Computer Vision and Image Understanding 136 (2015) 117–127

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Background modeling for generative image models

Sandro Schönborn *, Bernhard Egger, Andreas Forster, Thomas Vetter

Department of Mathematics and Computer Science, University of Basel, Spiegelgasse 1, 4051 Basel, Switzerland

ARTICLE INFO

Article history:
Received 15 May 2014
Accepted 20 January 2015

Keywords:
Generative models
Face model
Face analysis
Morphable Model
Bayesian model
Implicit background models

ABSTRACT

Face image interpretation with generative models is done by reconstructing the input image as well as possible. A comparison between the target and the model-generated image is complicated by the fact that faces are surrounded by background. The standard likelihood formulation only compares within the modeled face region. Through this restriction an unwanted but unavoidable background model appears in the likelihood. This implicitly present model is inappropriate for most backgrounds and leads to artifacts in the reconstruction, ranging from pose misalignment to shrinking of the face. We discuss the problem in detail for a probabilistic 3D Morphable Model and propose to use explicit image-based background models as a simple but fundamental solution. We also discuss common practical strategies which deal with the problem but suffer from a limited applicability which inhibits the fully automatic adaption of such models. We integrate the explicit background model through a likelihood ratio correction of the face model and thereby remove the need to evaluate the complete image. The background models are generic and do not need to model background specifics. The corrected 3D Morphable Model directly leads to more accurate pose estimation and image interpretations at large yaw angles with strong self-occlusion.

1. Introduction

A human face in a typical image is surrounded by arbitrary background. In Analysis-by-Synthesis settings, generative, parametric face models such as Active Shape Models, Active Appearance Models or Morphable Models, serve to reconstruct the input face as well as possible [5,4,2]. Depending on its parameter values, the model produces a synthetic image which is then compared to the input image through its likelihood under the model for a given set of parameter values. Since the face only occupies a part of the input image and it can appear in front of any background, one avoids to include background into the model likelihood. Consequently, the likelihood considers only the visible parts and ignores the rest of the image. But as we show in this article, even though background is ignored, it is still present in the model likelihood in the form of an implicit and usually wrong background model.

The wrong background model leads to a strong preference for background over the face. Wherever possible, the optimization algorithm will try to reduce the support of the face. This leads to unnatural optimal solutions which range from a strong shrinking effect to pose misalignment in non-frontal situations with self-occlusion (Fig. 1).

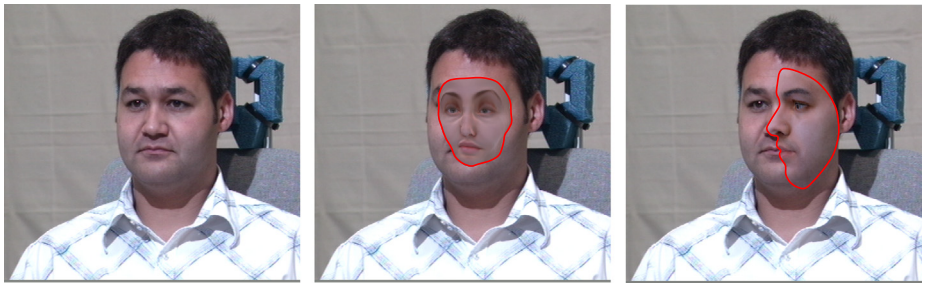
The issues with the implicit background model become evident as soon as the visibility of model parts can change with respect to the model parameters. So far, most model fitting methods kept the visibility constant, either by model restriction or determining it in advance. For a fully automatic model adaption in unconstrained situations, the full flexibility of the face is needed and the visibility cannot be fixed in advance.

Due to the unavoidable use of an implicit background model, we simply propose to use an explicitly controlled, image-based background model to resolve the problems. For practical implementations, we show how it is sufficient to correct the model likelihood for this background assumption without actually evaluating the whole background of the image. These minimal background models work by replacing the model likelihood by a likelihood ratio of model and background. The change in model likelihood is a fundamental change within the model likelihood and can be used to improve any fitting algorithm. It renders the desired interpretation more stable and leads to a likelihood maximum which is more consistent with the expectations of a face interpretation.

We focus the discussion of the problem mainly on the 3D Morphable Model (3DMM) [2] but in principle, our results apply to different generative models. We present an analysis of the problem within the probabilistic interpretation of the 3DMM fitting

* Corresponding author.
E-mail addresses: sandro.schoenborn@unibas.ch (S. Schönborn), bernhard.egger@unibas.ch (B. Egger), andreas.forster@unibas.ch (A. Forster), thomas.vetter@unibas.ch (T. Vetter).

http://dx.doi.org/10.1016/j.cviu.2015.01.008
1077-3142/



(a) Target (b) Shrinking (c) Pose Misalignment

Fig. 1. Wrong background models let the fit fail. Image-based evaluation leads to shrinking whereas model-based evaluation is prone to pose misalignments.

problem from Schönborn [15] which reveals a variable domain of evaluation as the underlying problem.

Model fitting at larger yaw angles is especially susceptible to effects of varying support size due to a large amount of self-occlusion and thus a changing visibility. We evaluate different background models with respect to their performance for pose estimation with the 3DMM. The empirical background model, based on a histogram of the image performs best in the evaluation. But all controlled background models lead to better results than the parasitic model arising from ignoring the effect.

Standard fitting algorithms for generative face models evaluate the image difference in the model domain. They project the target image back into the normalized model view and compare within the model, termed *model-based* evaluation. On the other side, the likelihood can also be evaluated directly within the target image, called *image-based* evaluation. Image-based evaluation leads to higher quality reconstructions and more accurate pose estimations. Contrary to model-based evaluation, it is even more susceptible to effects of varying visibility since the support of the face in the image changes strongly in accordance with the parameters.

We will first discuss the research background and theoretical setup followed by our analysis of the problem. The proposed solution using explicit background models is then discussed together with different choices of possible background models. In the evaluation part, we compare different models with respect to their performance solving a pose estimation problem, a face recognition problem and additionally, we qualitatively compare their performance on real-world images.

2. Background

The adaptation of a parametric, generative face model to an image is usually posed as an optimization problem, seeking the parameter values which explain a given image best. The two main types of generative face models, Active Appearance Models and 3D Morphable Models are discussed in the following [2,4].

Active Appearance Models (AAM) consist of a variable texture image which is deformed to match the target image [4]. The fitting algorithms warp back the target image to the model reference domain where it is compared to the model texture image. AAMs do not model self-occlusion and thus keep the model domain for evaluation constant [4,11]. They are not suited to capture the full 3D variability of face images, especially for large yaw angles with strong self-occlusion. There exist extensions of the model to deal with occlusions [6] but our focus lies on the 3DMM in this article.

The original 3D Morphable Model (3DMM) of Blanz and Vetter [2] reconstructs the image by rendering a deformable 3D template mesh into the image. The authors evaluate the likelihood per vertex of the underlying 3D mesh which forms a fixed evaluation domain. Since changing visibility due to self-occlusion directly affects the evaluation domain, the visibilities are kept constant for most of the optimization run. Romdhani and Vetter [14] extended the cost function to include more terms, especially contour costs. User-provided landmark positions are used to initialize the algorithm. Knothe [8] later extended the use of feature point positions to keep the model position and visibility fixed during optimization. All the methods keep the visibility of model parts constant during the optimization.[1]

Aldrian and Smith [1] took a fresh view on the fitting algorithm and proposed to use a mainly linear procedure. The feature points determine the camera and shape of the model while later steps, which depend on image intensities, only change normals, illumination and albedo.

We already presented a face recognition pipeline with the 3DMM, making practical use of a background model [15]. The pipeline is based on a probabilistic Data-Driven Markov Chain Monte Carlo sampler which can deal with unreliable input information. The present article focuses more on theoretical and implementation aspects of background models while the former puts an emphasis on a recognition application.

Summarizing, there are different strategies in use to prevent the shrinking problem in practice:

Model-based evaluation. Evaluating in the normalized model domain works especially well for 2D models and to some extent with 3D models if the visibility of face parts is fixed while it breaks down with variable visibilities due to self-occlusion.

Regularization. Penalizing parameter values can only work in restricted situations. Generally, neither the pose nor the face size in the image are known in advance and can thus not be restricted.

Landmarks. Predetermined feature point locations can prevent the problems through fixing the visibilities in advance or forcing the model to always match the given key points in the image. The information is generally not available or unreliable if obtained automatically.

Edge or contour terms. To force the model to match contour lines within the image does also depend on predetermined additional information. Just as with landmarks, contours are not generally available and hard to detect reliably.

Explicit Image Segmentation. In the case of image-based evaluation, image segmentation can fix the visibilities in advance. However, using a general-purpose image segmentation method is not straightforward as the face can change its appearance and size drastically in different illumination and pose settings. Therefore, the use of image segmentation methods needs careful design and is most successful if directly integrated into the fitting algorithm [9]. Such an approach renders the system much more complex and does not solve the problem for model-based evaluation.

[1] Some allow the visibility to change very slowly compared to the parameter values.

Image segmentation can motivate a fresh view on the problem. Especially region-based segmentation shares many similarities with image-based evaluation. For example, the concept of *region competition* [18], where two regions compete to explain an image pixel, is a formalization with many similarities to our problem. Contrary to most image segmentation methods, we use a very sophisticated foreground likelihood model with strong (marginal) correlations between different parts of the face region and have a high level prior through the 3DMM. The assumption that a foreground model cannot be adapted well without considering what is beyond is a fundamental insight from image segmentation [12].

In this article, we deal with large yaw angles and finally use a fully automatic pipeline [15] for face recognition which makes use of automatic detections. It is thus necessary to use a model which can support a large pose range and deal with uncertain input data. The unreliable nature of landmarks detection prevents early decisions and needs the model to explore many possible interpretations. Therefore, it is neither possible to fix the visibilities nor the location and size of the face in the image in advance.

We prefer image-based evaluation over model-based evaluation because its reconstruction quality is higher. Image-based evaluation makes a systematic treatment of the shrinking effect beyond the pragmatic solutions necessary.

3. Understanding shrinking

In this article, we interpret images of human faces with the generative 3DMM. Although we focus on the 3DMM of faces, the results conceptually apply to any generative model fitting problem.

In the probabilistic setting, the generative model will assign a likelihood value $\mathrm{L}(\theta|I)$ to each parameter setting θ given an input image I. The goal of inference is to find the posterior distribution of the parameters $\mathrm{P}(\theta|I) \propto \mathrm{L}(\theta|I)\mathrm{P}(\theta)$ or at least some measures of it. The prior $\mathrm{P}(\theta)$ captures the statistics of the face model in use and ensures a camera and illumination setup. Maximum-A-Posteriori (MAP) inference in the Bayesian framework is equivalent to regularized cost minimization, where the cost is given by $-\log \mathrm{L}$. But in the remainder of the article, we will use the probabilistic formulation.

The parameters θ of a deformable face model consist of transformation parameters which describe the positioning and alignment of the face either in 3D or directly on the image plane. θ also contains parameters describing the shape and appearance of the face itself, traditionally derived from a Probabilistic Principal Components Analysis (PPCA) or similar model. The 3D models also contain parameters describing the illumination setup. In this work we make use of the Basel Face Model (BFM) [13] which is publicly available. The actual choice of model and details is not crucial for our discussion, we mainly assume that the model is able to produce an image which can then be compared to the target image using the likelihood discussed below.

Image-based evaluation. A standard assumption of the model is to consider the pixels to be conditionally independent given the model parameters. The resulting likelihood is typically a large product of individual pixel likelihoods inside the face region

$$\mathrm{L}(\theta|I) = \prod_{x\in\mathcal{F}} \ell(\mathrm{M}(x;\theta)|I(x)). \tag{1}$$

In image-based evaluation, the index x runs over all pixels lying within the explained face region $\mathcal{F}$, and $\mathrm{M}(x;\theta)$ is the image rendered by the model. The individual pixel likelihood is usually Gaussian (see Fig. 2)

$$\ell(\mathrm{M}(x;\theta)|I(x)) = \mathcal{N}(I(x)|\mathrm{M}(x;\theta),\sigma^2).$$

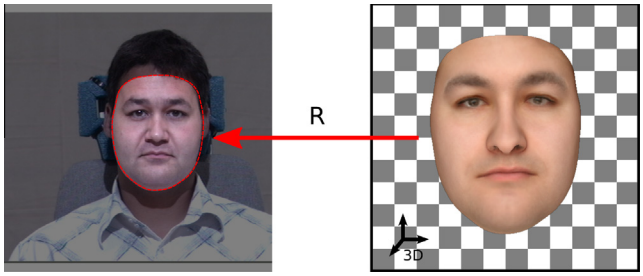


Fig. 2. Problem setup: The 3DMM reconstructs the input image (left) in 3D model space (right). Where the model is mapped to the image through the rendering function R, there is *foreground* $\mathcal{F}$ (inside dashed contour) surrounded by *background* $\mathcal{B}$ (grayed).

The size of the face region $\mathcal{F}$ within the image is not constant but depends on the parameter values.[2] A parameter change can thus alter the number of evaluated pixels. Consider a change $\theta \to \theta^-$ such that the likelihood values at individual pixels are approximately constant but the support size reduces by a single pixel p. The new likelihood value is given by removing ℓ_p from the product in L, where $\ell_p(\theta) = \ell(\theta|I(p))$ is the likelihood value at the pixel p (Fig. 3).

The likelihood ratio r^- is then given by the pixel's likelihood

$$r^- = \frac{\mathrm{L}^-}{\mathrm{L}} = \frac{1}{\ell_p}, \quad r^+ = \frac{\mathrm{L}^+}{\mathrm{L}} = \ell_p. \tag{2}$$

In the above equation, we also included r^+ which results from a similar argument for including an additional (different) pixel p.

The ratio r^- determines whether the pixel will likely be excluded during the inference run. The smaller region is preferred if $\ell_p < 1$. For many likelihood models, this is true for all possible values of $I(p)$, therefore the exclusion of p is preferred no matter how well the pixel actually fits the model assumptions. The likelihood ratio arising from different numbers of explained pixels leads to a net "force" towards smaller face explanations with fewer pixels to explain.

The formulation reveals the dependence of the algorithm on the absolute value of the likelihood ℓ_p. Since likelihoods can be arbitrarily scaled, this behavior is undesired.

Model-based evaluation. The evaluation of the likelihood on the model reference leads to

$$\mathrm{L}_\mathrm{M}(\theta|I) = \prod_{v\in\mathcal{V}} \ell(\mathrm{M}(x(v);\theta)|I(x(v))). \tag{3}$$

Here, the index v runs over all visible parts $\mathcal{V}$ of the model reference. $x(v) = \mathrm{P}(v;\theta)$ is the projection of v into the image plane. There are no image correspondences to compare the invisible model parts to.

The product also has a varying number of factors if the visibility of the model parts can change through self-occlusion. The exclusion of a single location v from the product due to a slight geometry change leads to the likelihood ratios of transitions

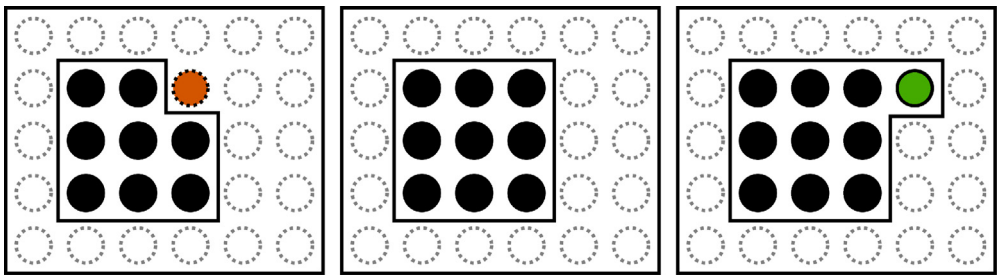
$$r^- = \frac{\mathrm{L}^-}{\mathrm{L}} = \frac{1}{\ell_v} \quad \text{and} \quad r^+ = \frac{\mathrm{L}^+}{\mathrm{L}} = \ell_v \tag{4}$$

with $\ell_v(\theta) = \ell(\mathrm{M}(x(v);\theta)|I(x(v)))$.

The ratios drive the model to remove parts from the evaluation if $\ell_v < 1$. The effect is strongest where there is much visibility change, for example around yaw angles above 45° (Fig. 4).

To get a meaningful gradient for use in an optimization algorithm, traditionally used fitting algorithms fixed the visibilities in advance to get a likelihood product with a constant number of factors.

[2] For example, moving away from the camera reduces the support due to a smaller projected area in the image.



(a) Exclusion θ^- (b) θ (c) Inclusion θ^+

Fig. 3. Model situations without considering background. Dotted pixels are implicitly present background pixels, filled circles are part of foreground. Removing a pixel from the foreground region (orange) also removes its contribution from the likelihood. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

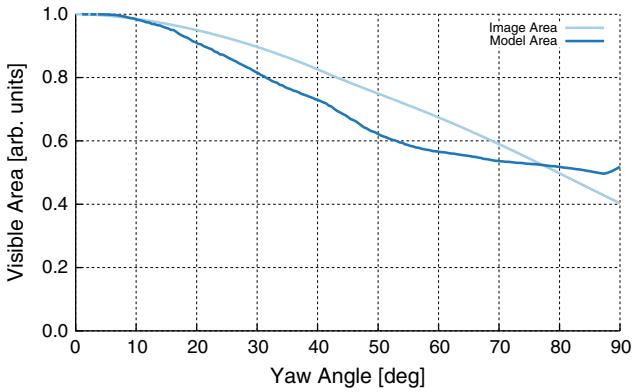


Fig. 4. The visible area of the face reduces considerably with increasing yaw angle for both, image-based and model-based evaluation.

4. Background models

The variable-length likelihoods above lead to a dependence on absolute likelihood values and also to problems with gradient calculations. A background model prevents the likelihood product from changes in the number of factors. It is an additional model b for background pixels $\mathcal{B}$. The total likelihood (1) becomes

$$\mathrm{L}(\theta|I) = \prod_{x\in\mathcal{F}} \ell(\theta|I(x)) \prod_{x'\in\mathcal{B}} b(I(x')). \tag{5}$$

Now, removing a pixel from the face region moves it into the background section (Fig. 5). The shrinking and growing ratios change accordingly to

$$r^- = \frac{b_p}{\ell_p}, \quad r^+ = \frac{\ell_p}{b_p}. \tag{6}$$

A comparison of the full background model (6) with the partial model (2) reveals the latter as a special case of a full model with a unit background likelihood

$$b = 1. \tag{7}$$

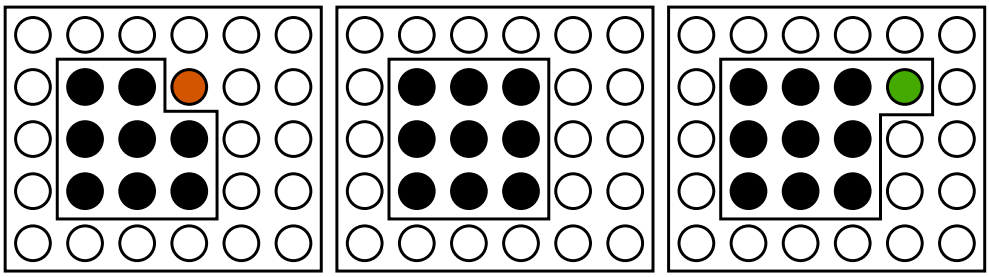
Even without considering background at all, the model always compares to an implicitly present unit likelihood background model. A unit likelihood for background is not appropriate for most model likelihoods and leads to shrinking, as explained above. It is reasonable to extend the model by a background model which can be controlled.

The values r^- and r^+ are ratios of likelihoods, they are more appropriate than absolute likelihood values. The exclusion of a pixel is favored whenever the background model assigns it a higher likelihood than the face model. And vice versa, a pixel is actively included into the face region if the face model can explain it better than the background model.

Adding an explicit background model is not easily possible for model-based evaluation. There is no image correspondence and therefore no image intensity value available for invisible parts of the model. A background model can only be defined per model location, independent of the actual image, but an implicit correction is still possible.

4.1. Background correction

The face model should be ignorant about specifics of backgrounds and concentrate its modeling capacity on the foreground to interpret. While the background model is conceptually necessary, it can be replaced by a much more lightweight mechanism which is sufficient to prevent shrinking. It is enough to get the exclusion and inclusion ratios r^- and r^+ right because they determine the behavior with respect to variable likelihood lengths.



(a) Exclusion θ^- (b) θ (c) Inclusion θ^+

Fig. 5. Model situations with an explicit background model. Background pixels are always evaluated (empty circles). Exclusion of pixel (orange) adds it to the background part. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

With a background correction, the model likelihood for a single pixel value ℓ_p is replaced by the likelihood ratio which compares this value with the background likelihood value b_p at the same location

$$\ell'_p = \frac{\ell_p}{b_p}. \tag{8}$$

The total image likelihood L′ is evaluated within the face region only

$$\mathrm{L}'(\theta|I) = \prod_{x\in\mathcal{F}} \frac{\ell_p}{b_p}. \tag{9}$$

The correction of the individual likelihoods leads to different absolute likelihood values L′ but shows the same exclusion and inclusion ratios $r^- = b_p/\ell_p$ and $r^+ = \ell_p/b_p$ as the full background model.

The correction only takes visible parts of the model into account. It is therefore also applicable to model-based evaluation which lacks the possibility to have an image-based background model. With model-based evaluation, the corrected likelihood becomes

$$\ell'_v = \frac{\ell_v}{b_{x(v)}}. \tag{10}$$

4.2. Different background likelihoods

The background likelihood reflects the assumptions taken about the image background color distribution. These can vary strongly and in extreme range up to a fully modeled background. We discuss rather simple and generic background models since we intend to keep the focus of our modeling power on the face in the foreground.

Uniform distribution. The uniform background model captures the assumption of a uniformly distributed background color with a constant likelihood. Through $b = 1$, this includes the ignored background from above.

The uniform likelihood leads to a background model where only the limits of the foreground model are relevant. The constant value determines where the face likelihood falls below acceptance. For example, a constant value of $b = \ell(|I(p) - \mathrm{M}(p;\theta)| = 2\sigma)$ prefers an explanation as background beyond two standard deviations away from the model prediction (Fig. 6).

The uniform model can thus be used to encode a completely background-ignorant model. The background likelihood just sets the bounds of what is still considered a "good" foreground explanation.

Gaussian. A Gaussian background likelihood is suited to model single mode color distributions without strong deviations. The assumption usually holds only in restricted setups, such as lab sessions. Mean and covariance of the Gaussian μ_{BG}, Σ_{BG} are specific to the image at hand.

$$b = \mathcal{N}(p|\mu_{\mathrm{BG}}, \Sigma_{\mathrm{BG}})$$

The parameter estimation follows a simple maximum likelihood approach using all image pixel colors with full covariance between color channels.

Histograms. The color histogram captures the color distribution of the whole image, including the face and thus represents the distribution of all the colors in the image. The model is image-specific and thus always adapts to the image at hand. It enforces a "contrast" between the face region and the rest of the image. The face model is better only where it leads to a more accurate prediction of a color value than the general image color distribution.

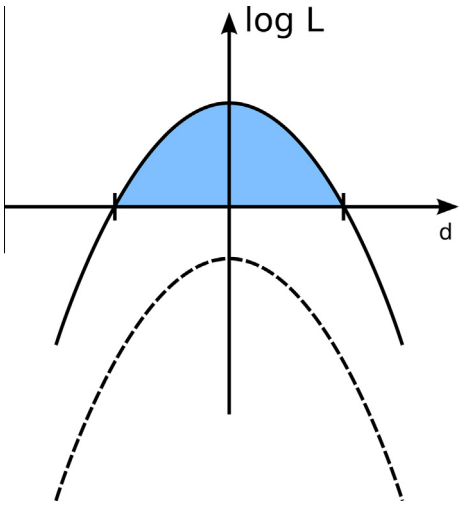


Fig. 6. A background correction shifts the log likelihood such that it can compete with the implicit background model $\log b = 0$ (blue shaded region). Without a background model, the foreground likelihood (dashed line) can never compete with the implicit background model. d is the color difference between target and reconstruction. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The background likelihood is derived from the histogram of relative color frequencies h with bin volume δ, interpreted as a probability density

$$b = \frac{1}{\delta} h(p).$$

The N bins have the same size each, in RGB space with values in $[0, M]^3$ their volume is $\delta = \left(\frac{M}{N}\right)^3$.

Filter responses. In place of a color distribution, we can also evaluate with respect to a collection of filter responses. Spatial filters gather information from the surroundings at each location. We directly implement the background model proposed by Wu and co-workers introduced with Active Basis models in [16]. They model a unit exponential distribution of the squared filter response values $r_i(x,y) = \langle I_m, B_{x,y,i}\rangle$, where $B_{x,y,i}$ is the filter i, centered at location (x,y). The scalar product is standard. For simplicity we only consider the filter responses on the intensity image.

We use the model as proposed in [16], including all the authors' choices, such as size and selection of Gabor filters as well as the whitening and heavy-tail correction transform F. The model, including whitening through the global response variance σ^2, essentially is a unit exponential distribution

$$-\log F\left(\frac{r_i^2}{\sigma_2}\right) \sim \exp(1).$$

The authors approximate the heavy tail correction through

$$-\log F(x) \approx G(x), \quad G(x) = \min(16, x).$$

They additionally propose more elaborate transforms G in [17] but we consider the actual choice rather uncritical in this context.

The background likelihood is

$$b(x,y) = N(x,y) \prod_{i=0}^{F} \exp\left(-G\left(\frac{r_i^2(x,y)}{\sigma^2}\right)\right),$$

where N is the normalization constant with respect to a varying background intensity at the current location. The normalization also turned-out to be uncritical in our experiments, the use of the unit exponential distribution $\exp(1)$ distribution as an approximation (intensity and thus response range is not infinite) leads to the same qualitative results.

Locally constant background. Locally constant color is a common pattern of generic background models, at least on a small scale. To model this, we introduce a Gaussian distribution on a local region around each pixel. The distribution is defined on each color channel but without covariance between them. Mean and variance are estimated with standard maximum likelihood, separately for each color channel. The free parameter of this model is the size of the rectangular local regions. We chose it to yield a local region having approximately 1% of the total image area. With this value of the window size, the average reconstruction probability of a complete natural image is similar to that using the histogram model. The background likelihood is then

$$b(x,y) = \prod_{c\in\{R,G,B\}} \mathcal{N}\left(p_c|\mu_c(x,y),\sigma_c^2(x,y)\right).$$

The locally constant model differs slightly from a piecewise-constant background model since it models each pixel's surroundings.

5. Evaluation

We compare the effect of different assumptions on background on the result of face model adaption. First, we will demonstrate the effects of the implicit unit likelihood background model which is present when background is simply ignored. Then, the background model correction is compared to the complete background model (5), where we can strengthen the theoretical considerations of equivalence. In a larger experiment, different background models are compared in a pose estimation task. Pose estimation is especially suitable to study the effects of background models since there is considerable self-occlusion and shrinking of the visible face area in the image above a yaw angle of 45° (Figs. 7 and 4). Finally, a fully automatic face recognition experiment underpins the need for a background model. To evaluate the performance in more natural situations, we add experimental results using images from the high-variation face database Annotated Facial Landmarks in the Wild (AFLW) [10].

General setup. We fit the 3DMM to the input image using the Data-Driven Markov Chain Monte Carlo adaption algorithm from Schönborn [15]. In most experiments user-provided landmarks are used (Fig. 8). They serve to initialize a fit but are "forgotten" afterward. The initialization just determines a pose setting to start from, nothing is kept fixed and during the inference run, only the image likelihood as discussed above is considered. In the fully automatic recognition experiment, nothing but the image itself is used as input.

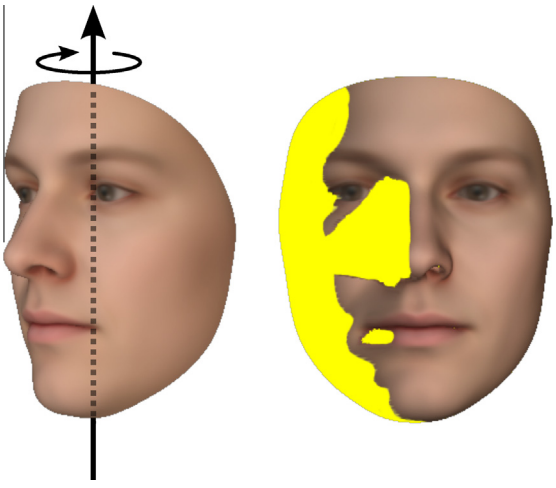


Fig. 7. Self-occlusion of the face at a yaw angle of 45°. The right image shows occluded regions in yellow. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

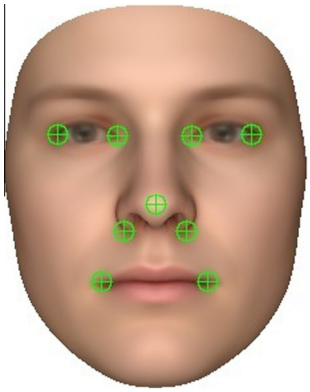


Fig. 8. The landmark points in use. They are user-provided as well as automatically detected, depending on the experiment.

The face model likelihood for each pixel p is an isotropic Gaussian distribution, centered at the currently predicted model color $\mathrm{M}(p;\theta)$ with a standard deviation of 6% of the intensity scale (11). All color values are 3-tuples in the RGB color space.

$$\ell(\theta|I(p)) = \mathcal{N}\left(I(p)|\mathrm{M}(p;\theta),\sigma^2\mathrm{I}_3\right). \tag{11}$$

5.1. Background model

A controlled background model resolves the problems arising from the implicit and inappropriate model above (Figs. 1 vs. 9). Properly setup, all five models presented in (4.2) can resolve the problems but mildly differ in performance in the large-scale experiments below.

In a small experiment, the background models are confronted with synthetically generated backgrounds which aim to mimic the face. The two background targets consist of the mean color of the face and of colors sampled from the color histogram of the face. The global and local Gaussian background models confuse the mean color with the face while the histogram model deals well with both cases, even though they share the same histogram (Fig. 10). None of the background models has troubles with a clearly distinct background, such as plain white or randomly colored.

5.2. Pose estimation

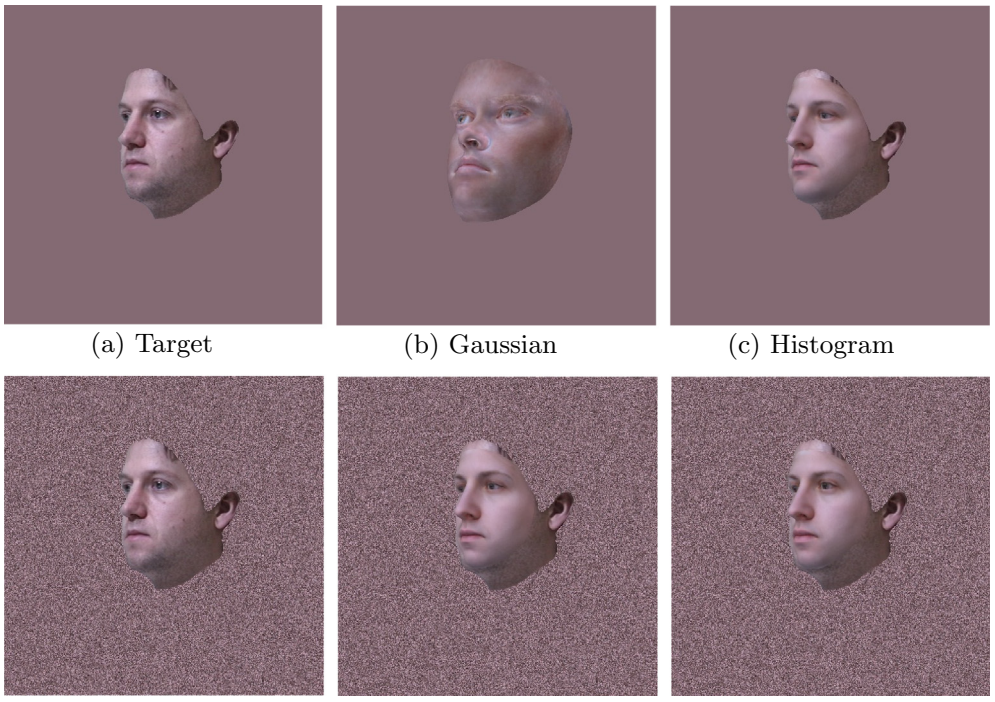
As an observation in practice, we can say that pose estimation in non-frontal face views is especially prone to background failures. In the relevant parameter regions, the visibility of pixels is heavily changing with respect to the rotation parameters.

We thus compare the finer differences among background models in a pose estimation problem on the Multi-PIE database [7]. The database contains images with strong but controlled pose variation. For the experiment, we selected neutral photographs of the 249 individuals in the first session with different yaw angles up to 90° (Table 1). Due to the lab setting, the background is very similar and rather controlled.

We use user-provided landmark information for initialization but the model fitting algorithm uses the fully flexible model with only image likelihood terms afterward, landmarks are ignored during the adaptation. The visibility is never fixed, it changes in each iteration. We chose the manual initialization to compare the performance of the free model after a certainly good initialization. We chose the yaw angle of the best reconstruction of 10,000 samples as a MAP estimate for yaw.

(a) Target (b) Image-based Fit (c) Model-based Fit

Fig. 9. The background model resolves the problems of both, image-based and model-based evaluation. The model fits are overlaid. Compare to Fig. 1 without correction. Also, the reconstruction quality of (b) is visually more satisfactory than (c).



(a) Target (b) Gaussian (c) Histogram

(d) Target (e) Gaussian (f) Histogram

Fig. 10. Synthetically generated backgrounds to mimic the face. In the first row, the background consists of the mean face color. In the lower row, the background is sampled from the face color histogram. The columns differ in the background likelihood model.

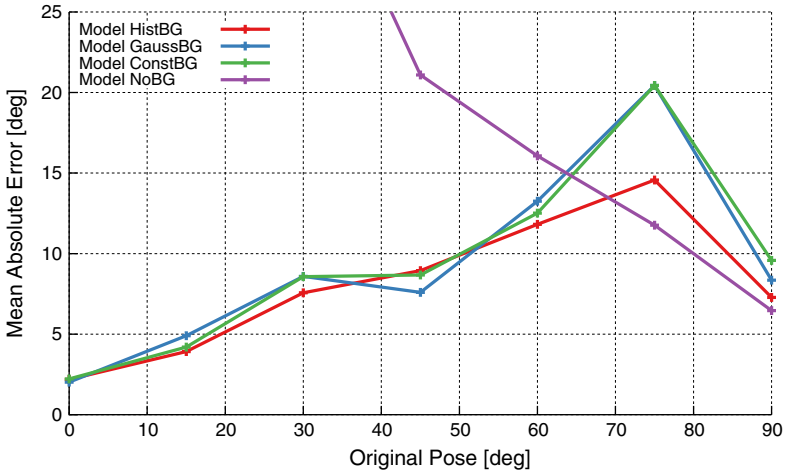
We compare both, image-based and model-based evaluation with all background models, including the unwanted implicit model $b = 1$ (called *NoBG*). We plot the resulting mean pose estimation (Fig. 11a), averaged for all 249 ids and also the Mean Absolute Error compared to the ground truth label as a measure of accuracy (Fig. 11b).

Table 1
Multi-PIE labels of images used in the experiments. *ID* is 001 . . . 250.

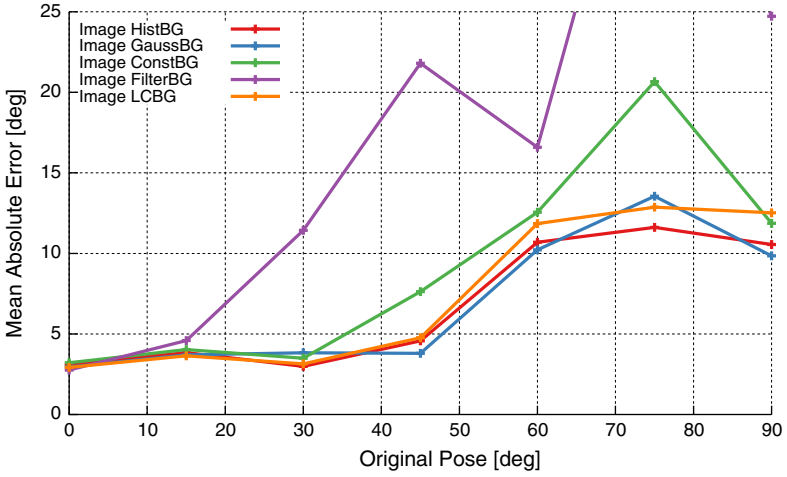
Yaw [deg]	Multi-PIE
0	*ID*_01_01_051_16
15	*ID*_01_01_140_16
30	*ID*_01_01_130_16
45	*ID*_01_01_080_16
60	*ID*_01_01_090_16
75	*ID*_01_01_120_16
90	*ID*_01_01_110_16

The experiments reveal the inappropriate setup without a background model as the worst of the runs. With image-based evaluation, there is not even a result to compare to due to failed fits. With model-based evaluation, the model constantly tries to occlude as much of the face as possible and therefore aligns the face in a strong side-view. The pose is severely overestimated for small angles whereas it actually fits for the side-view. Controlled background models resolve the problems and lead to acceptable results which are worst around angles with a high variability of visibility. There are only minor differences in performance between most choices of the background model.

The background model based on Gabor filter responses performs worst of all background models. A manual analysis of the results reveals a bad background reconstruction probability, foreground often "wins", even where not appropriate. This is probably due to a high complexity of this model, it is too general. In the original context in [16], this model is used in a more balanced setup with a foreground model which is also composed of



(a) Model-based evaluation. Ignoring background leads to strong misalignments with a preference for side-views.



(b) Image-based evaluation. The filter-based background model is too generic and cannot compete with the foreground model, it leads to strong misalignments with a preference for a frontal view.

Fig. 11. Pose estimation on Multi-PIE.

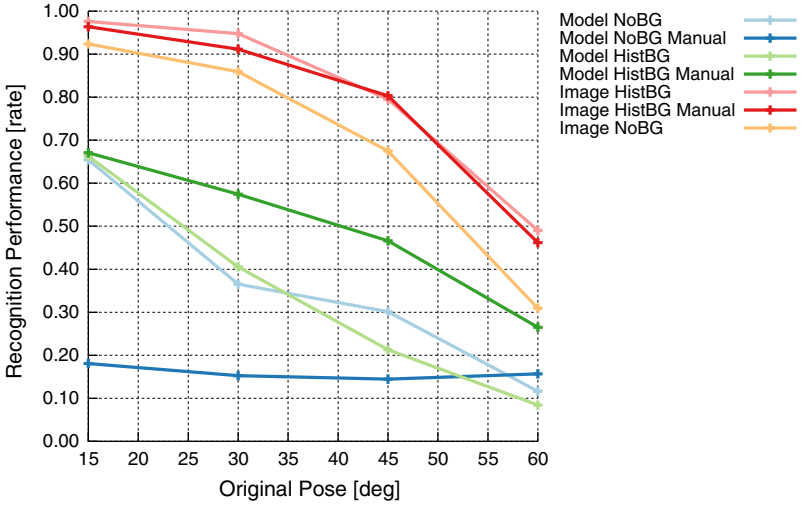


Fig. 12. Recognition performance on Multi-PIE. Background models improve the recognition performance.

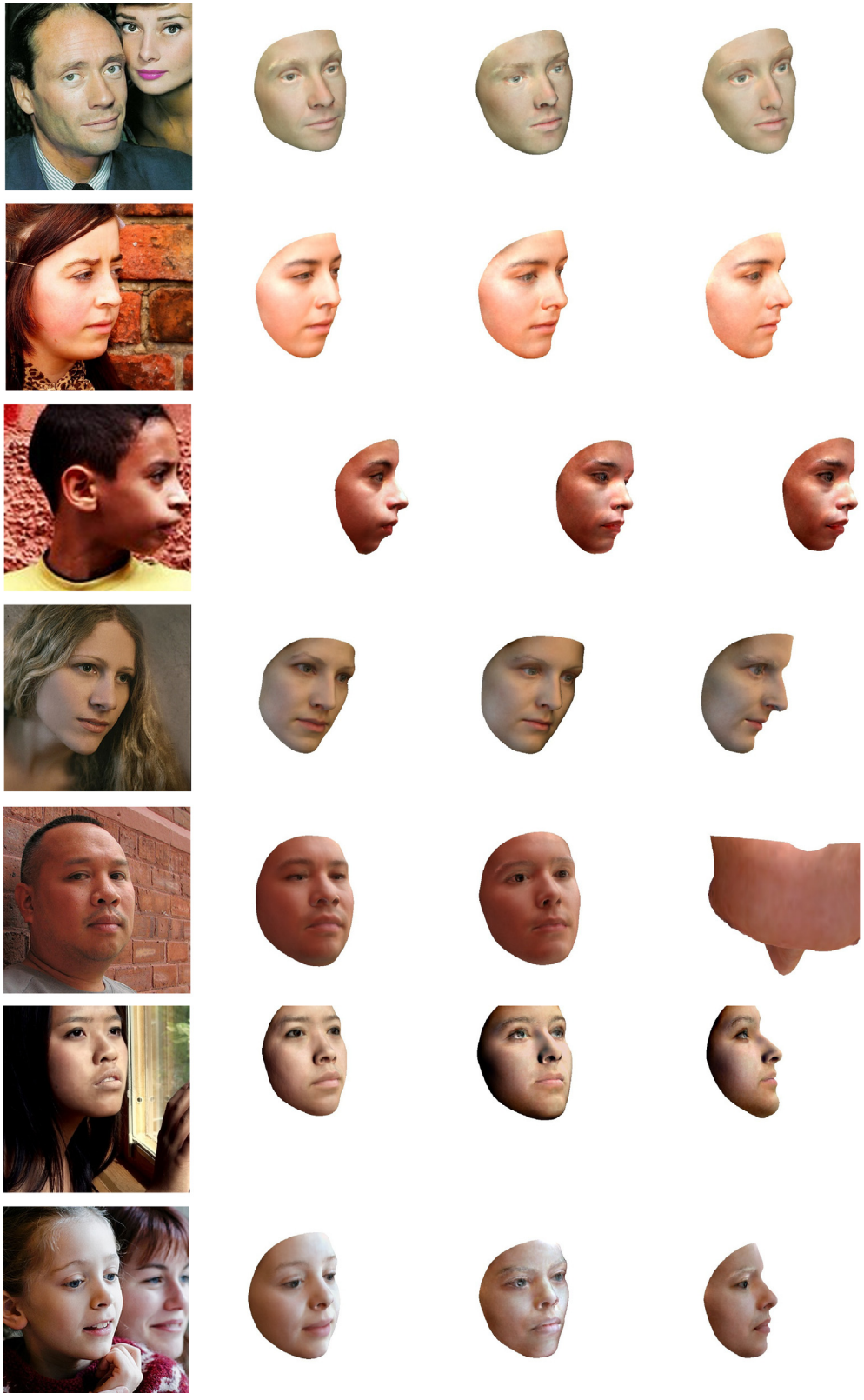


Fig. 13. Fitting results on a few selected images of the AFLW database. Each column has been created using a different background model. In (b), we used image-based evaluation while (c) and (d) are created using model-based evaluation (M). Background models are necessary but the exact type is uncritical. Face ids: 39370, 39796, 39853, 40199, 40505, 41008, 43311.

Gabor wavelets. In our context, this model is not useful in this form (see also Fig. 13).

We note that image-based evaluation yields higher quality reconstructions which also improve pose estimation over model-based evaluation.

The behavior of the complete background model (5) is equivalent to using the background correction (9). Both variants did not diverge in their results in this experiment.

5.3. Fully automatic recognition

Our fitting algorithm integrates unreliable detection information, resulting in a fully automatic face interpretation system. In this experiment, we test different background models for their performance when used in a fully automatic recognition pipeline [15].

Additionally to the pose estimation, we also perform a face recognition experiment with variable pose. The experiment setup is as described in [15] with a similarity measure S between two face representations f_1, f_2 from [3]:

$$S(f_1, f_2) = \frac{\langle f_1, f_2 \rangle}{\|f_1\| \|f_2\|}. \tag{12}$$

The image of a person in the frontal view serves as gallery whereas the non-frontal view is our probe. A recognition is successful if the most similar gallery image shows the same person. Contrary to the experiments above, the used fitting pipeline is fully automatic and we thus do not use any user-provided information. We do not evaluate with respect to all different background models since most of them perform very similarly.

The results show a decrease in performance with increasing yaw angle which is expected as the fitting quality deteriorates (Fig. 12). The automatic pipeline integrates unreliable detection through enforcing probabilistic consistency of the model with possible landmark detection candidates. Therefore, the solution has to adhere to detection results during the optimization. This forcing acts as a model restriction, keeping the model in place. The manually initialized runs (“Manual”) make use of the exact same setup as in the previous experiment. They rely on user-provided landmark positions for initialization only but do not restrict anything in the model.

The recognition experiments allow three main observations. First, image-based evaluation methods are superior to model-based evaluation. Second, an unconstrained model without a background model, even though properly initialized, is not able to reach useful recognition rates. The artifacts arising from the inadequate implicit background model are too strong (“Model NoBG Manual”, image-based failed). And third, the results improve by restriction through the automatic detections (“Model NoBG”, “Image NoBG”) but the best performance can be achieved with an explicit background model (“Model HistBG Manual”, “Image HistBG (Manual)”).

5.4. Qualitative evaluations on AFLW

The AFLW database [10] provides a high degree of variability of facial images taken in “real-world” scenarios. The Basel Face Model is a neutral face model without the possibility to deal with facial expression. Additionally, the fitting algorithm is not able to handle rough occlusions inside the face properly. These include strong make-up, beard or hair as well as various objects in front of the face. Therefore, we selected a few examples of these real-world photographs with considerable background variation to underpin the need for a background model when dealing with natural images of faces with real-world backgrounds.

The experiments are setup according to the manual initialization scenario, with manually labeled facial features. After initialization, nothing is kept fixed, as above.

The qualitative evaluation indicates the need of a background model (Fig. 13). However, the exact choice of background model is not critical to obtain a good fit. As long as a correction is made, the model adaptation converges. But the final quality of reconstruction then depends on the applicability of the background model for the current image. Without a background correction, there are the typical misalignment artifacts when using model-based evaluation while image-based evaluation completely fails.

5.5. Discussion

The experiments underpin the need for a controlled background model for both image-based and model-based likelihood evaluation. Without an explicit background model, the inappropriate implicit model with $b = 1$ becomes active with varying visibility. Variable visibility is not an exotic case but a realistic requirement for a face model without user-provided information. Even with user-provided initialization, the model can be expected to converge without keeping certain parameters or visibilities fixed during the optimization. Our experiments clearly show the danger the implicit background model brings in this case. The model diverges or seriously misaligns the face in the image.

The controlled models can be used either in their explicit form (5) or through the likelihood correction (9) which makes the model likelihood compete well with the implicit background model. We prefer the likelihood correction and suggest to use this formulation because it avoids evaluating the complete image.

The actually evaluated background models do not display many differences. The histogram model shows a stable performance in all experiments while the Gaussian model performs very well in the restricted Multi-PIE setup. The most important background model selection task is to actually use a background model while the concrete choice does not matter that much.

The fully automatic pipeline cannot yet deal with very large yaw angles well. For poses from frontal to 60° side views it already works as well as manually initialized models if image-based evaluation is used. Through the undetermined visibility, the background model becomes very important. An evaluation with the unconstrained implicit model severely breaks down with respect to recognition performance.

6. Conclusion

The face reconstruction of an image in unconstrained situations needs the 3DMM’s full flexibility, including variable visibilities.

The standard likelihood of generative models always includes an implicit background model if it is only evaluated on visible parts of the face. The parasitic model is equivalent to a unit likelihood background model which is rarely appropriate and leads to strong artifacts due to background preference. The effect reveals itself whenever the visibility of model parts can vary which is the case in unconstrained and uninformed situations like fully automatic model adaption.

While fitting algorithms so far fixed the visibility in advance or used additional information to constrain the model, we proposed and discussed the use of image-based background models to improve the likelihood fidelity. Our integration of background models through a simple likelihood ratio resolved the problem of evaluating the model on the complete image since the correction takes place completely within the foreground region. This implicit background model even made it possible to use an image-based background correction in model-based evaluation. A background

model with unit likelihood includes the standard likelihood approach above as a special case. Therefore, ignoring background in the standard likelihood formulation is equivalent to using an inappropriate background model.

Background modeling makes it possible to move to likelihood evaluation in the image domain which leads to face reconstructions of higher quality. Image-based evaluation severely suffers from variable visibilities and has not been possible without our background model before.

The discussion and evaluation of different background models promote the empirical model, based on a color histogram, as a useful background assumption. Generally, all evaluated background models led to fewer artifacts than the standard likelihood. Pose estimation at larger yaw angles with strong self-occlusion clearly benefits from a proper background correction.

The consequent future development direction of image-based model evaluation and background models is the step towards full image segmentation in conjunction with model adaption. Such a combination allows background inside the face region, a capability which can be used to remove face occlusion such as glasses, beards or hair.

References

[1] O. Aldrian, W. Smith, Inverse rendering of faces with a 3D morphable model, IEEE Trans. Pattern Anal. Mach. Intell. 35 (5) (2013) 1080–1093.

[2] V. Blanz, T. 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, ACM Press/Addison-Wesley Publishing Co., New York, USA, 1999, pp. 187–194.

[3] V. Blanz, T. Vetter, Face recognition based on fitting a 3D morphable model, IEEE Trans. Pattern Anal. Mach. Intell. 25 (9) (2003) 1063–1074.

[4] T. Cootes, G. Edwards, C. Taylor, Active appearance models, IEEE Trans. Pattern Anal. Mach. Intell. 23 (6) (2001) 681–685.

[5] T.F. Cootes, C.J. Taylor, D.H. Cooper, J. Graham, Active shape models-their training and application, Comput. Vis. Image Understand. 61 (1) (1995) 38–59.

[6] R. Gross, I. Matthews, S. Baker, Active appearance models with occlusion, Image Vis. Comput. 24 (6) (2006) 593–604.

[7] R. Gross, I. Matthews, J. Cohn, T. Kanade, S. Baker, Multi-PIE, Image Vis. Comput. 28 (5) (2010) 807–813.

[8] R. Knothe, A Global-to-Local Model for the Representation of Human Faces. PhD Thesis, University of Basel, Switzerland, 2009.

[9] I. Kokkinos, P. Maragos, Synergy between object recognition and image segmentation using the expectation–maximization algorithm, IEEE Trans. Pattern Anal. Mach. Intell. 31 (8) (2009) 1486–1501.

[10] M. Köstinger, P. Wohlhart, P.M. Roth, H. Bischof, Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization, in: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), 2011, pp. 2144–2151.

[11] I. Matthews, S. Baker, Active appearance models revisited, Int. J. Comput. Vis. 60 (2) (2004) 135–164.

[12] D. Mumford, J. Shah, Optimal approximations by piecewise smooth functions and associated variational problems, Commun. Pure Appl. Math. 42 (5) (1989) 577–685.

[13] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, T. Vetter, A 3D face model for pose and illumination invariant face recognition, in: 2009 Advanced Video and Signal Based Surveillance, 2009, pp. 296–301.

[14] S. Romdhani, 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, vol. 2, 2005, pp. 986–993.

[15] S. Schönborn, A. Forster, B. Egger, T. Vetter, A monte carlo strategy to integrate detection and model-based face analysis, in: J. Weickert, M. Hein, B. Schiele (Eds.), Pattern Recognition, Lecture Notes in Computer Science, 8142, Springer, Berlin, Heidelberg, 2013, pp. 101–110.

[16] Y.N. Wu, Z. Si, C. Fleming, S.-C. Zhu, Deformable template as active basis, in: IEEE 11th International Conference on Computer Vision, 2007. ICCV 2007, 2007, pp. 1–8.

[17] Y.N. Wu, Z. Si, H. Gong, S.-C. Zhu, Learning active basis model for object detection and recognition, Int. J. Comput. Vis. 90 (2) (2010) 198–235.

[18] S.C. Zhu, A. Yuille, Region competition: unifying snakes, region growing, and Bayes/MDL for multiband image segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 18 (9) (1996) 884–900.