Section 6.7.2 details regarding ranking and subsequent merging of plausible facial segments have been presented and based on Table 6.4 one specific example has been analyzed. In the following the additional tables which have also been mentioned in that context are attached.

| Unconstrained photo collection: Figure 5.8 (1st row) | | | | | |
|---|---|---|---|---|---|
| rank | id_eyes | id_mouth | id_nose | id_remainder | id_all |
| 1 | 4 | 4 | 4 | 4 | 4 |
| 2 | 15 | 15 | 15 | 18 | 18 |
| 3 | 18 | 21 | 1 | 21 | 21 |
| 4 | 20 | 5 | 5 | 23 | 23 |
| 5 | 21 | 6 | 20 | 6 | 6 |
| 6 | 23 | 20 | 6 | 9 | 9 |
| 7 | 6 | 1 | 18 | 16 | 16 |
| 8 | 5 | 18 | 23 | 15 | 15 |
| 9 | 2 | 14 | 21 | 5 | 5 |
| 10 | 1 | 23 | 14 | 20 | 20 |
| 11 | 24 | 2 | 22 | 14 | 1 |
| 12 | 14 | 24 | 24 | 1 | 14 |
| 13 | 16 | 22 | 2 | 8 | 8 |
| 14 | 22 | 16 | 12 | 22 | 22 |
| 15 | 9 | 9 | 16 | 2 | 2 |
| 16 | 12 | 12 | 9 | 24 | 12 |
| 17 | 8 | 8 | 3 | 12 | 24 |
| 18 | 3 | 3 | 8 | 25 | 25 |
| 19 | 25 | 10 | 19 | 7 | 7 |
| 20 | 10 | 19 | 10 | 3 | 3 |
| 21 | 13 | 25 | 25 | 19 | 19 |
| 22 | 19 | 13 | 13 | 10 | 13 |
| 23 | 7 | 7 | 7 | 13 | 10 |
| 24 | 11 | 11 | 11 | 11 | 11 |

**Table A.4:** Rankings of image ID's based on segments (Columns 2-5) or the entire face (Column 6) for the result in Figure 5.8, 1st row. Images with numbers in *black* color have been used, those with *red* numbers have been discarded.

| Photo collection: Figure 6.10 (top) | | | | | |
|---|---|---|---|---|---|
| rank | id_eyes | id_mouth | id_nose | id_remainder | id_all |
| 1 | 244 | 244 | 244 | 244 | 244 |
| 2 | 234 | 234 | 234 | 234 | 234 |
| 3 | 246 | 246 | 247 | 231 | 231 |
| 4 | 233 | 247 | 246 | 233 | 233 |
| 5 | 245 | 233 | 233 | 245 | 245 |
| 6 | 231 | 245 | 245 | 246 | 246 |
| 7 | 247 | 231 | 231 | 247 | 247 |
| 8 | 232 | 232 | 232 | 232 | 232 |

**Table A.5:** Rankings of image ID's based on segments (Columns 2-5) or the entire face (Column 6) for the result at the top of Figure 6.10. Images with numbers in *black* color have been used, those with *red* numbers have been discarded. Accordingly, in this example no images have been discarded.

| Photo collection: Figure 6.10 (bottom) | | | | | |
|---|---|---|---|---|---|
| rank | id_eyes | id_mouth | id_nose | id_remainder | id_all |
| 1 | 380 | 369 | 369 | 380 | 380 |
| 2 | 369 | 365 | 365 | 365 | 369 |
| 3 | 365 | 380 | 368 | 369 | 365 |
| 4 | 368 | 368 | 380 | 379 | 379 |
| 5 | 379 | 379 | 366 | 363 | 363 |
| 6 | 366 | 378 | 379 | 368 | 368 |
| 7 | 378 | 366 | 363 | 366 | 366 |
| 8 | 363 | 363 | 378 | 367 | 367 |
| 9 | 367 | 367 | 367 | 377 | 377 |
| 10 | 377 | 377 | 377 | 378 | 378 |
| 11 | 364 | 364 | 364 | 364 | 364 |

**Table A.6:** Rankings of image ID's based on segments (Columns 2-5) or the entire face (Column 6) for the result at the bottom of Figure 6.10. Images with numbers in *black* color have been used, those with *red* numbers have been discarded.

| Photo collection: Figure 6.11 (bottom) | | | | | |
|---|---|---|---|---|---|
| rank | id_eyes | id_mouth | id_nose | id_remainder | id_all |
| 1 | 319 | 319 | 319 | 334 | 334 |
| 2 | 314 | 312 | 325 | 319 | 319 |
| 3 | 312 | 328 | 314 | 325 | 325 |
| 4 | 331 | 314 | 315 | 315 | 315 |
| 5 | 328 | 310 | 312 | 331 | 314 |
| 6 | 333 | 327 | 327 | 310 | 331 |
| 7 | 334 | 331 | 328 | 314 | 310 |
| 8 | 310 | 325 | 331 | 333 | 333 |
| 9 | 315 | 315 | 310 | 312 | 312 |
| 10 | 327 | 334 | 333 | 327 | 320 |
| 11 | 325 | 333 | 334 | 320 | 327 |
| 12 | 313 | 311 | 311 | 311 | 311 |
| 13 | 311 | 313 | 313 | 313 | 313 |
| 14 | 320 | 320 | 320 | 328 | 328 |
| 15 | 323 | 323 | 323 | 323 | 323 |

**Table A.7:** Rankings of image ID's based on segments (Columns 2-5) or the entire face (Column 6) for the result at the bottom of Figure 6.11. Images with numbers in *black* color have been used, those with *red* numbers have been discarded.

# List of Figures

# List of Tables

# Bibliography

[AAB$^+$84] Edward H. Adelson, Charles H. Anderson, James R. Bergen, Peter J. Burt, and Joan M. Ogden. Pyramid Methods in Image Processing. *RCA engineer*, 29(6):33–41, 1984.

[ADW04] Frederik Anrys, Philip Dutré, and Y. D. Willems. Image Based Lighting Design. In *The 4th IASTED International Conference on Visualization, Imaging, and Image Processing*, volume 2, 2004.

[AFB$^+$13] Oleg Alexander, Graham Fyffe, Jay Busch, Xueming Yu, Ryosuke Ichikari, Andrew Jones, Paul Debevec, Jorge Jimenez, Etienne Danvoye, Bernardo Antionazzi, Mike Eheler, Zybnek Kysela, and von der Pahlen, Javier. Digital Ira: Creating a Real-Time Photoreal Digital Actor. *ACM SIGGRAPH Posters*, pages 1:1–1:1, 2013.

[AP08] Xiaobo An and Fabio Pellacini. AppProp: All-Pairs Appearance-Space Edit Propagation. *ACM Transactions on Graphics (TOG)*, 27(3):40, 2008.

[ARL$^+$09] Oleg Alexander, Mike Rogers, William Lambeth, Matt Chiang, and Paul Debevec. The Digital Emily Project: Photoreal Facial Modeling and Animation. *ACM SIGGRAPH Courses*, pages 12:1–12:15, 2009.

[AS10] Oswald Aldrian and William Smith. A Linear Approach to Face Shape and Texture Recovery using a 3D Morphable Model. In *British Machine Vision Conference (BMVC)*, pages 75.1–75.10. BMVA Press, 2010.

[AS13] Oswald Aldrian and William A. P. Smith. Inverse Rendering of Faces with a 3D Morphable Model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(5):1080–1093, 2013.

[AZCP13] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. Robust Discriminative Response Map Fitting with Constrained Local Models. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3444–3451, 2013.

[AZCP14] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. Incremental Face Alignment in the Wild. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[BAHH92] James R. Bergen, P. Anandan, Keith J. Hanna, and Rajesh Hingorani. Hierarchical Model-Based Motion Estimation. In *Proceedings of the Second European Conference on Computer Vision*, ECCV '92, pages 237–252, London, UK, UK, 1992. Springer-Verlag.

[BAP$^+$17] James Booth, Epameinondas Antonakos, Stylianos Ploumpis, George Trigeorgis, Yannis Panagakis, and Stefanos Zafeiriou. 3D Face Morphable Models "In-The-Wild". In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[BAPD13] Xavier P. Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust Face Landmark Estimation under Occlusion. In *2013 IEEE International Conference on Computer Vision*, pages 1513–1520, 2013.

[BB10] Pia Breuer and Volker Blanz. Self-Adapting Feature Layers. *European Conference on Computer Vision (ECCV)*, pages 299–312, 2010.

[BBB$^+$10] Thabo Beeler, Bernd Bickel, Paul Beardsley, Bob Sumner, and Markus Gross. High-Quality Single-Shot Capture of Facial Geometry. *ACM Transactions on Graphics*, 29(3):40:1–40:9, 2010.

[BBPV03] Volker Blanz, Curzio Basso, Tomaso Poggio, and Thomas Vetter. Reanimating Faces in Images and Video. *Computer Graphics Forum (EUROGRAPHICS)*, 22(3):641–650, 2003.

[BBZ96] Benedicte Bascle, Andrew Blake, and Andrew Zisserman. Motion Deblurring and Super-resolution from an Image Sequence. *European Conference on Computer Vision (ECCV)*, pages 573–582, 1996.

[BGPV05] Volker Blanz, Patrick Grother, P. Jonathon Phillips, and Thomas Vetter. Face Recognition Based on Frontal Views Generated from Non-Frontal Images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 454–461 vol. 2, 2005.

[BHB$^+$11] Thabo Beeler, Fabian Hahn, Derek Bradley, Bernd Bickel, Paul Beardsley, Craig Gotsman, Robert W. Sumner, and Markus Gross. High-Quality Passive Facial Performance Capture using Anchor Frames. *ACM Transactions on Graphics*, 30(4):75:1–75:10, 2011.

[BHPS10] Derek Bradley, Wolfgang Heidrich, Tiberiu Popa, and Alla Sheffer. High Resolution Passive Facial Performance Capture. *ACM Transactions on Graphics*, 29(4):41:1–41:10, 2010.

[BJK07] Ronen Basri, David Jacobs, and Ira Kemelmacher. Photometric Stereo with General, Unknown Lighting. *Int. J. Comput. Vision*, 72(3):239–257, 2007.

[BK98] Peter N. Belhumeur and David J. Kriegman. What Is the Set of Images of an Object Under All Possible Illumination Conditions? *International Journal of Computer Vision*, 28(3):245–260, 1998.

[BK00] Simon Baker and Takeo Kanade. Hallucinating Faces. *International Conference on Automatic Face and Gesture Recognition (FG)*, pages 83–88, 2000.

[BKK+08] Pia Breuer, Kwang-In Kim, Wolf Kienzle, Bernhard Schölkopf, and Volker Blanz. Automatic 3D Face Reconstruction from Single Images or Video. *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2008.

[Bla00] Volker Blanz. *Automatische Rekonstruktion der dreidimensionalen Form von Gesichtern aus einem Einzelbild.* PhD thesis, Universität Tübingen, 2000.

[Bla10] Volker Blanz. Simulated Aging of a Human, 2010. Retrieved from http://www.grk1564.uni-siegen.de/en/c1-face-recognition-2d3d-sensor-data (Accessed on 2018-03-28).

[BP95] D. Michael Burt and David I. Perrett. Perception of Age in Adult Caucasian Male Faces: Computer Graphic Manipulation of Shape and Colour Information. *Proceedings of the Royal Society of London B: Biological Sciences*, 259(1355):137–143, 1995.

[Bre10] Pia Breuer. *Automatic Model-based Face Reconstruction and Recognition.* Dissertation, University of Siegen, Germany, 2010.

[Bro14] Beckie Jane Brown. She Takes a Photo: 6.5 Years, 2014. Retrieved from https://www.youtube.com/watch?v=eRvk5UQY1Js (Accessed on 2016-09-03).

[BS81] Nancy Burson and Thomas D. Schneider. Method and Apparatus for Producing an Image of a Person's Face at a Different Age, 1981.

[BSBW16] Anil Bas, William A. P. Smith, Timo Bolkart, and Stefanie Wuhrer. Fitting a 3D Morphable Model to Edges: A Comparison Between Hard and Soft Correspondences. *CoRR*, abs/1602.01125, 2016.

[BSS07] Volker Blanz, Kristina Scherbaum, and Hans-Peter Seidel. Fitting a Morphable Model to 3D Scans of Faces. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, 2007.

[BSVS04] Volker Blanz, Kristina Scherbaum, Thomas Vetter, and Hans-Peter Seidel. Exchanging Faces in Images. In *Computer Graphics Forum*, volume 23, pages 669–676, 2004.

[BV99] Volker Blanz and Thomas Vetter. A Morphable Model for the Synthesis of 3D Faces. *Proceedings of SIGGRAPH*, pages 187–194, 1999.

[BV03] Volker Blanz and Thomas Vetter. Face Recognition Based on Fitting a 3D Morphable Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 25(9):1063–1074, 2003.

[Cat10] Jai Catalano. Father Takes a Photo Every Day for 1 Year, 2010. Retrieved from `https://www.youtube.com/watch?v=rvyxUfY-EUk` (Accessed on 2016-09-03).

[Cat12] Jai Catalano. Father Takes a Photo of his Daughter Every Day from Birth to 1st Year, 2012. Retrieved from `https://www.youtube.com/watch?v=jfCnYHLvhKI` (Accessed on 2016-09-03).

[CBZB15] Chen Cao, Derek Bradley, Kun Zhou, and Thabo Beeler. Real-Time High-fidelity Facial Performance Capture. *ACM Trans. Graph.*, 34(4):46:1–46:9, 2015.

[CCC$^+$18] Xuan Cao, Zhang Chen, Anpei Chen, Xin Chen, Shiying Li, and Jingyi Yu. Sparse Photometric 3D Face Reconstruction Guided by Morphable Models. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[CCH07] Biswarup Choudhury, Sharat Chandran, and Jens Herder. A Survey of Image-Based Relighting Techniques. *Journal of Virtual Reality and Broadcasting*, 4(7), 2007.

[CCH10] Kuang-Yu Chang, Chu-Song Chen, and Yi-Ping Hung. A Ranking Approach for Human Ages Estimation Based on Face Images. In *2010 20th International Conference on Pattern Recognition*, pages 3396–3399, 2010.

[CCH15] Bor-Chun Chen, Chu-Song Chen, and Winston H. Hsu. Face Recognition and Retrieval Using Cross-Age Reference Coding With Cross-Age Celebrity Dataset. *IEEE Transactions on Multimedia*, 17(6):804–815, 2015.

[CCNY11] Chris Columbus, Alfonso Cuaron, Mike Newell, and David Yates. Harry Potter (film series), Warner Bros. Pictures, 2001–2011.

[CET01] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active Appearance Models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(6):681–685, 2001.

[Cor16] Hugo Cornellier. Hugo Takes a Selfie, 2016. Retrieved from `https://www.youtube.com/watch?v=AfcDuysiej0` (Accessed on 2016-09-03).

[CSF99] António Cardoso Costa, António Augusto Sousa, and Fernando Nunes Ferreira. Lighting Design: A Goal Based Approach Using Optimisation. In *Rendering Techniques' 99*, pages 317–328. Springer, 1999.

[CWWS12] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by Explicit Shape Regression. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2887–2894, 2012.

[CWZ+14] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Face-Warehouse: A 3D Facial Expression Database for Visual Computing. *IEEE Transactions on Visualization and Computer Graphics*, 20(3):413–425, 2014.

[CYX04] Hong Chang, Dit-Yan Yeung, and Yimin Xiong. Super-Resolution Through Neighbor Embedding. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 1, page I, 2004.

[DB04] Roman Dovgard and Ronen Basri. Statistical Symmetric Shape from Shading for 3D Structure Recovery of Faces. In Tomás Pajdla and Matas Jiří, editors, *Computer Vision - ECCV 2004: 8th European Conference on Computer Vision, Prague, Czech Republic, May 11-14, 2004. Proceedings, Part II*, pages 99–113. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[DBK06] Göksel Dedeoglu, Simon Baker, and Takeo Kanade. Resolution-Aware Fitting of Active Appearance Models to Low Resolution Images. *European Conference on Computer Vision (ECCV)*, pages 83–97, 2006.

[DGFv12] Matthias Dantone, Jürgen Gall, Gabriele Fanelli, and Luc van Gool. Real-Time Facial Feature Detection using Conditional Regression Forests. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2578–2585, 2012.

[DHS00] Richard O. Duda, Peter E. Hart, and David G. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.

[DHT$^+$00] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the Reflectance Field of a Human Face. *Proc. 27th Annu. Conf. Comput. Graph. Interact. Tech. - SIGGRAPH '00*, pages 145–156, 2000.

[DKA04] Göksel Dedeoglu, Takeo Kanade, and Jonas August. High-Zoom Video Hallucination by Exploiting Spatio-Temporal Regularities. *International Conference on Pattern Recognition (ICPR)*, 2:151–158, 2004.

[DKN$^+$95] Yoshinori Dobashi, Kazufumi Kaneda, Hideki Nakatani, Hideo Yamashita, and Tomoyuki Nishita. A Quick Rendering Method using Basis Functions for Interactive Lighting Design. In *Computer Graphics Forum*, volume 14, pages 229–240, 1995.

[DSG12] Hamdi Dibeklioglu, Albert Ali Salah, and Theo Gevers. A Statistical Method for 2D Facial Landmarking. *IEEE Transactions on Image Processing*, 21(2):844–858, 2012.

[DWSK14] Pengfei Dou, Yuhang Wu, Shishir K. Shah, and Ioannis A. Kakadiaris. Robust 3D Face Shape Reconstruction from Single Images via Two-Fold Coupled Structure Learning and Off-the-Shelf Landmark Detectors. In *British Machine Vision Conference (BMVC)*, pages 1–5. BMVA Press, 2014.

[DWT$^+$02] Paul Debevec, Andreas Wenger, Chris Tchou, Andrew Gardner, Jamie Waese, and Tim Hawkins. *A Lighting Reproduction Approach to Live-Action Compositing*, volume 21. ACM, 2002.

[ECT98] Gareth J. Edwards, Timothy F. Cootes, and Christopher J. Taylor. Face Recognition using Active Appearance Models. *European Conference on Computer Vision (ECCV)*, pages 581–595, 1998.

[FBS05] Martin Fuchs, Volker Blanz, and Hans-Peter Seidel. Bayesian Relighting. In *Proceedings of the Sixteenth Eurographics Conference on Rendering Techniques*, EGSR'05, pages 157–164, Aire-la-Ville, Switzerland, Switzerland, 2005. Eurographics Association.

[FGH10] Yun Fu, Guodong Guo, and Thomas S. Huang. Age Synthesis and Estimation via Faces: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(11):1955–1976, 2010.

[FJA$^+$14] Graham Fyffe, Andrew Jones, Oleg Alexander, Ryosuke Ichikari, and Paul Debevec. Driving High-Resolution Facial Scans with Video Performance Capture. *ACM Trans. Graph.*, 34(1):8:1–8:14, 2014.

[FSHR06] Rob Fergus, Barun Singh, Aaron Hertzmann, and Sam T. Roweis. Removing Camera Shake from a Single Photograph. *ACM Transactions on Graphics*, 25:787–794, 2006.

[FZ16] Haoqiang Fan and Erjin Zhou. Approaching Human Level Facial Landmark Localization by Deep Learning. *Image Vision Comput*, 47(C):27–35, 2016.

[GF15] Golnaz Ghiasi and Charless C. Fowlkes. Occlusion Coherence: Detecting and Localizing Occluded Faces. *CoRR*, abs/1506.08347, 2015.

[GFT$^+$11] Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. Multiview Face Capture using Polarized Spherical Gradient Illumination. *ACM Transactions on Graphics*, 30(6):129:1–129:10, 2011.

[GH00] Reid Gershbein and Pat Hanrahan. A Fast Relighting Engine for Interactive Cinematic Lighting Design. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 353–358, 2000.

[GK08] Leon Gu and Takeo Kanade. A Generative Shape Regularization Model for Robust Face Alignment. In *Proceedings of the 10th European Conference on Computer Vision: Part I*, ECCV '08, pages 413–426, Berlin, Heidelberg, 2008. Springer-Verlag.

[GM11] Guodong Guo and Guowang Mu. Simultaneous Dimensionality Reduction and Human Age Estimation via Kernel Partial Least Squares Regression. In *CVPR 2011*, pages 657–664, 2011.

[GMC$^+$10] Ralph Gross, Iain Matthews, Jeffrey F. Cohn, Takeo Kanade, and Simon Baker. Multi-PIE. *Image and Vision Computing*, 28(5):807–813, 2010.

[GSC$^+$07] Michael Goesele, Noah Snavely, Brian Curless, Hugues Hoppe, and Steven M. Seitz. Multi-View Stereo for Community Photo Collections. *IEEE International Conference on Computer Vision (ICCV)*, pages 1–8, 2007.

[GVWT13] Pablo Garrido, Levi Valgaert, Chenglei Wu, and Christian Theobalt. Reconstructing Detailed Dynamic Face Geometry from Monocular Video. *ACM Trans. Graph.*, 32(6):158:1–158:10, 2013.

[GW12] Guodong Guo and Xiaolong Wang. A Study on Human Age Estimation under Facial Expression Changes. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2547–2553, 2012.

[GZC$^+$16] Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Pérez, and Christian Theobalt. Reconstruction of Personalized 3D Face Rigs from Monocular Video. *ACM Trans. Graph.*, 35(3):28:1–28:15, 2016.

[GZZ$^+$06] Xin Geng, Zhi-Hua Zhou, Yu Zhang, Gang Li, and Honghua Dai. Learning from Facial Aging Patterns for Automatic Age Estimation. In *Proceedings of the 14th ACM International Conference on Multimedia*, MM '06, pages 307–316, New York, NY, USA, 2006. ACM.

[Has13] Tal Hassner. Viewing Real-World Faces in 3D. *IEEE International Conference on Computer Vision (ICCV)*, pages 3607–3614, 2013.

[HASS10] Mark F. Hansen, Gary A. Atkinson, Lyndon N. Smith, and Melvyn L. Smith. 3D Face Reconstructions from Photometric Stereo using Near Infrared and Visible Light. *Computer Vision and Image Understanding*, 114(8):942–951, 2010.

[HBHP03] Tim J. Hutton, Bernard F. Buxton, Peter Hammond, and Henry W. W. Potts. Estimating Average Growth Trajectories in Shape-Space using Kernel Smoothing. *IEEE Transactions on Medical Imaging*, 22(6):747–753, 2003.

[HCH15] Gee-Sern Hsu, Kai-Hsiang Chang, and Shih-Chieh Huang. Regressive Tree Structured Model for Facial Landmark Localization. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3855–3861, 2015.

[HO15] Bernardo Henz and Manuel M. Oliveira. Artistic Relighting of Paintings and Drawings. *The Visual Computer*, pages 1–14, 2015.

[Hof12a] Frans Hofmeester. Lotte Time Lapse: Birth to 12 Years in 2 min. 45, 2012. Retrieved from https://vimeo.com/40448182 (Accessed on 2016-09-03).

[Hof12b] Frans Hofmeester. Vince Time Lapse: Birth to 9 Years in 2 min., 2012. Retrieved from https://vimeo.com/40613192 (Accessed on 2016-09-03).

[Hof15] Frans Hofmeester. Portrait of Lotte, 0 to 16 Years in 4 ½ Minutes, 2015. Retrieved from https://www.youtube.com/watch?v=-Plk7TLNmsU (Accessed on 2016-09-03).

[Hof17] Frans Hofmeester. Portrait of Lotte, 0 to 18 Years, 2017. Retrieved from https://www.youtube.com/watch?v=nPxdhnT4Ec8 (Accessed on 2018-07-23).

[HP16] Hewlett-Packard. 3D Structured Light Scanner Pro S3, 2016. Retrieved from http://www8.hp.com/us/en/campaign/3Dscanner/overview.html (Accessed on 2016-11-06).

[HRBLM] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments.

[HS81] Berthold K. P. Horn and Brian G. Schunck. Determining Optical Flow. *Artificial Intelligence*, 17:185–203, 1981.

[HSBL15] Miguel Heredia Conde, Davoud Shahlaei, Volker Blanz, and Otmar Loffeld. Efficient and Robust Inverse Lighting of a Single Face Image using Compressive Sensing. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, 2015.

[HVC08] Carlos Hernández, George Vogiatzis, and Roberto Cipolla. Multiview Photometric Stereo. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(3):548–554, 2008.

[HvM$^+$13] John F. Hughes, Andries van Dam, Morgan McGuire, David F. Sklar, James D. Foley, Steven K. Feiner, and Kurt Akeley. *Computer Graphics: Principles and Practice (3rd ed.).* Addison-Wesley Professional, Boston, MA, USA, 2013.

[HY12] Zhe Hu and Ming-Hsuan Yang. Good Regions to Deblur. *European Conference on Computer Vision (ECCV)*, pages 59–72, 2012.

[HZ04] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision.* Cambridge University Press, 2nd edition, 2004.

[IBP15] Alexandru Eugen Ichim, Sofien Bouaziz, and Mark Pauly. Dynamic 3D Avatar Creation from Hand-held Video Input. *ACM Trans. Graph.*, 34(4):45:1–45:14, 2015.

[Itz16] Dave Itzkoff. How 'Rogue One' Brought Back Familiar Faces, 2016. Retrieved from `http://www.nytimes.com/2016/12/27/movies/how-rogue-one-brought-back-grand-moff-tarkin.html` (Accessed on 2017-01-04).

[Jäh97] Bernd Jähne. *Digitale Bildverarbeitung.* Springer Berlin Heidelberg, 4. edition, 1997.

[JB02] Henrik Wann Jensen and Juan Buhler. A Rapid Hierarchical Rendering Technique for Translucent Materials. *ACM Trans. Graph.*, 21(3):576–581, 2002.

[JCK14] László A. Jeni, Jeffrey F. Cohn, and Takeo Kanade. Dense 3D Face Alignment from 2D Video for Analysis and Synthesis. *European Conference on Computer Vision (ECCV)*, 2014.

[JCK15] László A. Jeni, Jeffrey F. Cohn, and Takeo Kanade. Dense 3D Face Alignment from 2D Videos in Real-Time. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2015.

[JL15] Amin Jourabloo and Xiaoming Liu. Pose-Invariant 3D Face Alignment. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3694–3702, 2015.

[JL16] Amin Jourabloo and Xiaoming Liu. Large-Pose Face Alignment via CNN-Based Dense 3D Model Fitting. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[JTY$^+$16] László A. Jeni, Sergey Tulyakov, Lijun Yin, Nicu Sebe, and Jeffrey F. Cohn. The First 3D Face Alignment in the Wild (3DFAW) Challenge. In Gang Hua and Jégou Hervé, editors, *Computer Vision – ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II*, pages 511–520. Springer International Publishing, Cham, 2016.

[JWvB11] Rob Jenkins, David White, Xandra van Montfort, and A. Mike Burton. Variability in Photos of the same Face. *Cognition*, 121(3):313–323, 2011.

[KC16] Wei-Jen Ko and Shao-Yi Chien. Patch-based face hallucination with Multitask Deep Neural Network. In *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2016.

[KC18] Amit Kumar and Rama Chellappa. Disentangling 3D Pose in a Dendritic CNN for Unconstrained 2D Face Alignment. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[KD11] L. L. Gayani Kumari and Anuja Dharmaratne. Age Progression for Elderly People using Image Morphing. In *2011 International Conference on Advances in ICT for Emerging Regions (ICTer)*, pages 33–38, 2011.

[KGT$^+$18] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Nießner, Patrick Pérez, Christian Richardt, Michael Zollöfer, and Christian Theobalt. Deep Video Portraits. *ACM Transactions on Graphics (TOG)*, 37(4):163, 2018.

[KH96] Deepa Kundur and Dimitrios Hatzinakos. Blind Image Deconvolution. *IEEE Signal Processing Magazine*, 13(3):43–64, 1996.

[KH12] Martin Klaudiny and Adrian Hilton. High-Detail 3D Capture and Non-sequential Alignment of Facial Performance. In *3D Imaging, Modeling, Processing, Visualization and Transmission (3DIMPVT), 2012 Second International Conference on*, pages 17–24, 2012.

[Kon08] Konica Minolta. Non-Contact 3D Digitizer RANGE7, 2008. Retrieved from http://www.konicaminolta.com/instruments/download/instruction_manual/3d/pdf/range7-5_instruction_eng.pdf (Accessed on 2016-11-06).

[KP09] William B. Kerr and Fabio Pellacini. Toward Evaluating Lighting Design Interface Paradigms for Novice Users. *ACM Trans. Graph.*, 28(3):26:1–26:9, 2009.

[KRS+16] M. S. L. Khan, Shafiq Ur Réhman, Ulrik Söderström, Alaa Halawani, and Haibo Li. Face-Off: A Face Reconstruction Technique for Virtual Reality (VR) Scenarios. In Gang Hua and Jégou Hervé, editors, *Computer Vision – ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part I*, pages 490–503. Springer International Publishing, Cham, 2016.

[Kru09] Christian Krug. BMBF-Projekt: Interaktionsgesteuerte Bilddatenanalyse zur Interaktionsgesteuerte Bilddatenanalyse zur Bekämpfung von Kinderpornografie (INBEKI), 2009. Retrieved from http://www.sifo.de/de/inbeki-interaktionsgesteuerte-bilddatenanalyse-zur-bekaempfung-von-kinderpornografie-1949.html (Accessed on 2016-12-09).

[KS13] Ira Kemelmacher-Shlizerman. Internet-based Morphable Model. *IEEE International Conference on Computer Vision (ICCV)*, 2013.

[KS14] Vahid Kazemi and Josephine Sullivan. One Millisecond Face Alignment with an Ensemble of Regression Trees. In *CVPR*, 2014.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc, 2012.

[KSS11] Ira Kemelmacher-Shlizerman and Steven M. Seitz. Face Reconstruction in the Wild. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, pages 1746–1753, Washington, DC, USA, 2011. IEEE Computer Society.

[KSS12] Ira Kemelmacher-Shlizerman and Steven M. Seitz. Collection Flow. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1792–1799, 2012.

[KSSMB16] Ira Kemelmacher-Shlizerman, Steven M. Seitz, Daniel Miller, and Evan Brossard. The MegaFace Benchmark: 1 Million Faces for Recognition at Scale. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[KSSS14] Ira Kemelmacher-Shlizerman, Supasorn Suwajanakorn, and Steven M. Seitz. Illumination-Aware Age Progression. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014.

[LAV09] Marcel Lüthi, Thomas Albrecht, and Thomas Vetter. Probabilistic Modeling and Visualization of the Flexibility in Morphable Models. In Hancock, Edwin R. and Martin, Ralph R. and Sabin, Malcolm A, editor, *Mathematics of Surfaces XIII*, pages 251–264, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[LBA$^+$12] Marcel Lüthi, Remi Blanc, Thomas Albrecht, Tobias Gass, Orcun Goksel, Philippe Büchler, Michael Kistler, Habib Bousleiman, Mauricio Reyes, Philippe C. Cattin, and Thomas Vetter. Statismo - A Framework for PCA based Statistical Models. *The Insight Journal*, 2012.

[LBIC15] Claudia Lindner, Paul A. Bromiley, Mircea C. Ionita, and Tim F. Cootes. Robust and Accurate Shape Model Matching using Random Forest Regression-Voting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1862–1874, 2015.

[LC05] Xiaoming Liu and Tsuhan Chen. Pose-Robust Face Recognition using Geometry Assisted Probabilistic Modeling. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 502–509 vol. 1, 2005.

[LCS$^+$08] Bo Li, Hong Chang, Shiguang Shan, Xilin Chen, and Wen Gao. Hallucinating Facial Images and Features. *International Conference on Pattern Recognition (ICPR)*, pages 1–4, 2008.

[LDC04] Andreas Lanitis, Chrisina Draganova, and Chris Christodoulou. Comparing Different Classifiers for Automatic Age Estimation. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 34(1):621–628, 2004.

[LGLT15] Zhifeng Li, Dihong Gong, Xuelong Li, and Dacheng Tao. Learning Compact Feature Descriptor and Adaptive Matching Framework for Face Recognition. *IEEE Transactions on Image Processing*, 24(9):2736–2745, 2015.

[LGLT16] Zhifeng Li, Dihong Gong, Xuelong Li, and Dacheng Tao. Aging Face Recognition: A Hierarchical Learning Model Based on Local Patterns Selection. *IEEE Transactions on Image Processing*, 25(5):2146–2154, 2016.

[LHK01] Kuang-Chih Lee, Jeffrey Ho, and David Kriegman. Nine Points of Light: Acquiring Subspaces for Face Recognition under Variable Lighting. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–519–I–526 vol.1, 2001.

[Liu07] Xiaoming Liu. Generic Face Alignment using Boosted Appearance Model. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8, 2007.

[LK81] Bruce D. Lucas and Takeo Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'81, pages 674–679, San Francisco, CA, USA, 1981. Morgan Kaufmann Publishers Inc.

[LL04] Yang Li and Xueyin Lin. Face Hallucination with Pose Variation. *International Conference on Automatic Face and Gesture Recognition (FG)*, pages 723–728, 2004.

[LLP$^+$12] Youn J. Lee, Sung J. Lee, Kang R. Park, Jaeik Jo, and Jaihie Kim. Single View-Based 3D Face Reconstruction Robust to Self-Occlusion. *EURASIP Journal on Advances in Signal Processing*, 2012(1):176+, 2012.

[LLZC12] Yan Liang, Jian-Huang Lai, Wei-Shi Zheng, and Zemin Cai. A Survey of Face Hallucination. *Chinese Conference on Biometric Recognition (CCBR)*, 7701:83–93, 2012.

[LPJ11] Zhifeng Li, Unsang Park, and Anil K. Jain. A Discriminative Model for Age Invariant Face Recognition. *IEEE Transactions on Information Forensics and Security*, 6(3):1028–1037, 2011.

[LSF07] Ce Liu, Heung-Yeung Shum, and William T. Freeman. Face Hallucination: Theory and Practice. *International Journal of Computer Vision (IJCV)*, 75(1):115–134, 2007.

[LSKS16] Shu Liang, Linda G. Shapiro, and Ira Kemelmacher-Shlizerman. Head Reconstruction from Internet Photos. In *European Conference on Computer Vision*, pages 360–374, 2016.

[LSRJ10] Haibin Ling, Stefano Soatto, Narayanan Ramanathan, and David W. Jacobs. Face Verification Across Age Progression using Discriminative Methods. *IEEE Transactions on Information Forensics and Security*, 5(1):82–91, 2010.

[LSZ01] Ce Liu, Heung-Yeung Shum, and Chang-Shui Zhang. A Two-Step Approach to Hallucinating Faces: Global Parametric Model and Local Monparametric Model. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1:192–198, 2001.

[LTC02] Andreas Lanitis, Chris J. Taylor, and Timothy F. Cootes. Toward Automatic Simulation of Aging Effects on Face Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(4):442–455, 2002.

[LTS$^+$15] Andreas Lanitis, Nicolas Tsapatsoulis, Kleanthis Soteriou, Daiki Kuwahara, and Shigeo Morishima. FG2015 Age Progression Evaluation. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pages 1–6, 2015.

[LZL14] Chen Li, Kun Zhou, and Stephen Lin. Intrinsic Face Image Decomposition with Human Face Priors. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, volume 8693 of *Lecture Notes in Computer Science*, pages 218–233. Springer, 2014.

[LZZL16] Feng Liu, Dan Zeng, Qijun Zhao, and Xiaoming Liu. Joint Face Alignment and 3D Face Reconstruction. In *Proc. European Conference on Computer Vision*, Amsterdam, The Netherlands, 2016.

[MDWE04] Pina Marziliano, Frederic Dufaux, Stefan Winkler, and Touradj Ebrahimi. Perceptual Blur and Ringing Metrics: Application to JPEG2000. *Signal Processing: Image Communication*, 19(2):163–172, 2004.

[Mil17] Jeff Miller. Earliest Known Uses of some of the Words of Mathematics: "Simplex", 2017. Retrieved from `http://jeff560.tripod.com/s.html` (Accessed on 2018-09-30).

[ML04] Claus B. Madsen and Rune Laursen. Image Relighting: Getting the Sun to Set in an Image Taken at Noon. In *In 13th Danish Conference on Pattern Recognition and Image Analysis*, pages 13–20, 2004.

[NFS15] Richard A. Newcombe, Dieter Fox, and Steven M. Seitz. DynamicFusion: Reconstruction and Tracking of Non-Rigid Scenes in Real-Time. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[NLGB16] Chi Nhan Duong, Khoa Luu, Kha Gia Quach, and Tien D. Bui. Longitudinal Face Modeling via Temporal Deep Restricted Boltzmann Machines. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[NOA10] Tomo Nakai, Taro Okakura, and Kaoru Arakawa. Face Recognition Across Age Progression using Block Matching Method. In *2010 10th International Symposium on Communications and Information Technologies*, pages 620–625, 2010.

[NSD95] Jeffry S. Nimeroff, Eero Simoncelli, and Julie Dorsey. Efficient Re-Rendering of Naturally Illuminated Environments. In *Photorealistic Rendering Techniques*, pages 373–388. Springer, 1995.

[NZL$^+$16] Zhenxing Niu, Mo Zhou, Le Wang, Xinbo Gao, and Gang Hua. Ordinal Regression With Multiple Output CNN for Age Estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[OHJ12] Charles Otto, Hu Han, and Anil Jain. How Does Aging Affect Facial Components? In *Proceedings of the 12th International Conference on Computer Vision - Volume 2*, ECCV'12, pages 189–198, Berlin, Heidelberg, 2012. Springer-Verlag.

[OMSI07] Makoto Okabe, Yasuyuki Matsushita, Li Shen, and Takeo Igarashi. Illumination Brush: Interactive Design of All-Frequency Lighting. In *Computer Graphics and Applications, 2007. PG '07. 15th Pacific Conference on*, pages 171–180, 2007.

[OZM$^+$06] Makoto Okabe, Gang Zeng, Yasuyuki Matsushita, Takeo Igarashi, Long Quan, and Heung-Yeung Shum. Single-View Relighting with Normal Map Painting. In *Proc. Pacific Graphics*, pages 27–34, 2006.

[PB16] Marcel Piotraschke and Volker Blanz. Automated 3D Face Reconstruction from Multiple Images using Quality Measures. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[PBMF07] Fabio Pellacini, Frank Battaglia, R. Keith Morley, and Adam Finkelstein. Lighting with Paint. *ACM Trans. Graph.*, 26(2), 2007.

[PCF06] Nikos Paragios, Yunmei Chen, and Olivier D. Faugeras, editors. *Handbook of Mathematical Models in Computer Vision*. Springer, 2006.

[PF92] Pierre Poulin and Alain Fournier. Lights from Highlights and Shadows. In *Proceedings of the 1992 symposium on Interactive 3D graphics*, pages 31–38, 1992.

[PF05] Emmanuel Prados and Olivier Faugeras. A Generic and Provably Convergent Shape-from-Shading Method for Orthographic and Pinhole Cameras. *International Journal of Computer Vision*, 65(1):97–125, 2005.

[PFWM16] Xi Peng, Rogerio Feris, Xiaoyu Wang, and Dimitris Metaxas. A Recurrent Encoder-Decoder Network for Sequential Face Alignment. In Bastian Leibe, Matas Jiri, Sebe Nicu, and Welling Max, editors, *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I*, pages 38–56. Springer International Publishing, Cham, 2016.

[PGB03] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson Image Editing. *ACM Transactions on Graphics*, 22(3):313–318, 2003.

[Pho75] Bui Tuong Phong. Illumination for Computer Generated Pictures. *Commun. ACM*, 18(6):311–317, 1975.

[PHS08] Sung Won Park, Jingu Heo, and Marios Savvides. 3D Face Reconstruction from a Single 2D Face Image. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

[PHSC18] Hongyu Pan, Hu Han, Shiguang Shan, and Xilin Chen. Mean-Variance Loss for Deep Age Estimation from a Face. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[PHW08] Gang Pan, Song Han, and Zhaohui Wu. Hallucinating 3D Facial Shapes. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–8, 2008.

[PKA$^+$09] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3D Face Model for Pose and Illumination Invariant Face Recognition. In *Advanced Video and Signal Based Surveillance, 2009. AVSS '09. Sixth IEEE International Conference on*, pages 296–301, 2009.

[PKv99] Marc Pollefeys, Reinhard Koch, and Luc van Gool. Self-Calibration and Metric Reconstruction Inspite of Varying and Unknown Intrinsic Camera Parameters. *International Journal of Computer Vision*, 32(1):7–25, 1999.

[PPTK13] Panagiotis Perakis, Georgios Passalis, Theoharis Theoharis, and Ioannis A. Kakadiaris. 3D Facial Landmark Detection under Large Yaw and Expression Variations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1552–1564, 2013.

[PRJ97] Pierre Poulin, Karim Ratib, and Marco Jacques. Sketching Shadows and Highlights to Position Lights. In *Computer Graphics International, 1997. Proceedings*, pages 56–63, 1997.

[PS09] Ankur Patel and William A. P. Smith. 3D Morphable Face Models Revisited. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1327–1334, 2009.

[PS12] Ankur Patel and William A. P. Smith. Driving 3D Morphable Models using Shading Cues. *Pattern Recogn*, 45(5):1993–2004, 2012.

[PSRB09] Eric Patterson, Amrutha Sethuram, Karl Ricanek, and Frederick Bingham. Improvements in Active Appearance Model Based Synthetic Age Progression for Adult Aging. In *2009 IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems*, pages 1–5, 2009.

[PSS14] Eric Patterson, Devin Simpson, and Arnrutha Sethuram. Establishing a Test Set and Initial Comparisons for Quantitatively Evaluating Synthetic Age Progression for Adult Aging. In *IEEE International Joint Conference on Biometrics*, pages 1–8, 2014.

[PTG02] Fabio Pellacini, Parag Tole, and Donald P. Greenberg. A User Interface for Interactive Cinematic Shadow Design. In *ACM Transactions on Graphics (TOG)*, volume 21, pages 563–566, 2002.

[PTJ10] Unsang Park, Yiying Tong, and Anil K. Jain. Age-Invariant Face Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):947–954, 2010.

[PTMD07] Pieter Peers, Naoki Tamura, Wojciech Matusik, and Paul Debevec. Post-production Facial Performance Relighting Using Reflectance Transfer. *ACM Trans. Graph.*, 26(3), 2007.

[PTVF92] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C (2nd Edition): The Art of Scientific Computing.* Cambridge University Press, New York, NY, USA, 1992.

[PXF16] Weilong Peng, Chao Xu, and Zhiyong Feng. 3D Face Modeling based on Structure Optimization and Surface Reconstruction with B-Spline. *Neurocomputing*, 179:228–237, 2016.

[QMSB15] Chengchao Qu, Eduardo Monari, Tobias Schuchert, and Jürgen Beyerer. Adaptive Contour Fitting for Pose-Invariant 3D Face Shape Reconstruction. In Mark W. Jones Xianghua Xie and Gary K. L. Tam, editors, *Proceedings of the British Machine Vision Conference (BMVC)*, pages 87.1–87.12. BMVA Press, 2015.

[RBS15] Karl Ricanek, Shivani Bhardwaj, and Michael Sodomsky. A Review of Face Recognition against Longitudinal Child Faces. In *BIOSIG*, 2015.

[RC06] Narayanan Ramanathan and Rama Chellappa. Modeling Age Progression in Young Faces. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 1, pages 387–394, 2006.

[RC08] Narayanan Ramanathan and Rama Chellappa. Modeling Shape and Textural Variations in Aging Faces. In *8th IEEE International Conference on Automatic Face Gesture Recognition*, pages 1–8, 2008.

[RCB09] Narayanan Ramanathan, Rama Chellapa, and Soma Biswas. Age Progression in Human Faces: A Survey. *Journal of Visual Languages and Computing*, 15:3349–3361, 2009.

[RDL$^+$15] Peiran Ren, Yue Dong, Stephen Lin, Xin Tong, and Baining Guo. Image Based Relighting using Neural Networks. *ACM Trans. Graph.*, 34(4):111:1, 2015.

[RSK16] Elad Richardson, Matan Sela, and Ron Kimmel. 3D Face Reconstruction by Learning from Synthetic Data. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 460–469, 2016.

[RT06] Karl Ricanek and Tamirat Tesafaye. MORPH: A Longitudinal Image Database of Normal Adult Age-Progression. In *2006 7th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 341–345, 2006.

[RTL15] Joseph Roth, Yiying Tong, and Xiaoming Liu. Unconstrained 3D Face Reconstruction. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[RTL16] Joseph Roth, Yiying Tong, and Xiaoming Liu. Adaptive 3D Face Reconstruction from Unconstrained Photo Collections. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[RTL17] Joseph Roth, Yiying Tong, and Xiaoming Liu. Adaptive 3D Face Reconstruction from Unconstrained Photo Collections. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(11):2127–2141, 2017.

[SB12] Matthaeus Schumacher and Volker Blanz. Which Facial Profile Do Humans Expect After Seeing a Frontal View? A Comparison with a Linear Face Model. *ACM Transactions on Applied Perception*, 9(3):1–16, 2012.

[SB15] Davoud Shahlaei and Volker Blanz. Realistic Inverse Lighting from a Single 2D Image of a Face, Taken under Unknown and Complex Lighting. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pages 1–8, 2015.

[SBOR06] Pawan Sinha, Benjamin Balas, Yuri Ostrovsky, and Richard Russell. Face Recognition by Humans: Nineteen Results All Computer Vision Researchers Should Know About. *Proceedings of the IEEE*, 94(11):1948–1962, 2006.

[Sch02] Pieter H. Schoute. *Mehrdimensionale Geometrie*. Cornell University Library Historical Math Monographs. G.J. Göschen, 1902.

[SCS$^+$12] Jinli Suo, Xilin Chen, Shiguang Shan, Wen Gao, and Qionghai Dai. A Concatenational Graph Evolution Aging Model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11):2083–2096, 2012.

[SDS$^+$93] Chris Schoeneman, Julie Dorsey, Brian Smits, James Arvo, and Donald Greenberg. Painting with Light. In *Proceedings of the 20th annual conference on Computer graphics and interactive techniques*, pages 143–146, 1993.

[SEMFV17] Sandro Schönborn, Bernhard Egger, Andreas Morel-Forster, and Thomas Vetter. Markov Chain Monte Carlo for Automated Face Image Analysis. *International Journal of Computer Vision*, 123(2):160–183, 2017.

[SEV18] Andreas Schneider, Bernhard Egger, and Thomas Vetter. A Parametric Freckle Model for Faces. In *2018 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 431–435, 2018.

[SH06] William A. P. Smith and Edwin R. Hancock. Recovering Facial Shape using a Statistical Model of Surface Normal Direction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):1914–1930, 2006.

[SKSS14] Supasorn Suwajanakorn, Ira Kemelmacher-Shlizerman, and Steven M. Seitz. Total Moving Face Reconstruction. *European Conference on Computer Vision (ECCV)*, pages 796–812, 2014.

[SL01] Ram Shacked and Dani Lischinski. Automatic Lighting Design using a Perceptual Quality Metric. In *Computer Graphics Forum*, volume 20, pages 215–227, 2001.

[SLSL11] Cheng-Ta Shen, Wan-Hua Lu, Sheng-Wen Shih, and Hong-Yuan Mark Liao. Exemplar-based Age Progression Prediction in Children Faces. In *2011 IEEE International Symposium on Multimedia*, pages 123–128, 2011.

[SMP07] Sudipta N. Sinha, Philippos Mordohai, and Marc Pollefeys. Multi-View Stereo via Graph Cuts on the Dual of an Adaptive Tetrahedral Mesh. *IEEE International Conference on Computer Vision (ICCV)*, pages 1–8, 2007.

[Smy14] Glenn Smyth. Avie Aged 0-5 Years Timelapse, 2014. Retrieved from https://www.youtube.com/watch?v=E3O53cN3N7U (Accessed on 2016-09-03).

[SPB15] Matthaeus Schumacher, Marcel Piotraschke, and Volker Blanz. Hallucination of Facial Details from Degraded Images using 3D Face Models. *Image and Vision Computing*, 40:49–64, 2015.

[SPB16] Davoud Shahlaei, Marcel Piotraschke, and Volker Blanz. Lighting Design for Portraits with a Virtual Light Stage. In *IEEE International Conference on Image Processing (ICIP)*, pages 1579–1583, 2016.

[SRH$^+$11] Kristina Scherbaum, Tobias Ritschel, Matthias Hullin, Thorsten Thormählen, Volker Blanz, and Hans-Peter Seidel. Computer-Suggested Facial Makeup. *Computer Graphics Forum (EUROGRAPHICS)*, 30(2):485–492, 2011.

[SRK11] Gowri Somanath, M. V. Rohith, and Chandra Kambhamettu. VADANA: A Dense Dataset for Facial Image Analysis. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 2175–2182, 2011.

[SS96] Richard R. Schultz and Robert L. Stevenson. Extraction of High-Resolution Frames from Video Sequences. *IEEE Transactions on Image Processing*, 5:996–1011, 1996.

[SSG06a] Catherine M. Scandrett, Christopher J. Solomon, and Stuart J. Gibson. A Person-Specific, Rigorous Aging Model of the Human Face. *Pattern Recognition Letters*, 27(15):1776–1787, 2006.

[SSG06b] Catherine M. Scandrett, Christopher J. Solomon, and Stuart J. Gibson. Towards a Semi-Automatic Method for the Statistically Rigorous Ageing of the Human Face. *IEE Proceedings - Vision, Image and Signal Processing*, 153(5):639–649, 2006.

[SSSB07] Kristina Scherbaum, Martin Sunkel, Hans-Peter Seidel, and Volker Blanz. Prediction of Individual Non-Linear Aging Trajectories of Faces. *Computer Graphics Forum (EUROGRAPHICS)*, 26(3):285–294, 2007.

[Ste10] Mark Stead. Child Growth Face Morph Time-Lapse, 2010. Retrieved from `https://www.youtube.com/watch?v=ZTjHLF3xKWo` (Accessed on 2016-09-03).

[STL$^+$15] Xiangbo Shu, Jinhui Tang, Hanjiang Lai, Luoqi Liu, and Shuicheng Yan. Personalized Age Progression with Aging Dictionary. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3970–3978, 2015.

[STZP13] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 Faces in-the-Wild Challenge: The First Facial Landmark Localization Challenge. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 397–403, 2013.

[SWH$^+$16] Shunsuke Saito, Lingyu Wei, Liwen Hu, Koki Nagano, and Hao Li. Photorealistic Facial Texture Inference using Deep Neural Networks. *ArXiv e-prints*, 2016.

[SWT13] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep Convolutional Network Cascade for Facial Point Detection. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3476–3483, 2013.

[SWT15] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deeply Learned Face Representations are Sparse, Selective, and Robust. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2892–2900, 2015.

[SWTC14] Fuhao Shi, Hsiang-Tao Wu, Xin Tong, and Jinxiang Chai. Automatic Acquisition of High-fidelity Facial Performances using Monocular Videos. *ACM Trans. Graph.*, 33(6):222:1–222:13, 2014.

[TB99] Michael E. Tipping and Chris M. Bishop. Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society*, 61(3):611–622, 1999.

[THM$^+$18] Anh Tuan Tran, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gérard Medioni. Extreme 3D Face Reconstruction: Seeing Through Occlusions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[THMM17] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. Regressing Robust and Discriminative 3D Morphable Models with a Very Deep Neural Network. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1493–1502, 2017.

[TL12] Marshall F. Tappen and Ce Liu. A Bayesian Approach to Alignment-based Image Hallucination. *European Conference on Computer Vision (ECCV)*, pages 236–249, 2012.

[TL18] Luan Tran and Xiaoming Liu. Nonlinear 3D Face Morphable Model. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[TLL14] Ming-Han Tsai, Yen-Kai Liao, and I-Chen Lin. Human Face Aging with Guided Prediction and Detail Synthesis. *Multimedia Tools and Applications*, 72(1):801–824, 2014.

[TMSP80] James T. Todd, Leonard S. Mark, Robert E. Shaw, and John B. Pittenger. The Perception of Human Growth. *Scientific American*, 242:132–144, 1980.

[TMT16] Oncel Tuzel, Tim K. Marks, and Salil Tambe. Robust Face Alignment using a Mixture of Invariant Experts. In Bastian Leibe, Matas Jiri, Sebe Nicu, and Welling Max, editors, *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, pages 825–841. Springer International Publishing, Cham, 2016.

[TP14] Georgios Tzimiropoulos and Maya Pantic. Gauss-Newton Deformable Part Models for Face Alignment In-the-Wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1851–1858, 2014.

[TSKZ17] George Trigeorgis, Patrick Snape, Iasonas Kokkinos, and Stefanos Zafeiriou. Face Normals "In-The-Wild" using Fully Convolutional Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[TSN+16] George Trigeorgis, Patrick Snape, Mihalis A. Nicolaou, Epameinondas Antonakos, and Stefanos Zafeiriou. Mnemonic Descent Method: A Recurrent Process Applied for End-To-End Face Alignment. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[TZS+16] Justus Thies, Michael Zollhöfer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2Face: Real-time Face Capture and Reenactment of RGB Videos. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[VJ04] Paul Viola and Michael J. Jones. Robust Real-Time Face Detection. *Int. J. Comput. Vision*, 57(2):137–154, 2004.

[VMBP10] Michel Valstar, Brais Martinez, Xavier Binefa, and Maja Pantic. Facial Point Detection using Boosted Regression and Graph Models. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 2729–2736, 2010.

[WAB$^+$11] Lun Wu, Ganesh Arvind, Shi Boxin, Matsushita Yasuyuki, Wang Yongtian, and Ma Yi. Robust Photometric Stereo via Low-Rank Matrix Completion and Recovery. In Ron Kimmel, Klette Reinhard, and Sugimoto Akihiro, editors, *Computer Vision – ACCV 2010: 10th Asian Conference on Computer Vision, Queenstown, New Zealand, November 8-12, 2010, Revised Selected Papers, Part III*, pages 703–717. Springer Berlin Heidelberg, Berlin, Heidelberg, 2011.

[WCY$^+$16] Wei Wang, Zhen Cui, Yan Yan, Jiashi Feng, Shuicheng Yan, Xiangbo Shu, and Nicu Sebe. Recurrent Face Aging. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[WDC$^+$08] Oliver Wang, James Davis, Erika Chuang, Ian Rickard, Krystle de Mesa, and Chirag Dave. Video Relighting using Infrared Illumination. In *Computer Graphics Forum*, volume 27, pages 271–280, 2008.

[Wil17] Jordan Wilson. Photo a Day for 10 Years, 2017. Retrieved from https://www.youtube.com/watch?v=zuRd_Eneuk8 (Accessed on 2018-02-04).

[WJ15] Yue Wu and Qiang Ji. Robust Facial Landmark Detection under Significant Head Poses and Occlusion. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3658–3666, 2015.

[WJ16] Yue Wu and Qiang Ji. Constrained Joint Cascade Regression Framework for Simultaneous Facial Action Unit Recognition and Facial Landmark Detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[WLQ16] Yandong Wen, Zhifeng Li, and Yu Qiao. Latent Factor Guided Convolutional Neural Networks for Age-Invariant Face Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[WMP$^+$06] Tim Weyrich, Wojciech Matusik, Hanspeter Pfister, Bernd Bickel, Craig Donner, Chien Tu, Janet McAndless, Jinho Lee, Addy Ngan, Henrik Wann Jensen, and Markus Gross. Analysis of Human Faces Using a Measurement-based Skin Reflectance Model. *ACM Trans. Graph.*, 25(3):1013–1024, 2006.

[Woo89] Robert J. Woodham. Photometric Method for Determining Surface Orientation from Multiple Images: Shape from Shading. In Berthold K. P. Horn and

Michael J. Brooks, editors, *Shape from shading*, pages 513–531. MIT Press, Cambridge, MA, USA, 1989.

[WPLN99] Yin Wu, Kalra Prem, Moccozet Laurent, and Magnenat-Thalmann Nadia. Simulating Wrinkles and Skin Aging. *The Visual Computer*, 15(4):183–198, 1999.

[WTLG18] Zongwei Wang, Xu Tang, Weixin Luo, and Shenghua Gao. Face Aging with Identity-Preserved Conditional Generative Adversarial Networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[WZS05] Sen Wang, Lei Zhang, and Dimitris Samaras. Face Reconstruction Across Different Poses and Arbitrary Illumination Conditions. *Lecture Notes in Computer Science*, pages 91–101, 2005.

[XD13] Xuehan Xiong and De la Torre, Fernando. Supervised Descent Method and Its Applications to Face Alignment. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 532–539, 2013.

[XD15] Xuehan Xiong and De la Torre, Fernando. Global Supervised Descent Method. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2664–2673, 2015.

[YCS$^+$08] Lijun Yin, Xiaochen Chen, Yi Sun, T. Worm, and M. Reale. A High-Resolution 3D Dynamic Facial Expression Database. In *Automatic Face & Gesture Recognition, 2008. FG '08. 8th IEEE International Conference on*, pages 1–6. IEEE, 2008.

[YCSS14] Chang Yang, Jiansheng Chen, Nan Su, and Guangda Su. Improving 3D Face Details based on Normal Map of Hetero-source Images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2014.

[YDWJ18] Hongyu Yang, Di Huang, Yunhong Wang, and Anil K. Jain. Learning Face Age Progression: A Pyramid Architecture of GANs. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[YHZ$^+$13] Xiang Yu, Junzhou Huang, Shaoting Zhang, Wang Yan, and Dimitris N. Metaxas. Pose-Free Facial Landmark Fitting via Optimized Part Mixtures and Cascaded Deformable Shape Model. In *2013 IEEE International Conference on Computer Vision*, pages 1944–1951, 2013.

[YLYL13] Junjie Yan, Zhen Lei, Dong Yi, and Stan Z. Li. Learn to Combine Multiple Hypotheses for Accurate Face Alignment. In *Proceedings of the 2013 IEEE International Conference on Computer Vision Workshops*, ICCVW '13, pages 392–396, Washington, DC, USA, 2013. IEEE Computer Society.

[YP13] Heng Yang and Ioannis Patras. Sieving Regression Forest Votes for Facial Feature Detection in the Wild. In *2013 IEEE International Conference on Computer Vision*, pages 1936–1943, 2013.

[YSEB99] Alan L. Yuille, Daniel Snow, Epstein Russell, and Peter N. Belhumeur. Determining Generative Models of Objects under Varying Illumination: Shape and Albedo from Multiple Images using SVD and Integrability. *International Journal of Computer Vision*, 35(3):203–222, 1999.

[ZBL13] Feng Zhou, Jonathan Brandt, and Zhe Lin. Exemplar-based Graph Matching for Robust Facial Landmark Localization. *IEEE International Conference on Computer Vision (ICCV)*, pages 1025–1032, 2013.

[ZC01] Wen Yi Zhao and Rama Chellappa. Symmetric Shape-from-Shading using Self-ratio Image. *International Journal of Computer Vision*, 45(1):55–75, 2001.

[ZFC$^+$13] Erjin Zhou, Haoqiang Fan, Zhimin Cao, Yuning Jiang, and Qi Yin. Extensive Facial Landmark Localization with Coarse-to-Fine Convolutional Network Cascade. In *2013 IEEE International Conference on Computer Vision Workshops*, pages 386–391, 2013.

[ZKSC15] Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Leveraging Datasets with Varying Annotations for Face Alignment via Deep Regression Network. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3801–3809, 2015.

[ZKSC16] Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Occlusion-Free Face Alignment: Deep Regression Networks Coupled With De-Corrupt AutoEncoders. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[ZLL$^+$16] Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z. Li. Face Alignment Across Large Poses: A 3D Solution. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[ZLLL17] Xiangyu Zhu, Xiaoming Liu, Zhen Lei, and Stan Z. Li. Face Alignment in Full Pose Range: A 3D Total Solution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page 1, 2017.

[ZLLT14] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial Landmark Detection by Deep Multi-task Learning. In David Fleet, Pajdla Tomas, Schiele Bernt, and Tuytelaars Tinne, editors, *Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI*, pages 94–108. Springer International Publishing, Cham, 2014.

[ZLLT16a] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Learning Deep Representation for Face Alignment with Auxiliary Attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(5):918–930, 2016.

[ZLLT16b] Shizhan Zhu, Cheng Li, Chen-Change Loy, and Xiaoou Tang. Unconstrained Face Alignment via Cascaded Compositional Learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[ZLLT16c] Shizhan Zhu, Sifei Liu, Chen Change Loy, and Xiaoou Tang. Deep Cascaded Bi-Network for Face Hallucination. In Bastian Leibe, Matas Jiri, Sebe Nicu, and Welling Max, editors, *Computer Vision – ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V*, pages 614–630. Springer International Publishing, Cham, 2016.

[ZLY+15] Xiangyu Zhu, Z. Lei, Junjie Yan, D. Yi, and S. Z. Li. High-fidelity Pose and Expression Normalization for Face Recognition in the Wild. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 787–796, 2015.

[ZR12] Xiangxin Zhu and Deva Ramanan. Face Detection, Pose Estimation, and Landmark Localization in the Wild. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2879–2886, 2012.

[ZS06] Lei Zhang and D. Samaras. Face Recognition from a Single Training Image under Arbitrary Unknown Lighting using Spherical Harmonics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(3):351–363, 2006.

[ZSKC14] Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen. Coarse-to-Fine Auto-Encoder Networks (CFAN) for Real-Time Face Alignment. In David Fleet, Pajdla Tomas, Schiele Bernt, and Tuytelaars Tinne, editors, *Computer Vision – ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part II*, pages 1–16. Springer International Publishing, Cham, 2014.

[ZYY$^+$15] Xiangyu Zhu, Junjie Yan, Dong Yi, Zhen Lei, and Stan Z. Li. Discriminative 3D Morphable Model Fitting. *Automatic Face and Gesture Recognition (FG)*, 2015.