Probabilistic Morphable Models

Bernhard Egger, Sandro Schönborn, Clemens Blumer, Thomas Vetter

Department of Mathematics and Computer Science

University of Basel, Switzerland

Abstract

3D Morphable Face Models have been introduced for the analysis of 2D face photographs. The analysis is performed by actively reconstructing the three-dimensional face from the image in an Analysis-by-Synthesis loop, exploring statistical models for shape and appearance. Here we follow a probabilistic approach to acquire a robust and automatic model adaptation.

The probabilistic formulation helps to overcome two main limitations of the classical approach. First, Morphable Model adaptation is highly depending on a good initialization. The initial position of landmark points and face pose was given by manual annotation in previous approaches. Our fully probabilistic formulation allows us to integrate unreliable Bottom-Up cues from face and feature point detectors. This integration is superior to the classical feed-forward approach, which is prone to early and possibly wrong decisions. The integration of uncertain Bottom-Up detectors leads to a fully automatic model adaptation process. Second, the probabilistic framework gives us a natural way to handle outliers and occlusions. Face images are recorded in highly unconstrained settings. Often parts of the face are occluded by various objects. Unhandled occlusions can mislead the model adaptation process. The probabilistic interpretation of our model makes possible to detect and segment occluded parts of the image and leads to robust model adaptation.

Throughout this chapter we develop a fully probabilistic framework for image interpretation. We start by reformulating the Morphable Model as a probabilistic model in a fully Bayesian framework. Given an image, we search for a posterior distribution of possible image explanations. The integration of Bottom-Up information and the model parameters adaptation is performed using a Data Driven Markov Chain Monte Carlo approach. The face model is extended to be occlusion-aware and explicitly segments the image into face and non-face regions during the model adaptation process. The segmentation and model adaptation is performed in an Expectation-Maximization-style algorithm utilizing a robust illumination estimation method.

The presented fully automatic face model adaptation can be used in a wide range of applications like face analysis, face recognition or face image manipulation. Our framework is able to handle images containing strong outliers, occlusions and facial expressions under arbitrary poses and illuminations. Furthermore, the fully probabilistic embedding has the additional advantage that it also delivers the uncertainty of the resulting image interpretation.

1 Introduction

In this chapter, we present Probabilistic Morphable Models - a fully probabilistic framework to interpret face images. We reconstruct 3D faces from 2D images with a statistical model in an Analysis-by-Synthesis setting. A given target face is represented by model instances which are similar to the target image. So far, most Morphable Model adaptation techniques relied on manual initialization and were prone to outliers and occlusions. We urge to use a fully probabilistic framework to obtain an automated and occlusion-aware system.

Statistical models have been applied for segmentation in CT, MRI or 2D photographs. Usually the model adaptation is initialized manually and solved by optimization techniques. This approach is sensitive to initialization and prone to occlusions and outliers. The optimization often leads to local minima. In our probabilistic setting, we do not aim for a single best solution through optimization, but we search for the posterior probability distribution of possible model explanations of the input image. During the model adaptation process we only have uncertain correspondence between the target image and the face model. The aim of model adaptation is to find the location and orientation of the face (pose), statistical model parameters and the illumination condition. The likelihood functions are highly non-convex as the dependence on the input image renders the adaptation rough and highly nonlinear. Occlusions make it even harder to target this problem with standard optimization techniques. In order to handle occlusions and to include uncertain information our framework is fully probabilistic.

We present two challenges to highlight the benefit of using a fully probabilistic framework. First, our model adaptation process includes uncertain detection results for feature points, such as eye or mouth corners. The localization of such feature points is still a challenge in computer vision and produces unreliable results. In a classical feed-forward optimization procedure, the uncertainty is ignored. This leads to pipelines which take early and possibly wrong decisions that can not be reconsidered in later steps. Our probabilistic approach enables us to integrate this uncertain information source and guide the overall adaptation process. Second, faces are often occluded by various objects like glasses, hands or microphones. Our probabilistic setting makes possible to build an occlusion-aware adaptation framework. We detect occlusions using our strong appearance prior of the statistical model, combined with knowledge arising from classical image segmentation. The segmentation enforces smoothness of the labels assigned to neighboring pixels and is not a simple pixel-wise thresholding. The face model is adapted to the image and its uncertainty guides the segmentation of occlusion. The segmentation then drives the model adaptation to explain contiguous regions and guides it to explain as much as possible by the face model.

We use a 3D Morphable Model (3DMM) [5] as appearance prior for faces. The 3DMM is a Parametric Appearance Model (PAM). PAMs are able to generate images controlled by parameters θ. PAMs are widely used in generative setups, especially in the field of face image analysis. In medical image analysis, related models, namely Active Shape Models [7] are prominent. Instead of modeling appearance, they define gradients or higher level image features. An application to face photographs brings some additional challenges compared to medical data. Whilst medical data are most often recorded in a controlled setting, facial photography is highly unconstrained. Be-

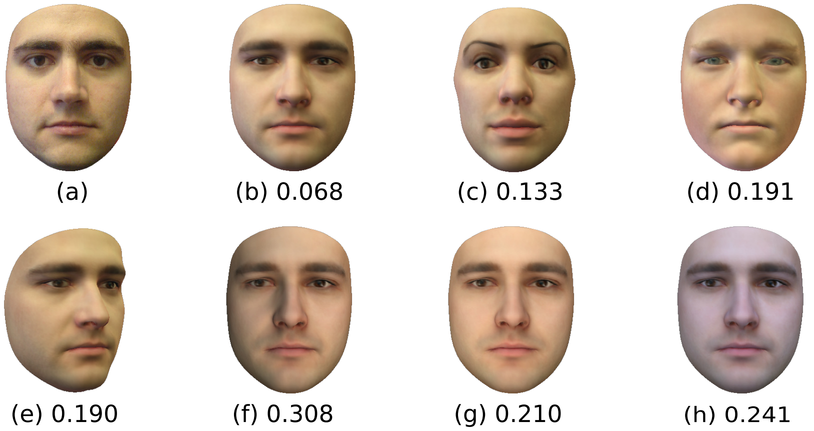

Figure 1: We illustrate the dominance of illumination effects on facial appearance. We show the target image (a) and its best fitting model instance (b). We manipulated a block of parameters to obtain the other images (c-h). We indicate the RMS-distance to the target image (a) in color space for each rendered image. We inverted shape (c) and color (d) parameters and also changed the yaw angle to 45 degree (e). All those changes significantly influence the facial appearance. However, the illumination changes to an illumination from the side (f), the front (g) or a real world illumination from another image (h) have a higher influence on the RMS-distance. When adapting the model parameters and searching for the best instance (b), the illumination is therefore dominant.

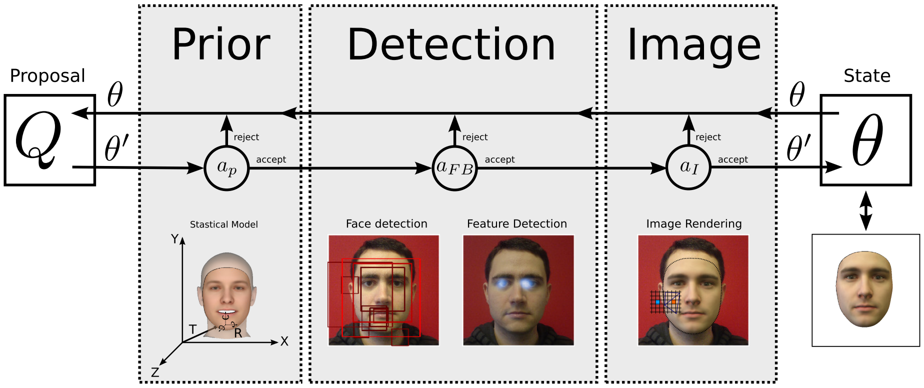

Figure 2: Our sampling framework is built on a Metropolis-Hastings algorithm. A parameter update θ' is drawn from the proposal generator Q and evaluated through three filters. At each filtering stage a_x a sample can be accepted or rejected by its corresponding likelihood function following the Metropolis-Hasting acceptance rule. If a sample is accepted from the prior, detection and image filtering stage, it builds the new state θ of the Markov Chain. The sequence of states in the Markov chain builds the posterior distribution over θ.

sides the shape, we also need to estimate facial color, pose and illumination from the single input image. Together with a camera model, 3DMMs can synthesize new face images. Due to pose and shape variations, facial parts can be invisible by self-occlusion effects. Especially the effects of a color model and illumination add additional challenges as presented in Figure 1.

We reinterpret the 3DMM to build a fully probabilistic framework. The model consists of a shape and a color model. We build both models in the Gaussian Process framework proposed by Lüthi et al. [20]. It is the basis for our statistical prior on face shape and color appearance. We use a multi-linear face model to handle expressions.

Model adaptation is the most challenging part of face image analysis. Through the model adaptation process, we search model instances which match the input image. This process is called fitting. It is solved with different methods which we summarize in the related work section. We infer the posterior distribution of possible image explanations by our face model. Our approach is a Data Driven Markov Chain Monte Carlo (DDMCMC) sampling technique. In contrast to other 3DMM adaptation techniques, it does not aim for a single model instance as output but approximates the posterior distribution of possible solutions. The distribution carries information about the certainty of a fit.

As this posterior distribution can not be computed analytically, we use the Metropolis-Hastings algorithm to generate samples from the posterior distribution. The algorithm consists of a proposal and a verification step. It is based on a proposal distribution to generate samples and enforces consistency to the observed data and the model in the verification step. This partitioning is a key feature of the algorithm and represents a propose-and-verify architecture. All proposals are evaluated in the verification steps, therefore they can be explorative and do not have to always improve the result. The verification step accepts and rejects proposals based on their likelihood. In the verification steps, we use a filtering strategy. Different evaluation criteria, like the L_2 image difference or detection responses, are integrated by cascading Metropolis-Hastings acceptance stages with the respective likelihoods, see Figure 2. For example, a proposal is evaluated against a feature point likelihood in an early stage where bad proposals can be filtered out quickly. Filtering allows us to focus computing time on promising regions which are more expensive to evaluate, like the image difference.

We explicitly distinguish between face and non-face regions in the target image to handle occlusions. The image is segmented into regions which can be explained by the face model and regions which are explained by a simple background color model. This segmentation is defined on the 2D image plane and integrated into the model likelihood.

During model adaptation, we have to find face and non-face regions simultaneously. The strong appearance prior from the statistical face model is used to decide whether a pixel is considered face or background. The likelihood of a pixel being part of the face region changes during fitting. Therefore, we constantly reestimate the segmentation in an Expectation-Maximization (EM) procedure during the whole model adaptation process.

The illumination conditions in unconstrained face photographs vary heavily. In the beginning of the model adaptation, the distance between the current estimation and the target image is large and dominated by illumination mismatch, see Figure 1. Measuring only color distance, e.g. shadowed regions can differ stronger than occlusions. Especially under occlusion, illumination has to be estimated in a robust way. Occlusions would mislead the illumination estimation strongly. We exploit the illumination by using a RANSAC-based robust illumination estimation technique. This gives us a proper initialization of the illumination conditions and a first guess of occluded pixels in the image.

The following paragraphs are organized as follows: First we give a short overview over related work in the field of 3DMMs in Section 1.1. We then present our probabilistic face model (Section 2.1). In Section 2.2 we present how inference can be performed given a target image and integrate detection information (Section 2.3). Finally we extend the framework to handle background (Section 2.4) and become aware of occlusions (Section 2.5). We include most experiments directly in the corresponding methods parts and show the performance of the full framework in Section 3.

1.1 Related Work

PAMs are common in computer vision. The first successful PAM was the Eigenfaces approach [15, 29]. Principal Component Analysis (PCA) was performed on pixel intensity values of roughly aligned face images. This led to a parametric and generative representation of face images. The models were successful in a strongly constrained face recognition task. The next step in parametric face modeling have been Active Appearance Models [8], which combine Active Shape Models [7] and the idea of the Eigenfaces approach. Whilst the Eigenfaces approach assumes the images to have pixel-wise correspondence, AAMs add a shape model to handle different face shapes.

Shape deformations are modeled separately from the appearance. The shape model is learned from a set of 2D correspondences while the appearance model is restricted to shape-normalized images. Both models use PCA to find an efficient parameterization. Active Appearance Models became successful by the availability of specialized fast fitting algorithms [21, 3]. However, those models are defined in 2D and therefore cannot handle strong pose variation with self-occlusion. The next development for parametric modeling of faces are 3DMMs. The 3DMM models a face as a 3D object. The image formation process is explicitly modeled using a pinhole camera and a Phong reflectance model. The model is built on 3D face scans which are in dense correspondence and combines separate color and shape models. Contrary to the AAM, the shape and color models only describe face variation while the rendering part is handled by the explicit camera and illumination models. 3DMMs can therefore handle self-occlusion and head pose in a very natural way.

There are different approaches for adaptation of Morphable Models to images. The original approach by Blanz and Vetter [5] used a stochastic gradient descent method. Romdhani et al. [25, 24] presented a multiple-features fitting approach. Aldrian et al. presented a fast model adaptation method based on inverse rendering [2]. Recently, we presented an approach based on sampling which is able to include unreliable information sources in a probabilistic way [27]. Current developments in machine learning also investigate Supervised Descent Methods [30, 14] and probabilistic methods combined with deep learning [17]. Those learning methods achieve promising results for shape model adaptation with limited pose. However, they do not include a color or illumination model and are therefore not fully generative. All those model adaptation techniques rely on good initialization and are characterized by standard optimization techniques which are prone to local minima.

Occlusions can severely mislead the model adaptation process. Only few generative model adaptation approaches are able to handle occlusions. There is previous work, for all kind of PAMs, on how to handle occlusions using robust error measures or robust strategies like RANSAC [12]. In contrast, for 3DMM adaptation only few robust approaches exist. Most of them rely on manual labeling of occlusions or knowledge about how much of the face is occluded. Other approaches implement robust error measures. As appearance is dominated by illumination, robust error measures work only for almost ambient illumination settings. Those approaches tend to exclude facial parts which can be explained by complex illumination. Another main source of error are regions which are difficult to explain by the face model [25, 9, 23]. Examples for such regions are the eye, eyebrow, nose and mouth region, they vary much stronger in color appearance than e.g. the cheek. The pixels in those regions are harder to fit by the model but crucial for representing facial characteristics. Note that previous works on occlusion handling using a 3DMM focused on databases with artificial and homogeneous, frontal illumination settings. Our approach includes a robust illumination estimation which allows us to adapt the model in presence of occlusions even under complex illumination settings.

2 Methods

We give an overview over all components of our Analysis-by-Synthesis approach. We present our generative probabilistic face model and explain the image formation process. The most challenging part of the overall process is the inference for model adaptation to a given target image. We show how DDMCMC sampling can be used for this inference task and how it allows us to integrate various sources of unreliable information into the model adaptation process. The face model is only defined in the face region, to explain the whole image, we need a model for the background. We present our approach of background modeling and explain why it is important for generative models. Finally we want to be able to adapt our model to unconstrained face images. Occlusions are a challenge in those settings and can mislead the Analysis-by-Synthesis process. Therefore we extend the model and its adaptation to simultaneously segment the target image into face and non-face regions and become occlusion-aware.

2.1 Probabilistic Morphable Model

Our generative 3D Morphable Face Model is a variant of the Basel Face Model (BFM) [22], built from face scans of 200 people taken with a structured light 3D scanner. We extend the original face model to a multi-linear model to handle expressions as described in [4]. The multi-linear statistical model consists of two independent Probabilistic PCA (PPCA) models for face shape and face color as well as an additional model for the deformations by facial expression. Facial expression is modeled as a difference to the neutral face shape. The PPCA models are learned on the aligned vertex positions $\vec{s}$ and color $\vec{c}$. Each face mesh consists of 21 662 vertices. This leads to a distribution over facial color $P(\vec{c} \mid \vec{\theta})$ and shape $P(\vec{s} \mid \vec{\theta})$ and also handles observation noise in the training data:

$$P(\vec{s} \mid \vec{\theta}) = \mathcal{N}\left(\vec{s} \mid \vec{\mu}_S + U_S D_S \vec{\theta}_S + \vec{\mu}_E + U_E D_E \vec{\theta}_E, \sigma_S^2 I\right) \tag{1}$$

$$P(\vec{c} \mid \vec{\theta}) = \mathcal{N}\left(\vec{c} \mid \vec{\mu}_C + U_C D_C \vec{\theta}_C, \sigma_C^2 I\right) \tag{2}$$

All the models (subscripts S for shape, C for color and E for expression) consist of a mean $\vec{\mu}$, the principal components in matrix U and the variances along each component in diagonal matrix D. The additive Gaussian noise is only added for shape and for color. The parameters $\vec{\theta}$ follow a standard normal distribution in latent space:

$$P(\theta_x) = \mathcal{N}\left(\vec{\theta}_x \mid \vec{0}, I\right) \tag{3}$$

Previous ad hoc probabilistic interpretations of the 3D Morphable Model have been introduced for shape reconstruction in [6, 1, 18]. We resort to the recently introduced Gaussian Process Morphable Model of Lüthi et al. [20] as a consistent and clean framework to understand and create probabilistic shape models. The PPCA prior can be understood as a Gaussian Process with a covariance function consisting of a statistical kernel of N samples $\vec{s}_i(\vec{x})$ and independent Gaussian noise with variance σ^2 (only shown for shape):

$$K(\vec{x},\vec{y}) = \frac{1}{N}\sum_{i=1}^{N}\left(\vec{s}_i(\vec{x}) - \vec{\mu}(\vec{x})\right)\left(\vec{s}_i(\vec{y}) - \vec{\mu}(\vec{y})\right)^T + \sigma^2\delta(\vec{x},\vec{y})I_3 \tag{4}$$

Additionally to the shape process, we add an expression deformation process and also define a color appearance process analogously. The Gaussian Process model is discretized on the vertices of our reference mesh and parametrized through a low-rank expansion of the statistical part of the kernel (corresponds to PCA). The independent Gaussian kernel is handled without approximation (corresponds to PPCA). These Gaussian Process models are identical to the PPCA models described above.

The 3DMM also defines a rendering process $\Re$ to generate synthetic face images I. We use a pinhole camera model and a spherical harmonics illumination model [28]:

$$I(\theta) = \Re\left(M(\theta_S, \theta_C, \theta_E); \theta_P, \theta_L\right) \tag{5}$$

The shape θ_S, color θ_C and expression θ_E parameters are coupled to the principal components of the statistical PPCA prior, the camera θ_P and illumination parameters θ_L are handled separately. We assume a multivariate Gaussian distribution as a prior for the spherical harmonics expansion parameters and a uniform distribution on the camera parameters. The distribution of the illumination parameters is empirically estimated from successful fittings of the AFLW [16] database. This leads to a full face model prior distribution, including shape, color, expression, camera and illumination.

2.2 MCMC Sampling for Inference

Standard optimization for adapting the model to a given target image leads to a single local optimal parameter set. In our probabilistic setting, we perform probabilistic inference instead of optimization. The result of inference is not a single maximum, but a posterior distribution of the model parameters θ conditioned on the target image $\tilde{I}$:

$$P(\theta \mid \tilde{I}) = \frac{P(\tilde{I} \mid \theta)P(\theta)}{\int P(\tilde{I} \mid \theta)P(\theta)\mathrm{d}\theta} \tag{6}$$

The posterior is intractable and can only be evaluated point-wise with respect to an unknown multiplicative constant. The involvement of the target image in the likelihood with unknown correspondence leads to a highly non-convex distribution without a closed-form representation. A single evaluation involves rendering an image and comparing it to the target.

We use the Metropolis-Hastings algorithm for inference. The algorithm defines a Markov Chain which delivers samples approximately from the posterior distribution $P(\theta \mid \tilde{I})$. The algorithm achieves this by stochastic acceptance of random samples from a simpler probability distribution, called the proposal distribution $Q(\theta' \mid \theta)$. The probability of accepting a new sample is given by

$$a = \min\left\{\frac{P(\theta' \mid \tilde{I})}{P(\theta \mid \tilde{I})}\frac{Q(\theta \mid \theta')}{Q(\theta' \mid \theta)}, 1\right\}. \tag{7}$$

To calculate the acceptance probability, it is sufficient to evaluate posterior ratios which allow for an unnormalized evaluation. If a proposal is rejected the last sample in the chain remains as the state of the chain. The algorithm defines a formal propose-and-verify structure. It decouples finding solutions from validating them which allows us to also use uncertain and unreliable proposals. Verification with the model likelihood and prior ensures a consistent sample.

The proposal distribution has to be designed carefully. On one hand, it has to be easy to sample from, on the other hand, it has to be complex enough to explore the high dimensional parameter space. In our case, we use a mixture of a set of different proposals. We group logical parts together and only adapt a single block at once. A proposal consists of a shape, color, illumination or camera parameter update. The proposal can be unreliable - the verification step enforces consistency by evaluating the proposal against a likelihood function and the model assumptions. Most proposals are simple normally distributed random walks of fine to coarse scale.

During the verification steps, the proposals are evaluated according to their likelihood e.g. the likelihood given the target image. When adapting to an image, our main likelihood measures the color distance between the rendering of the current model estimate $I_i(\theta)$ at each pixel i and the raw pixels of the target image $\tilde{I}_i$:

$$\ell_{\text{face}}(\theta; \tilde{I}_i) = \frac{1}{N} \exp\left(-\frac{1}{2\sigma^2} \left\|\tilde{I}_i - I_i(\theta)\right\|^2\right) \tag{8}$$

In practice, the verification step also includes information about face and feature point detections and is split into different filtering stages described in Section 2.3.

We draw samples from our Markov Chain to obtain a posterior distribution. The Markov Chain needs to be initialized by a specific set of parameters. In the beginning, the chain is depends on the initialization. After a so called burn-in phase, we draw samples from the Markov Chain which are independent of the initialization. Those samples are used to approximate the posterior distribution.

Some applications do not need a posterior distribution but a single optimal parameter estimate. Using our Metropolis-Hastings algorithm, the best observed sample can be used as MAP estimate. The longer the sampling runs, the better the approximation gets. It can be stopped when the desired precision is obtained.

The proposed Metropolis-Hastings algorithm has some additional advantages: Due to the verification step it can easily integrate various sources of information. It is extendable to proposals which e.g. include gradients or parameter updates from learning techniques (e.g. cascaded regression techniques). The proposals are allowed to be imperfect as they are validated in the verification step. Last but not least, the algorithm is simple and easy to implement.

2.3 Integration of Bottom-Up Cues by Filtering

The ability to integrate Bottom-Up cues is a key strength of the proposed MCMC inference method. It is open to include information from arbitrary, unreliable Bottom-Up methods, even from different sources at the same time. We give an overview on how

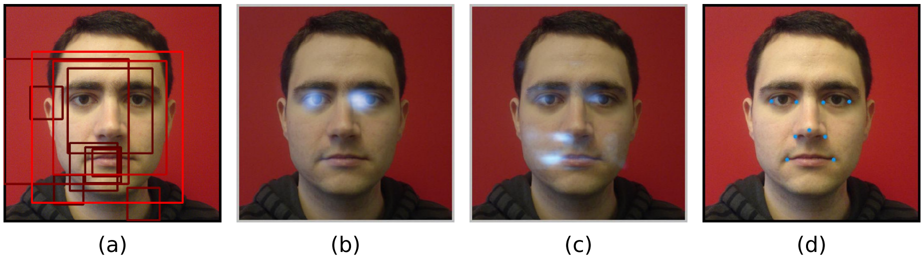

Figure 3: Face and feature point detections: In (a) the 10 strongest face detections are visualized. The probability of the detections is indicated by light red color. In the middle, we show the detection probability of the left inner eye corner (b) and for the right lip corner (c). We can observe that the lip corner detection is much more distributed over the image whilst the eye corner is more precisely located. However for both points we also get strong wrong detections at other facial features locations. In (d) we show all the nine feature points we trained detectors for.

those Bottom-Up cues can be integrated in the Metropolis-Hasting algorithm and then concrete examples including face and facial feature point detectors.

In non-probabilistic algorithms, Bottom-Up cues, e.g. detections, are usually included in a feed-forward manner. This leads to early and possibly wrong decisions which cannot be reviewed in later steps. In our fully probabilistic setting, we can use unreliable input of Bottom-Up methods with uncertainty and include this input directly in the inference process.

The integration of detection methods is simple and takes part in the evaluation step of the Metropolis-Hastings algorithm. Thus, we need a likelihood function for measuring how likely the observed data are under the current model estimate. The different likelihoods of Bottom-Up cues are considered in a step-by-step manner. We cascade multiple stochastic acceptance steps of the Metropolis-Hastings algorithm to integrate one likelihood after the other, see Figure 2. Only a proposal which is compatible with the respective likelihood at each stage is accepted. This process is called filtering. Filtering is very flexible and allows us to combine various information sources in a single algorithm efficiently. The cascade prevents a computationally expensive evaluation with respect to all likelihood functions if an early stage does not agree with the sample. This leads to early rejection of bad samples and improves the performance of the fitting procedure.

We include face and facial feature point detection into our model adaptation process to demonstrate the filtering chains. We search for the 10 best face detections in the target image and run facial feature detection for prominent landmark points like the eye or mouth corners, see Figure 3. Instead of using the strongest detection, we integrate the detection evidence as a likelihood model as described above. The exact procedure of detection integration is described in [28].

The likelihood of a face detection candidate considers landmarks detection maps $\mathcal{D}_l$ and the face box $\mathcal{B}$ with location $\vec{p}$ and scale s. The face box likelihood ℓ_B (Equation 9) consists of a Gaussian likelihood with respect to its center position and a log-normal distribution for face scale. The landmark map likelihood ℓ_{LM} (Equation 10) combines a noisy Gaussian landmarks observation model with the detection map.

$$\ell_B(\theta;\mathcal{B}_i) = \mathcal{LN}\left(s\left(\theta\right) \mid \mathcal{B}_i.s, \sigma_{\text{bs}}\right) \mathcal{N}\left(p\left(\theta\right) \mid \mathcal{B}_i.\vec{p}, \sigma_{\text{bp}}\right) \tag{9}$$

$$\ell_{\text{LM}}(\vec{x};\mathcal{D}) = \max_{\vec{t}} \mathcal{N}\left(\vec{t} \mid \vec{x}, \sigma^2_{\text{LM}}\right) \mathcal{D}\left(\vec{t}\right) \tag{10}$$

We perform a max convolution to find the best possible combination of detection and distance at each point in the image (as described in [27]). We combine the corresponding face and feature point detections:

$$\ell_i\left(\theta;\mathcal{B}_i,\mathcal{D}_i\right) = \ell_B\left(\theta;\mathcal{B}_i\right) \prod_l \ell_{\text{LM}}\left(\vec{x}\left(\theta\right);\mathcal{D}_l\right) \tag{11}$$

Our assumption is that the image contains exactly one face, therefore we optimize for the best candidate i of face and feature point detection:

$$\ell_{\text{FB}}\left(\theta;\mathcal{B},\mathcal{D}\right) = \max_i \ell_i\left(\theta;\mathcal{B}_i,\mathcal{D}_i\right) \tag{12}$$

We roughly initialize the 3D pose of the face by evaluating only with the detection or landmark likelihoods in the beginning. In later model adaptation steps, the consistency to the detections through filtering guides the adaptation procedure, see Figure 2. A proposal is first evaluated by its likelihood given the detections ℓ_{FB} and then also against the image ℓ_{I}:

$$P(\theta) \xrightarrow{\ell_{\text{FB}}(\theta;\mathcal{B},\mathcal{D})} P(\theta \mid \mathcal{B},\mathcal{D}) \xrightarrow{\ell_I(\theta;\tilde{I})} P(\theta \mid \mathcal{B},\mathcal{D},\tilde{I}) \tag{13}$$

2.4 Explicit Background Modelling

Most PAMs do not model the full image but solely the object of interest. Therefore, those generative models synthesize the foreground $\mathcal{F}$ of the image containing the face. The parts that cannot be modeled are referred to as background and are ignored by most adaptation methods. Ignoring the background leads to problems during model adaptation. The most prominent problem is shrinking of the face during the fitting process. Shrinking can be overcome by evaluating in normalized model space of 2D models rather than on the image. This solution does not work for 2D analysis with 3D models. Parts of the face are not always visible, depending on the 3D pose, due to self-occlusion. In a frontal view, different parts of the face are visible than in a profile view. If change in visibility is ignored, the pose cannot be adapted properly during the model adaptation. Earlier approaches handled this by assuming a fixed and predetermined visibility for each point on the face. This requires an accurate initialization of pose and strongly restricts the flexibility of the model with respect to pose and shape.

The effects of ignoring background in generative modeling are independent of the concrete optimization mechanism and should be handled on a model level to retain the full model flexibility. Our completely probabilistic approach allows us to include a background model directly into the likelihood function

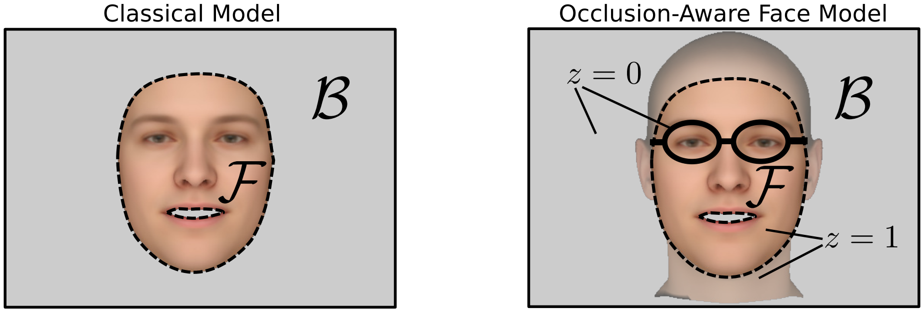

Figure 4: On the left, we show the classical model presented in Equation 14. The classical model distinguishes pixels in the foreground $\mathcal{F}$, where the face model is defined, from pixels in the background $\mathcal{B}$. The occlusion-aware model adds an additional variable z which is a label for face ($z = 1$) or non-face ($z = 0$), compare Equation 15. This additional label allows to exclude pixels in the foreground $\mathcal{F}$ from the face model explanation.

$$\ell(\theta; \tilde{I}) = \prod_{i \in \mathcal{F}} \ell_{\text{face}}(\theta; \tilde{I}_i) \prod_{i' \in \mathcal{B}} b(\tilde{I}_{i'}). \tag{14}$$

The idea is to distinguish between foreground pixels $\mathcal{F}$ which are covered by the rendered face and background pixels $\mathcal{B}$, compare Figure 4. The pixels in the background are then covered by an explicit background model with likelihood b. In some settings, it makes sense to build a complex background model which is able to adapt to the expected background appearance. In medical image analysis, the surroundings of objects in the body (e.g. organs or bones) are often similar and can be modeled. In the setting of face image analysis, background is unconstrained and a concrete background model is infeasible. In order to counteract the negative effects of background, it is already sufficient to use a simple background model. We resort to simple global color models. Such models are either general, e.g. a uniform color distribution, or adapted to the concrete target, e.g. a histogram. In Schönborn et al. [26] we discuss the problem and different background models in detail. However, the actual choice of background model is not very critical.

2.5 Occlusion-aware Morphable Model

Images of faces often contain occlusions of the face by other objects. The most common sources are not only unrelated objects in front of the face but also glasses, beards and hair. In an Analysis-by-Synthesis setting, occlusions act as outliers which render the fitting problem even harder. The appearance prior of our model and the ability to generate images is successfully applied to handle occlusions. Unhandled occlusions can disrupt the fitting process due to their unpredictable color and location, see Figure 5. The strong appearance prior is used to detect and segment occlusions in the

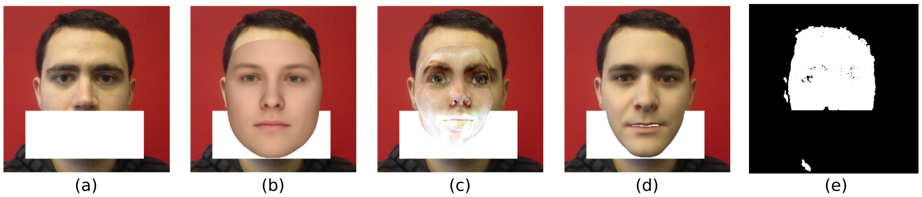

Figure 5: We use a partially occluded face as target image (a). Even the model adaptation is properly initialized manually (b), when using the likelihood from Equation 14, the adaptation is strongly misled by the occluded part (c). The model is able to deform color appearance strongly. With our occlusion-aware likelihood from Equation 15 and the segmentation approach, we can properly fit the target face (d) and segment face versus non-face (e). The fitting of the not-occluded target image is contained in Figure 2

image as those parts which cannot be explained by the model. Our probabilistic approach allows us to build an occlusion-aware model which deals with uncertainty of occlusion segmentation and the adapted face model.

We propose to segment the image into face and non-face parts. The image likelihood is therefore extended with an additional binary label z to mark face pixels (compare Figure 4):

$$\ell\left(\theta; \tilde{I}, z\right) = \prod_i \ell'_{\text{face}}\left(\theta; \tilde{I}_i\right)^z \cdot \ell_{\text{non-face}}\left(\theta; \tilde{I}_i\right)^{1-z} \tag{15}$$

The general idea is very similar to background modeling. Contrary to the background $\mathcal{B}$, occlusion can also appear within the foreground region $\mathcal{F}$. Occluded parts are excluded from model explanation and are not evaluated with the standard foreground image likelihood. The occlusion segmentation becomes part of the model parameters and needs to be included in model adaptation.

We choose the variational framework suggested by Chan-Vese as a basis for segmentation. We replace the plain average color models by our more complicated face model likelihoods for foreground and background to adapt it to our problem and model. The segmentation algorithm then correctly evaluates with respect to the likelihood of model fit rather than using simply its mean color. The original Chan-Vese algorithm is formulated with an energy term E. We reformulate it as posterior to obtain the label z for a given parameter set θ and the target image $\tilde{I}$:

$$-\log \mathrm{p}\left(z|\tilde{I}, \theta\right) = E = \Psi + \int_\Omega z(x) \log \ell'_{\text{face}}(\theta; \tilde{I}(x)) + (1 - z(x)) \log \ell_{\text{non-face}}(\theta; \tilde{I}(x))\, \mathrm{d}x \tag{16}$$

Ψ is the length term of the classical Chan-Vese formulation and regularizes the boundary of the segmentation.

The segmentation relies on the model adaptation and vice-versa. Therefore both parts have to be optimized together. We choose an EM-like procedure to optimize

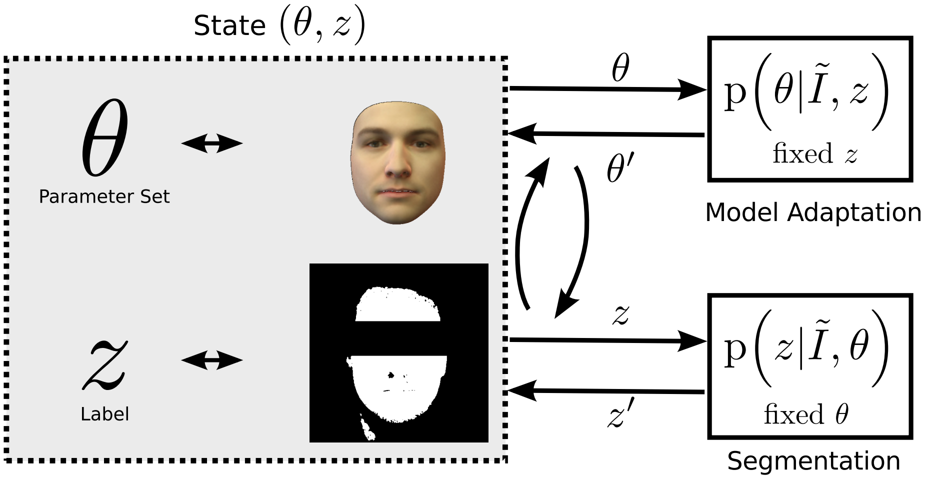

Figure 6: The occlusion handling is integrated into the probabilistic framework. A state consists of both, model parameters θ and the segmentation label z. During the model adaptation we assume a fixed segmentation and during segmentation we assume fixed model parameters. Inference of the parameter set and the segmentation is performed in an EM-style manner in alternation.

them in close alternation. During segmentation, we assume fixed model parameters and during fitting we assume a given segmentation, see Figure 6. However, in both steps we assume an uncertainty of the other estimate. By considering also the neighboring pixels N in the segmentation likelihood, we assume the fit not to be perfect yet:

$$\ell'_{\text{face}}\left(\theta;\tilde{I}_i\right)=\begin{cases}\frac{1}{N}\exp\left(-\frac{1}{2\sigma^2}\min_{n\in N(i)}\left\|\tilde{I}_i-I_{i,n}(\theta)\right\|^2\right) & \text{if } i\in\mathcal{F}\\ b_{\mathcal{F},z=1}\left(\tilde{I}_i\right) & \text{if } i\in\mathcal{B}\end{cases} \tag{17}$$

Taking the neighborhood into the likelihood function considers small misfits of the position on the whole face. Since the likelihood has to be defined on the whole image, we extend the likelihood to cover also pixels in the background. Therefore, we use a color model learned on foreground pixels labeled as face. We respect the uncertainty of segmentation by allowing the face model to include pixels which it can explain better than the background model:

$$\ell_{\text{non-face}}\left(\theta,\tilde{I}_i\right)=\begin{cases}\max\left(\ell_{\text{face}}\left(\theta,\tilde{I}_i\right),b\left(\tilde{I}_i\right)\right) & \text{if } i\in\mathcal{F}\\ b\left(\tilde{I}_i\right) & \text{if } i\in\mathcal{B}\end{cases} \tag{18}$$

As presented in Figure 1 the appearance of faces is dominated by illumination. Therefore we initialize the segmentation and fitting, by a robust illumination estimation with a RANSAC-like procedure as described in [10]. This gives us a first guess of the illumination conditions and the pixels which probably belong to the face region see Figure 7.

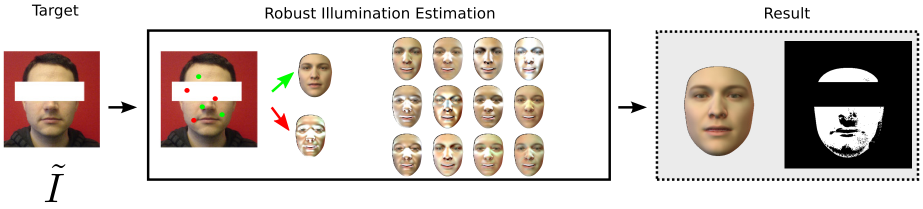

Figure 7: We randomly select points in the foreground $\mathcal{F}$ for robust illumination estimation. Those points are used to estimate the spherical harmonics illumination. Depending on the selection of points we get different illumination estimations. We added exemplary estimations on the three points which are suited for illumination estimation (green), three points on occluded parts (red) and some more on other random points. The iterative RANSAC-like algorithm chooses the most probable illumination setting. The results are illumination parameters θ_L and a mask of pixels which can be explained by this illumination setting. The mask is used as starting point for the label z in our occlusion-aware model adaptation.

3 Applications and Results

In this section, we present experiments and practical application of our framework. We do not strive for a complete evaluation because all proposed elements of the probabilistic Morphable Model image analysis framework, including model fitting, are already thoroughly validated elsewhere. Instead, we highlight the power of the method and quality of our results with a few rather different example applications.

In the following we make use of the elements described above. Technical details, such as the final proposal distribution, feature point and face detectors and the histogram background model are described in [27, 10, 28, 11]. The software implementation is based on the Statismo [19] and Scalismo [1] software frameworks. Parts of our implemented Probabilistic Morphable Model framework are contributed to both frameworks and therefore available open source.

In our first experiment, we explore the posterior distribution under occlusion. The posterior distribution is the result of our image analysis process conditioned on detections and the target image. The posterior distribution can be visualized to analyze the uncertainty of the model adaptation. After a burn-in phase of 50’000 samples we draw further 50’000 samples to estimate the posterior distribution. We use a collective likelihood for this experiment as described in [27]. With the posterior distribution, we investigate the remaining shape flexibility under occlusion, see Figure 8. We observe that parts of the model can be recovered with high certainty whilst other parts are more flexible and therefore can only be recovered with low certainty.

In a second experiment our model adaptation recovers from wrong initialization as seen in Figure 9. We therefore show an example where the strongest face detection is the wrong one. Our cascaded filtering presented in Equation 13 and Figure 2 succeeds

[1] Scalismo - A Scalable Image Analysis and Shape Modelling Software Framework available as Open Source under `https://github.com/unibas-gravis/scalismo`

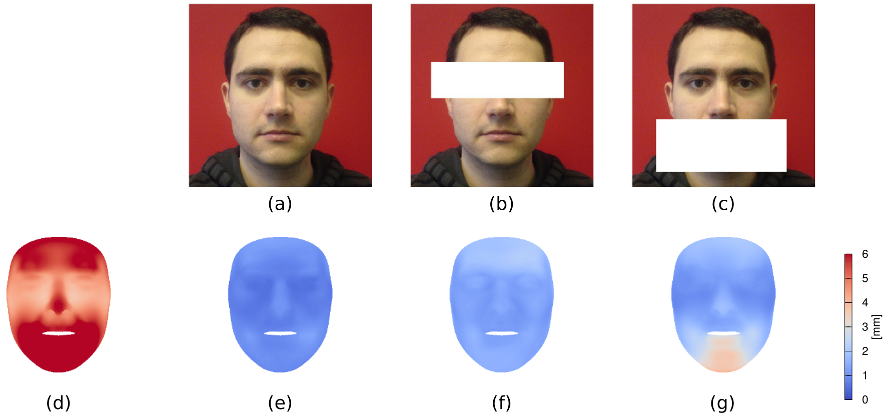

Figure 8: Posterior variability of the face shape for different target images. The shape standard deviation of the model prior is shown in (d). Conditioning on a target image without occlusion (a) we get a small posterior variance (e) for the whole face region. With occlusion in the eye region (b), the model gets less certain about the shape, but can narrow down the variability by the strong appearance prior of the statistical model (f). Under occlusion of the chin (c) the variation in the chin region increases strongly (g), this uncertainty is based on the strong influence of expressions which lead to high variability in this region.

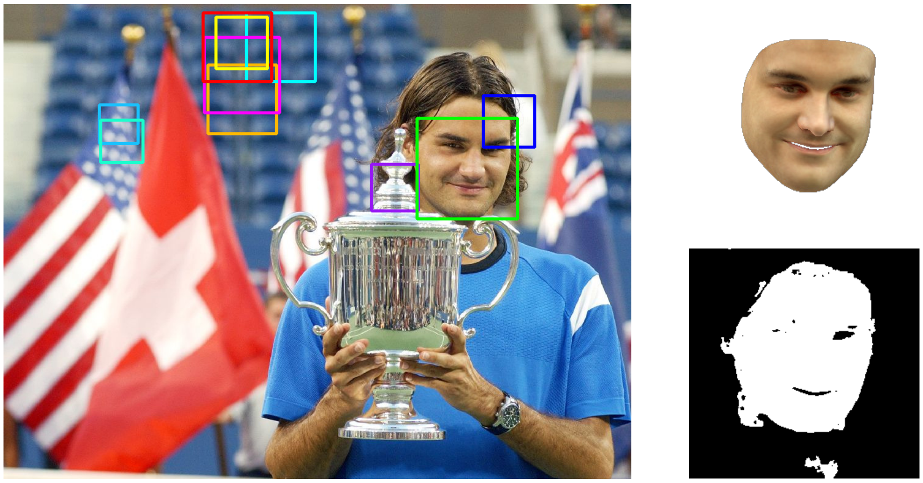

Figure 9: We show the 10 strongest face detection results for this real world target image. The yellow facebox is the strongest face detection. Our fully probabilistic framework succeeds to recover from this wrong strongest detection and adapts to the most consistent detection. On the right we show the best fitting result and the label map. Image: KEYSTONE/AP/Richard Drew

to find the solution with the highest probability which is in this case the correct face. We draw 10’000 samples for the model adaptation conditioned on the image and the detections.

In our third experiment we perform a qualitative evaluation on the Labeled Faces in the Wild database (LFW) [13]. It contains occlusions, expressions and real world illumination settings. We again draw 10’000 samples and present the best fitting result. The LFW database is challenging for generative models. The presented images depict the variability of the database. Pose, illumination, expressions and occlusions are challenging variations. By using a Morphable Model including facial expressions, we add an additional challenge for the background model. The 3DMM does not contain the inner mouth region which is explained by the background model. We use feature point detections at the mouth corner (compare Figure 3). The expression, respectively the opening of the mouth is solely inferred in the Analysis-by-Synthesis loop from the image. With our explicit background model we succeed to open and close the mouth matching the target image. The most prominent occlusion in the LFW database are glasses. We present results for a variety of faces (neutral, with expression, glasses, sunglasses, occluding hair and other occlusion). Selected results are shown in Figure 10.

In a fourth experiment we present a straightforward application. The fit of our model to an image infers correspondence information of every face pixel to a vertex point in the model space. The correspondence information can be used for face image manipulation, see Figure 11. We manipulate parameters and therefore facial appearance in model space and render the changes back into the face image. The transfer from the manipulation in 3D is rendered in a 2D warp field on the image. The result is a photo-realistic face image manipulation. The presented manipulation is achieved by removing the expression mean $\vec{\mu}_E$ and the expression deformation $\vec{\theta}_E$ from the face model to neutralize facial expressions from the target image.

All parts of the proposed framework have been published and evaluated individually. Additional qualitative experiments on different tasks are presented in [27, 28, 11, 26]. Those publications contain extensive experimental evaluation on pose-invariant face recognition, pose estimation, attribute description, eye gaze estimation, shape reconstruction and also experiments to investigate characteristics of the DDMCMC sampling strategy.

4 Conclusion

We present an unprecedented completely automatic fitting framework for 3DMM adaptation including detection and occlusions. It is generic and based on a fully probabilistic approach. Our probabilistic setting is capable of robust and fully automatic face image analysis. We demonstrated the flexibility of our approach by integrating discriminative Bottom-Up cues and segmentation of occlusion. The Bottom-Up integration is superior to feed-forward methods and can recover from unreliable detection results. We build an occlusion-aware model by exploiting the uncertainty of the face model to segment occlusions which cannot be explained otherwise. Both, the occlusion awareness and the Bottom-Up integration can be naturally integrated in the likelihood functions

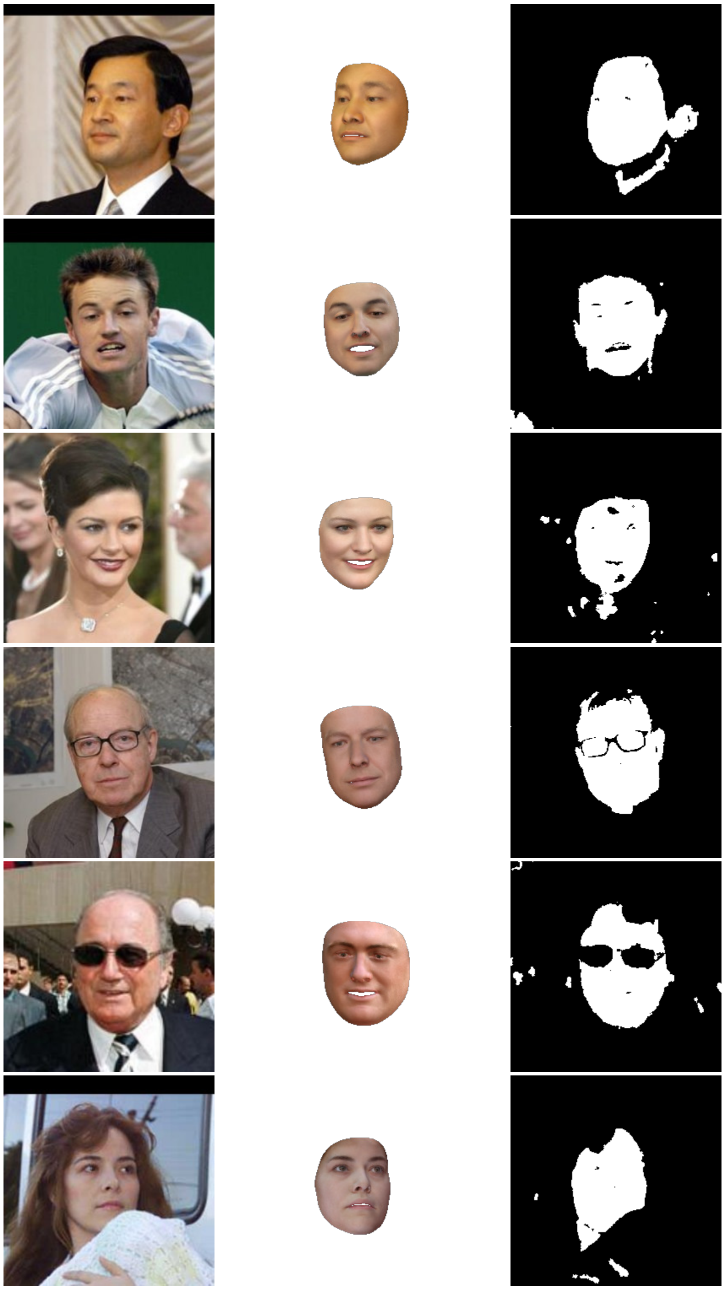

Figure 10: Exemplary fitting results of the fully automatic probabilistic framework on the LFW database. The target image is shown on the left, the best fit (MAP estimate) in the middle and the segmentation label z on the right.

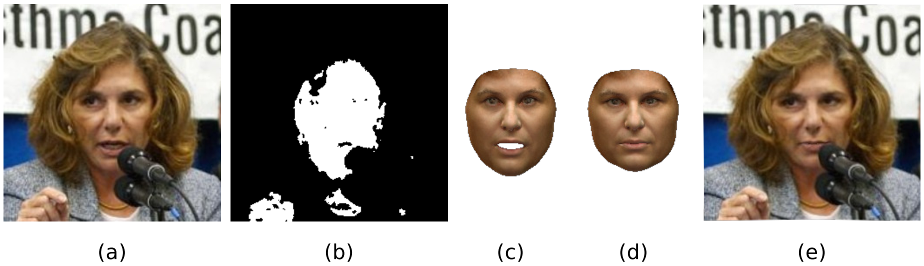

Figure 11: Photo-realistic image manipulation becomes possible through the high-quality results of our probabilistic analysis. The target image (a) is fitted using our fully automatic and robust model adaptation framework. The occlusion label (b) and fitting result (c) hold correspondence information from the face model to the target image. The expression manipulation is performed in model space by removing the expression components (d). We obtain the final manipulation result (e) by rendering the 3D manipulation back into a warp-field for the target image.

of the probabilistic framework. We extended our 3DMM with facial expressions in order to adapt it to unconstrained face photographs. We highlight the importance of background modeling for Analysis-by-Synthesis settings. The Metropolis-Hastings algorithm builds the core of our probabilistic framework and makes future extensions easy. Our approach is promising for image analysis and scene understanding. The resulting framework is generic and flexible. The proposed approach is not specific for faces but can also be applied for occlusion handling or initialization of other statistical shape models. The showcased integration of Bottom-Up cues and segmentation are only two examples of extensions. The model adaptation process would highly profit from more Bottom-Up cues like edge or pose information. Due to its propose-and-verify architecture, our framework is open to include arbitrary uncertain information sources.

REFERENCE

References

[1] Thomas Albrecht, Reinhard Knothe, and Thomas Vetter. Modeling the remaining flexibility of partially fixed statistical shape models. In *2nd MICCAI Workshop on Mathematical Foundations of Computational Anatomy*, pages 160–169, 2008.

[2] Oswald Aldrian and William A.P. Smith. Inverse rendering of faces with a 3d morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(5):1080–1093, May 2013.

[3] Brian Amberg, Andrew Blake, and Thomas Vetter. On compositional image alignment, with an application to active appearance models. In *Computer Vi-*

sion and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 1714–1721. IEEE, 2009.

[4] Brian Amberg, Reinhard Knothe, and Thomas Vetter. Expression invariant 3d face recognition with a morphable model. In *Automatic Face & Gesture Recognition, 2008. FG'08. 8th IEEE International Conference on*, pages 1–6. IEEE, 2008.

[5] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '99, pages 187–194, New York, NY, USA, 1999. ACM Press/Addison-Wesley Publishing Co.

[6] Volker Blanz and Thomas Vetter. Reconstructing the complete 3d shape of faces from partial information (rekonstruktion der dreidimensionalen form von gesichtern aus partieller information). *it-Information Technology (vormals it+ ti) Methoden und innovative Anwendungen der Informatik und Informationstechnik*, 44(6/2002):295, 2002.

[7] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, January 1995.

[8] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, June 2001.

[9] Michael De Smet, Rik Fransens, and Luc 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.

[10] Bernhard Egger, Andreas Schneider, Clemens Blumer, Andreas Morel-Forster, Sandro Schönborn, and Thomas Vetter. Occlusion-aware 3d morphable face models. In *Proceedings of the British Machine Vision Conference (BMVC)*, September 2016.

[11] Bernhard Egger, Sandro Schönborn, Andreas Forster, and Thomas 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.

[12] Martin A. Fischler and Robert 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.

[13] Gary B. Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forStudying face recognition in unconstrained environments. 2008.

[14] Patrik Huber, Zhen-Hua Feng, William Christmas, Josef Kittler, and Matthias Rätsch. Fitting 3d morphable face models using local features. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 1195–1199. IEEE, 2015.

[15] Michael Kirby and Lawrence Sirovich. Application of the karhunen-loeve procedure for the characterization of human faces. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 12(1):103–108, 1990.

[16] Martin Köstinger, Paul Wohlhart, Peter M. Roth, and Horst 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.

[17] Tejas D Kulkarni, Pushmeet Kohli, Joshua B Tenenbaum, and Vikash 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.

[18] Marcel Lüthi, Thomas Albrecht, and Thomas Vetter. Probabilistic modeling and visualization of the flexibility in morphable models. In *IMA International Conference on Mathematics of Surfaces*, pages 251–264. Springer, 2009.

[19] Marcel Lüthi, Remi Blanc, Thomas Albrecht, Tobias Gass, Orcun Goksel, Philippe Buchler, Michael Kistler, Habib Bousleiman, Mauricio Reyes, Philippe C Cattin, and others. Statismo-a framework for PCA based statistical models. *The Insight Journal*, pages 1–18, 2012.

[20] Marcel Lüthi, Christoph Jud, Thomas Gerig, and Thomas Vetter. Gaussian process morphable models. *CoRR*, abs/1603.07254, 2016.

[21] Iain Matthews and Simon Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164, 2004.

[22] Paysan Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. *2009 Advanced Video and Signal Based Surveillance*, pages 296–301, 2009.

[23] Jean-Sébastien Pierrard and Thomas Vetter. Skin detail analysis for face recognition. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[24] S. Romdhani and T. Vetter. Estimating 3d shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005*, volume 2, pages 986–993 vol. 2, June 2005.

[25] Sami Romdhani and Thomas Vetter. Efficient, robust and accurate fitting of a 3d morphable model. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 59–66, 2003.

[26] Sandro Schönborn, Bernhard Egger, Andreas Forster, and Thomas Vetter. Background modeling for generative image models. *Computer Vision and Image Understanding*, 136:117–127, July 2015.

[27] Sandro Schönborn, Bernhard Egger, Andreas Morel-Forster, and Thomas Vetter. Markov chain monte carlo for automated face image analysis. *International Journal of Computer Vision*, pages 1–24, 2016.

[28] Sandro Schönborn, Andreas Forster, Bernhard Egger, and Thomas Vetter. A monte carlo strategy to integrate detection and model-based face analysis. In *Pattern Recognition*, number 8142 in Lecture Notes in Computer Science, pages 101–110. Springer Berlin Heidelberg, 2013.

[29] Matthew Turk and Alex Pentland. Eigenfaces for recognition. *Journal of cognitive neuroscience*, 3(1):71–86, 1991.

[30] Xuehan Xiong and F. De La Torre. Supervised descent method and its applications to face alignment. In *2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 532–539, June 2013.