Semantic Morphable Models

Inauguraldissertation

zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von

Bernhard Egger

von Sattel, Schwyz

Basel, 2017

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät

auf Antrag von

Prof. Dr. Thomas Vetter, Universität Basel, Schweiz, Dissertationsleiter

Dr. Will Smith, University of York, United Kingdom, Korreferent

Basel, den 20. Juni 2017

Prof. Dr. Martin Spiess, Dekan

Abstract

In this thesis we discuss how computers can automatically interpret images of human faces. The applications of face image analysis systems range from image description, face analysis, interpretation, human-computer interaction, forensics to image manipulation. The analysis of faces in unconstrained scenes is a challenging task. Faces appear in images in a high variety of shape and texture and factors influencing the image formation process like illumination, 3D pose and the scene itself. A face is only a component of a scene and can be occluded by glasses or various other objects in front of the face.

We propose an attribute-based image description framework for the analysis of unconstrained face images. The core of the framework are copula Morphable Models to jointly model facial shape, color and attributes in a generative statistical way. A set of model parameters for a face image directly holds facial attributes as image description. We estimate the model parameters for a new image in an Analysis-by-Synthesis setting. In this process, we include a semantic segmentation of the target image into semantic regions to be targeted by their associated models. Different models compete to explain the image pixels. We focus on face image analysis and use a face, a beard and a non-face model to explain different parts of input images. This semantic Morphable Model framework leads to better face explanation since only pixels belonging to the face have to be explained by the face model. We include occlusions or beards as semantic regions and model them as separated classes in the implemented application of the proposed framework. A main cornerstone for the Analysis-by-Synthesis process is illumination estimation. Illumination dominates facial appearance and varies strongly in natural images. We explicitly estimate the illumination condition robust to occlusions and outliers.

This thesis combines copula Morphable Models, semantic model adaptation, image segmentation and robust illumination estimation which are necessary to build the overall semantic Morphable Model framework.

Acknowledgments

I would like to thank the following individuals for their support:

All members of the the Graphics and Vision Research Group - I have enjoyed all scientific and non-scientific discussions over the past years.

Prof. Thomas Vetter for the opportunity to be part of his team, and for providing guidance and a free rein during my PhD studies.

Prof. Volker Roth provided valuable inputs during committee meetings and the copula project.

Dr. Will Smith for reviewing my thesis as co-referee and providing valuable feedback.

Sandro Schönborn for mentoring and motivating me during my research.

A PhD is like a marathon - thank you, Marcel Lüthi, for motivating me for both.

Dinu Kaufmann for the pleasant collaboration on the copula project.

Andreas Forster for his help in automatic hair detection for the purpose of beard segmentation.

Andreas Schneider for the evaluation of segmentation results.

Clemens Blumer, Anna Engler and Sandro Schönborn for valuable feedback and proofreading of this thesis.

All contributors of Scalismo and Scalismo-Faces for providing a great framework to conduct research in.

My friends for great leisure activities and providing an environment that made my time as a PhD student more enjoyable.

Marie-Claire und Hermann-Josef Egger für ihre lebenslange, uneingeschränkte Unterstützung und Liebe.

Meinrad Egger for his invaluable support and all the beautiful memories over the years.

Anna Engler for being my ally in both science and life. Countless joyful moments lie behind us and we will master new challenges and passion in the future. I love you with all my heart.

Lastly, I would like to express my gratitude to the Swiss National Science Foundation for partially funding this project (SNF DORE 13DPD3 136840 “Visual Search +”).

Contents

Chapter 1

Introduction

Scene understanding is the guiding principle of computer vision. To fully understand what we see or what is in a photograph, every single visible component and their interactions have to be parsed and described. Such analysis is performed using visual cues and prior knowledge. In this thesis, we focus on the analysis of photographs of faces in unconstrained scenes. To understand and interpret a face in a scene, we search for a description of the face and its setting, in our case, we search for a parametrized one. Faces not only vary in shape and texture but those variations are also coupled to attributes like sex or age. A parametrized face description should capture those facial characteristics and should also hold a human-understandable face description based on attributes. As part of a scene, a face is a 3 dimensional object and this property should be accounted. The position and pose can vary to all extents and parts of the face can be (self-)occluded. The illumination condition plays a major role in the image formation process and dominates facial appearance. To fully interpret a face in a scene we need to be aware of all those factors.

We follow an Analysis-by-Synthesis approach to analyze face images. We use a 3D Morphable Model (3DMM, [Blanz and Vetter, 1999]) which is a parameterized face model. The model is generative and can synthesize face images for a set of parameters. In the analysis process, we infer the parameters given a new unseen image (target). The model parameters provide a model based image description. We propose an extension of the classical 3DMM where this parametric description directly leads to an attribute-based face description. The idea of Analysis-by-Synthesis is to use computer graphics to produce parametrized renderings of a face similar to the target image. This inverse-rendering process is ill-posed since pixel appearance could be

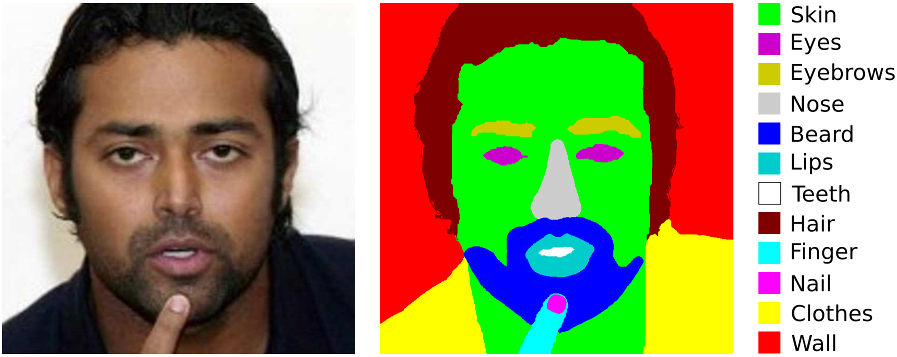


Figure 1.1: Human faces can be subdivided in semantic parts. The division itself is diverse and can be more or less detailed - the eye for example can be seen as one semantic region or refined further into sclera, iris and pupil. We propose semantic Morphable Models to provide a framework to use separate models to analyze and synthesize different semantic regions. The result of model adaptation to a target image is then a set of parameters for all involved models, as well as an image segmentation into semantic regions.

explained by various effects. The analysis procedure therefore builds on a generative parametric model as a strong prior for facial appearance. The search space over all poses, illumination conditions, facial shapes and textures is immense. Inferring those parameters from a 2D image is a highly non-convex task and can not be solved by simple optimization algorithms. For model adaptation, this work builds on the recently proposed 3DMM adaptation framework ([Schönborn et al., 2016]). We infer the posterior distribution of parameters for an observed target image. The framework is fully probabilistic and therefore able to include uncertain information e.g. feature point detections.

We extend the 3DMM adaptation framework to semantic Morphable Models. A face consists of multiple parts which are all complex in appearance itself. The key idea is to parse the face in an image and segment it into parts which are explained by separate models. The ideal case would be specific models for all regions of the face as depicted in Figure 1.1. Our analysis framework aims at 2D image analysis and therefore the semantic Morphable Model can be combined of different models generating 2D images. Different models compete to explain each pixel of a target image. The 3DMM is in the center of the proposed framework and enriched by additional models to explain a complete image. The main goal of using semantics is to improve the quality of the face model adaptation by relieving it from pixels which are out of the scope of the face model or which are not modeled at the desired degree of detail. Beards for example are not represented in our 3DMM - in an Analysis-by-Synthesis setting it is e.g. not helpful to compare the cheek with a beard. In the eye region we have similar challenges with eye gaze or eye closing which are not modeled by the 3DMM - comparing a closed eye with an open one is again not suitable. The additional models are more specific for a certain face region. Specific and local models for facial regions are coupled by the strong shape prior of the 3DMM, leading to a coarse-to-fine model adaptation strategy.

The semantic Morphable Model framework is open to various specific models for face and non-face regions. The presented framework segments the target image into face, beard and occlusion/background regions as the first implementation of a semantic Morphable Model (see Figure 1.2). Those additional models aim to overcome common limitations of the classical approach.

A main drawback of classical 3DMM adaptation is the lack of robustness against occlusions, they strongly mislead the model adaptation process. Occlusions are caused by various objects between the camera and the face (see Figure 1.3). Those objects hide parts of the face. The semantic framework models occlusions separately and excludes affected regions from the face model explanation.

Facial hair, like beards, are another limitation of current 3DMM adaptation. They are not contained in the training data and therefore the model can not properly adapt to them. Extending the training data to cover beards is not trivial: scanning and modeling beards is a challenge on its own. We model the beard region separately to overcome this limitation. Beards are itself very complex and can be modeled in different degrees of detail ([Beeler et al., 2012; Echevarria et al., 2014]). We do not aim to model each hair since this degree of detail is not available in most images. The main focus of our work is the analysis of the face, therefore we decided to use a less complex beard model compared to the face model. Beards can be grouped into different categories from full beards to mustaches. We propose a prototype-based shape prior to handle different beard types.

We compare two different beard appearance models. The first is a color-based appearance model estimated on the target image and estimated during the analysis procedure.

The second is a detection-based model incorporating discriminative methods to locate the beard region. We use strong prior knowledge for the location of the beard - the beard model is coupled to the location of the face model. The coupling of the beard to the face model is valuable in two ways: to posi-

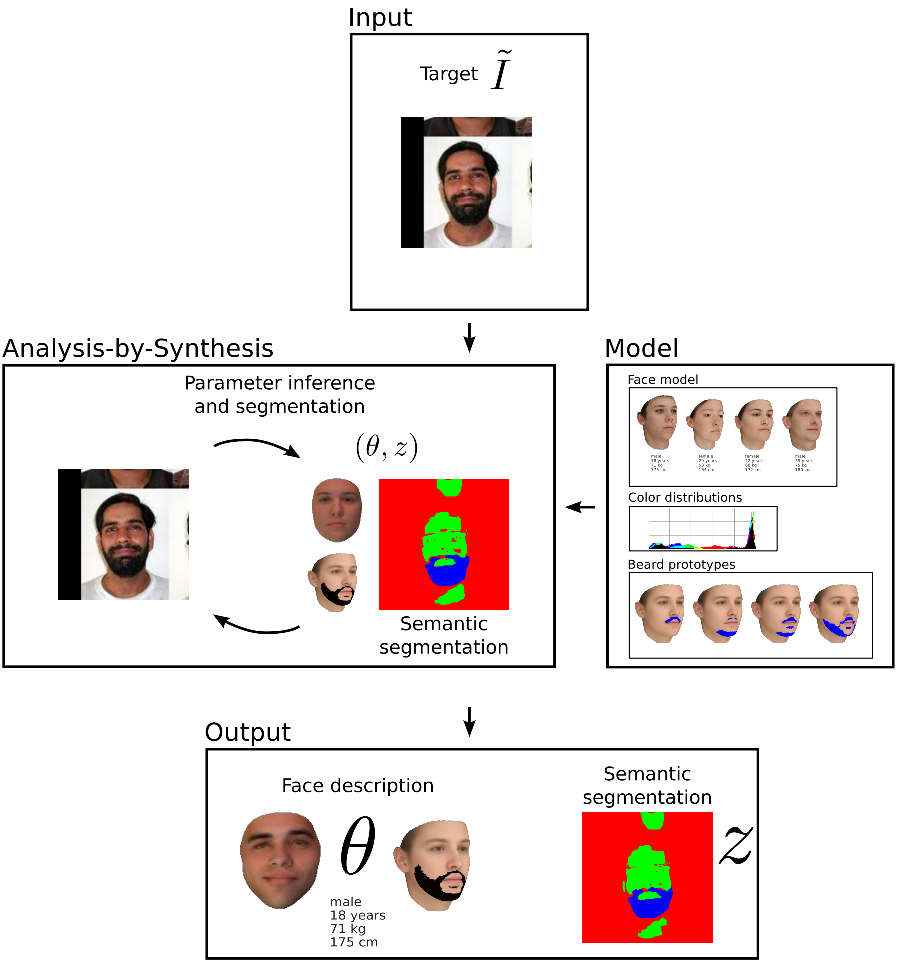


Figure 1.2: Semantic Morphable Models Overview: The target image is segmented into semantic regions which are explained by separate parametric models. The model inference of model parameters θ and semantic segmentation label z is performed simultaneously in an Analysis-by-Synthesis manner. The parameter inference is based on the synthetic image generated by all involved models and based on the semantic segmentation. The final set of model parameters and segmentation (fit) holds the model-based image interpretation.

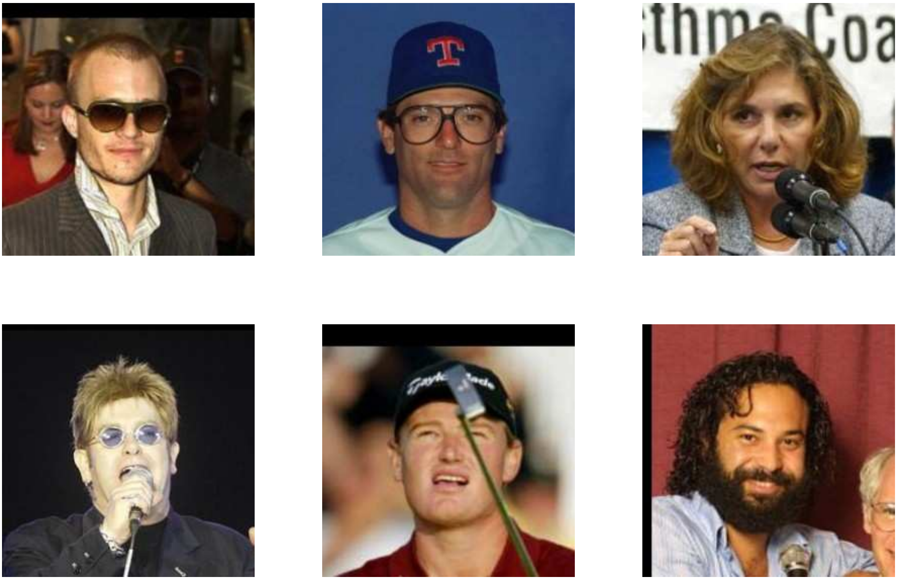


Figure 1.3: Occlusions by various objects appear frequently in real world face images. Occlusions range from face-related objects like glasses and unrelated objects like microphones or tools. When analyzing face images, occlusions should be kept out of the analysis procedure. Beards are also not included in 3DMMs and must be modeled separately or as occlusions. The images are from the LFW database ([Huang et al., 2007]).

tion the face correctly behind the occluding beard and to guide the position of the beard by the face model.

Different models compete to explain every pixel in the target image. Face pixels are explained by the face model, beard pixels by the beard model and occluding or background pixels by a general color model. The used models are of different complexity and level of detail. We are interested mainly in the face region and therefore use a detailed and parametrized generative model. The beard model is a medium complexity model which is only based on a simple shape and appearance prior. For occlusion and background, we use a very simple color model which is suitable to explain the region but does not hold much information for image description. Taking the semantics into account leads to a better model explanation and allows us to use more specific models for regions of the face. The result of the proposed semantic Morphable Model adaptation is a segmentation of the target image and the posterior distribution of model parameters.

Optimally, the segmentation can rely on a set of good face model parameters and the face model adaptation can be performed with a perfect segmentation. Both are not known at the beginning of the inference process and can not be derived directly by bottom-up methods. Since the segmentation influences the adaptation - and vice-versa - it would be optimal to infer them simultaneously. However, simultaneous inference of the parameters and the segmentation is infeasible. Therefore we propose an EM-like algorithm for semantic model adaptation which provides a good trade-off between accuracy and computational complexity.

A major challenge for the inference and segmentation is illumination. Facial appearance is dominated by the illumination condition (see Figure 1.4). Illumination can strongly guide and mislead the model adaptation. This becomes especially crucial under occlusion - effects from complex illumination conditions can easily be confoundedare in with occlusions. We propose a robust illumination estimation technique which leads to a reasonable illumination condition as well as to a first guess of the present occlusions. Robust illumination estimation can be applied to a wide range of unconstrained face images. We estimated the illumination condition on 15’000 images of the Annotated Facial Landmarks in the Wild (AFLW) ([Köstinger et al., 2011]) face database. This database contains face images taken in various settings and under various in and outdoor illumination settings. From this dataset we derive the first illumination prior estimated from real world images.

The 3DMM consists of a separate statistical model for shape and color. The shape and color parameters of a 3DMM hold the face interpretation, but

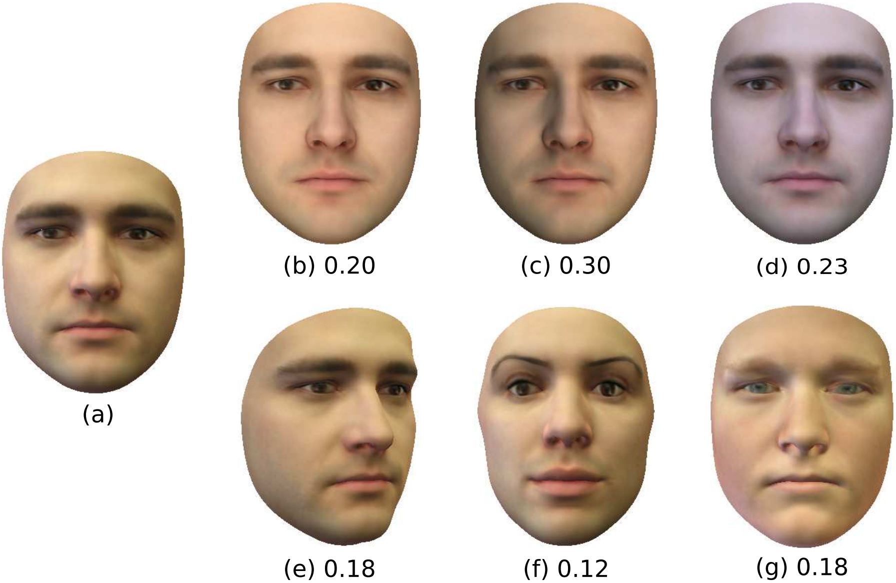


Figure 1.4: Illumination dominates facial appearance. We indicate the RMS-distance in color space of different renderings to a target rendering (a). We rendered the same target under new illumination conditions (b-d), compared to other changes (e-g). We present a frontal illumination (b), an illumination from the side (c) and a real world illumination (d). For comparison, we rendered the original face (a) under the original illumination conditions with strong changes in pose (e), shape (f) and texture (g). All those changes (e-g) are influencing the distance to the target image less than changes in illumination (b-d). The shown RMS distances caused by illumination are on average 50% higher than those caused by varying other parameters.

this is not interpretable for humans. Shape, color and attributes are often modeled separately because they are not scaled in the same range and live in different spaces. Copulas allow us to decouple the marginal distributions from the dependency structure. This decoupling leads also to scale-invariant analysis of the dependency structure which enables us to learn a combined shape, color and attribute model and even integrate continuous and noncontinuous attributes in the statistics. By combining shape, color and attributes, the resulting model can encode correlations between different modalities and gets more specific to faces. We propose to use a copula Morphable Model to integrate attributes for description directly into the statistical model.

The goal of our full analysis framework (see Figure 1.2) is an image description. Whilst classical 3DMM parameters do not hold a human-understandable image description, the copula extension leads to an integrated and understandable description by attributes. We perform attribute based description of single face images as a straightforward application of the proposed occlusion-aware and semantic copula Morphable Model adaptation framework.

The software implementation is based on the Statismo ([Lüthi et al., 2012]), Scalismo[1] and Scalismo-Faces [2] software frameworks.

1.1 Contribution

- We introduce semantic Morphable Models which enable us to model parts of the face separately and lead to an occlusion-aware analysis framework.
- We present a segmentation strategy including model-based and detection-based cues. This merges ideas from Conditional and Markov Random Field segmentation approaches.
- We propose a robust illumination estimation method which is key for robust model based face image analysis.
- We build an illumination prior built on real world illumination conditions.
- We present copula Morphable Models which allow us to learn a combined shape, color and attribute model and respect non-Gaussian marginal distributions.

[1] Scalismo - A Scalable Image Analysis and Shape Modeling Software Framework `https://github.com/unibas-gravis/scalismo`

[2] Scalismo-Faces - Module to work with 2D images, with a focus on face images `https://github.com/unibas-gravis/scalismo-faces`

1.2 Organization

The thesis proposes an attribute based image description framework and contains three main parts: Copula Morphable Models, semantic Morphable Models and robust illumination estimation. Each part can be read, understood and implemented separately - however, they are unified in the proposed framework for face image description and each part is necessary for the proposed attribute-based image description framework. We first summarize the related work of the individual parts of the thesis and the overall ideas in Chapter 2. We then introduce the copula extension of 3DMM to build our appearance prior for faces and include human-understandable attributes in Chapter 3. Then the semantic and occlusion-aware model adaptation framework in Chapter 4 builds the main part of this thesis. Robust illumination estimation is introduced in Chapter 5 and is necessary to adapt the model to images under unconstrained settings including occlusions. The parts are explained and evaluated in separate chapters and the complete framework is evaluated in an attribute description task and discussed in Chapter 6. Big parts of this thesis were already published or submitted to international conferences or journals ([Egger et al., 2014, 2016a,b, 2017a,b,c]). The thesis is concluded by ideas to further develop the framework in Chapter 7, some with preliminary results. In Chapter 8 the outcome of the thesis is summarized and evaluated.

Chapter 2

Related Work

The overall idea of semantic Morphable Models is based on several components which are discussed in this thesis. We provide an overview of the related work for all components our contributions enter.

2.1 Copula Morphable Models

The Eigenfaces approach ([Sirovich and Kirby, 1987], [Turk et al., 1991]) was a first parametric model for faces. They used PCA on facial images to analyze and synthesize faces. It performed well on images which where already aligned and did not contain pose variations. The next step in parametric appearance modeling for faces were Active Appearance Models ([Cootes et al., 1998]). They add a shape component which allows to model the shape independently from the appearance. This extension enables the model to adapt to stronger shape variations and to a certain degree of pose variation. As soon as self-occlusion arises those 2D methods fail. The 3DMM ([Blanz and Vetter, 1999]) uses a dense registration, extends the shape model to 3D and adds a camera and illumination model. The 3DMM allows handling appearance independently from pose, illumination and shape. This model can handle faces in all pose angles and isolates facial texture from the illumination. Through 3D modeling standard computer graphic techniques can be applied for rendering and simulation of illumination.

The initial motivation behind 3DMMs was 3D reconstruction from 2D images and this is still a wide field of research. The different optimization methods range from stochastic gradient descent ([Blanz and Vetter, 1999]), multi-feature gradient descent ([Romdhani and Vetter, 2003]), fast multi-step model adaptation ([Aldrian and Smith, 2013]), sampling based model adaptation ([Schönborn et al., 2013], cascaded regression techniques ([Zhu et al., 2015; Huber et al., 2015]) and recently deep learning techniques ([Tewari et al., 2017]).

Our work is based on the recent work on 3DMM adaptation ([Schönborn et al., 2016]) which frame all the ideas and the model adaptation into a fully probabilistic framework. The model parameter adaptation is performed with a sampling algorithm to infer the posterior distribution of suitable model parameters.

Color appearance and shape are modeled independently in AAMs and 3DMMs. Recently, it was demonstrated that facial shape and appearance are correlated ([Schumacher and Blanz, 2015]) and those correlations were investigated using Canonical Correlation Analysis on separate shape and appearance PCA models. Attributes like age, weight, height, sex are often added to the PCA models as additional linear vectors ([Paysan et al., 2009]) or with limitations to Gaussian marginal distributions ([Blanc et al., 2012]).

For face image analysis attributes estimation is mainly explored with discriminative approaches ([Kumar et al., 2011]). Model-based approaches lack a direct attribute-based description and therefore attributes are estimated as post-processing steps ([Egger et al., 2014]).

The main reason to build separate models is a practical one – shape, color and attribute values are neither in the same space, nor scaled in the same range. Attributes are not even always continuous. Some methods approach this issue by normalization and combine color and shape models ([Edwards et al., 1998; Castelan et al., 2007]). With our copula Morphable Model we are the first to build a joint attribute, shape and color model. By integrating this additional dependency information, the model becomes more specific. However, this normalization does not allow us to include categorical attributes, is highly sensitive to outliers and not suitable to compare those different modalities.

2.2 Semantic Morphable Models

Semantic segmentation is a recognized cornerstone of computer vision. Segmentation is often performed as a pre-processing step for image analysis pipelines. Most approaches for semantic segmentation are discriminative (e.g. [Khan et al., 2015]). The idea of having different generative models in competition to explain different regions of the image is related to image parsing framework proposed by [Tu et al., 2005] and is unique for face image analysis. In an Analysis-by-Synthesis setting segmentation is uncommon.

Whilst the 3DMM is a global model for face appearance there are approaches for hierarchical models ([Jones and Poggio, 1998; Paysan et al., 2009]). The classical 3DMM lacks shape and textural details - this limitations are overcome by specific models for specific regions of the face. There are convincing results for model-based eye ([Bérard et al., 2016; Wood et al., 2016]), teeth ([Wu et al., 2016]) and hair reconstruction ([Chai et al., 2016]) from single images. Such models would be excellent to be used in our semantic Morphable Model framework.

Recently semantic segmentation was proposed for model based analysis for 3D input data by [Maninchedda et al., 2016]. Similar to our work segmentation and model adaptation is performed jointly. 3D data provides more reliable bottom-up cues than 2D images. This allows for better segmentation of e.g. glasses from 3D images. The semantic segmentation is also used to improve the quality of face reconstruction. The general challenges are related but the used depth information, which is not available in our setting, strongly helps when searching for occlusions, beards or glasses.

For the generative analysis of 2D images [Morel-Forster, 2017] detected hair to be excluded during the model adaptation to 2D images. This approach does not include segmentation methods, relies on working hair detection and is limited to two classes (face and non-face). We integrate the proposed hair detections in our semantic Morphable Model as additional bottom-up cue to guide the segmentation of beards.

The work of [Huang et al., 2004] is not related to faces but combines deformable models with Markov random fields for segmentation of digits. The beard prior proposed in our work is integrated in a similar way as they incorporate a prior from deformable models.

The closest method to the proposed one is the image parsing framework proposed by [Tu et al., 2005]. A similar model has recently been proposed in the medical imaging community for atlas-based segmentation of leukoaraiosis and strokes from MRI brain images ([Dalca et al., 2014]) and for model-based forensic shoe-print recognition from highly cluttered images ([Kortylewski, 2017]).

Occlusion-aware Morphable Models ([Egger et al., 2016b]) are excluding non-face pixels from the face model adaptation and represent a first step towards Semantic Morphable Models. Although occlusions are omnipresent in face images, most research using 3DMMs relies on occlusion-free data. There exist only few approaches for fitting a 3DMM under occlusion.

Standard robust error measures are not sufficient for generative face image analysis. Areas like mouth or eye regions tend to be excluded from the fitting because of their strong variability in appearance ([Romdhani and Vetter, 2003; De Smet et al., 2006]), and robust error measures like applied in [Pierrard and Vetter, 2007] are highly sensitive to illumination. Therefore, we explicitly aim to cover the whole face region in the image by the face model explanation and only exclude occlusions or outliers from the model adaptation.

[De Smet et al., 2006] learned an appearance distribution of the observed occlusion per image. This approach focuses on large-area occlusions like sunglasses and scarves. However it is sensitive to appearance changes due to illumination and cannot handle thin occlusions.

[Yildirim et al., 2017] presents a generative model including occlusions by various objects. 3D occlusions are included in the training data. During inference the input image is decomposed into face and occluding object and occlusions are excluded for face model adaptation. The performance is comparable to human performance on a recognition task on synthetic images.

The above mentioned works on occlusion handling use a 3DMM focused on synthetic data or databases with artificial and homogeneous, frontal illumination settings. We present a model which can handle occlusions during 3DMM adaptation under illumination conditions arising in “in the wild” databases.

2.3 Robust Illumination Estimation

Robust illumination estimation or inverse lighting is an important cornerstone of our approach. Inverse lighting ([Marschner and Greenberg, 1997]) is an inverse rendering technique trying to reconstruct the illumination condition. Inverse rendering is applied for scenes ([Barron and Malik, 2015]) or specific objects. For faces, 3DMMs are the most prominent technique used in inverse rendering settings. The recent work of [Shahlaei and Blanz, 2015] focuses on illumination estimation and provides a detailed overview of face specific inverse lighting techniques. The main focus of the presented methods is face model adaptation in an Analysis-by-Synthesis setting. Those methods are limited either to near-ambient illumination conditions ([De Smet et al., 2006; Pierrard and Vetter, 2007]) or cannot handle occlusions ([Romdhani and Vetter, 2003; Aldrian and Smith, 2013; Schönborn et al., 2016]). Even the most recent deep learning based methods suffer from occlusions when estimating illumination ([Tewari et al., 2017]).

Our robust illumination estimation technique handles both, occlusions and complex illumination conditions by approximating the environment map using a spherical harmonics illumination model. Few methods incorporate prior knowledge of illumination conditions. The most sophisticated priors are multivariate normal distributions learned on spherical harmonics parameters estimated from data as proposed in [Schönborn et al., 2016] and [Barron and Malik, 2015]. Those priors are less general and not available to the research community. Our robust estimation method enables us to estimate an illumination prior from available real world face databases. This illumination prior fills a gap for generative models.

Chapter 3

Copula Morphable Model

Parametric Appearance Models (PAM) build the basis for most generative image analysis methods. Objects are described in terms of pixel intensities. In the context of faces, Active Appearance Models [Cootes et al., 1998] and 3DMMs [Blanz and Vetter, 1999] are established PAMs to model appearance and shape. The dominant method for learning the parameters of a PAM is Principal Component Analysis (PCA) [Jolliffe, 2002] or Probabilistic PCA (PPCA) [Tipping and Bishop, 1999]. (P)PCA is used to describe the variance and dependency in the data. Due to the sensitivity of (P)PCA to space and scaling, separate models are learned for shape and appearance.

We propose a method based on copula to build joint models of shape and color and even integrate continuous and categorical attributes. We use a semi-parametric Gaussian copula model, where dependency and variance are modeled separately. This model enables us to use arbitrary marginal distributions. Moreover, facial color, shape and continuous or categorical attributes can be analyzed in an unified way. Accounting for the joint dependency between all those facial components leads to a more specific and joint face model.

Copula methods are based on Sklar's theorem which allows the decomposition of every continuous and multivariate distribution function into its marginal distributions and a copula [Sklar, 1959]. A copula model provides the decomposition of the dependency and the marginal distributions such that the copula contains the dependency structure only. In general, separating all marginals from the dependency structure leads to a scale invariant description of the underlying dependency. This is desired when working with data from different modalities, arising from different spaces. Scale invariance enables us to learn a combined dependency structure of shape, color and attributes.

We use the observed empirical marginal distributions and keep the parametric dependency structure; in particular, we chose a Gaussian copula because of its inherent Gaussian latent space. PCA can then be applied in the latent Gaussian space to learn the dependencies of the data independently from the marginal distribution. The method is proposed and evaluated in [Han and Liu, 2012] and is called Copula Component Analysis (COCA). Samples drawn from a COCA model follow the empirical marginal distribution of the training data and are, more specific to the modeled object.

In the previous work on Copula Eigenfaces ([Egger et al., 2016a, 2017a]), we focused on artifacts arising in the color model. This is due to the assumption that the color intensities or, in other words, the marginals at each vertex are Gaussian-distributed. This approximation is far from the actual observed distribution of the training data (see Figure 3.1), and leads to unnatural artifacts in samples from the generative model. Those artifacts are removed using COCA instead of PCA.

In this work we focus more on building a joint model incorporating shape, color and attributes and adapt it in an image analysis task. Scale-invariance and decoupling of the dependency structure from the marginals enable us to include multi-modal data in a common statistical model. In an Analysis-by-Synthesis setting, we search for model parameters reconstructing the image. In the case of the classical 3DMM there are separate shape and color parameters, for our copula Morphable Model, the joint model parameters directly lead to an attribute based image description since attributes are an integrated component of the model.

This Chapter is based on research in close collaboration with Dinu Kaufmann ([Egger et al., 2016a, 2017a]). Chapter 3.1 to 3.4.1 contain excerpts of those works and summarize the relevant parts for this thesis.

3.1 Morphable Face Models

Let $x \in \mathbb{R}^{3n}$ describe a zero-mean vector representing 3 color channels (RGB color space) or the 3 dimensions of a shape coordinate for n vertex points of a 3D scan. The color channels are vectorized such that

$$x_{color} = (r_1, g_1, b_1, r_2, b_2, b_3, \dots, r_n, g_n, b_n)^T \tag{3.1}$$

and vertex points such that

$$x_{shape} = (x_1, y_1, z_1, x_2, y_2, z_3, \dots, x_n, y_n, z_n)^T \tag{3.2}$$

respectively. The set of m face scans in dense correspondence is arranged as the data matrix $X \in \mathbb{R}^{3n \times m}$ separately for shape and color.

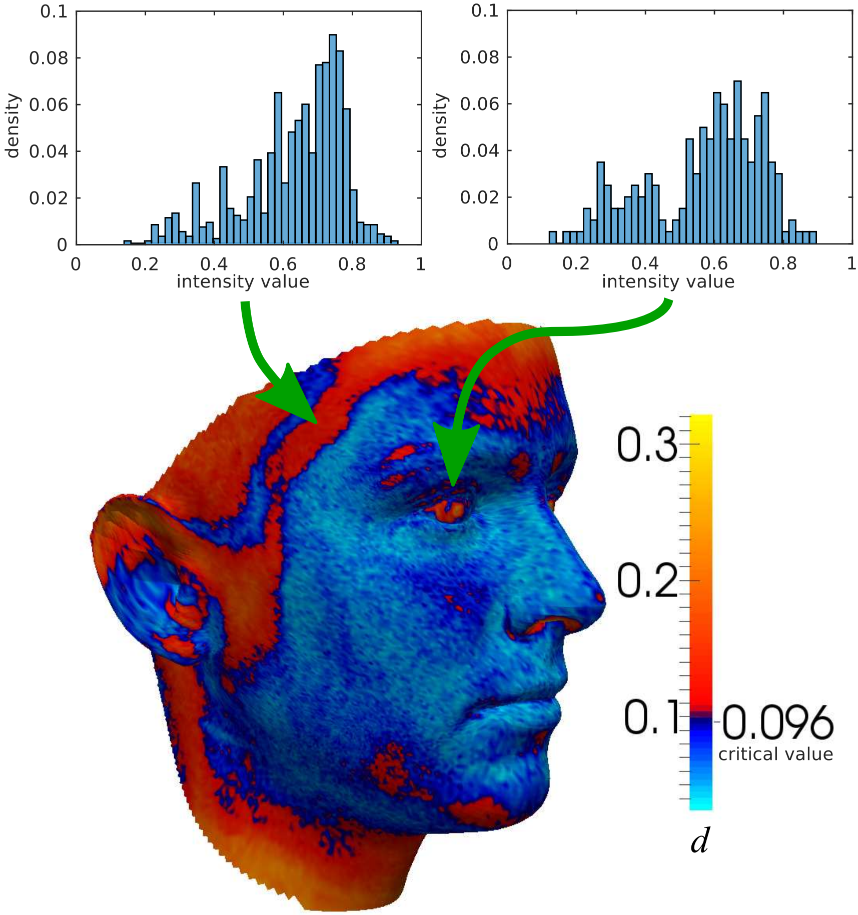


Figure 3.1: The result of the Kolmogorov-Smirnov Test ([Massey Jr, 1951]) to compare the empirical marginal distributions of color values from our 200 face scans with a Gaussian-reference probability distribution. We plot the highest value of the three color channels per vertex, because the values for the individual components are very similar. The Gaussian assumption does not hold for the color marginals. We show two exemplary marginal distributions in the eye and temple region. They are not only non-Gaussian but also not similar. The critical value assumes a significance level of $1 - \alpha = 0.05$

PCA [Jolliffe, 2002] aims at diagonalizing the sample covariance $\Sigma = \frac{1}{m}XX^T$, such that

$$\Sigma = \tfrac{1}{m}US^2U^T \tag{3.3}$$

where S is a diagonal matrix and U contains the transformation to the new basis. The columns of matrix U are the eigenvectors of Σ and the corresponding eigenvalues are on the diagonal of S.

PCA is usually computed by a singular value decomposition (SVD). In case of a rank-deficient sample covariance with rank $m < 3n$ we cannot calculate U^{-1}. Therefore, SVD leads to a compressed representation with a maximum of $m - 1$ dimensions. The eigenvectors in the transformation matrix U are ordered by the magnitude of the corresponding eigenvalues.

3.2 Copula Extension

While the variance in the data captures the scattering of the values, the covariance describes the underlying dependency structure. When computing PCA, the principal components are guided by the variance as well as the covariance in the data. This mingling of factors leads to results which are sensitive to different scales and to outliers in the training set. Regions with large variance and outliers influence the direction of the resulting principal components in an undesired manner.

We uncouple variance and dependency structure such that PCA only captures the dependency in the data. Our approach for uncoupling is a copula model which provides an analytical decomposition of the aforementioned factors.

Copulas ([Nelsen, 2013],[Joe, 1997]) allow us a detached analysis of the marginals and the dependency pattern. We consider a semiparametric Gaussian copula model ([Genest et al., 1995], [Tsukahara, 2005]). We keep the Gaussian copula for describing the dependency pattern, but we allow non-parametric marginals.

Let $x \in \mathbb{R}^{3n}$ describe the same zero-mean vector as used for PCA, representing 3 color channels or 3D coordinates of n vertices of a 3D scan. Sklar's theorem allows the decomposition of every continuous and multivariate cumulative distribution function (CDF) into its marginals $F_i(X_i), i = 1, \ldots, 3n$ and a copula C. The copula comprises the dependency structure, such that

$$F(X_1, \cdots, X_{3n}) = C\left(W_1, \ldots, W_{3n}\right) \tag{3.4}$$

where $W_i = F_i(X_i)$. W_i are uniformly distributed and generated by the probability integral transformation.

For our application, we consider the Gaussian copula because of its inherently implied latent space

$$\tilde{X}_i = \Phi^{-1}\left(W_i\right), \quad i = 1, \dots, 3n \tag{3.5}$$

where Φ is the standard normal CDF.

The multivariate latent space is standard normal-distributed and fully parametrized by the sample correlation matrix $\tilde{\Sigma} = \frac{1}{m}\tilde{X}\tilde{X}^T$ only. PCA is then applied on the sample correlation in the latent space $\tilde{X}$.

The separation of dependency pattern and marginals has multiple benefits: First, the Gaussian copula captures the dependency pattern separated from variance of color, shape and attributes. Second, whilst PCA is distorted by outliers, the semi-parametric copula extension solves this problem ([Han and Liu, 2012]). Third, the nonparametric marginals maintain the non-Gaussian nature of the color distribution and allow us to integrate attributes into the model.

3.3 Inference

We learn the latent sample correlation matrix $\tilde{\Sigma} = \frac{1}{m}\tilde{X}\tilde{X}^T$ in a semi-parametric fashion using nonparametric marginals and a parametric Gaussian copula. We compute $\hat{w}_{ij} = \hat{F}_{\mathrm{emp},i}(x_{ij}) = \frac{r_{ij}(x_{ij})}{m+1}$ using empirical marginals $\hat{F}_{\mathrm{emp},i}$, where $r_{ij}(x_{ij})$ is the rank of the data x_{ij} within the set $\{x_{i\bullet}\}$. Then, $\tilde{\Sigma}$ is simply the sample covariance of the normal scores

$$\tilde{x}_{ij} = \Phi^{-1}\left(\frac{r_{ij}(x_{ij})}{m+1}\right), \quad i = 1, \dots, 3n, \quad j = 1, \dots, m. \tag{3.6}$$

Above equation contains the nonparametric part, since $\tilde{\Sigma}$ is computed from the ranks $r_{ij}(x_{ij})$ solely and contains no information about the marginal distribution of the x's. Note, $\tilde{x} \sim \mathcal{N}(0, \tilde{\Sigma})$ is standard normal distributed with correlation matrix $\tilde{\Sigma}$. Subsequently, an eigen-decomposition is applied on the latent correlation matrix $\tilde{\Sigma}$.

Generating a sample using PCA then simply requires a sample from the model parameters

$$h \sim \mathcal{N}\left(0, I\right) \tag{3.7}$$

which is projected to the latent space

$$\tilde{x} = \tilde{U}\frac{\tilde{S}}{\sqrt{m}}h \tag{3.8}$$

Algorithm 3.1: Learning.

```
Input: Training set $\{X\}$
Output: Projection matrices $U$, $S$
for all dimensions $i$ do
    for all samples $j$ do
        $\tilde{x}_{ij} = \Phi^{-1}\left(\frac{r_{ij}(x_{ij})}{m+1}\right)$
find $\tilde{U}, \tilde{S}$ such that $\tilde{\Sigma} = \frac{1}{m}\tilde{U}\tilde{S}^2\tilde{U}^T$ (via SVD)
```

Algorithm 3.2: Sampling.

```
Output: Random sample $x$
$h \sim \mathcal{N}(0, I)$
$\tilde{x} = \tilde{U}\frac{\tilde{S}}{\sqrt{m}}h$
for all dimensions $i$ do
    $w_i = \Phi(\tilde{x}_i)$
    $x_i = \hat{F}_{\text{emp},i}(w_i)$
```

and further projected component-wise to

$$w_i = \Phi(\tilde{x}_i), \quad i = 1, \ldots, 3n. \tag{3.9}$$

Finally, the projection to the color, shape and attribute space of faces requires the interpolated empirical marginals

$$x_i = \hat{F}_{\text{emp},i}(w_i), \quad i = 1, \ldots, 3n. \tag{3.10}$$

All necessary steps are summarized in Algorithms 3.1 and 3.2 and visualized in Figure 3.2.

3.4 Implementation

The additional steps for using COCA can be implemented as simple pre- and post-processing before applying PCA. Basically, the data is mapped into a latent space where all marginals are Gaussian-distributed. The mapping is performed in two steps. First, the data is transformed to a uniform distribution by ranking the intensity values. Then it is transformed to a standard normal distribution. On the transformed data, we perform PCA to learn the dependency structure in the data.

To generate new instances from the model, all steps have to be reversed. Figure 3.2 gives an overview of all necessary transformations. These are the additional steps which have to be performed as pre- and post-processing for the analysis of the data and the synthesis of new random samples. In terms of computing resources we have to consider the following: The empirical marginal distributions F_{emp} are now part of the model and have to be kept in memory. In the learning part, the complexity of sorting the input data is added. In the sampling part, we have to transform the data back by looking up their values in the empirical distribution. The model needs almost double the memory of a PCA model whilst the additional computational effort is negligible.

The copula extension comes with low additional effort: it is easy to implement and has only slightly higher computing costs. We encourage the reader to implement these few steps since the increased flexibility in the modeling provides a valuable extension. We provide a MATLAB implementation to calculate COCA in Listing 3.1 and 3.2

```
% calculate empirical cdf
[empCDFs, indexX] = sort(X, 2);

% transform emp. cdf to uniform
[~, rank] = sort(indexX, 2);
uniformCDFs = rank / (size(rank, 2)+1);

% transform uni. cdf to std. normal cdf
normCDFs = norminv(uniformCDFs',0,1)';

% calculate PCA
[U,S,V] = svd(normCDFs, 'econ');
```

Listing 3.1: Learning

```
% random sample
m = size(normCDFs, 2);
h = random('norm' ,0 ,1 ,m ,1);
sample = U * S / sqrt(m) * h;

% std. normal to uniform
uniformSample = normcdf(sample, 0, 1) * (m - 1) + 1;

% uniform to emp. cdf
empSample = empCDFs(sub2ind(size(empCDFs), 1:size(data, 1), ...
    round(uniformSample')))';
```

Listing 3.2: Sampling

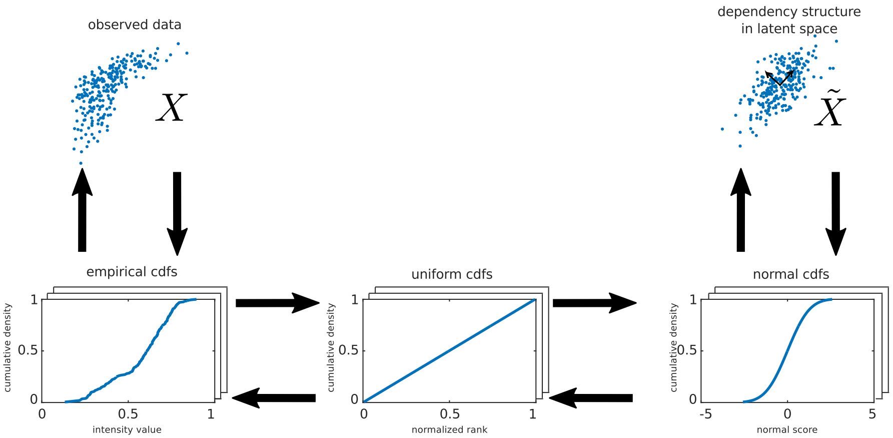


Figure 3.2: This figure shows the pre- and post-processing steps necessary to use a Gaussian copula before calculating PCA (toy data).

3.4.1 Discrete Ordinal Marginals

The formulation of the copula framework as above works with arbitrary continuous marginals. We extend the copula model for attributes, which follow discrete ordinal marginals. With this extension, we can even augment our model with attributes following binary distribution, such as sex. The underlying generative model assumes a continuous latent space, which is identified with the latent space $\tilde{X}$ of the copula. From this space, we observe the measurements via a discretization, which is related to the marginal distribution containing discontinuities. Using the CDFs of these marginals, for inferring the latent space as in the previous sections, causes problems. This is because the CDF transformations $\Phi^{-1} \circ \hat{F}_{\mathrm{emp},i} : X_i \to \tilde{X}_i$ do not change the marginal data distribution to be uniform and hence do not recover the continuous latent space. Instead, these CDF transformations only change the sample space. This leads to an invalid distribution of the copula and subsequently also of the latent space.

In order to resolve this problem, we follow the approach of the extended rank likelihood ([Hoff, 2007]). This provides us with an association-preserving mapping between measurement x_{ij} and latent observation $\tilde{x}_{ij}$. The essential idea behind this approach is, that the rank relation from the observations are preserved in the latent space.

In our case, we want to include a binary variable (sex). Note, that a binary variable can always be considered as an ordinal variable, since the ordering of the encoding does not matter. We replace the label x_{sex} with logistic regression in a preprocessing step. Specifically, logistic regression provides us a (continuous) score $x'_{\text{sex}} = E(x_{\text{sex}}|x_{-\text{sex}})$, which is the conditional expectation over (a low rank approximation of) the remaining variables $x_{-\text{sex}}$. Since the score constitutes of the conditional expectation, it relates to an approximation of the conditional posterior distribution in the latent space. The variable can then be treated as a continuous variable.

3.5 Combined Model

We learned a COCA model combining color, shape and attribute information of the face (see Figure 3.3 and Figure 3.4). Shape, color and attributes are combined by simply concatenating them:

$$x_{coca} = (x_{shape}^T, x_{color}^T, sex, age, weight, height)^T \tag{3.11}$$

Age, weight and height are continuous attributes and can therefore directly be integrated by concatenation in the COCA model. We added sex as a binary attribute and used the strategy presented in Section 3.4.1, where we replaced the binary labels with scores, obtained by logistic regression on the covariates.

The combined model allows us to generate random samples with consistent and correlated facial features. In Figure 3.5 we present how different modalities are correlated in the first parameters. By integrating this additional dependency information, the model becomes more specific ([Edwards et al., 1998]).

3.6 Model Adaptation

The main task we target in this thesis is the analysis of new images. We therefore search for model parameters which can reconstruct the target image well. The copula Morphable Model is highly related to the 3DMM. The main difference, relevant for model adaptation, is that shape and color are modeled jointly and therefore share a joint set of parameters. To infer the model parameters from a new image we adapt the model adaptation framework of [Schönborn et al., 2016] to apply it in combination with our copula Morphable Model. The framework is very flexible and can handle the novel model with few adaptations.

The only adaptation to deal with the joint set of parameters is in the proposal distribution. Instead of proposing separate steps for color and shape parameter updates, we changed the proposals to update the joint COCA

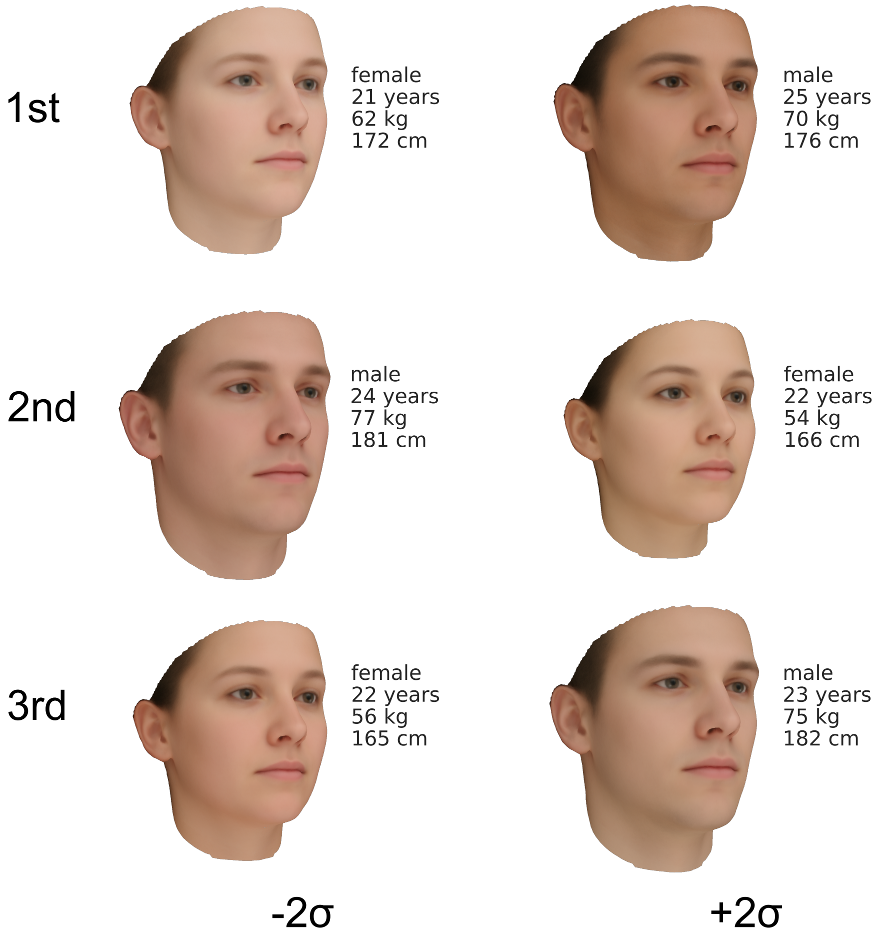


Figure 3.3: We learned a common shape, color and attribute model using COCA. We visualize the first eigenvectors with 2 standard deviations, which show the strongest dependencies in our training data. Whilst the first parameter is strongly dominated by color the latter parameters are targeting shape, color and attributes (compare Figure 3.5). Since the model is built from 100 females and 100 males, the first components are strongly connected to sex. The small range in age is caused by the training data which mainly consists of people with similar age.

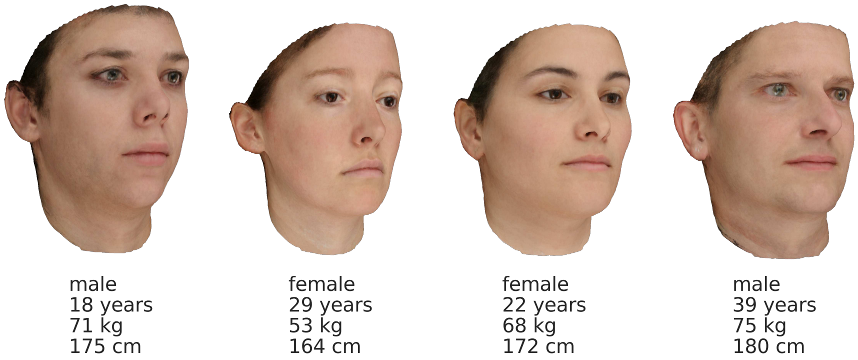


Figure 3.4: Random samples projected by a common shape, color and attribute model using COCA. Our model leads to samples with consistent appearance and attributes.

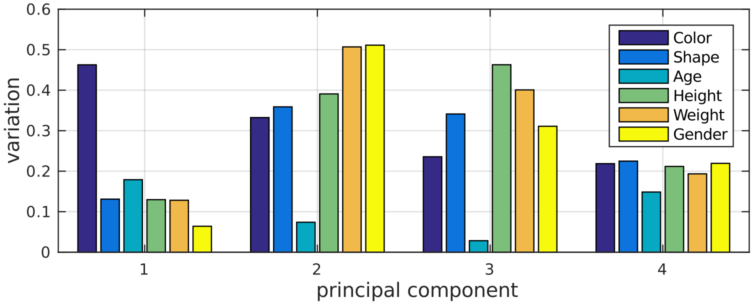


Figure 3.5: The influence of the first principal components on the different modalities of our model is shown. The variation is shown as the RMS distance of the normalized attributes in the covariance matrix. Whilst the first parameter is strongly dominated by color the later parameters are targeting shape, color and attributes (compare Figure 3.3). We observe strong correlations between the different modalities and attributes.

Table 3.1: Random walk proposals for color, shape and COCA-parameters. The shape and color parameter proposals correspond to the proposals in [Schönborn et al., 2016]. In our proposal distribution, the COCA-Proposal is designed according to the shape and color proposals and replaces them. σ is the standard deviation of the normal distribution, centered at the current location. λ designates mixture coefficients of the different scales coarse (C), intermediate (I) and fine (F).

Parameter	**Mixture**					
	σ_C	σ_I	σ_F	λ_C	λ_I	λ_F
Shape, $\vec{q}_S$	0.2	0.1	0.025	0.1	0.5	0.2
Radial Shape, $\|\vec{q}_S\|$			0.2			0.2
Color, $\vec{q}_C$	0.2	0.1	0.025	0.1	0.5	0.2
Radial Color, $\|\vec{q}_C\|$			0.2			0.2
Coeffs, $\vec{q}_{Coeffs}$	0.2	0.1	0.025	0.1	0.5	0.2
Radial Coeffs, $\|\vec{q}_{Coeffs}\|$			0.2			0.2

parameters, see Table 3.1. There are two types of proposals, a random walk proposal (Coeffs, $\vec{q}_{Coeffs}$) and a caricature proposal multiplying the current parameter set with a constant (Radial Coeffs, $\|\vec{q}_{Coeffs}\|$). We keep all other components and parameters of the model adaptation process fixed to make the results more comparable. The COCA-parameters directly map to color, shape and attributes and generate a complete face instance.

3.7 Experiments

To build our copula Morphable Model, we use the 200 face scans with attribute information used for building the Basel Face Model (BFM) ([Paysan et al., 2009]). The scans are in dense correspondence and were captured under an identical illumination setting. The specificity and generalization ability of the resulting model was evaluated in [Egger et al., 2016a] with a focus on the color model. We observed the specificity of the resulting model instances is higher, the generalization ability is slightly worse (measurable but not visible), see Figure 3.6 and 3.7. To compare the joint model against the separate model for color and shape we perform specific tasks like 3D reconstruction and attribute estimation.

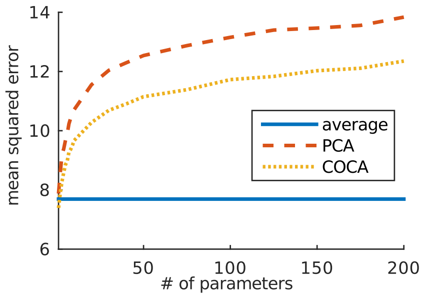


Figure 3.6: The specificity shows how close generated instances are to instances in the training data. The average distance of 1000 random samples to the training set (mean squared error per pixel and color channel) is shown. A model is more specific if the distance of the generated samples to the training set is smaller. We observe that COCA is more specific to faces (lower is better).

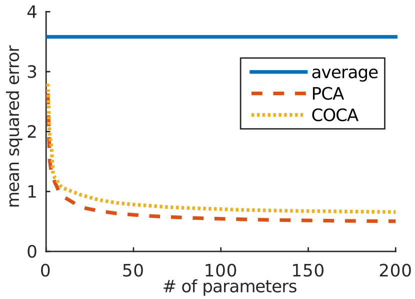


Figure 3.7: The generalization ability shows how exactly unseen instances can be represented by a model. The lower the error, the better a model generalizes. As a baseline, we present the generalization ability of the average face. We observe that PCA generalizes slightly better (lower is better).

Table 3.2: Shape reconstruction error (RMSD) in mm of our copula Morphable Model (COCA), a 3DMM built on the exact same data and evaluated in the exact same setting (PCA) and the result obtained by the mean face shape (mean-only).

Model	COCA	PCA	mean-only
RMSD in mm	5.68	5.78	6.79

3.7.1 3D Reconstruction

The main task of 3DMMs is 3D reconstruction of a face from a 2D image. To measure the eligibility of our copula Morphable Model for this task, we compare it to a classical 3DMM on the BU-3DFE face database ([Yin et al., 2006]). We render frontal images from the 100 individuals in the database and compare the shape reconstruction performance as proposed in [Schönborn et al., 2016]. Initialization was performed using 23 landmarks and the best sample reconstructing the target image (out of 10'000 samples) is taken for evaluation. We kept all 199 parameters for the model adaptation to keep the full flexibility of the model (for PCA 199 for shape and 199 for color). The resulting 3D reconstruction results are close to the results of the classical 3DMM, see Table 3.2.

3.7.2 Attribute Prediction

We perform an attribute prediction task on the Multi-PIE database ([Gross et al., 2010]). The COCA-parameters directly map to color, shape and attributes and generate a complete face instance. A copula Morphable Model instance contains the attribute prediction directly for a set of model parameters. We choose the task of sex prediction from a 2D target image.

The sampling method is performed in the same setting as in [Egger et al., 2014]. We choose a histogram background model and draw 10'000 samples for model adaptation. The initialization was performed on 9 manually annotated landmarks and only the first 50 COCA-parameters were adapted. The experiment was performed on all individuals of the first session of the Multi-PIE database under a frontal illumination and with poses between 0° and 60° of yaw angle.

For our experiments, we did not adapt any component of the copula Morphable Model analysis framework to the Multi-PIE database whilst our pre-

Table 3.3: Prediction performance of sex attribute classifiers (SPP) on PCA coefficients and on the pose-normalized representation using HOG features and color intensities from [Egger et al., 2014] compared to the result obtained by our copula Morphable Model. 69.9% of the individuals in the database are male.

Model / Feature	PCA	HOG	COCA
SPP	76.2 %	76.0%	82.5%

Table 3.4: Sex prediction performance of copula Morphable Model over different pose angles.

Pose	0°	15°	30°	45°	60°
Multi-PIE label	051_16	140_16	130_16	080_16	090_16
SPP	82.7 %	81.9 %	83.5 %	82.3 %	81.9 %

vious approach ([Egger et al., 2014]) used a part of the database for training. In our previous work we estimated attributes based on the estimated model parameters or the obtained pose-normalized face texture using Histogram of Oriented Gradients features (HOG, [Dalal and Triggs, 2005]). For both, the model parameters and HOG features we trained a classifier to predict sex. With our copula Morphable Model we outperform both, our generative and discriminative approaches, see Table 3.3.

We further analyzed sex prediction performance over the different pose angles (see Table 3.4) and over different ethnic groups (see Table 3.5). Whilst the performance does not vary over different pose angles, there are strong differences for different ethnic groups. The data used for building the face model has a strong bias to Caucasian faces. The observed performance for the ethnic groups incorporated in the model is much better than for those which are under-represented in the face scans.

The performance obtained in this setting is comparable to state of the art discriminative techniques on “in the wild” images [Kumar et al., 2011]. Our approach is unique when analyzing facial attributes with a fully generative model without post-processing. The integrated modeling of attributes enables us to estimate real conditional models and also include the uncertainty in the attribute prediction. Our model does not incorporate hair or other

Table 3.5: Sex prediction performance of copula Morphable Model itemized by ethnic groups. The used data has a strong bias to caucasian faces and therefore the performance is much better for them.

Ethnic Group	Caucasian	Asian	Indian	African-American
SPP	88.7%	70.3%	85.8%	63.3%

surrounding information in the image whilst most discriminative approaches incorporate this information as well.

3.8 Limitations

The main advantage of decoupling the marginal distributions comes with additional flexibility. Those marginal distributions can be handled in various ways - we chose to model them empirically. Modeling them empirically can lead to noisy model samples, especially when building models with few training examples. For many applications, it makes sense to take assumptions on the marginal distributions and model them parametrically. It is alternatively possible to smoothen the empirical marginals with a kernel k and replace (3.10) by $x_i = k(w_i, X_{i\bullet}), \quad i = 1, \ldots, 3n$.

Whilst specificity of the arising face model is higher, generalization drops due to the empirical marginal distributions and especially due to the coupling of shape and color model. Depending on the application, good generalization is important, we e.g. receive slightly worse image reconstruction results.

For non-continuous or categorical attributes, an ordering has to be derivable - if there is no natural ordering possible (like ethnic groups) an artificial ordering has to be defined or the categories have to be mapped to binary attributes. If the binary or categorical attributes are not balances in the training data, sampling strategies as described in [Hoff, 2007] have to be applied during model building.

The Multi-PIE database provides age annotation. We were not able to predict the age above chance rate. Elderly people are underrepresented in the data we built our model from, and our model based approach misses textural details like wrinkles which are important for age estimation.

3.9 Conclusion

In this work, we present a first step for copula-based parametric appearance models. Copulas itself are a huge field of research, we collect some ideas which could be interesting for statistical modeling of faces in the future work Section 7.2.

The main advantage we explore in this thesis is a joint model which includes facial attributes. Whilst the model parameters of the 3DMM do not directly lead to an attribute-based face description the copula Morphable Model allows us to integrate the attributes of interest directly into the face model. The model adaptation leads not only to a 3D reconstruction, illumination and color estimation but also an attribute-based face description.

Chapter 4

Semantic Morphable Models

A face image contains different semantic regions like skin, eyes, mouth, hair, background and various objects in the scene. Our face analysis is mainly focused on the face region but background and especially occlusions have to be taken into account during analysis. Ignoring the background and especially ignoring objects occluding the face leads to wrong image interpretation results. Not only the image, but also the face itself contains different regions. The eye region is e.g. complex in appearance, texture and movement. Other regions in the face, like beards are also complex in all those categories but are different from the eye region. A face is a highly complicated object with parts which should be aimed by highly specific models. We propose a semantic Morphable Model framework for combining segmentation of the target image and model adaptation. The basic idea is to segment a face image into different regions which are explained by models specific for those regions. Local models which are very specific for a part of the face are coupled by the 3DMM which builds the cornerstone of semantic Morphable Models. The 3DMM is coupled to the local models and guides them by a strong global shape and appearance prior.

We propose a very general and extensible framework together with a concrete implementation of a semantic Morphable Model. Our implementation focuses on occlusion-awareness to enable 3DMM adaptation on “in the wild” face images. In generative face image analysis, occlusions are a major challenge. Model adaptation is misled by occlusions if they are not taken into account, see Figure 4.1. We argue to handle and segment occlusions explicitly in the target image. Our implemented semantic Morphable Model combines

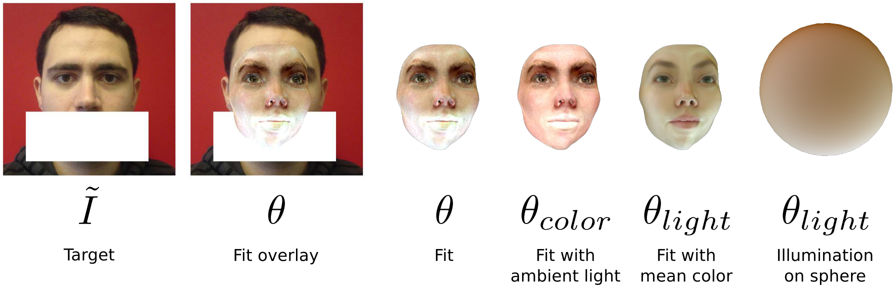


Figure 4.1: A fitting result of classical Morphable Model adaptation under occlusion. To analyze the composition of the fit we rendered the individual parts of θ separately. Occluding regions tend to be explained by both, the illumination and color coefficients.

a face, a beard and a non-face model. The target image is segmented in those three semantic regions. The beard model is an example for a model coupled to the face model by its location. The parameters of all those models are adapted to the target image and simultaneously the image is semantically segmented. The resulting framework leads to semantic model adaptation and occlusion-awareness. During inference we rely on a strong initialization of the segmentation which is explained in Chapter 5.

Semantic morphable models are based on six main ideas:

1. Pixels can be explained by different models. The separate models are adapted only to pixels assigned to them. Beard and non-face pixels arising from background or occlusions are excluded from face model adaptation.

2. We semantically segment the target image into regions. In our case we segment for occlusions, beards and the face. We pose segmentation as a Markov random field (MRF) with a beard prior.

3. Models are coupled. The beard model is explicitly coupled to the face shape and position. The coupling works bi-directionally: The face model parameters guide the beard segmentation and the segmentation guides face model adaptation.

4. Models can be of different complexity and this is explored in our implementation. Whilst our 3DMM is complex, the beard model modeling shape and appearance is less complex and the non-face model is a simple color model.

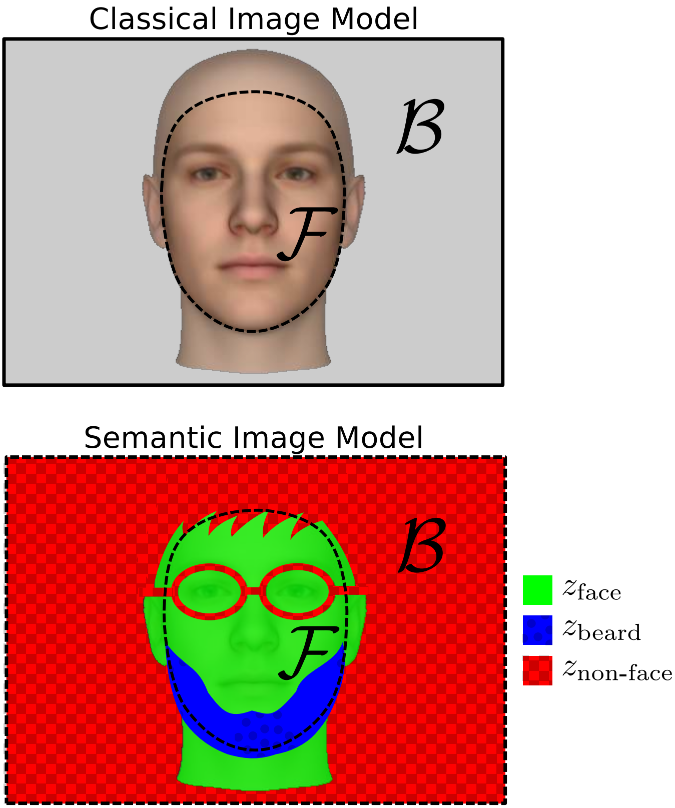


Figure 4.2: The regions used by the likelihood model by [Schönborn et al., 2016] (top). Each pixel belongs to the face model region $\mathcal{F}$ or the background model region $\mathcal{B}$. Assignment to foreground or background is based on the face model visibility only. In the proposed framework we have the same labels $\mathcal{F}$ and $\mathcal{B}$ but additional segmentation variables z to integrate occlusions (bottom). We assign a label z indicating if the pixel belongs to face, beard or non-face. Occlusions in the face model region $\mathcal{F}$ (in this case glasses) can hereby be excluded from the face model adaptation. Beards are handled explicitly and labeled separately.

5. We perform model adaptation and segmentation at the same time using an EM-like procedure. Model adaptation assumes a fixed segmentation and vice-versa.

6. We robustly estimate illumination for initialization (Chapter 5). Illumination is dominating facial appearance and has to be estimated to find occlusions.

4.1 Image Model

Our image formation model is based on the 3DMM interpreted in a Bayesian framework by [Schönborn et al., 2016]. The aim of face model adaptation (fitting) is to find model parameters generating a synthetic face image which is as similar to the face in the target image as possible. A likelihood model is used to rate parameters given a target image. The likelihood model is a product over the pixels i of the target image $\tilde{I}_i$, assuming conditional independence between all pixel observations. In the formulation of [Schönborn et al., 2016], pixels belong to the face model ($\mathcal{F}$) or the background model ($\mathcal{B}$). The foreground and background likelihoods (ℓ_{face}, b) compete to explain pixels in the image. The full likelihood model covering all pixels i in the image is

$$\ell\left(\theta; \tilde{I}\right) = \prod_{i \in \mathcal{F}} \ell_{\text{face}}\left(\theta; \tilde{I}_i\right) \prod_{i' \in \mathcal{B}} b\left(\tilde{I}_{i'}\right). \tag{4.1}$$

The foreground F is defined solely by the position of the face model (see Figure 4.2) and therefore this formulation cannot handle occlusions.

4.2 Semantic Image Model

We extend (4.1) to handle multiple models. Therefore, we introduce a random vector z containing a random variable z_i for each pixel i, indicating the class k it belongs to. The standard likelihood model (4.1) is extended to incorporate different classes:

$$\ell\left(\theta; \tilde{I}, z\right) = \prod_i \prod_k \ell_k\left(\theta; \tilde{I}_i\right)^{z_{ik}} \tag{4.2}$$

with $\sum_k z_{ik} = 1 \; \forall i$ and $z_{ik} \in \{0, 1\}$.

The likelihood model is open for various models for different parts of the image. In this work we use three classes k, namely face (z_{face}), beard (z_{beard})

and non-face ($z_{\text{non-face}}$). In Figure 4.2 we present all different labels and regions.

The main difference to the formulation by [Schönborn et al., 2016] is that the face model does not have to fit all pixels in the face region. Pixels in the image are evaluated by different likelihoods ℓ_k for the respective class models k. For our implementation those likelihoods are ℓ_{face}, ℓ_{beard} and $\ell_{\text{non-face}}$. They are explained in more detail in Section 4.2.2.

To select the likelihood per pixel during face model adaptation, we choose the strongest label z for every pixel $\max_k P(z_{ik})$. The generative face model with the likelihood ℓ_{face} is adapted to pixels with the label z_{face} only, according to (4.2). Beard and other non-face pixels are handled by separate likelihoods during face model adaptation. Non-face pixels are only characterized by a low likelihood of the face and beard model. Thus, they can be outliers, occlusions or background pixels.

4.2.1 Segmentation

To estimate the label z for a given parameter set θ we use an extension of the classical MRF segmentation technique including a beard prior similar as in [Huang et al., 2004], see Figure 4.4.

The MRF is formulated in the following form:

$$P(z|\tilde{I},\theta) \propto \prod_i \prod_k \ell_k \left(\theta; \tilde{I}_i\right)^{z_{ik}} P(z_{ik}|\theta)P(c) \prod_{j \in n(i)} P(z_{ik}, z_{jk}). \tag{4.3}$$

The data-term is built from the likelihoods for all classes k and over all pixels i and combined with the beard prior. The smoothness assumption $P(z_{ik}, z_{jk})$ enforces spatial contiguity of all pixels j which are neighbors $n(i)$ of i.

The beard prior is a prior on the labels z:

$$P(z|\theta, c) \tag{4.4}$$

The prior on the label z per pixel is defined by marginalizing over all m prototype shapes $l \in \{1..m\}$ defined on the face surface (see Figure 4.3):

$$P(z_i|\theta) = \sum_l P(z_i|c_l, \theta)P(c_l). \tag{4.5}$$

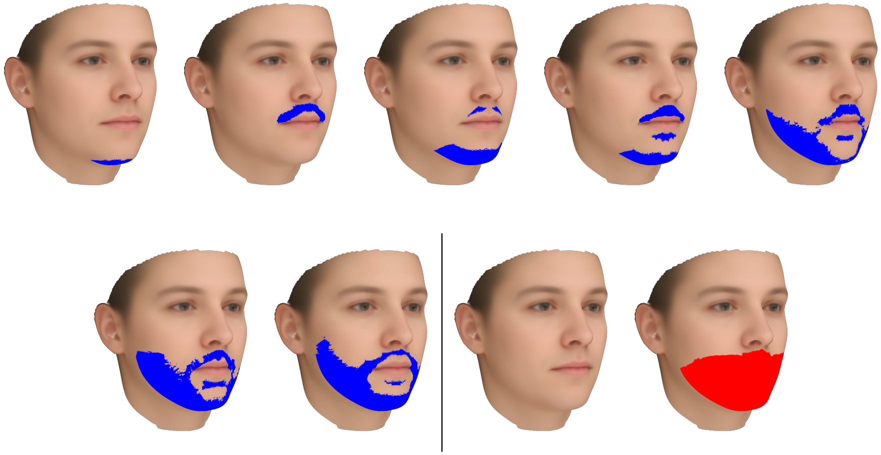


Figure 4.3: The seven beard prototypes derived from k-means^{++} clustering on manual beard segmentations on the Multi-PIE database (blue labels). We manually added a prototype for no-beard and one to handle occlusions over the complete beard region (bottom right, red). The prototypes are defined on the 3D face model and can be rendered to the image according to the face model parameters θ.

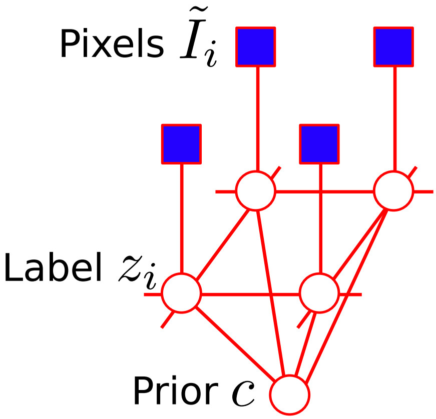


Figure 4.4: Graphical model of our MRF with beard prior c. We are interested in the segmentation labels z_i and observe the image pixels $\tilde{I}_i$. Every label z_i at position i is connected to the prior c and its neighbours.

The label z is depending on the beard prior $P(z|\theta, c)$. For the non-face and face label we use a uniform prior. The beard prior is depending on the current set of face model parameters θ since the pose and shape of the face influence position and shape of the beard in the image. We derived the prototype from manual beard segmentations labeled on the Multi-PIE database ([Gross et al., 2010]). We used k-means^{++} clustering technique as proposed in [Arthur and Vassilvitskii, 2007] to derive a small set of prototypes. The resulting prototypes are shown in Figure 4.3. We manually added a prototype for no-beard and another one to handle occlusion of the complete beard region. Those priors vote for non-face respectively face in the beard region. Those additional prototypes allow us to consider all possible labels in the beard region of the face. Large occlusions in the beard region with similar color appearance or detection responses would vote for a full beard, since the full beard prior covers most pixels in this regions. By adding a prior covering a bigger region than possibly covered by beard, we allow to label those regions correctly as non-face. The additional prior for no-beard is necessary since our prototypes are learned on male individuals with beard, which does not reflect all male and female individuals. All prototypes are defined on the face surface and their position in the image is depending on the current pose and face model parameters in θ.

4.2.2 Likelihood Models

Depending on the label z we apply different likelihood models for each pixel.

Face Likelihood

The likelihood of pixels to be explained by the face model is the following:

$$\ell_{\text{face}}\left(\theta; \tilde{I}_i\right) = \begin{cases} \frac{1}{N} \exp\left(-\frac{1}{2\sigma^2} \left\| \tilde{I}_i - I_i(\theta) \right\|^2\right) & \text{if } i \in \mathcal{F} \\ \frac{1}{\delta} h_f(\tilde{I}_i, \theta) & \text{if } i \in \mathcal{B}. \end{cases} \tag{4.6}$$

Pixels are evaluated by the face model if they are labeled as face (z_{face}) and are located in face region $\mathcal{F}$. The rendering function I generates a face for given parameters θ. This synthesized image $I(\theta)$ is compared to the observed image $\tilde{I}$. The likelihood model for pixels in the face region $\mathcal{F}$ is assuming per-pixel Gaussian noise. The likelihood ℓ_{face} is defined over the whole image and therefore also in the non-face region $\mathcal{B}$. For those pixels that are not in the region of the generative face model, we use a simple color model to compute the likelihood. We use a color histogram h_f with δ bins estimated on all pixels in $\mathcal{F}$ labeled as face (z_{face}).

Occlusion and Background Likelihood

For background modeling, we use a color model based on the observed image $\tilde{I}$. An overview over the necessity of a background model and different possible models can be found in [Schönborn et al., 2015]. The likelihood of the non-face model to describe occluding and background pixels is the following:

$$\ell_{\text{non-face}}\left(\theta; \tilde{I}_i\right) = b\left(\tilde{I}_i\right) = \frac{1}{\delta} h_{\tilde{I}}(\tilde{I}_i) \tag{4.7}$$

where δ is the bin volume and $h_{\tilde{I}}(\tilde{I}_i)$ is the relative frequency of the observed color value $\tilde{I}_i$ in the input image $\tilde{I}$.

We use a simple color histogram estimated on the whole image $\tilde{I}$ to estimate the background likelihood as proposed in [Egger et al., 2014].

4.2.3 Beard Model

For the beard model, we compare different likelihoods based on appearance estimated on the target image and detection learned from a database.

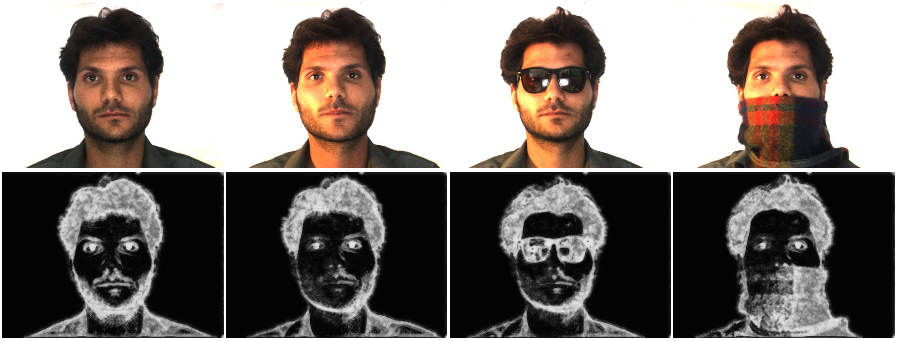


Figure 4.5: Hair detection results for one individual of the AR face database.

The appearance-based likelihood can directly be integrated into the proposed framework. The detection has to be interpreted as a likelihood to make it fit into the framework.

Beard Appearance Likelihood

The beard appearance likelihood is a simple histogram color model:

$$\ell^{a}_{\text{beard}}\left(\theta;\tilde{I}_i\right)=\frac{1}{\delta}h_{beard}(\tilde{I}_i,\theta), \tag{4.8}$$

The histogram h_{beard}, is estimated on the current segmentation of z_{beard}, where δ is the bin volume and $h_{beard}(\tilde{I}_i,\theta)$ is the relative frequency of the color value $\tilde{I}_i$ conditioned on the position of the beard prototype defined by position of the face model θ.

Beard Detection Likelihood

In contrast to the beard appearance likelihood ℓ^{a}_{beard} we also formulate a likelihood based on hair detection. The idea is to use bottom-up cues to derive information of possible beard candidate positions directly from the image and use this information during segmentation. Combining the likelihood based on detection with appearance likelihoods of the other models merges ideas from MRF and CRF segmentation and was so far not explored in the literature. We show how we can combine those modalities in our probabilistic framework.

Whilst the appearance is estimated on the target image based on the current model estimate, the likelihood based on hair detection ℓ^{d}_{beard} is derived from a hair classifier trained on the FERET database ([Phillips et al., 1998, 2000]). The hair classifier was proposed and trained in [Morel-Forster, 2017]. HOG ([Dalal and Triggs, 2005]) and Gabor ([Daugman, 1985]) features were used to classify hair using random forests ([Breiman, 2001]). The detection result yields a probability $\tilde{P}(\text{beard}_i|I_i)$ for every pixel i. In Figure 4.5 we present some hair detection results.

Detecting hair in a sliding window approach leads to noisy results and false detections $\tilde{P}$. As proposed by [Morel-Forster, 2017], we assume an uncertainty of the hair detection algorithm by incorporating a false positive (fp) and false negative rate (fn) of 5% in the following way:

$$P\Big(\text{beard}_i|\tilde{I}_i\Big) = \tilde{P}\left(\text{beard}_i|\tilde{I}_i\right)(1-(fn+fp))+fn. \tag{4.9}$$

The appearance likelihood fits well in a MRF, whilst a detection would fit in a Conditional Random Field (CRF). To use the detection result, we interpret the detection result as posterior and find an equivalent likelihood ℓ^d_{beard}:

$$P\Big(\text{beard}_i|\tilde{I}_i\Big) = \frac{\ell^d_{\text{beard}}\left(\theta;\tilde{I}_i\right)}{\ell_{\text{face}}\left(\theta;\tilde{I}_i\right)\ell_{\text{non-face}}\left(\theta;\tilde{I}_i\right)\ell^d_{\text{beard}}\left(\theta;\tilde{I}_i\right)}. \tag{4.10}$$

For simplicity we are assuming a uniform prior on the different class labels and we can therefore omit the class priors $P(\text{beard}_i)$, $P(\text{face}_i)$ and $P(\text{non-face}_i)$ in above equation. The likelihood can be directly derived from (4.10) by conversion:

$$\ell^d_{\text{beard}}\left(\theta;\tilde{I}_i\right) = \frac{P\Big(\text{beard}_i|\tilde{I}_i\Big)\left(\ell_{\text{face}}\left(\theta;\tilde{I}_i\right)+\ell_{\text{non-face}}\left(\theta;\tilde{I}_i\right)\right)}{1-P\Big(\text{beard}_i|\tilde{I}_i\Big)}. \tag{4.11}$$

4.2.4 Inference

The full model consists of the likelihoods for face model adaptation shown in (4.2) and segmentation from (4.3). Those equations depend on each other. In the fully probabilistic setting, we would have to include the uncertainty of the current parameter estimate θ' and the estimate on the segmentation label z' by estimating $P(z|\theta)P(\theta')$ and at the same time $P(\theta|z)P(z')$. Taking those uncertainties into account renders inference infeasible in practice. We therefore estimate the segmentation label z assuming a given set of face model parameters θ. And vice versa, we assume a given segmentation label z when adapting the face model parameters θ. Both are not known in advance and are adapted during the inference process to get a joint MAP-estimate of face model parameters and segmentation. We use an EM-like algorithm ([Dempster et al., 1977]) for alternating inference of the full model. In the expectation step, we update the label z of the segmentation. In the maximization step, we adapt the face model parameters θ. The choice of this procedure is motivated by convergence analysis in Section 4.4.1. In practice, this approximative inference leads to good results. An overview of the alternating steps is illustrated in Figure 4.6.

Face model adaptation is performed by a Markov Chain Monte Carlo strategy ([Schönborn et al., 2016]) with our extended likelihood from (4.2). MRF segmentation is performed using Loopy Belief Propagation with a sum product algorithm as proposed by [Murphy et al., 1999].

The histogram-based appearance model of beards is adapted during segmentation and fitting. During segmentation, the appearance is updated respecting the current belief on z_i. During fitting, the beard appearance is also updated due to the spatial coupling with the face model. When the shape or camera parameters of the face model change, the beard model has to be updated.

During segmentation, we assume fixed parameters θ and during fitting we assume given labels $\max_k z_{ik}$. Since the fixed values are only an approximation during the optimization process and fully probabilistic inference including the real uncertainty is infeasible, we account for those uncertainties by adapting the likelihoods. The uncertainty arises as misalignments and mislabeling, especially in important regions like the eye, nose and mouth. These regions are often mislabeled as occlusion due to their high variability in appearance when using other robust error measures. In the inference process, those regions are automatically incorporated gradually by adapting the face and non-face likelihood to incorporate this uncertainty. To account for the uncertainty of the face model parameters θ during segmentation, we adapt the face model likelihood for segmentation by taking neighboring pixels n into account (compare to (4.6)):

$$\ell'_{\text{face}}(\theta; \tilde{I}_i) = \frac{1}{N} \exp\left(- \frac{1}{2\sigma^2} \min_{j \in n(i)} \left\| \tilde{I}_i - I_{i,n}(\theta) \right\|^2 \right). \tag{4.12}$$

The small misalignment of the current state of the fit is taken into account by the neighboring pixels j in the target image. In our case we take the minimum over a patch of the 9 × 9 neighboring pixels direction (interpupillary distance is ~120 pixels).

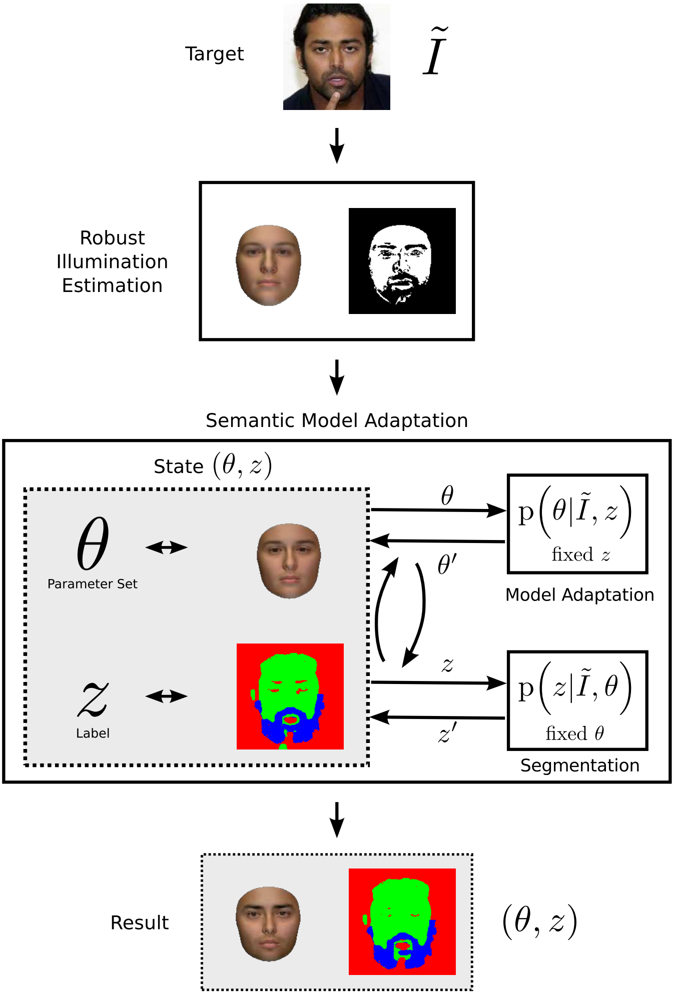


Figure 4.6: Algorithm overview: We start with an initial face model fit of our average face with a pose estimation. Then we perform a RANSAC-like robust illumination estimation for initialization of the segmentation label z and the illumination setting (for more details see Chapter 5). Then our face model and the segmentation are simultaneously adapted to the target image $\tilde{I}$. The result is a set of face model parameters θ and a segmentation into face and non-face regions. The presented target image is from the LFW face database ([Huang et al., 2007]).

To account for the uncertainty of the segmentation label z for face model adaptation, we adapt the likelihood of the non-face during face model adaptation. Pixels which are masked as non-face can be explained by the face model if it can do better (compare to (4.7)):

$$\ell'_{\text{non-face}}\left(\theta;\tilde{I}_i\right) = \max\left(\ell_{\text{face}}\left(\theta;\tilde{I}_i\right), b\left(\tilde{I}_i\right)\right)\text{if } i \in \mathcal{F}. \tag{4.13}$$

Both modifications more likely label pixels as face and this leads to consider them during face model adaptation.

4.3 Initialization

Our robust illumination estimation described in Chapter 5 gives a rough estimate of the illumination and segmentation. However, the obtained mask is underestimating the face region. Especially the eye, eyebrow and mouth regions are not included in this first estimate. Those regions differ from the skin regions of the face by their higher variance in appearance, they will be gradually incorporated during the full model inference.

The initialization of the beard model is derived from the segmentation obtained by robust illumination estimation. The prior is initialized by the mean of all beard prototypes. The appearance is estimated from the pixels in the prototype region segmented as non-face by the initial segmentation.

4.4 Experiments

For the model adaptation experiments, we perform alternating 2,000 Metropolis-Hastings sampling steps (best sample is taken to proceed) followed by a segmentation step with five iterations and repeat this procedure five times. This amounts to a total of 10,000 samples and 25 segmentation iterations.

For the 3DMM adaptation, the 3D pose has to be initialized. In the literature, this is performed manually ([Blanz and Vetter, 1999; Romdhani and Vetter, 2003; Aldrian and Smith, 2013]) or by using fiducial point detections ([Schönborn et al., 2013]). For all our experiments, we use automatic fiducial point detection results from the CLandmark Library made available by [Uřičář et al., 2015]. Our method is therefore fully automatic and does not need manual input.

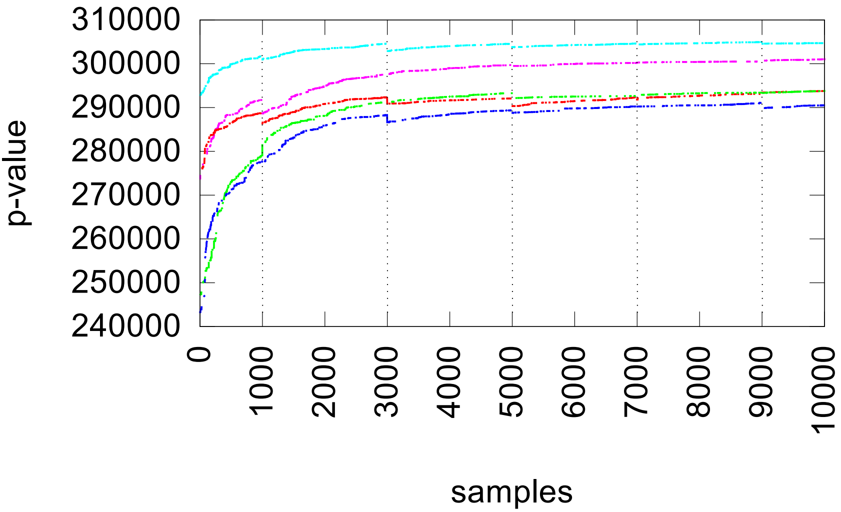


Figure 4.7: The image-likelihood (4.2) evaluated for all accepted samples of independent sampling runs on different target images. A segmentation step is performed at iteration 1000 and then every 2000th iteration (indicated by dotted line).

4.4.1 Convergence Analysis

In this Section we motivate the choice of the EM-like inference method based on convergence analysis. Our sampling-based model adaptation framework proposes parameter updates and verifies them according to (4.2). Those updates are typically small steps which draw after a burn-in phase samples from the posterior distribution of suitable model parameters. The segmentation could be directly integrated as such a proposal into the model adaptation. However, the segmentation changes the pixels which are evaluated by the different components of the likelihood. Applying a segmentation-step therefore often leads to a smaller likelihood according to (4.2). In practice, the model parameters of the corresponding models have to adapt to those new pixels. We visualize the image-likelihood of semantic model adaptations in Figure 4.7. It shows the jumps in the likelihood caused by segmentation.

The EM-like inference method can balance segmentation and model adaptation. By allowing the model parameters to adapt to the current segmentation, the model parameters can converge to the current segmentation setting.

We choose the number of samples such that we observe this convergence. After most E- and M-steps the image-likelihood improves over the last combination of segmentation and model parameters.

4.4.2 Segmentation

To evaluate the segmentation performance of the proposed algorithm, we manually segmented occlusions and beards in the AR face dataset ([Martinez and Benavente, 1998]). We took a subset of images consisting of the first 10 male and female participants appearing in both sessions. Unlike other evaluations, we include images under illuminations from the front and the side. We selected the neutral (containing glasses and beards) images from the first session and images with sunglasses and scarves from the second session. The total set consists of 120 images. We had to exclude five images because the fiducial point detection (four images) or the illumination estimation (one image) failed for them (m-009-25, w-009-25, w-002-22, w-012-24, w-012-25). For evaluation we labeled an elliptical face region, beards and occlusions manually. Evaluation was done within the elliptical face region only. Our manual annotations, used in this evaluation, are available under http://gravis.cs.unibas.ch/publications/2017/2017_Occlusion-aware_3D_Morphable_Models.zip.

In our previous work ([Egger et al., 2016b]), we compared our method for occlusion-aware model adaptation to a standard technique to handle outliers, namely a trimmed estimator including only $n\%$ of the pixels which are best explained by the face model. In this work we present the segmentation result including beards as an additional label. We present the simple matching coefficient (SMC) and the F1-Score for detailed analysis in Table 4.1. In our experiments we distinguish the three image settings: neutral, sunglasses and scarves. We include the result of the initialization to depict its contribution and show that the fitting improves the segmentation even more. We provide a separate evaluation of the different proposed beard likelihoods. The appearance and detection based likelihood lead to similar segmentation performance. The only outlier is the detection-based beard segmentation - the strong hair detection on the sunglasses (compare Figure 4.5) misleads votes for strong occlusions and select the no-beard prototype and misleads the segmentation.

4.4.3 Quality of Fit

We present qualitative results of our fitting quality on the AR face database ([Martinez and Benavente, 1998]) and the Labeled Faces in the Wild database (LFW) ([Huang et al., 2007]). In our results in Figure 4.9, the images include

Table 4.1: Comparison of segmentation performance in SMC and in brackets the F1-Scores (class|rest) for all labels on the AR face database ([Martinez and Benavente, 1998]). We present separate results for our initialization using robust illumination estimation (line 1-3). The evaluation of the full model is split into the appearance likelihood ℓ^a_{beard} (line 4-6) and the detection likelihood ℓ^d_{beard} (line 7-9).

Method	**Neutral**	**Glasses**	**Scarf**
Initialization z_{face}	0.78 (0.86|0.41)	0.81 (0.83|0.77)	0.73 (0.73|0.73)
Initialization z_{beard}	0.97 (-|0.99)	0.95 (-|0.98)	1.00 (-|1.00)
Initialization $z_{\text{non-face}}$	0.71 (0.09|0.83)	0.75 (0.67|0.80)	0.69 (0.66|0.69)
Full model ℓ^a_{beard}, z_{face}	0.85 (0.91|0.51)	0.85 (0.87|0.82)	0.84 (0.85|0.82)
Full model ℓ^a_{beard}, z_{beard}	0.98 (0.63|0.99)	0.96 (0.53|0.98)	0.97 (-|0.98)
Full model ℓ^a_{beard}, $z_{\text{non-face}}$	0.80 (0.89|0.15)	0.81 (0.86|0.72)	0.76 (0.80|0.69)
Full model ℓ^d_{beard}, z_{face}	0.85 (0.91|0.52)	0.85 (0.87|0.81)	0.86 (0.87|0.84)
Full model ℓ^d_{beard}, z_{beard}	0.98 (0.45|0.99)	0.95 (0.04|0.97)	0.98 (-|0.99)
Full model ℓ^d_{beard}, $z_{\text{non-face}}$	0.80 (0.88|0.14)	0.79 (0.83|0.70)	0.79 (0.82|0.74)

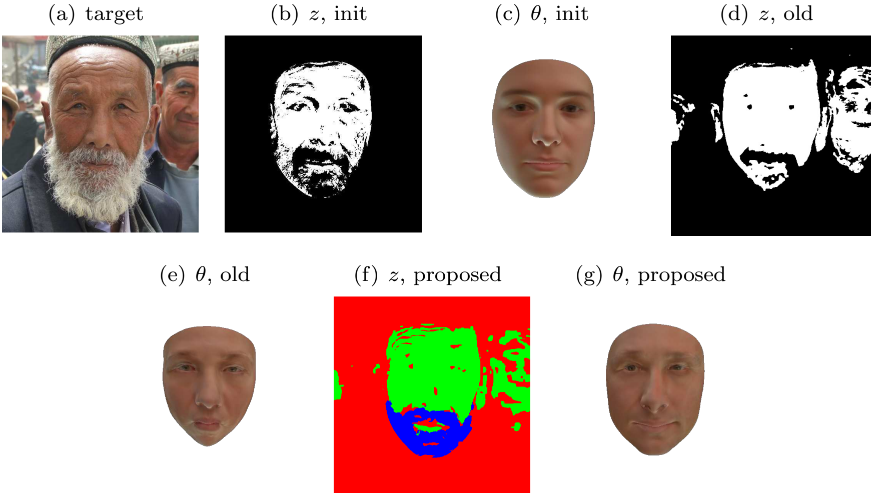
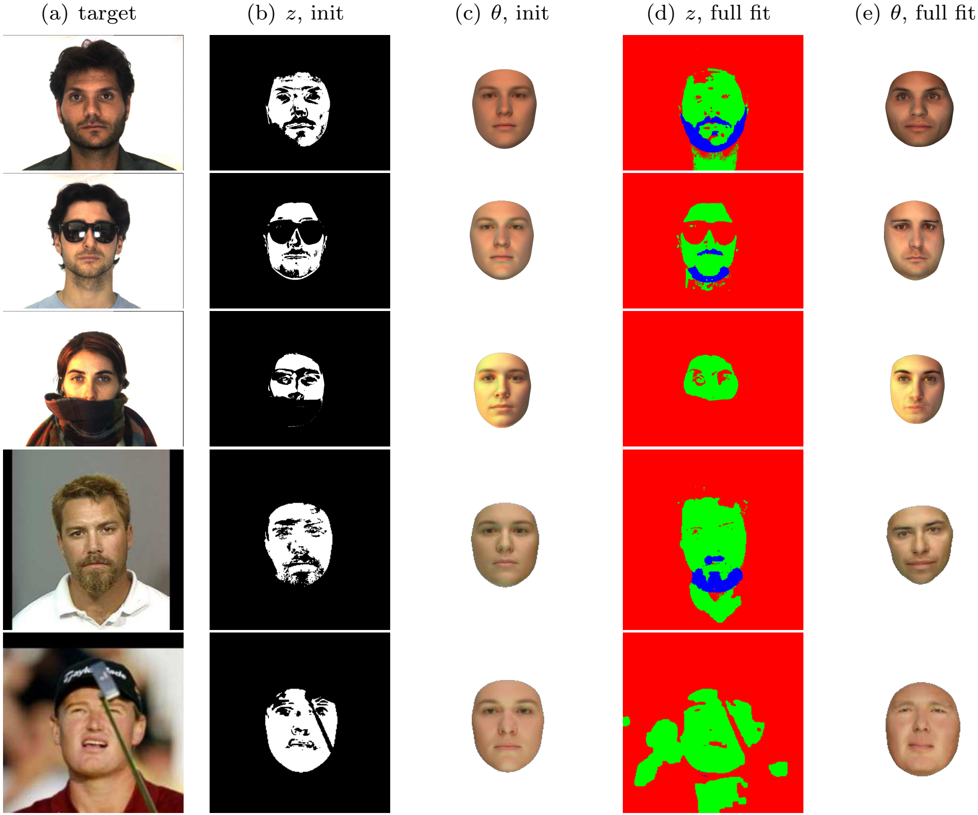


Figure 4.8: The benefit of explicitly modeling beards. The beards can not be reliably excluded during illumination estimation since they can partially be explained by illumination effects (b, c). We compare our results to the same approach without modeling beards explicitly (d, e, [Egger et al., 2016b]). By explicitly modeling beards the face model adaptation is not mislead by the beard region (f). Through the coupling of the beard model with the face model, the underlying face is kept at the correct position (g). The target image (a) arises from the LFW database ([Huang et al., 2007]).



(a) target

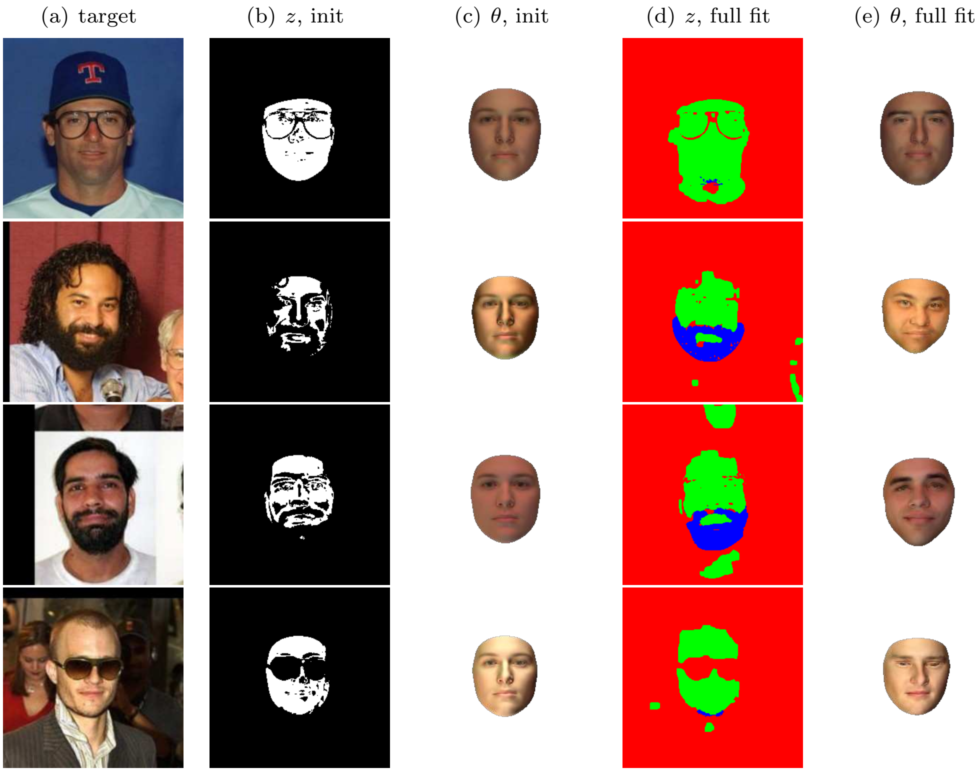
(b) z, init

(c) θ, init

(d) z, full fit

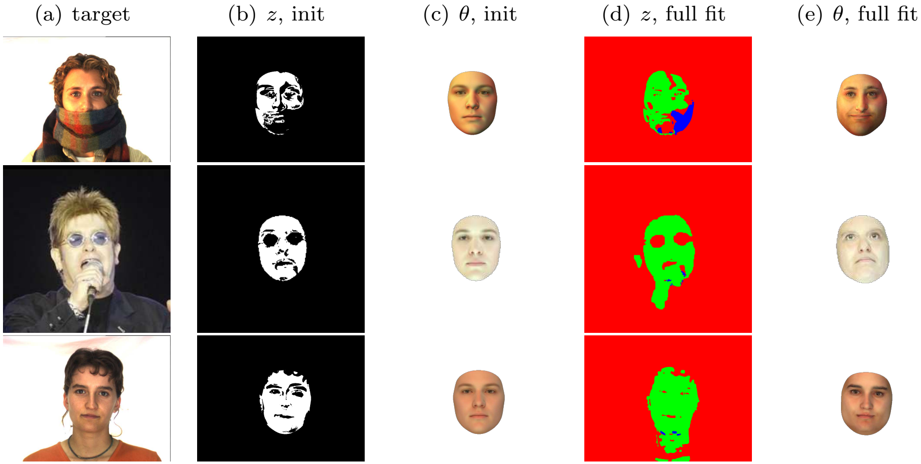
(e) θ, full fit

This Figure proceeds on the next page.



(a) target (b) z, init (c) θ, init (d) z, full fit (e) θ, full fit

Figure 4.9: (a) Target images from the AR face database (first three) [Martinez and Benavente, 1998] and the LFW database ([Huang et al., 2007]). (b) and (c) depict our initialization arising from the robust illumination estimation, (d) and (e) present the final results. Our final segmentation and synthesized face includes much more information of the eye, eyebrow, nose and mouth regions than the initialization. More examples on the previous page.



(a) target (b) z, init (c) θ, init (d) z, full fit (e) θ, full fit

Figure 4.10: Usual cases of failure: scarves can be explained by the light and color model and are therefore mislabeled. Hands have similar color appearance and do not distort the face model adaptation but lead to a wrong segmentation. The prototype for chin-beards is mislead by shadows under the chin which are not modeled in our illumination model. Note that our method is not adapted to a specific kind of outlier. The first and last target is from the AR face database ([Martinez and Benavente, 1998]), the middle one from the LFW database ([Huang et al., 2007]).

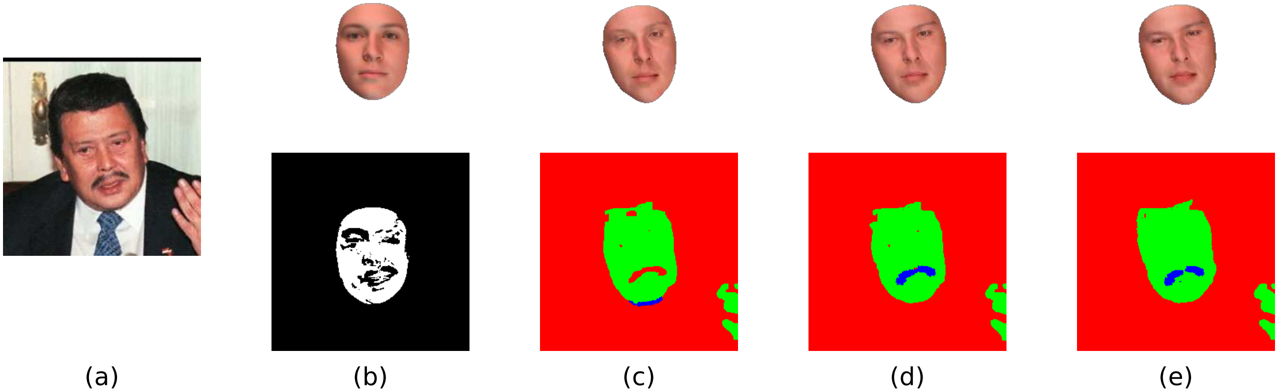


Figure 4.11: The results at different steps in our framework including the EM-strategy. The target image from the LFW database ([Huang et al., 2007]) is shown in (a). The result of the robust illumination estimation is shown in (b), we observe a strong pose misalignment in the roll angle. After the first 1000 samples of our model adaptation process the pose was adapted to the image (c) during the later model adaptation the correspondence gets better and the beard and face region are segmented better. We present the result after 3000 samples (d) and after the full 10'000 samples (e).

beards, occlusions and un-occluded face images. In Figure 4.10 we also include results where our method fails. Our method detects occlusions by an appearance prior from the face model. If occlusions can be explained by the color or illumination model, the segmentation will be wrong. Through the explicit modeling of beards the errors in the beard region are reduced and the face model does not drift away anymore in this region (see Figure 4.8). For this experiment we used the appearance based beard likelihood ℓ^{a}_{beard}. An additional result of the inference process is an explicit beard type estimation. The quality of the model adaptation on the AFLW database show almost the same performance as we obtain on data without occlusions. Interesting parts of the face are included gradually during the fitting process, see Figure 4.11. Our method also performs well on images without occlusions and does not tend to exclude parts of the face.

4.5 Limitations

The presented model is a first instance of a semantic Morphable Model. The approach is limited to the flexibility of the individual models and is complex to adapt to images. The current implementation excludes everything which can not be properly explained by the face model from the face model explanation. The texture model is very coarse and the model explanation misses textural details like wrinkles. Eye gaze is excluded in the 3DMM and most often handled as occlusion in the face model explanation. Those textural flaws could be eliminated by higher quality models for specific regions.

4.6 Conclusion

We proposed semantic Morphable Models and implemented a concrete instance. Our model distinguishes faces from beards and non-face pixels. This semantic segmentation of the target image leads to occlusion-awareness and better face reconstruction under occlusions. Our implementation is a first semantic Morphable Model - the idea of different models competing to explain different semantic regions in the image is open to more specific models to explain parts of the face. A high quality face model could be combined from individual components like an eye model ([Bérard et al., 2016; Wood et al., 2016]), a teeth model ([Wu et al., 2016]) and a hair model ([Chai et al., 2016]). We presented how different models can be coupled and how appearance based and detection based methods can be incorporated for segmentation. The specific models can be regionally coupled to the coarser face model as we have done for the beard model. The resulting framework is a hierarchical set of models for a coarse to fine image explanation framework.

Chapter 5

Robust Illumination Estimation

Illumination is a crucial part of every image formation process. In order to comprehend scenes from images, we need to understand the actual illumination setting. Illumination is dominating facial appearance, see Figure 1.4. Changes in texture or shape strongly influence our perception of identity, whilst our perception is invariant against illumination change. However, illumination changes are much stronger than changes in identity measured by standard Euclidean distance in RGB color space. Handling and estimating illumination is crucial for generative face image analysis.

Most approaches in computer vision do not explicitly model illumination but aim to be locally robust against illumination variations. Explicit illumination estimation is avoided because it is an ill-posed problem. The observed appearance arises from a multiplication of albedo and illumination and allows ambiguous explanations. For estimation of illumination we need prior knowledge about the shape and albedo of the observed object. With a given shape and albedo, illumination can be approximated directly. Uncertainty on shape or albedo are propagated and lead to uncertainty in the resulting illumination estimation. Occlusions and outliers render illumination estimation additionally complex in the analysis of real world face images.

Ignoring occlusions strongly misleads illumination estimation as shown in Figure 5.1. In the analysis process we search for the model instance which is most consistent with the target image despite occlusions and outliers. Face shape, albedo and pose also influence appearance and cannot be estimated independently. The novelty of our estimation method is handling of occlusions

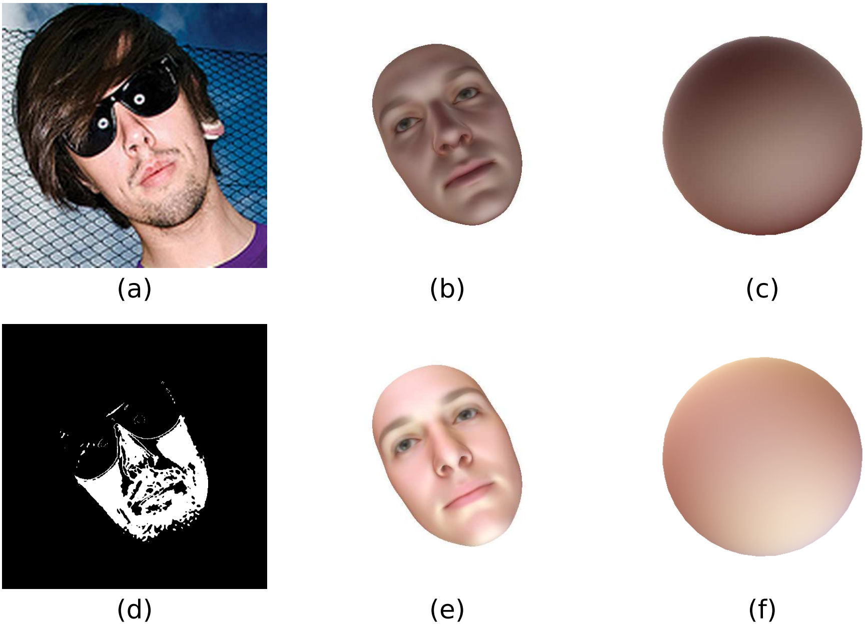


Figure 5.1: The target image (a) contains strong occlusions through sunglasses and facial hair. Non-robust illumination estimation techniques ([Schönborn et al., 2016]) lead to wrong illumination parameters under those occlusions. Non-robustly estimated illumination rendered on the mean face (b) and on a sphere with average face albedo (c). The sphere provides a normalized rendering of the illumination condition. The result obtained with our robust illumination estimation technique (e) and (f). The white pixels in (d) are pixels selected for illumination estimation by our robust approach. The target image is from the AFLW database ([Köstinger et al., 2011]).

like glasses or facial hair which are omnipresent in face images. We build our algorithm on the concept of robust random sample consensus algorithms (RANSAC, [Fischler and Bolles, 1981]). To estimate the illumination, the algorithm needs a model to generate face images with an arbitrary parametric illumination model. We use the mean face of the BFM, as prior of face shape and albedo. The shape and appearance prior is rendered under a spherical harmonics illumination model and a pinhole camera. The spherical harmonics illumination model efficiently parametrizes an environment map and is able to represent complex illumination conditions.

The proposed robust illumination estimation needs a rough pose estimation to start from. The output of the algorithm is a set of illumination parameters and a set of pixels used for the estimation (consensus set). Both outputs can be used for further analysis in our Analysis-by-Synthesis setting. In our semantic Morphable Model framework, we need a robust initialization of the illumination condition. The proposed method not only provides this initialization but also an initial label z to exclude occlusions and outliers from the model adaptation. Occlusions are however hard to determine in the beginning of the face model adaptation due to the strong influence of illumination on facial appearance. The estimated illumination is integrated into the model adaptation process and the consensus set is used to initialize the label z (see Chapter 4). Alternatively, the algorithm can be applied for illumination normalization or relighting as proposed in [Shahlaei et al., 2016].

The AFLW face database provides "in the wild" photographs under diverse illumination settings. We estimate the illumination conditions on this database to obtain an unprecedented prior on natural illumination conditions. The obtained prior is made publicly available.

The proposed prior closes a gap in generative modeling and applies to a wide range of applications. It can be integrated in probabilistic image analysis frameworks like [Schönborn et al., 2016] or [Kulkarni et al., 2015]. Furthermore, the resulting illumination prior can improve discriminative methods which aim to be robust against illumination. This is especially helpful for data-greedy methods like deep learning. Those methods are already including a 3DMM as prior for face shape and texture to augment ([Jourabloo and Liu, 2016; Zhu et al., 2016]) or synthesize ([Richardson et al., 2016; Kulkarni et al., 2015]) training data and could profit from using the proposed illumination prior. Currently, no illumination prior learned on real world data is available. The proposed illumination prior is an ideal companion of the 3DMM and allows the synthesis of more realistic images.

5.1 Illumination Estimation

The main requirement for our illumination estimation method is robustness against occlusions. The idea is to find the illumination setting which most consistently explains the observed face in the image. On a set of points with known albedo and shape, the illumination condition can be estimated. Uncertain shape and albedo as well as occlusions and outliers render this task ill-posed. Non-robust illumination techniques are misled since their observed color is not consistent to the object to analyze. The selection of points used for estimation shall not contain outliers or occlusions. We use an adapted RANSAC algorithm which adapts a generative model to the target image and is specially designed to handle outliers. We synthesize illumination conditions estimated on randomly sampled point sets to find the illumination parameters most consistent to the target image. The following steps of our procedure are visualized in Figure 5.2 and written in Pseudo-Code in Algorithm 5.1.

The idea of the RANSAC algorithm is to iteratively find a set of points which generalizes well to the observed target image. In each iteration we randomly select a set of points on the surface, which are visible in the target image $\tilde{I}$, and estimate the illumination parameters L_{lm} from the observed color values of those points (step 1 and 2). The quality of the estimated illumination is then evaluated on all available points (step 3). The full set of points consistent with this estimation is called the consensus set. Consistency is measured by counting the pixels of the target image which are explained well by the current illumination setting. If the consensus set is large enough the illumination is re-estimated on all points from the full consensus set for a better approximation. If this illumination estimation is better than the last best estimation according to a quality measure, it is set as the current best estimation. At the end the algorithm holds a set of points best approximating the observed illumination as well as a consensus set of points which can be explained by the resulting illumination estimation.

We calculate how well a set of illumination parameters reconstructs the target image $\tilde{I}$, to measure the quality of this estimate. We measure the color likelihood of the rendered face under the estimated illumination parameters $I(L_{lm})$ at each pixel i :

$$\ell_{\mathrm{L}}(L_{lm};\tilde{I}_i) = \frac{1}{N}\exp\left(-\frac{1}{2\sigma^2}\left\|\tilde{I}_i - I_i(L_{lm})\right\|^2\right) \tag{5.1}$$

We adapted the sampling of points on the face surface (step 1) by adding domain knowledge. There are some occlusions which often occur on faces like facial hair or glasses. Therefore, we include a region prior to sample more efficiently in suitable regions which arise at face regions which have low variance in appearance. Details on how to obtain this prior are found in Section 5.3.

Most "in the wild" face images contain compression artifacts (e.g. from the jpeg file format). To reduce those artifacts and noise we blur the image in a first step with a Gaussian kernel ($r = 4$ pixels with an image resolution of $512 * 512$ pixels).

Algorithm 5.1: Robust Illumination Estimation.

```
Input: Target image $\tilde{I}$, surface normals $\vec{n}$, albedo $a^c$, iterations $m$,
       number of points for illumination estimation $n$, threshold $t$,
       minimal size $k$ of consensus set, rendering function $I$
Output: Illumination parameters $L_{lm}$, consensus set $c$

$c = \varnothing$
for $m$ iterations do
    1. Draw $n$ random surface points $p$
    2. Estimate $L_{lm}$ on $p, \vec{n}$, $a^c$ and $\tilde{I}$ (Eq. 5.2)
    3. Compare $I(L_{lm})$ to $\tilde{I}$ (Eq. 4.6). Pixels consistent with $L_{lm}$
    ($\ell_{\text{L}}(L_{lm}; \tilde{I}_i) > t$), build $c$.
    4. if $|c| > k$ then
        Estimate illumination on $c$
        Save $c$ if $L_{lm}$ is better than previous best
```

5.2 Illumination Model

The proposed algorithm is not limited to a specific illumination model. The main requirement is, that it should be possible to estimate the illumination from a few points with given shape, albedo and appearance. Using the spherical harmonics illumination model has two main advantages. First, it is able to render natural illumination conditions by approximating the full environment map. Second, illumination estimation from a set of points corresponds to solving a system of linear equations.

Spherical harmonics allow an efficient representation of an environment map with a small set of parameters L_{lm} ([Ramamoorthi and Hanrahan, 2001; Basri and Jacobs, 2003]). This leads to an expressive illumination model

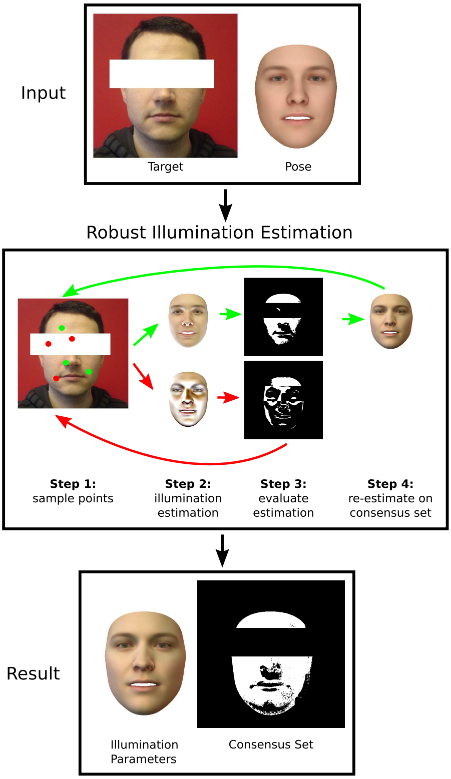


Figure 5.2: Robust illumination estimation: Our RANSAC-like algorithm (compare Algorithm 5.1) takes the target image and a pose estimation as input. We added strong occlusion (white bar) for better visualization. The algorithm iteratively samples points on the face surface (step 1). We estimate illumination from the appearance of those points in the target image (step 2 and 5.2). The estimation is then evaluated by comparing the model color to the target image (step 3 and 5.1). The illumination is mislead by including the occluded regions (red points). Choosing good points (green points) leads to a reasonable estimation. For good estimates we re-estimate the illumination on the full consensus set. We repeat the estimation on random point sets and choose the most consistent one as a result.

which is able to approximate complex environment maps in a parametric model and therefore suitable in our generative and parametric setting. The radiance function is parametrized through real spherical harmonics basis functions Y_{lm}. The radiance p_j^c per color channel c and for every point j on the surface is calculated from its albedo a_j, surface normal $\vec{n}_j$ and the illumination parameters L_{lm}:

$$p_j^c = a_j^c \sum_{l=0}^{2} \sum_{m=-l}^{l} Y_{lm}(\vec{n}_j) L_{lm}^c \alpha_l. \tag{5.2}$$

The expansion of the convolution with the Lambert reflectance kernel is given by α_l, for details, refer to [Basri and Jacobs, 2003]. We use Phong shading and interpolate the intensities at each pixel. Because the light model is linear (5.2), the illumination expansion coefficients L_{lm}^c are estimated directly by solving a linear system (least squares) with given geometry, albedo and observed radiance as described by [Zivanov et al., 2013]. The system of linear equations is solved during the RANSAC algorithm using a set of points.

5.3 Region Prior

Facial regions differ in appearance and elicit strong variations. Whilst some regions like the eyebrows or the beard region vary strongly between different faces, other regions like the cheek are more constant. Also, common occlusions through glasses or beards strongly influence facial appearance. Regions with low appearance variation are more suitable for illumination estimation than those showing stronger variation. We restrict the samples in the first step of the RANSAC algorithm to the most constant regions. The regions with strong variation are excluded from sample generation (step 1) but included in all other steps.

We estimate texture variation on the Multi-PIE database ([Gross et al., 2010]). It contains faces with glasses and beards under controlled illumination. We select images with frontal pose (camera 051) and with frontal, almost ambient illumination (flash 16) from the first session. With this subset we exclude all variation in illumination and pose (which we model explicitly) from our prior. The variation is estimated on all 330 identities. The images are brought into correspondence to the face model surface by adapting the BFM to each image with the approach by [Schönborn et al., 2016]. For the first step of the illumination estimation algorithm, we use the regions of the

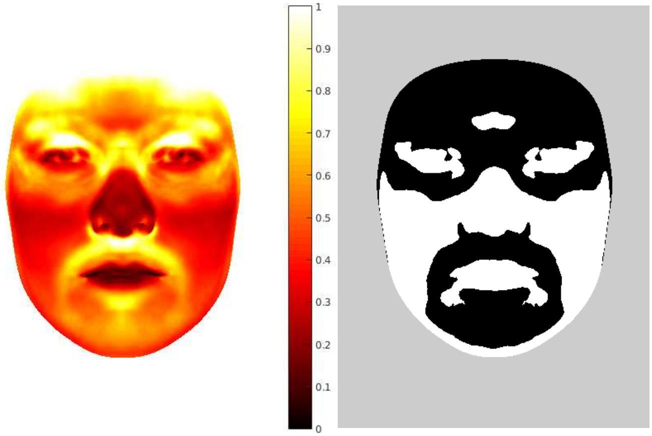


Figure 5.3: We derive a region prior from the average variance of appearance over all color channels per vertex on the face surface (on the left). Scaling is normalized by the maximal observed variance. On the right, the mask obtained by thresholding at half of the maximal observed variance. We use the white regions to sample points in the first step of our algorithm. Note that especially multi-modal regions potentially containing facial hair or glasses are excluded.

face where the texture variation is below one half of the strongest variation. The variation and the resulting constant region are depicted in Figure 5.3.

5.4 Illumination Prior

The idea of our illumination prior is to learn a distribution of natural illumination conditions from "in the wild" face images. There are various areas of application for such a statistical prior. It can be directly integrated in generative approaches for image analysis or be used to synthesize training data for discriminative approaches.

For faces, the 3DMM provides a prior distribution for shape and color but does not contain a prior on illumination. We therefore estimate an illumination prior learned on real world illumination conditions. In [Egger et al., 2017c], we publish the raw data of our illumination prior as well as the estimated multivariate normal distribution of all 27 concatenated spherical harmonic parameters L_{lm}^{c} of the first three bands and color channels. The prior allows us to generate realistic random illumination settings from the prior distribution.

5.5 Experiments

To evaluate the quality and robustness of the proposed illumination estimation, we performed our experiments on faces with a spherical harmonics illumination model. We use synthetic data to evaluate the sensitivity against occlusion. Our algorithm assumes a given pose, we investigate the errors introduced by this pose estimation. We exclude appearance variations by using the mean face shape and texture. We also examine the sensitivity to shape and texture changes. We show the performance of our method on real world face images in a qualitative experiment. And last, we present a novel illumination prior learned on empirical data.

We rely on the mean face of the BFM ([Paysan et al., 2009]) as face shape and texture for all experiments. We estimate the illumination parameters on $n = 30$ points (step 2). We use $\sigma = 0.043$ estimated on how well the BFM is able to explain a face image ([Schönborn et al., 2016]) and threshold the points for the consensus set at $t = 2\sigma$. We estimate the illumination on the full consensus set if the consensus set contains more than $x = 40\%$ of the surface points visible in the rendering. We stop the algorithm after $m = 500$ iterations.

On synthetic data, we measure how robust our algorithm is against occlusions. We also investigate how robust the algorithm is against pose misalignments and how much our simplified shape and texture prior influences the result. We need ground truth albedo and illumination, since there is a lack of a database providing this, we generate synthetic data. We use the mean shape and texture from the BFM as object and render it under 50 random illumination conditions. We randomly generate spherical harmonics illumination parameters L_{lm} according to a uniform distribution between -1 and 1.

5.5.1 Robustness against Occlusions

For the first experiment, we add synthetic random occlusion to this data. The random occlusion is a block with a random color. For the proposed RANSAC algorithm, those occlusions by large blocks of uniform color depict a worst-case scenario, since they consistently vote for wrong illumination parameters. Synthetic occlusions are positioned randomly on the face. An example of the synthesized data is depicted in Figure 5.4. We estimate the illumination

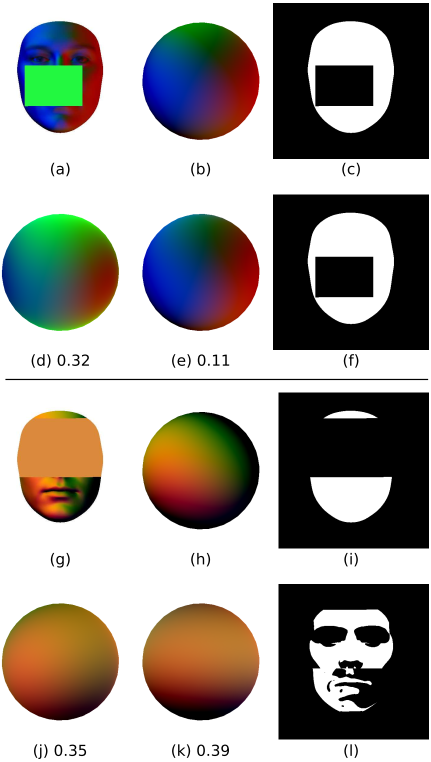


Figure 5.4: Two examples of our synthetic data to measure the robustness against occlusions of our approach. The target image is shown in (a, g) and the ground truth illumination for comparison (b, h). The ground truth occlusion map is rendered in (c, i). The baseline illumination estimation estimated on 1000 random points is shown in (d, j). Our robust illumination estimation result (e, k) as well as the estimated mask is shown in (f, l). Together with the visual result, we indicate the measured RMS-distance on the rendered sphere in color space. In the successful case (a-f), the consensus set is perfect and the occlusion is excluded from illumination estimation. In the failure case (g-l), the chosen occlusion is similar in color appearance to the observed face appearance. This leads the best consensus set to fail in explaining the correct region of the image. The first example (a-f) is a synthesized target image with 30% of the face occluded. The second example is with 60% occlusion (g-l).

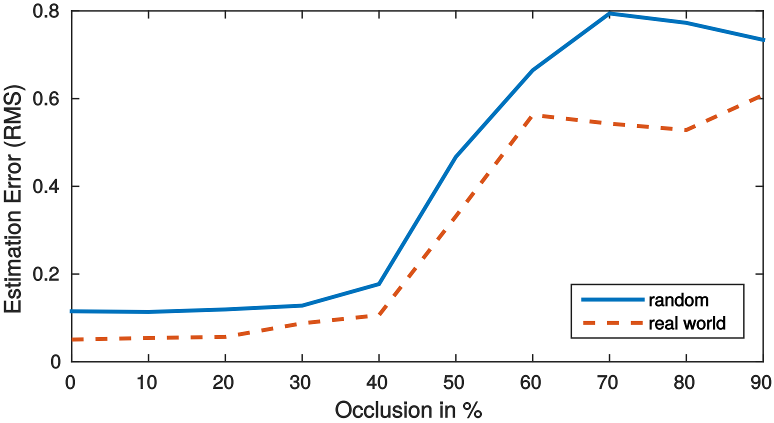


Figure 5.5: We measured the illumination estimation error of our approach for different degrees of occlusion. We compare randomly generated illumination conditions with those arising from real world settings. We observe that our algorithm is robust up to 40% of occlusion.

condition on this data using our robust illumination estimation technique to measure the robustness. We measured the approximation error by measuring RMS-distance in color space of the sphere rendered under the estimated illumination condition and the sphere with the ground truth illumination condition, as proposed by [Barron and Malik, 2015].

We cope with up to 40% of occlusion and reach a constantly good estimation, see Figure 5.5. Occlusions which surpass 40%, and those which can partially be explained by illumination(see Figure 5.6), are not properly estimated by our algorithm, see Figure 5.4. The generated illumination conditions are unnatural, therefore we also evaluated the robustness against occlusion on observed real world illumination conditions (Section 5.5.4). Every measurement is the average over 50 estimations.

5.5.2 Robustness against Pose Estimation Error

Our algorithm relies on a given pose estimation. Pose estimation is a problem which can only be solved approximatively. We therefore show how robust our algorithm is against small misalignments in pose. We again generate synthetic data with random illumination parameters and manipulate the pose before we estimate the illumination. The results are shown in Figure 5.7. We present

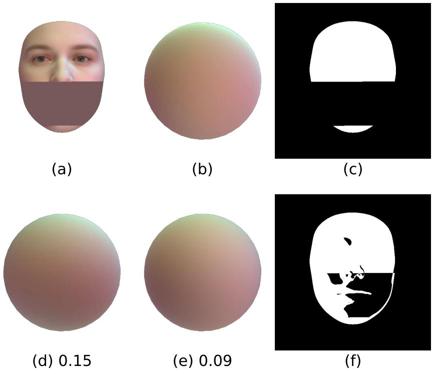


Figure 5.6: An example of wrong occlusion estimation. Whilst the illumination estimate is close to the ground truth, the occluded region can partially be explained by the illumination condition and the color of the mean face. The arrangement is the same as in Figure 5.4.

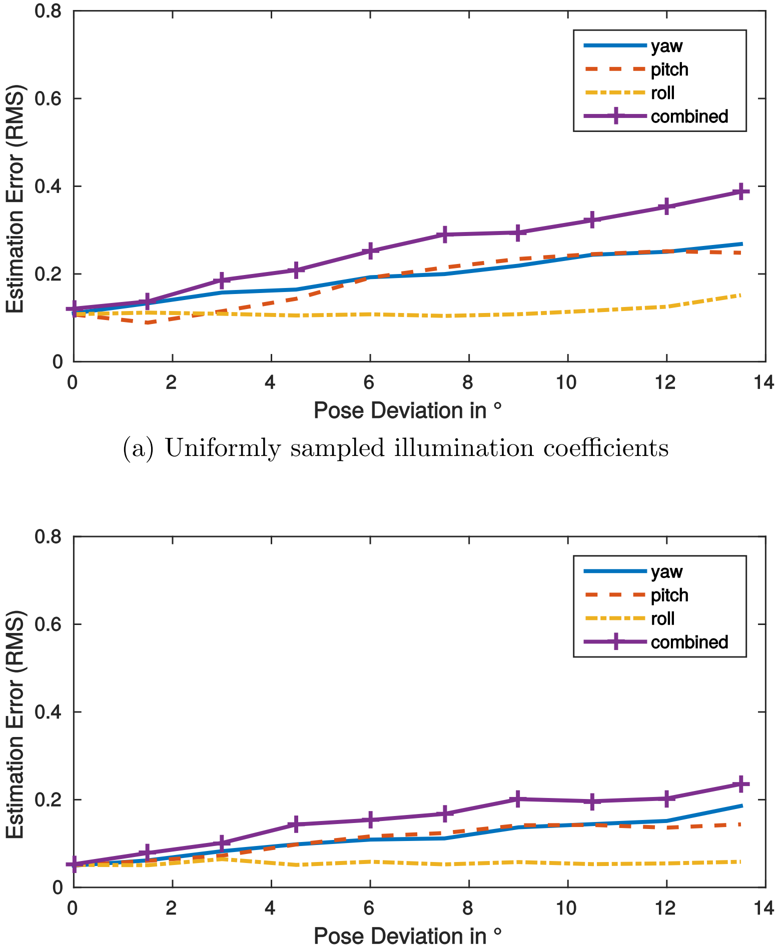
separate effects of yaw, pitch and roll angle as well as the effect of combining all three error sources. Small pose estimation errors still lead to good illumination estimations. As expected, they grow with stronger pose deviations. We have to expect errors smaller than 10 degrees from pose estimation methods (compare [Murphy-Chutorian and Trivedi, 2009]).

5.5.3 Robustness against Shape and Texture Variation

With the mean of the BFM, we use a simple prior for shape and texture for the proposed illumination estimation. In this experiment, we want to measure how shape and texture changes influence the illumination estimation. We therefore modify all shape and color parameters gradually. The result is presented in Figure 5.8. Under artificial illumination conditions we get good estimations of the true illumination condition even for stronger changes. For real world illumination conditions the ambiguity of color and illumination leads to wrong illumination estimates. We observe that small variations influence the illumination estimation but do not break it.

5.5.4 Illumination Estimation “in the wild”

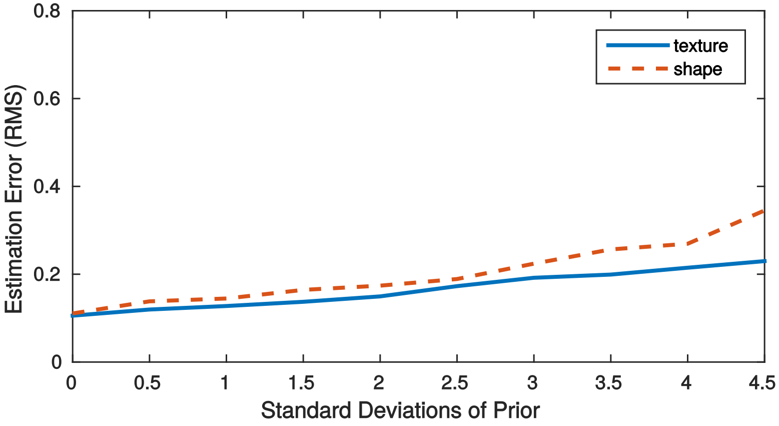
We applied our robust illumination estimation method on the AFLW database ([Köstinger et al., 2011]) containing 25’993 images with a high variety of pose



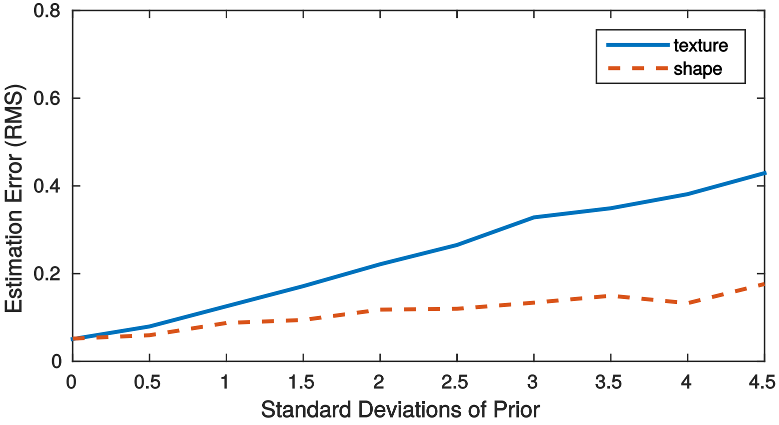
(a) Uniformly sampled illumination coefficients

(b) Illumination coefficients sampled from Prior

Figure 5.7: We measured the illumination estimation error of our approach related to pose estimation errors. Our algorithm handles pose deviations which arise by wrong pose estimation input. We compare the result for synthetic random illumination conditions (a) as described and illumination conditions from our illumination prior (b). We observe that illumination estimation on real world illumination conditions is less sensitive to pose estimation errors.



(a) Uniformly sampled illumination coefficients



(b) Illumination coefficients sampled from Prior

Figure 5.8: We measured the illumination estimation error of our approach related to shape and texture changes. Even with the mean of the BFM as simple appearance prior, we reasonably estimate the illumination condition. We compare the result for synthetic random illumination conditions (a) as described and illumination conditions from our illumination prior (b). We observe real world illumination conditions to be much more sensitive to changes in facial texture and less sensitive to changes in shape than purely synthetic samples.

and illumination. The provided landmarks were used for a rough pose estimation following the algorithm proposed by [Schönborn et al., 2016]. The illumination conditions are not limited to lab settings but are complex and highly heterogeneous. We observe that small misalignments due to pose estimation still lead to a reasonable illumination estimation. The affected pixels, e.g. in the nose region, are automatically discarded by the algorithm. Both, estimated illumination and occlusion mask arising from the consensus set can be integrated for further image analysis as described in Chapter 4. We present a selection of images under a variety of illumination conditions with and without occlusions in Figure 5.10. The illumination estimation results demonstrate robustness of our approach against occlusions like facial hair, glasses and sunglasses in real world images.

5.5.5 Illumination Prior

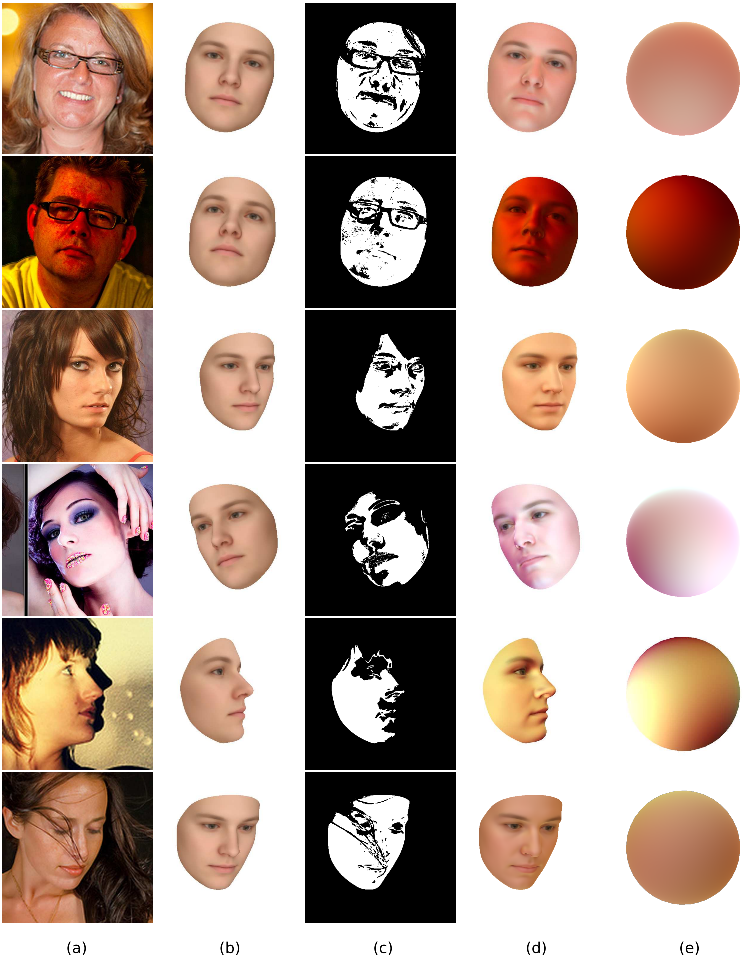
To derive an illumination prior, we again chose the AFLW database in the same experimental setting as described before. It contains a high variety of illumination settings. We excluded gray-scale images and faces which do not match our face model prior (strong make-up, dark skin, strong filter effects). We manually excluded images where the estimation failed and used the remaining 14'348 images as training data.

The obtained illumination estimations depict an empirical illumination distribution. We estimate a multivariate normal distribution to get a parametrized representation and present the first eigenmodes applying PCA in Figure 5.11. We also generate some new unseen random illumination conditions in Figure 5.12.

5.6 Limitations

The limitations of this illumination prior and the robust illumination estimation are a direct consequence of the used spherical harmonics illumination model. We did not incorporate specular highlights or effects of self-occlusion explicitly to keep the model simple. This simplification is not critical, since regions which are sensitive to self-occlusion or contain specular highlights are excluded from the illumination estimation during our robust approach (see Figure 5.10).

The main limitation of our model arises from using the mean face of the BFM as very simple prior for facial color. Everything that can be explained by the illumination model is explained by it using the proposed algorithm.



This Figure proceeds on the next page.

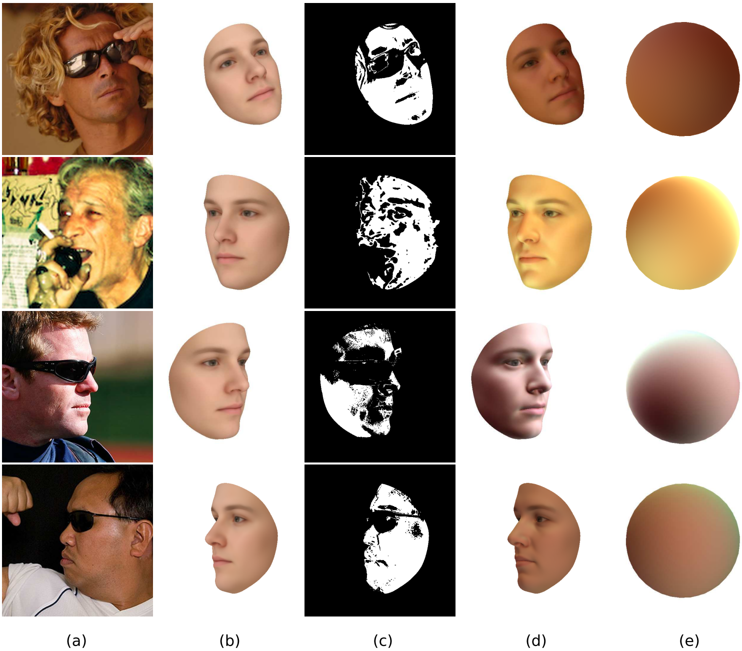


Figure 5.10: A qualitative evaluation of the illumination estimation on the AFLW database ([Köstinger et al., 2011]). We show the target image (a), pose initialization (b), the consensus set of the RANSAC algorithm (c), the mean face of the BFM rendered under the estimated illumination condition (d) and our normalized representation (e). We observe that glasses, facial hair and various other occlusions are excluded from illumination estimation. At the same time, we see minor limitations: things that are not well explained by our simplified illumination model, like specular highlights and cast shadows or strong variations in facial appearance, e.g. in the mouth, eye or eyebrow region. Affected regions do not mislead the illumination estimation but are excluded by our robust approach. More examples on the previous page.

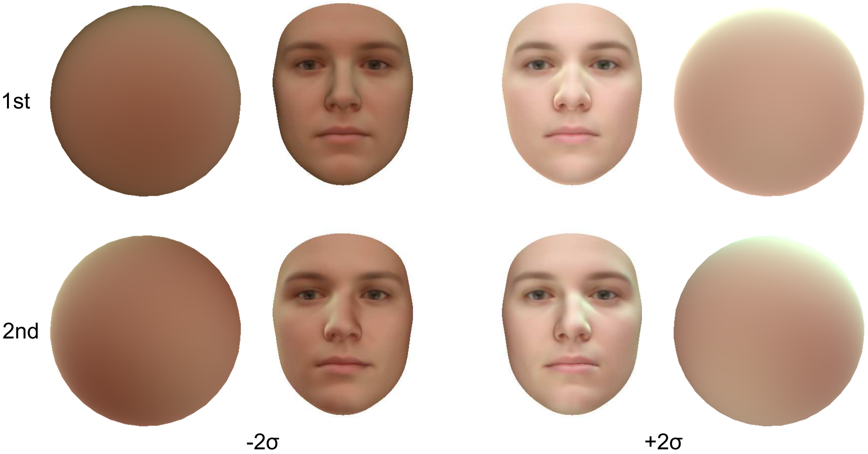


Figure 5.11: The first two eigenmodes of our illumination prior. The first parameter represents luminosity. From the second eigenmode we see that illuminations from above are very prominent in the dataset. The illumination conditions are rendered on the mean face of the BFM and a normalized representation.

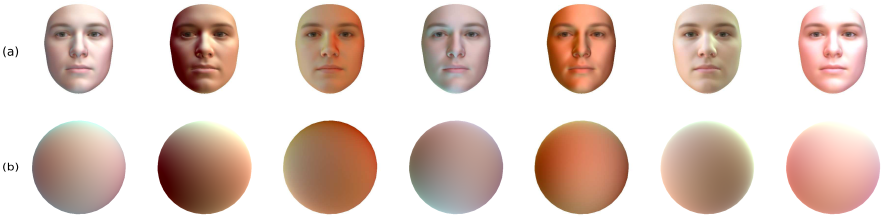


Figure 5.12: Random samples from our illumination prior represent real world illumination conditions. The proposed prior represents a wide range of different illumination conditions. The samples are rendered with the mean face of the BFM (a) and a normalized representation (b).

Mainly the global skin tone is therefore explained by the illumination components and not by the color model of the face. The problem arises from an ambiguity of color and illumination discussed in the next section. The same limitation affects all previous 3DMM adaptation frameworks and is not specific to our approach.

5.6.1 Color-Illumination Ambiguity

Generative face image analysis is limited by some prominent ambiguities. The recent work of [Smith, 2016] investigates the perspective face shape ambiguity and gives an overview over the relevant ones for the 3D reconstruction task.

We here illustrate the color-illumination ambiguity since it is strongly connected to the task of illumination estimation. In an image, albedo and illumination appear as a product which makes them indistinguishable. The global skin-tone of a face can well be explained by illumination and the near-ambient part of illumination can be explained by color. This renders the Analysis-by-Synthesis setting problem ill-posed. It is not possible to strictly distinguish between both components of appearance. The arising image explanation is in our case a mixture of illumination and color and not necessarily the correct or even close to the correct one.

To demonstrate this ambiguity, we performed several small experiments based on our model adaptation framework focusing on color and illumination parameters. In a first experiment, we generate synthetic images with a fixed spherical harmonics illumination or fixed color model components. We then try to explain the generated image completely by the other component as shown in Figure 5.13 and Figure 5.14. For the color parameters adaptation we run 10'000 random color proposals to adapt to the target image and kept the illumination ambient. For illumination estimation the color parameters stay constant and the illumination was directly estimated from all visible pixels.

In a second experiment, we start from a reasonable fitting result estimated from a synthetic target image and draw model instances from the posterior of suitable model instances changing only color and illumination parameters. To explore the ambiguity we use decorrelated combined color and illumination proposals as introduced by [Maggi, 2014]. The proposal first changes the color parameters and then updates the illumination parameters optimally. We chose the collective image likelihood for drawing real samples from the posterior as described in [Schönborn et al., 2016]. The resulting samples in Figure 5.15 visualize similar reconstructions of the target image with entirely different combination of shape and texture.

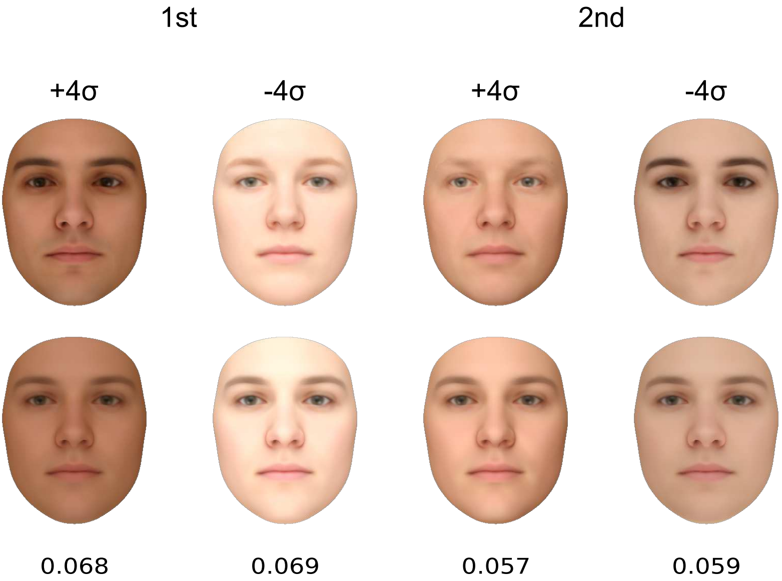


Figure 5.13: Color explained by illumination: The first row shows the first two principal components of the BFM color model rendered under ambient illumination. The second row shows the mean color of the BFM rendered under the illumination condition best reconstructing above image. We observe the global skin-tone to be well explained by adapting the illumination condition, whilst details like the eyebrows are not explained well. The values under the reconstruction depict the RMS distance of the reconstruction by illumination to the original, the values are directly comparable to Figure 1.4. The observed instances appear similar and also the measured distances are small.

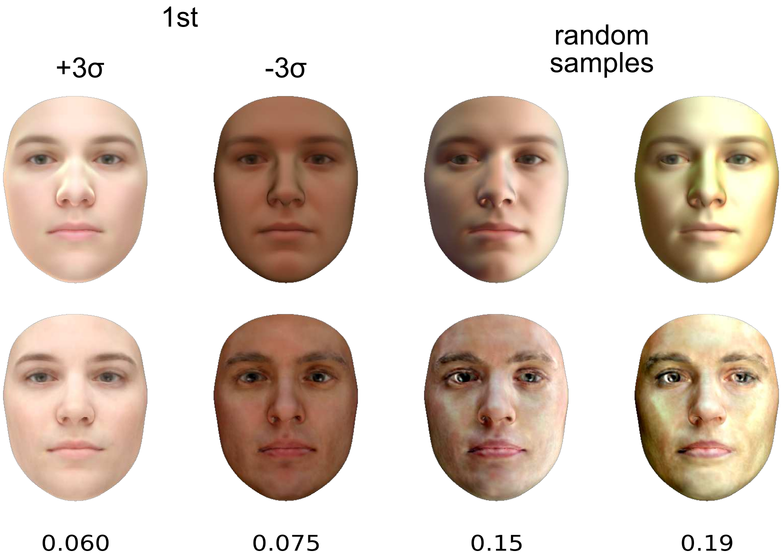


Figure 5.14: Illumination explained by color: The first row shows the first principal component and two random samples of our illumination prior described in Section 5.5.5, as texture we used the mean color of the BFM. The second row shows how well the color model can reconstruct the above illumination conditions rendered under ambient illumination. We observe the global color of the illumination to be well explained by the facial color model, complex illumination as in the random samples can not be reconstructed as nicely by the color model. The values under the reconstruction depict the RMS distance of the reconstruction by illumination to the original, the values are directly comparable to Figure 1.4.

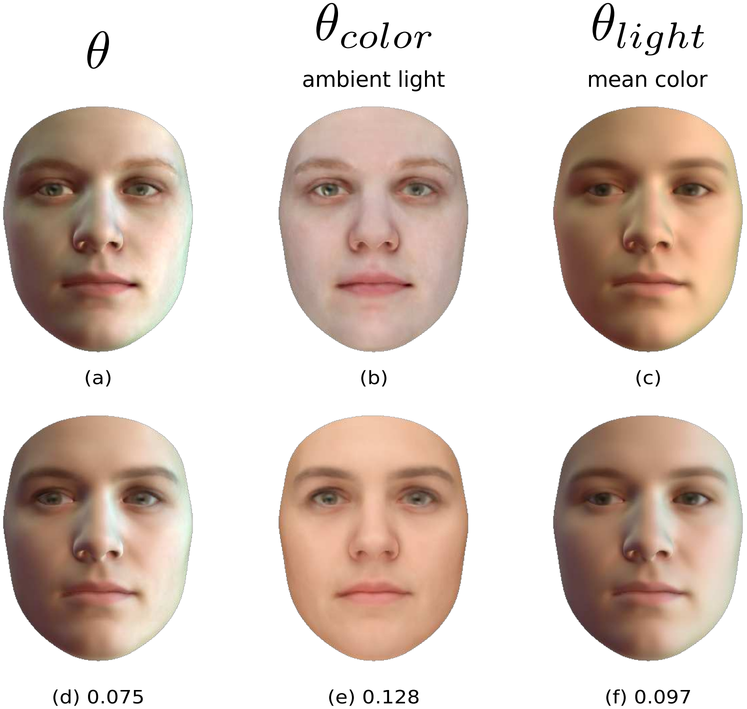


Figure 5.15: The target image in the first row (a), its color rendered under ambient illumination (b) and its illumination rendered on the mean color of the BFM (c). The second row shows a model instance with different color (e) and illumination (f) parameters but very similar appearance (d). The value under the reconstruction (d-f) depict the RMS distance of the reconstruction to the upper images (a-c), the values are directly comparable to Figure 1.4.

The only approach for such ambiguities is strong prior knowledge, but even with a perfect prior multiple explanations can be equally likely. In our case, we have the BFM as a strong prior for the facial color model and the proposed illumination prior. The illumination prior is however estimated on real world images using the mean color of the BFM. The illumination prior therefore is prone to the ambiguity itself and will not resolve it. We weaken this limitation of the prior by including only illumination conditions estimated on skin tones near the mean skin tone of the BFM.

5.7 Conclusion

We demonstrate that a simple prior for shape and texture is sufficient to acquire a useful estimation of illumination from a single image. We show in qualitative and quantitative experiments, that our approach is robust against small pose misalignments, changes in shape and texture and handles up to 40% of occlusion. The proposed occlusion-aware illumination estimation is not limited to faces or the chosen illumination model. It can be applied on various objects and in combination with different parametric illumination models.

We apply our algorithm on the AFLW database containing faces in a huge variety of scenes and under arbitrary illumination conditions. The resulting prior is highly applicable for probabilistic frameworks as well as data-greedy algorithms like deep learning methods for augmenting or generating data under unconstrained illumination conditions. Our illumination prior, from a broad range of real world photographs, is the first, publicly available.

Chapter 6

Application

We combine all the components presented in the previous chapters to apply our semantic Morphable Model to an “in the wild” attribute estimation task. The robust illumination estimation enables us to handle images under challenging illumination conditions and give a first estimation of occlusions. The semantic model adaptation gives face model parameters and segments the image into the semantic regions. The result by using a copula Morphable Model holds a 3D face reconstruction, the semantic image segmentation as well as an attribute based image description. An overview of the full application is depicted in Figure 6.1.

We present intermediate and full results of the framework on a variety of target images in Figure 6.2. We selected the images to contain various challenges for classical 3DMM adaptations. The images contain different kind of occlusions, illumination settings and beards. We also include images without occlusions since those should also be segmented correctly.

We use an external library for feature detection, namely the CLandmark Library made available by [Uřičář et al., 2015]. The detection result is not perfect and is integrated by taking this uncertainty into account as described in [Schönborn et al., 2013]. We chose this library, because it is able to still detect features even if the full face is not visible. The detector works for near-frontal poses. The landmarks are only used for initialization. During the semantic model adaption, they are not integrated anymore. This enables to refine the pose and feature positions during the Analysis-by-Synthesis process. If the feature point detection fails, it leads to a far off pose estimation and it is hard to recover during model adaptation. For robust illumination estimation we use the experimental setting presented in Chapter 5.5 and for the semantic

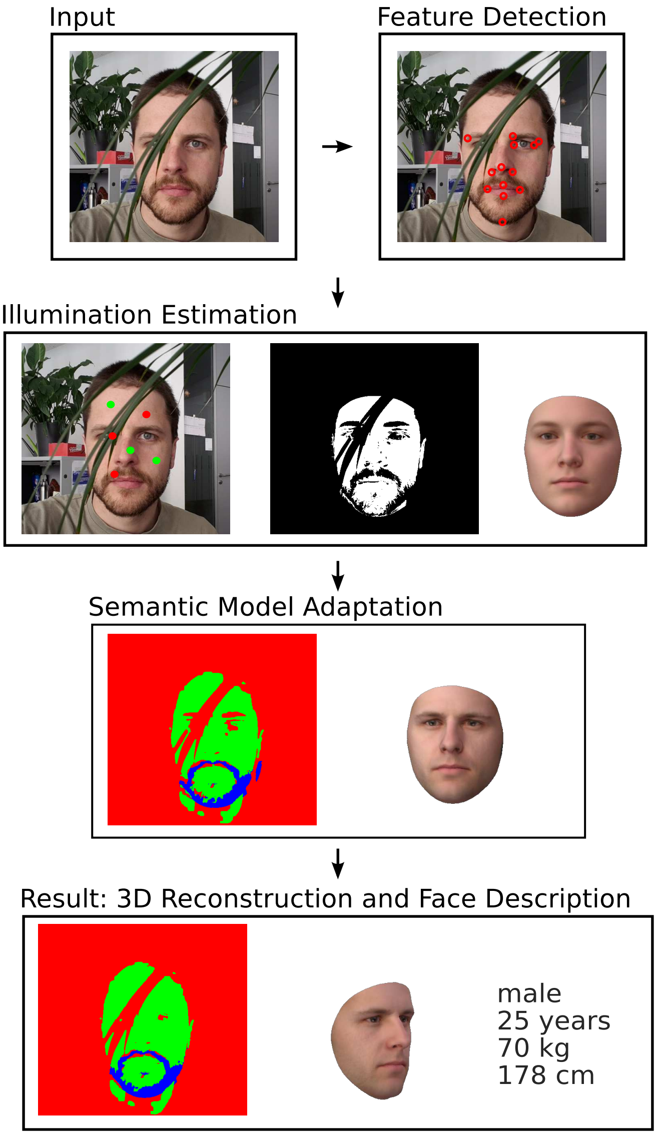

Input
Feature Detection
Illumination Estimation
Semantic Model Adaptation
Result: 3D Reconstruction and Face Description
male
25 years
70 kg
178 cm

Figure 6.1: A fully automatic image analysis pipeline built from all components proposed in this thesis. The input is a single still image containing a face and the output is an image segmentation, a 3D reconstruction of the face as well as an attribute-based face description.

Morphable Model we apply the setting presented in Chapter 4.4. As face model we use the copula Morphable Model presented in Chapter 3.7.2 to get an attribute-based face description.

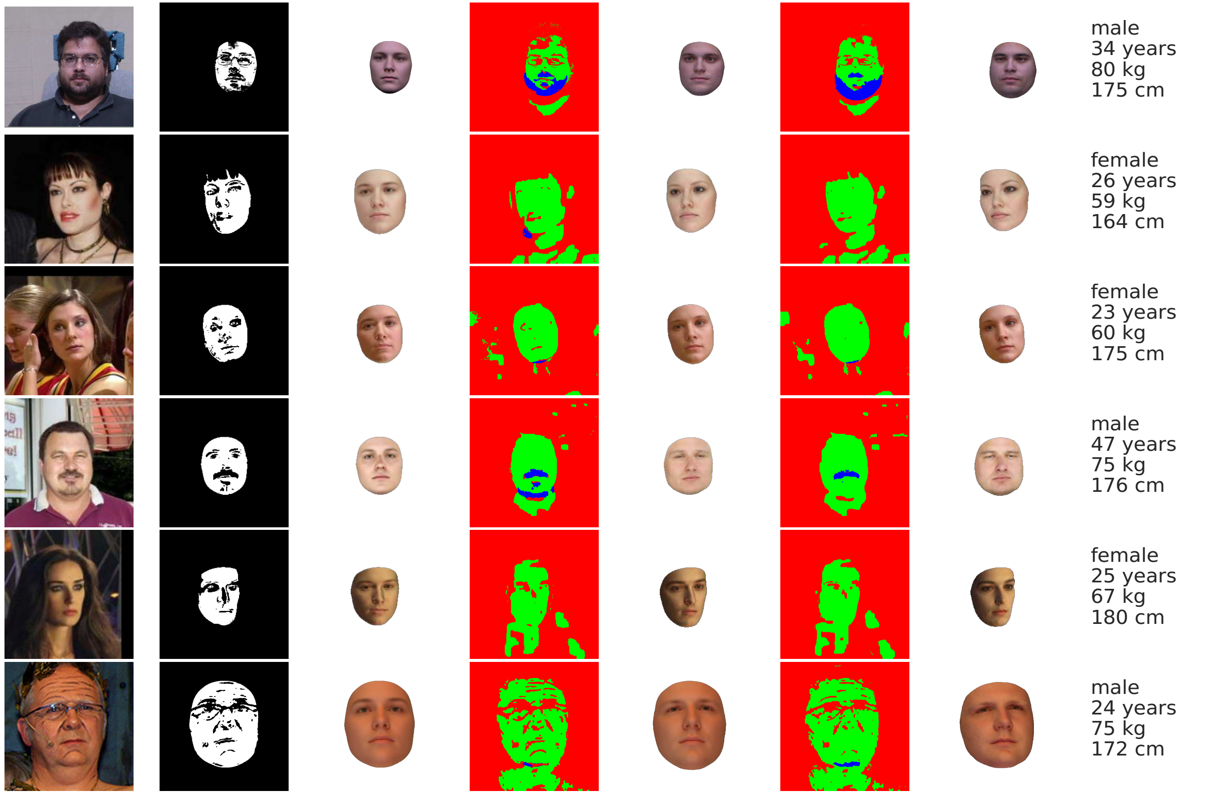
In the results presented in Figure 6.2 we observe the robust illumination estimation to give a good starting point for model adaptation. The result of the illumination estimation builds the initialization of the semantic model adaptation. It tends to exclude more than only occlusions and outliers. During the integrated segmentation and model adaptation the segmentation is improved, especially regions which are hard to explain by the face model.

The occlusion and background model covers all parts which can not be explained by the face and beard models. Some of those are real occlusions by various objects, others are facial details which are not contained in our face model like wrinkles or eye gaze. Those facial details are missed by our model explanation and could hold important information e.g. on the age of the face. Make-up and specular highlights are also not part of our face and illumination model and therefore covered by the non-face model.

The attribute estimation leads to results of different quality. Whilst the sex estimation results are reasonable and competitive, the age estimation is not accurate (compare Chapter 3.7.2 and Chapter 3.8). Elderly people are underrepresented in the data the model is built on and this leads to weak age estimation results. We assume the main cause for bad age estimation results are textural details like wrinkles, which are missed by our model and give a strong cue on age.

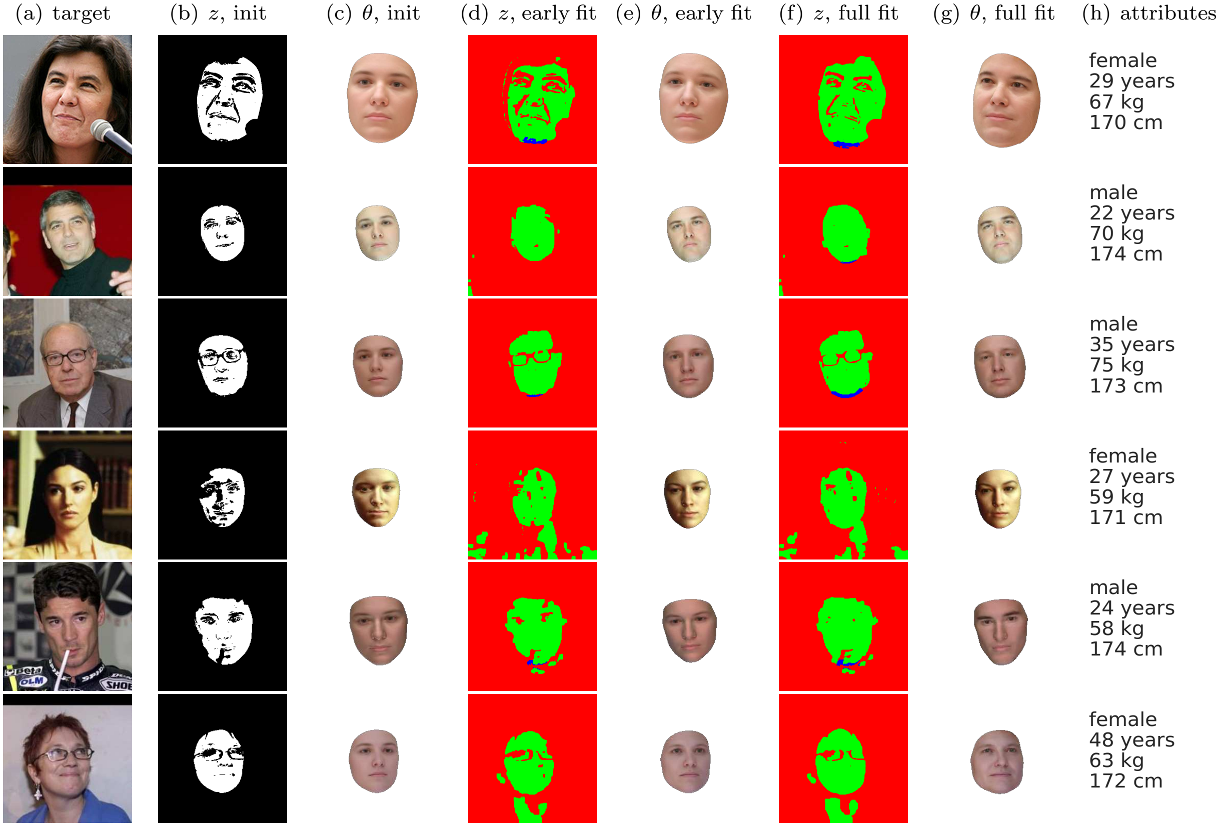
In this work we excluded facial expressions and assume neutral facial expressions. In [Egger et al., 2017b] we presented 3DMM adaptation with an expression model under occlusion. If the expressions are not explicitly handled they are excluded as outlier during model adaptation as observed in some of the shown results.

A main limitation of our method arises from the strong dependence on reasonable initialization. If feature point detection, pose estimation or illumination estimation fails, the semantic Morphable Model adaptation can not recover from this early wrong decision. In the proposed setting including semantic segmentation it is hard to avoid those early decisions. By stronger and more cues from detections, like bottom-up pose estimation or segmentation cues, the framework could highly profit (more details in Chapter 7.4).

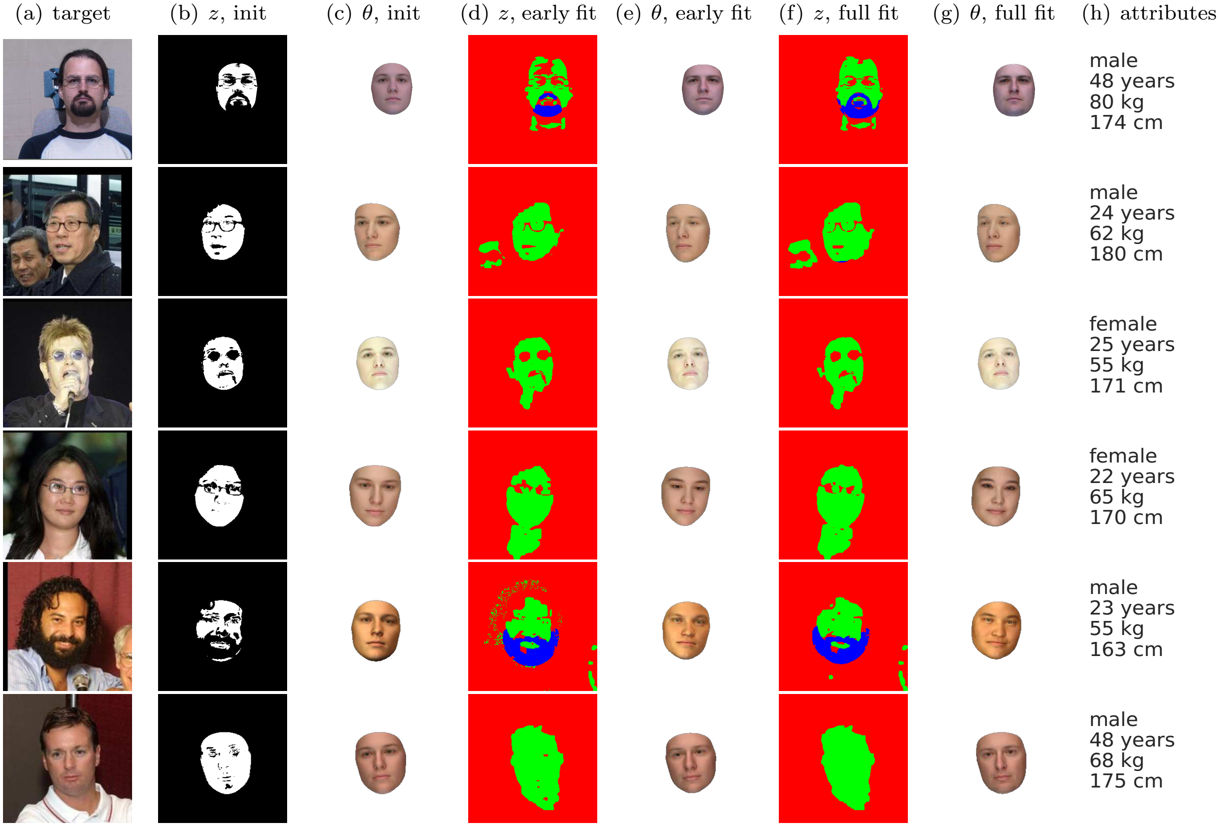


(a) target (b) z, init (c) θ, init (d) z, early fit (e) θ, early fit (f) z, full fit (g) θ, full fit (h) attributes

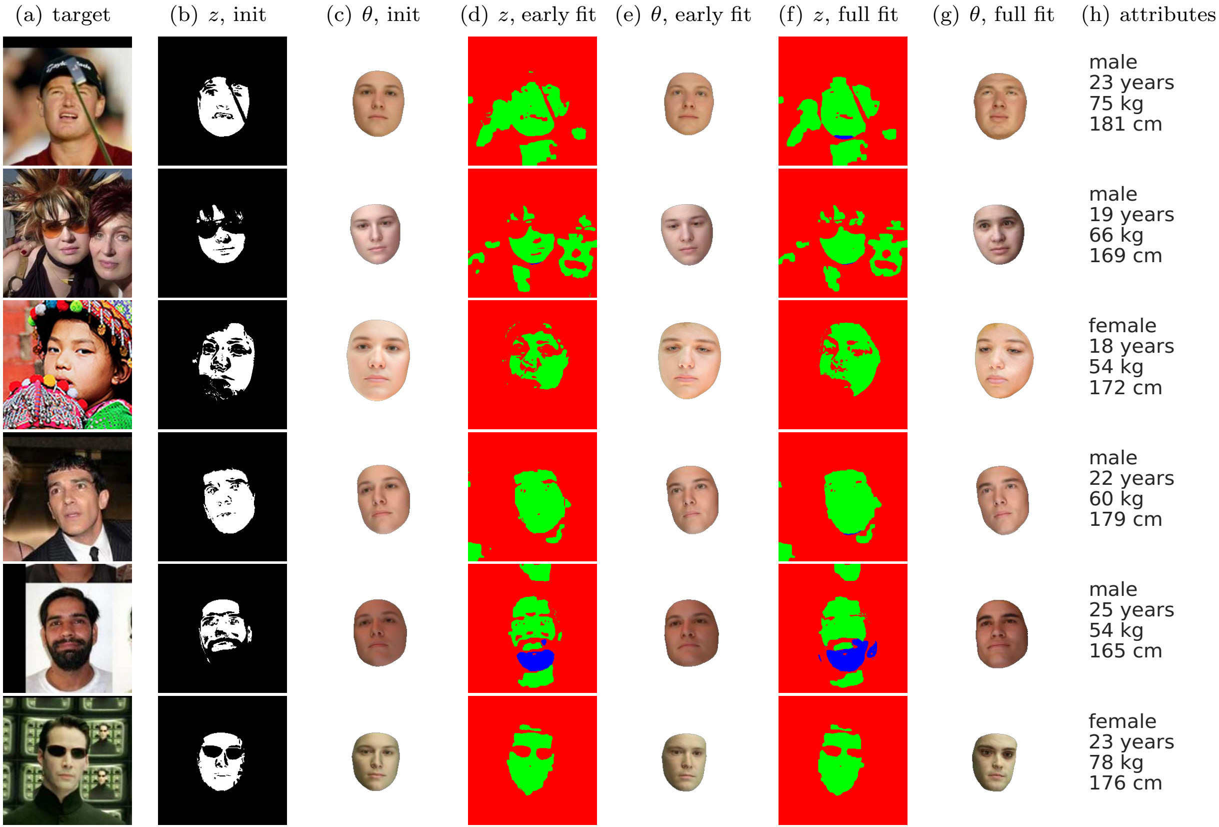
This Figure proceeds on the next page.


(a) target (b) z, init (c) θ, init (d) z, early fit (e) θ, early fit (f) z, full fit (g) θ, full fit (h) attributes

This Figure proceeds on the next page.



This Figure proceeds on the next page.



This Figure proceeds on the next page.

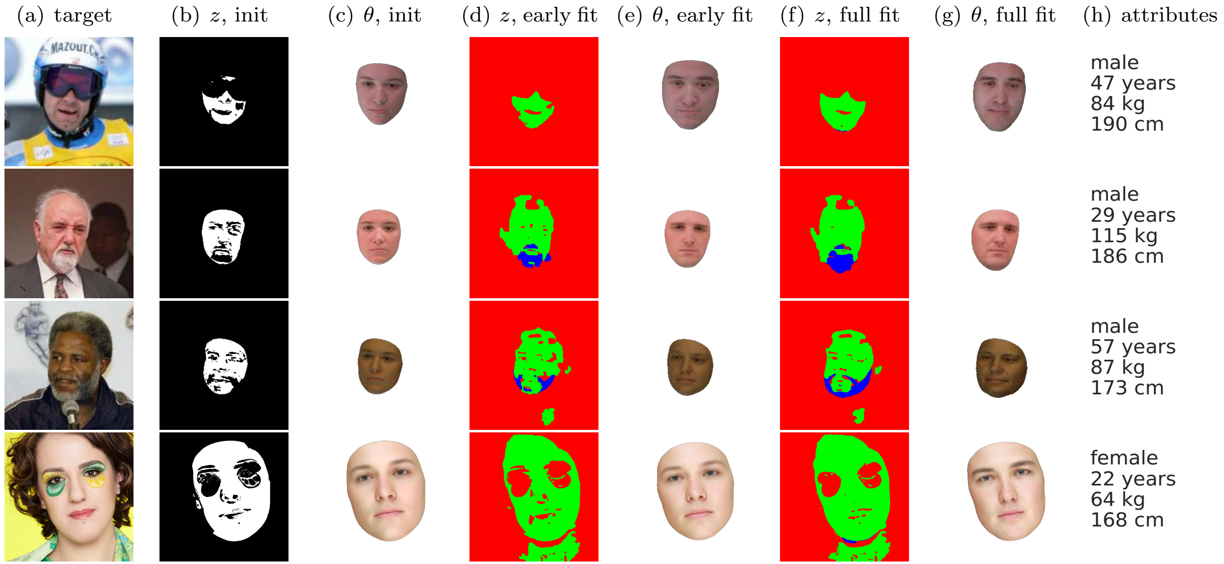


Figure 6.2: Results of our full semantic Morphable Model image analysis framework. The target image (a), the result of robust illumination estimation (b,c), the model parameters and segmentation after 1'000 sampling steps (d,e) and 10'000 sampling steps (f,g,h). The images are from the Multi-PIE ([Gross et al., 2010]), AFLW ([Köstinger et al., 2011]) and mainly LFW ([Huang et al., 2007]) face databases. More examples on the previous pages.

Chapter 7

Future Extensions

In this chapter we present ideas for future research in the field of face image analysis in an Analysis-by-Synthesis setting. The collected ideas are to overcome current limitations of state of the art research or to further investigate interesting findings of this PhD thesis. The first two ideas are focusing on modeling whilst the other two are centered in the context of model adaptation. Additional ideas for future research directions which are directly connected to one of the previous chapters, are mentioned in the corresponding chapter.

7.1 Texture Modeling

The 3DMM is strong in modeling facial shape but weak in modeling texture. Texture models based on PCA or COCA miss a lot of skin details like wrinkles, moles or freckles. Such details are not only necessary for photo-realistic rendering, but are also important for face analysis. Age, for example, is highly encoded in such textural details. Whilst state of the art rendering methods can produce astonishing facial renderings there is a lack of a parametric generative model which captures skin texture.

There are first approaches to improve the facial texture of face models for Eigenfaces or 3DMM based textures. The approach of [Mohammed et al., 2009] leads to photo-realistic facial portraits by a patch based approach. A similar method was used to post-process 3DMM instances to produce high resolution textures in [Dessein et al., 2015].

We applied a discriminative approach as post-processing after 3DMM adaptation to improve the face analysis in [Egger et al., 2014]. Texture sensitive features were used to capture the information the generative model misses in the image. [Pierrard and Vetter, 2007] showed unique face recognition results by solely performing recognition based on moles. This was again performed as post-processing after 3DMM adaptation since strong correspondence is necessary for this approach. To capture all the information encoded in a face with a generative model, a parametric and detailed texture model should be added to replace the simple PCA based model in the 3DMM.

7.2 Copula Morphable Models

Scale invariance is the main advantage of COCA over PCA for statistical modeling of faces. Scale invariance enables us to build a joint model of shape, color, attributes and other sources of facial data. This joint model is highly interesting for analysis and synthesis of faces. Whilst current posterior models as proposed in [Albrecht et al., 2013] are limited to a single modality, the copula enables posterior models over multi-modal data. We can explore posterior models for all involved parts of the model - e.g. we can build posterior models on attributes. We present posterior models for male and female faces in Figure 7.1.

Such posterior models can be applied in various applications. For face image analysis we could integrate knowledge on the individual or bottom-up detections to build a posterior model. This leads to a more specific face model for analysis. Whilst such information to build a posterior model is often not given or hard to detect in face image analysis tasks, this feature could be highly interesting in medical image analysis. When analyzing medical data there is often ground truth meta-data like sex, age, size and weight available, which could be used to build a patient-specific shape model for analysis.

Besides the posterior models, the multi-modal models allow a joint analysis of the underlying dependency structure. Analyzing the covariance matrix enables to investigate which parts of a face are influenced by which attributes or which parts of the face lead to certain ascribed attributes.

7.3 Color vs. Illumination

In Section 5.6.1 we discussed the color-illumination ambiguity. The human visual system can however distinguish between certain degrees of skin tone and effects caused by illumination. The necessary hints are encoded in both shape and texture. Textural cues like the reflection properties of skin are neglected by our approach but important for skin tone estimation.

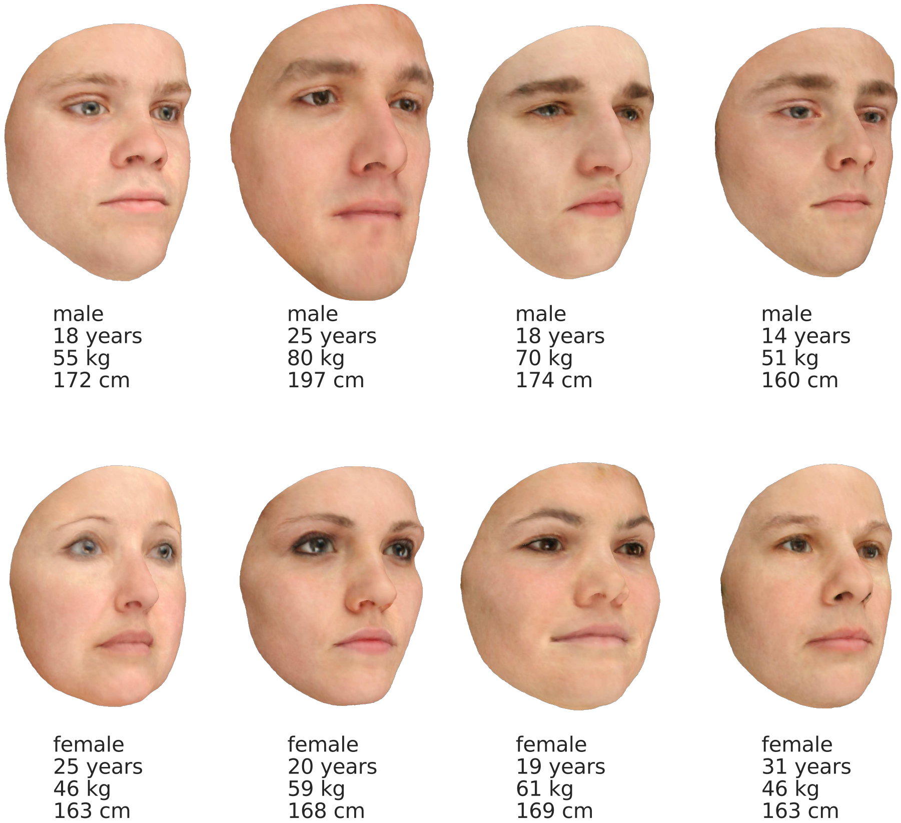


Figure 7.1: Random samples from posterior models constrained on male (top) respectively female (bottom).

One reason for this limitation arises directly from the BFM. It is built from mainly caucasian faces with similar skin tone and therefore under-represents other ethnic groups. Recently other face models where made publicly available which better represent the variety of different ethnic groups ([Huber et al., 2016] [Booth et al., 2016]).

Our work could contribute in several ways to better estimate illumination and skin-tone. First, the copula Morphable Model allows to explore the dependencies between color and shape. Second, the proposed illumination prior fills a gap for generative modeling, necessary to overcome this limitation. And third the semantic Morphable Model enables to include local models which could be useful to estimate the illumination locally from the eyes ([Nishino and Nayar, 2006]) or the sclera ([Do et al., 2006]).

7.4 Bottom-up Cues

The main limitation of our current approach on "in the wild" face images are missing image-related cues. The proposed model is dominated by top-down knowledge arising from the generative model. Bottom-up cues like detections are underrepresented in the implementation of the framework. Recently new methods like cascaded regression techniques ([Zhu et al., 2015; Huber et al., 2015]) or deep learning methods ([Tewari et al., 2017]) were proposed for 3DMM adaptation. Our model adaptation and segmentation algorithm is open to those techniques and could integrate them in a proposal step or into the segmentation.

In the current framework we draw a lot of samples in a random walk fashion following always the same proposal distribution. With strong bottom-up cues e.g. for the pose we could restrict the proposal distribution strongly. The proposed copula Morphable Model even enables us to integrate cues like bottom-up attribute prediction directly into the model adaptation through posterior models. Cascaded regression techniques or deep learning methods could be integrated as fast-forward proposals and profit from the verification step in the Metropolis-Hastings algorithm.

Bottom-up cues are not only helpful on the proposal side of the Metropolis-Hastings, but also during the verification step itself. At the moment the verification is performed using the prior from the face model and measuring the image difference in color space. The evaluation step would profit from additional likelihoods, e.g. to take edges and contours into account (compare [Bas et al., 2016]).

Besides from model adaptation, also the segmentation can profit from bottom-up cues. For some occlusions we could train detectors, like we explored for facial hair. Glasses for example are frequent occluders and could not only be modeled but also detected using bottom-up detector techniques. Recently a semantic face parsing based on deep learning was proposed and could be integrated as a segmentation proposal ([Liu et al., 2015]). The strong prior knowledge arising from the face model could then guide the detectors to focus on the correct regions.

Chapter 8

Conclusion

In this thesis, we proposed a framework for semantic and attribute-based face image interpretation. A face consists of different semantic regions like e.g. eyes, eyebrows, mouth or beard. During model-based image analysis those regions should not be confounded. Our semantic Morphable Model framework enables us to model different facial regions separately by different and specific models.

A lot of facial images do not fulfill model assumptions of the classical 3DMM and contain glasses, beards, make-up or various other occlusions. The semantic Morphable Model framework enables to incorporate different models to compete in explaining semantically connected parts of the image. This enables to reveal the face model adaptation from regions which can not be explained by the face. Such regions are covered by another more specific model or a model for background or occlusions. The framework is open for various models, our implementation is an example with a face model, a beard model and a non-face model. We integrated segmentation of the target image into semantic regions directly by a joint likelihood into the model adaptation process. The segmentation, the model parameters and attributes build the image description. The framework results in enhanced face model adaptations for “in the wild” face images outside the scope of previous 3DMM adaptation frameworks.

Semantic Morphable Model adaptation relies on robust illumination estimation in early steps of the adaptation process. We demonstrated the strong influence of illumination on model adaptation and illustrated the need for reliable and robust illumination estimation. We propose to use a simple but effective algorithm for robust illumination estimation and initialization of the semantic segmentation. This illumination estimation procedure was applied to an “in the wild” face database, leading to an illumination prior which fills a gap for generative models.

To obtain a attribute-based image-description, we built a copula Morphable Model which not only covers shape and color but also includes attributes. Finding the model instance for a target face directly leads to an attribute-based face description. The attribute-based face description is more comprehensible for humans than the description based on statistical parameters used by the classical 3DMM. We presented competitive results on sex estimation from 2D images. The copula Morphable Model fits well in current 3DMM frameworks, can be implemented easily and is a powerful extension of the underlying statistical model.

The overall framework combines all those parts and enables to analyze “in the wild” face images and describe them in an attribute-based and human understandable way. We discussed current limitations of our proposed implementation and suggest ideas to further develop the framework.

Appendix A

List of Abbreviations

3DMM	3D Morphable Model
AFLW	Annotated Facial Landmarks in the Wild
BFM	Basel Face Model
CDF	Cumulative Distribution Function
CRF	Conditional Random Field
COCA	Copula Component Analysis
EM	Expectation-Maximization
FERET	Face Recognition Technology
HOG	Histogram of Oriented Gradients
LFW	Labeled Faces in the Wild
MAP	Maximum-A-Posteriori
MRF	Markov Random Field
MRI	Magnetic Resonance Imaging
PCA	Principal Components Analysis
PPCA	Probabilistic Principal Components Analysis
RANSAC	Random Sample Consensus
RMS	Root Mean Square
RMSD	Root Mean Square Distance
SMC	Simple Matching Coefficient
SPP	Sex Prediction Performance
SVD	Singular Value Decomposition

Appendix B

Smiley Model

Computer Science consists of a lot of interesting challenges which are easy to understand but hard to solve. Unfortunately only few reach the audience at high schools. Our motivation is to attract more students to the university studies in Computer Science by catching their interest with scientific challenges. We do not only want to astonish with nice presentations, but also to communicate the challenges behind our research in an understandable manner. Therefore we condensed the problem of facial portrait manipulation omitting all the technical details. We ended up with a statistical smiley model which is very intuitive to demonstrate the challenges when manipulating e.g. the happiness of a smiley (see Figure B.1). The smiley modeling tool and all the teaching material is made publicly available under `http://gravis.dmi.unibas.ch/smiley/`.

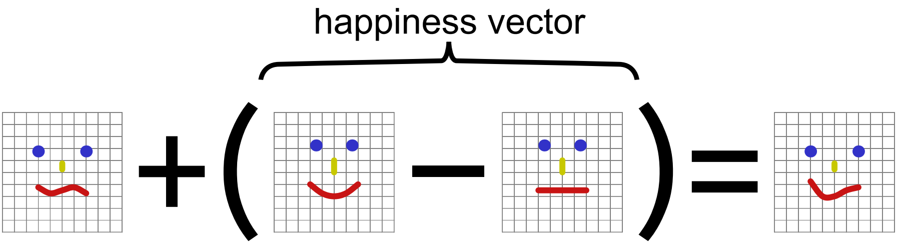


Figure B.1: Facial image manipulation is simplified using a smiley model. The necessary steps are the same as for manipulation of a real facial portrait but all calculations can be performed by hand.

Bibliography

T. Albrecht, M. Lüthi, T. Gerig, and T. Vetter. Posterior shape models. *Medical image analysis*, 17(8):959–973, 2013. 92

O. Aldrian and W. A. Smith. Inverse rendering of faces with a 3d morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(5):1080–1093, 2013. 12, 14, 47

D. Arthur and S. Vassilvitskii. k-means^{++}: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007. 41

J. T. Barron and J. Malik. Shape, illumination, and reflectance from shading. *IEEE transactions on pattern analysis and machine intelligence*, 37(8): 1670–1687, 2015. 14, 15, 69

A. Bas, W. A. Smith, T. Bolkart, and S. Wuhrer. Fitting a 3d morphable model to edges: A comparison between hard and soft correspondences. *ACCV Workshop on Facial Informatics*, LNCS vol. 10117:pp. 377–391, 2016. 94

R. Basri and D. W. Jacobs. Lambertian reflectance and linear subspaces. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(2): 218–233, 2003. 63, 65

T. Beeler, B. Bickel, G. Noris, P. Beardsley, S. Marschner, R. W. Sumner, and M. Gross. Coupled 3d reconstruction of sparse facial hair and skin. *ACM Transactions on Graphics (ToG)*, 31(4):117, 2012. 3

P. Bérard, D. Bradley, M. Gross, and T. Beeler. Lightweight eye capture using a parametric model. *ACM Transactions on Graphics (TOG)*, 35(4): 117, 2016. 13, 56

R. Blanc, C. Seiler, G. Székely, L.-P. Nolte, and M. Reyes. Statistical model based shape prediction from a combination of direct observations and various surrogates: application to orthopaedic research. *Medical image analysis*, 16(6):1156–1166, 2012. 12

V. Blanz and T. Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH'99 Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194. ACM Press, 1999. 1, 11, 17, 47

J. Booth, A. Roussos, S. Zafeiriou, A. Ponniah, and D. Dunaway. A 3d morphable model learnt from 10,000 faces. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 94

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. 44

M. Castelan, W. A. Smith, and E. R. Hancock. A coupled statistical model for face shape recovery from brightness images. *IEEE Transactions on Image Processing*, 16(4):1139–1151, 2007. 12

M. Chai, T. Shao, H. Wu, Y. Weng, and K. Zhou. Autohair: Fully automatic hair modeling from a single image. *ACM Transactions on Graphics (TOG)*, 35(4):116, 2016. 13, 56

T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *Computer Vision—ECCV'98*, pages 484–498. Springer, 1998. 11, 17

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005. 31, 44

A. V. Dalca, R. Sridharan, L. Cloonan, K. M. Fitzpatrick, A. Kanakis, K. L. Furie, J. Rosand, O. Wu, M. Sabuncu, N. S. Rost, et al. Segmentation of cerebrovascular pathologies in stroke patients with spatial and shape priors. In *Medical image computing and computer-assisted intervention: MICCAI International Conference on Medical Image Computing and Computer-Assisted Intervention*, volume 17, page 773. NIH Public Access, 2014. 13

J. G. Daugman. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. *Journal of the Optical society of America A*, 2(7):1160–1169, 1985. 44

M. De Smet, R. Fransens, and L. Van Gool. A generalized em approach for 3d model based face recognition under occlusions. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 1423–1430. IEEE, 2006. 14

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977. 45

A. Dessein, W. A. Smith, R. C. Wilson, and E. R. Hancock. Example-based modeling of facial texture from deficient data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3898–3906, 2015. 91

H.-C. Do, J.-Y. You, and S.-I. Chien. Skin color detection through estimation and conversion of illuminant color using sclera region of eye under varying illumination. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 1, pages 327–330. IEEE, 2006. 94

J. I. Echevarria, D. Bradley, D. Gutierrez, and T. Beeler. Capturing and stylizing hair for 3d fabrication. *ACM Transactions on Graphics (ToG)*, 33 (4):125, 2014. 3

G. J. Edwards, A. Lanitis, C. J. Taylor, and T. F. Cootes. Statistical models of face images—improving specificity. *Image and Vision Computing*, 16(3): 203–211, 1998. 12, 25

B. Egger, S. Schönborn, A. Forster, and T. Vetter. Pose normalization for eye gaze estimation and facial attribute description from still images. In *German Conference on Pattern Recognition*, pages 317–327. Springer, 2014. 9, 12, 30, 31, 42, 91

B. Egger, D. Kaufmann, S. Schönborn, V. Roth, and T. Vetter. Copula eigenfaces - semiparametric principal component analysis for facial appearance modeling. In *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 1: GRAPP*, pages 50–58, 2016a. 9, 18, 28

B. Egger, A. Schneider, C. Blumer, A. Forster, S. Schönborn, and T. Vetter. Occlusion-aware 3d morphable face models. In *British Machine Vision Conference (BMVC)*, 2016b. 9, 13, 49, 51

B. Egger, D. Kaufmann, S. Schönborn, V. Roth, and T. Vetter. Copula eigenfaces with attributes. In *under review CCIS on VISIGRAPP 2016*, 2017a. 9, 18

B. Egger, S. Schönborn, C. Blumer, and T. Vetter. Probabilistic morphable models. In *Statistical Shape and Deformation Analysis*. Elsevier, 2017b. 9, 85

B. Egger, S. Schönborn, A. Schneider, A. Kortylewski, A. Morel-Forster, C. Blumer, and T. Vetter. Occlusion-aware 3d morphable models and an illumination prior for face image analysis. In *submitted to IJCV Special Issue on BMVC 2016*, 2017c. 9, 66

M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 61

C. Genest, K. Ghoudi, and L.-P. Rivest. A semiparametric estimation procedure of dependence parameters in multivariate families of distributions. *Biometrika*, 82(3):543–552, 1995. 20

R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-PIE. *Image and Vision Computing*, 28(5):807–813, 2010. 30, 41, 65, 90

F. Han and H. Liu. Semiparametric principal component analysis. In *Advances in Neural Information Processing Systems*, pages 171–179, 2012. 18, 21

P. D. Hoff. Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics*, pages 265–283, 2007. 24, 32

G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 2007. 5, 46, 49, 51, 53, 54, 55, 90

R. Huang, V. Pavlovic, and D. N. Metaxas. A graphical model framework for coupling mrfs and deformable models. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–739. IEEE, 2004. 13, 39

P. Huber, Z.-H. Feng, W. Christmas, J. Kittler, and M. Rätsch. Fitting 3d morphable face models using local features. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 1195–1199. IEEE, 2015. 12, 94

P. Huber, G. Hu, R. Tena, P. Mortazavian, P. Koppen, W. Christmas, M. Ratsch, and J. Kittler. A multiresolution 3d morphable face model and fitting framework. In *Proceedings of the 11th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 2016. 94

H. Joe. *Multivariate models and multivariate dependence concepts.* CRC Press, 1997. 20

I. Jolliffe. *Principal component analysis.* Wiley Online Library, 2002. 17, 20

M. J. Jones and T. Poggio. Hierarchical morphable models. In *Computer Vision and Pattern Recognition, 1998. Proceedings. 1998 IEEE Computer Society Conference on*, pages 820–826. IEEE, 1998. 13

A. Jourabloo and X. Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *Computer Vision and Pattern Recognition, 2016. Proceedings CVPR*, 2016. 61

K. Khan, M. Mauro, and R. Leonardi. Multi-class semantic segmentation of faces. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 827–831. IEEE, 2015. 12

A. Kortylewski. Model-based image analysis for forensic shoe print recognition. *PhD Thesis*, 2017. 13

M. Köstinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 2144–2151, 2011. 6, 60, 70, 75, 90

T. D. Kulkarni, P. Kohli, J. B. Tenenbaum, and V. Mansinghka. Picture: A probabilistic programming language for scene perception. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4390–4399, 2015. 61

N. Kumar, A. Berg, P. Belhumeur, and S. Nayar. Describable visual attributes for face verification and image search. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(10):1962–1977, 2011. 12, 31

S. Liu, J. Yang, C. Huang, and M.-H. Yang. Multi-objective convolutional learning for face labeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3451–3459, 2015. 95

M. Lüthi, R. Blanc, T. Albrecht, T. Gass, O. Goksel, P. Buchler, M. Kistler, H. Bousleiman, M. Reyes, P. C. Cattin, and others. Statismo-a framework for PCA based statistical models. *The Insight Journal*, pages 1–18, 2012. 8

D. Maggi. Dekorrelation von Licht und Textur im 3DMM. *Bachelor Thesis, not publically available*, 2014. 77

F. Maninchedda, C. Häne, B. Jacquet, A. Delaunoy, and M. Pollefeys. Semantic 3d reconstruction of heads. In *European Conference on Computer Vision*, pages 667–683. Springer, 2016. 13

S. R. Marschner and D. P. Greenberg. Inverse lighting for photography. In *Color and Imaging Conference*, volume 1997, pages 262–265. Society for Imaging Science and Technology, 1997. 14

A. M. Martinez and R. Benavente. The ar face database. *CVC Technical Report*, 24, 1998. 49, 50, 53, 54

F. J. Massey Jr. The kolmogorov-smirnov test for goodness of fit. *Journal of the American statistical Association*, 46(253):68–78, 1951. 19

U. Mohammed, S. J. Prince, and J. Kautz. Visio-lization: generating novel facial images. *ACM Transactions on Graphics (TOG)*, 28(3):57, 2009. 91

A. Morel-Forster. Generative shape and image analysis by combining gaussian processes and mcmc sampling. *PhD Thesis*, 2017. 13, 44

K. P. Murphy, Y. Weiss, and M. I. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 467–475. Morgan Kaufmann Publishers Inc., 1999. 45

E. Murphy-Chutorian and M. M. Trivedi. Head pose estimation in computer vision: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 31(4):607–626, 2009. 70

R. B. Nelsen. *An introduction to copulas*, volume 139. Springer Science & Business Media, 2013. 20

K. Nishino and S. K. Nayar. Corneal imaging system: Environment from eyes. *International Journal of Computer Vision*, 70(1):23–40, 2006. 94

P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter. A 3d face model for pose and illumination invariant face recognition. In *Proceedings of the 6th IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS)*, pages 296–301. IEEE, 2009. 12, 13, 28, 67

P. J. Phillips, H. Wechsler, J. Huang, and P. J. Rauss. The feret database and evaluation procedure for face-recognition algorithms. *Image and vision computing*, 16(5):295–306, 1998. 44

P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss. The feret evaluation methodology for face-recognition algorithms. *IEEE Transactions on pattern analysis and machine intelligence*, 22(10):1090–1104, 2000. 44

J.-S. Pierrard and T. Vetter. Skin detail analysis for face recognition. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007. 14, 92

R. Ramamoorthi and P. Hanrahan. An efficient representation for irradiance environment maps. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 497–500. ACM, 2001. 63

E. Richardson, M. Sela, and R. Kimmel. 3d face reconstruction by learning from synthetic data, 2016. 61

S. Romdhani and T. Vetter. Efficient, robust and accurate fitting of a 3d morphable model. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 59–66. IEEE, 2003. 11, 14, 47

S. Schönborn, A. Forster, B. Egger, and T. Vetter. A monte carlo strategy to integrate detection and model-based face analysis. In *Pattern Recognition*, pages 101–110. Springer, 2013. 12, 47, 83

S. Schönborn, B. Egger, A. Forster, and T. Vetter. Background modeling for generative image models. *Computer Vision and Image Understanding*, 136: 117–127, 2015. 42

S. Schönborn, B. Egger, A. Morel-Forster, and T. Vetter. Markov chain monte carlo for automated face image analysis. In *International Journal of Computer Vision*. Springer, 2016. 2, 12, 14, 15, 25, 28, 30, 37, 38, 39, 45, 60, 61, 65, 67, 73, 77

M. Schumacher and V. Blanz. Exploration of the correlations of attributes and features in faces. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, pages 1–8. IEEE, 2015. 12

D. Shahlaei and V. Blanz. Realistic inverse lighting from a single 2d image of a face, taken under unknown and complex lighting. In *Automatic Face and*

Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on, volume 1, pages 1–8. IEEE, 2015. 14

D. Shahlaei, M. Piotraschke, and V. Blanz. Lighting design for portraits with a virtual light stage. In *Image Processing (ICIP), 2016 IEEE International Conference on*, pages 1579–1583. IEEE, 2016. 61

L. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces. *Journal of the Optical society of America A*, 4(3):519–524, 1987. 11

M. Sklar. *Fonctions de répartition à n dimensions et leurs marges.* Université Paris 8, 1959. 17

W. A. P. Smith. *The Perspective Face Shape Ambiguity*, pages 299–319. Springer International Publishing, Cham, 2016. 77

A. Tewari, M. Zollhöfer, H. Kim, P. Garrido, F. Bernard, P. Pérez, and C. Theobalt. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. *arXiv*, (1703.10580), 2017. 12, 14, 94

M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999. 17

H. Tsukahara. Semiparametric estimation in copula models. *Canadian Journal of Statistics*, 33(3):357–375, 2005. 20

Z. Tu, X. Chen, A. L. Yuille, and S.-C. Zhu. Image parsing: Unifying segmentation, detection, and recognition. *International Journal of Computer Vision*, 63(2):113–140, 2005. 13

M. Turk, A. P. Pentland, et al. Face recognition using eigenfaces. In *Computer Vision and Pattern Recognition, 1991. Proceedings CVPR'91., IEEE Computer Society Conference on*, pages 586–591. IEEE, 1991. 11

M. Uřičář, V. Franc, D. Thomas, S. Akihiro, and V. Hlaváč. Real-time Multi-view Facial Landmark Detector Learned by the Structured Output SVM. In *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2015*, volume 02, pages 1–8, 2015. 47, 83

E. Wood, T. Baltrušaitis, L.-P. Morency, P. Robinson, and A. Bulling. A 3d morphable eye region model for gaze estimation. In *European Conference on Computer Vision*, pages 297–313. Springer, 2016. 13, 56

C. Wu, D. Bradley, P. Garrido, M. Zollhöfer, C. Theobalt, M. Gross, and T. Beeler. Model-based teeth reconstruction. *ACM Transactions on Graphics (TOG)*, 35(6):220, 2016. 13, 56

I. Yildirim, M. Janner, M. Belledonne, C. Wallraven, W. A. Freiwald, and J. B. Tenenbaum. Causal and compositional generative models in online perception. In *to be published at 39th Annual Conference of the Cognitive Science Society*, 2017. 14

L. Yin, X. Wei, Y. Sun, J. Wang, and M. Rosato. A 3d facial expression database for facial behavior research. In *7th International Conference on Automatic Face and Gesture Recognition, 2006. FGR 2006*, pages 211–216, 2006. 30

X. Zhu, J. Yan, D. Yi, Z. Lei, and S. Li. Discriminative 3d morphable model fitting. In *Proceedings of 11th IEEE International Conference on Automatic Face and Gesture Recognition FG2015*, Ljubljana, Slovenia, 2015. 12, 94

X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face alignment across large poses: A 3d solution. In *Computer Vision and Pattern Recognition, 2016. CVPR 2016.*, 2016. 61

J. Zivanov, A. Forster, S. Schönborn, and T. Vetter. Human face shape analysis under spherical harmonics illumination considering self occlusion. In *ICB-2013, 6th International Conference on Biometrics*, Madrid, 2013. 65