Occlusion-aware 3D Morphable Face Models

Bernhard Egger
bernhard.egger@unibas.ch
Andreas Schneider
andreas.schneider@unibas.ch
Clemens Blumer
clemens.blumer@unibas.ch
Andreas Morel-Forster
andreas.forster@unibas.ch
Sandro Schönborn
sandro.schoenborn@unibas.ch
Thomas Vetter
thomas.vetter@unibas.ch

Department of Mathematics and
Computer Science
University of Basel
Basel Switzerland
http://gravis.cs.unibas.ch

Abstract

We propose a probabilistic occlusion-aware 3D Morphable Face Model adaptation framework for face image analysis based on the Analysis-by-Synthesis setup. In natural images, parts of the face are often occluded by a variety of objects. Such occlusions are a challenge for face model adaptation. We propose to segment the image into face and non-face regions and model them separately. The segmentation and the face model parameters are not known in advance and have to be adapted to the target image. A good segmentation is necessary to obtain a good face model fit and vice-versa. Therefore, face model adaptation and segmentation are solved together using an EM-like procedure. We use a stochastic sampling strategy based on the Metropolis-Hastings algorithm for face model parameter adaptation and a modified Chan-Vese segmentation for face region segmentation. Previous robust methods are limited to homogeneous, controlled illumination settings and tend to fail for important regions such as the eyes or mouth. We propose a RANSAC-based robust illumination estimation technique to handle complex illumination conditions. We do not use any manual annotation and the algorithm is not optimised to any specific kind of occlusion or database. We evaluate our method on a controlled and an "in the wild" database.

1 Introduction

Face image analysis is a major field in computer vision. We focus on 3D reconstruction of a face given a single still image. Since the problem of reconstructing a 3D shape from a single 2D image is inherently ill-posed, a strong object prior is required. Our approach builds on the Analysis-by-Synthesis strategy. We work with a 3D Morphable Face Model (3DMM) [2] representing the face by shape and colour parameters. With illumination and rendering parameters we can synthesise facial images. We adapt all parameters to render an image which is as similar as possible to the target. This process is called *fitting*.

Pages 64.1-64.11
DOI: https://dx.doi.org/10.5244/C.30.64

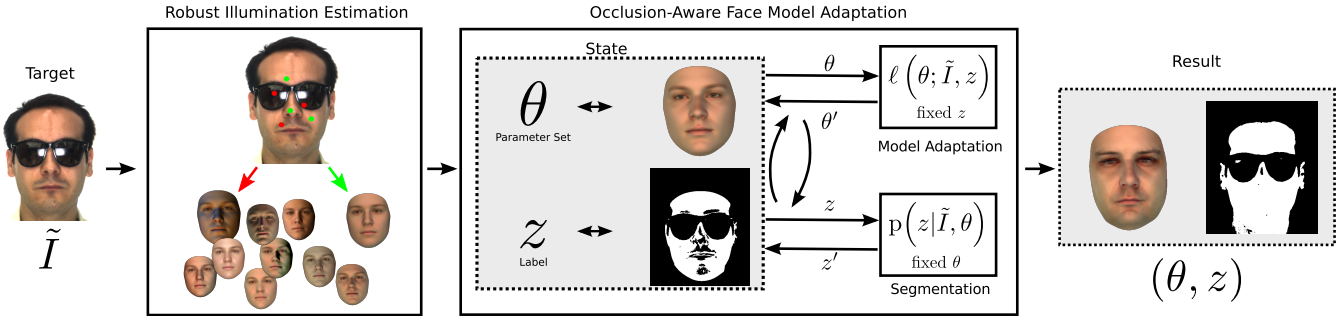

Figure 1: Algorithm overview: First we perform a RANSAC-like robust illumination estimation for initialisation of the segmentation label z and the illumination setting. Then our face model and the segmentation are simultaneously adapted to the target image $\tilde{I}$. The result is a set of face model parameters θ and a segmentation into face and non-face regions. The presented target image is from the AR face database [9].

Most "in the wild" face images contain occlusions like glasses, hands, facial hair, microphones and various other things [8]. Occlusions are a challenge in an Analysis-by-Synthesis setting. The fitting procedure of 3DMMs is misled by occlusions. Occlusions which differ from the face in colour appearance make the face model drift away, or shape and colour parameters are distorted because the face model adapts to those occlusions. At the same time, it is not possible to detect occlusions without strong prior knowledge.

Our approach is to segment the image into face and non-face regions. Non-face regions can arise from occlusions or outliers in the face or background region. We formulate an extended model that deals explicitly with occlusions. This enables us to adapt the face model only to regions labelled as face. During face model adaptation we also update the segmentation into face and non-face regions. Segmentation and parameter adaptation can not be performed separately. We require a set of face model parameters as prior for segmentation and a given segmentation for parameter adaptation. We therefore use an EM-like algorithm to perform segmentation and parameter adaptation in alternation.

Illumination determines facial appearance. Previous works on occlusion handling with 3DMMs were evaluated on controlled, homogeneous illumination only. However, complex illuminations are omnipresent in real photographs. We show the corrupting effect of illumination using a standard robust technique in Figure 2. We analyse "in the wild" facial images. They contain complex illumination settings and their occlusions cannot be handled with previous approaches. Therefore, we propose a robust illumination estimation. We incorporate the face shape prior in a novel RANSAC-like algorithm for robust illumination estimation. The results of this illumination estimation are illumination parameters and an initial segmentation into face and non-face regions. We use this robust illumination estimation to get a good initialisation for our face model adaptation and the segmentation.

1.1 Related Work

3DMMs are widely applied for 3D reconstruction from single 2D images [1, 2, 7, 12, 13, 16]. Although occlusions are omnipresent in face images, most research using 3DMMs relies on occlusion-free data. There exists only few approaches for fitting a 3DMM under occlusion. Standard robust error measures are not sufficient for generative face image analysis. Areas like mouth or eye regions tend to be excluded from the fitting because of their strong variability in appearance [4, 12], and robust error measures are highly sensitive to illumination

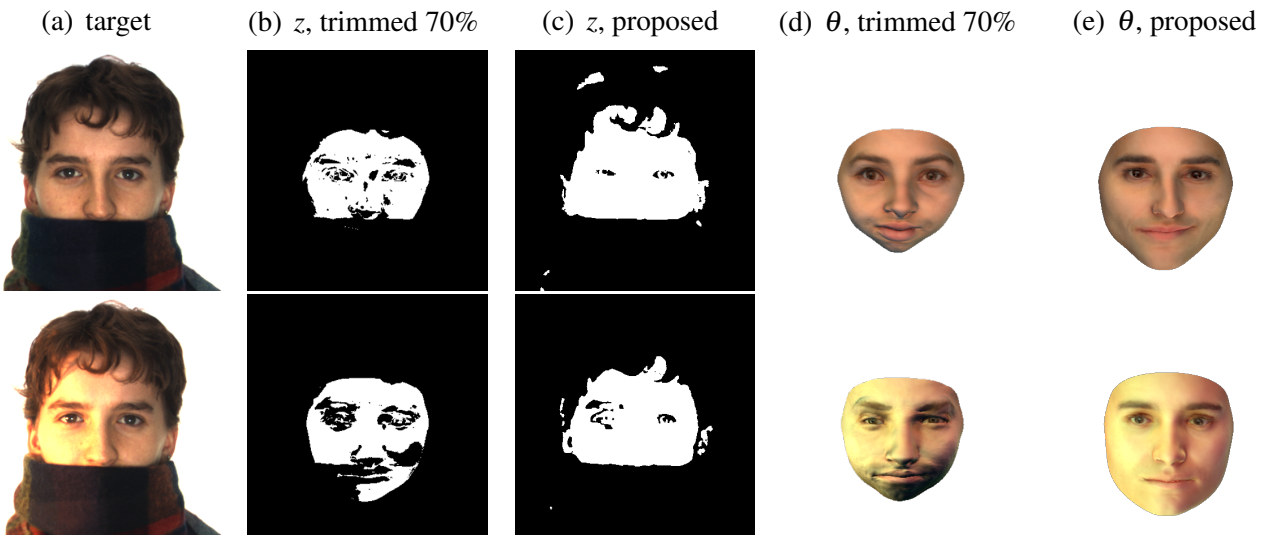

Figure 2: We present the effect of illumination on a standard robust technique. (a) depicts the target image from the AR face database [9] under homogeneous and inhomogeneous illumination. (b) and (d) are showing the segmentation and fitting result using a trimmed evaluator (70%, explained in Section 3.1). (c) and (e) depicts the segmentation and fitting result of the proposed method. Whilst the segmentation using the trimmed version succeeds only with homogeneous frontal illumination and fails with illumination from the side, our approach is illumination-invariant and succeeds in both cases for segmentation and fitting.

[11]. Therefore, we explicitly cover as many pixels as possible by the face model.

De Smet *et al*. [4] learned an appearance distribution of the observed occlusion per image. This approach focuses on large-area occlusions like sunglasses and scarves. However it is sensitive to appearance changes due to illumination and cannot handle thin occlusions. Additionally the work uses manually labeled fiducial points.

Note that previous works on occlusion handling using a 3DMM focussed on databases with artificial and homogeneous, frontal illumination settings. To the best of our knowledge we present the first model which can handle occlusion during 3DMM adaptation on illumination conditions arising in "in the wild" databases.

2 Methods

We propose to combine occlusion segmentation and 3DMM adaptation into an occlusion-aware face analysis framework (Figure 1). Additionally, the challenge of initialisation of the segmentation is solved by a RANSAC strategy for robust illumination estimation.

Our approach is based on four main ideas. First, occlusions should be excluded during the face model adaptation. The face model should be adapted only to pixels belonging to the face. We propose an extended face model likelihood to exclude pixels which are labeled as non-face and treated by a non-face model. For the adaptation of the face model parameters we use a Metropolis-Hastings algorithm [13]. Second, we need to segment the image into face and non-face regions. This is performed using the facial appearance prior arising from the statistical face model. For segmentation, we use a variant of the algorithm by Chan-Vese which leads to contiguous regions [3]. Third, illumination determines facial appearance. We therefore use a RANSAC-like algorithm [6] to get a robust segmentation and illumination estimation as initialisation. Fourth, face model adaptation and segmentation has to

be performed at the same time. The face model adaptation assumes a given segmentation and vice-versa. Combined face model adaptation and segmentation is performed using an EM-like procedure [5].

2.1 Face Model

The 3D Morphable Face Model was first described by Blanz and Vetter [2]. Principal Component Analysis is applied to build a statistical model of face shape and colour. Faces are synthesised by rendering them with an illumination and camera model. We work with the publicly available Basel Face Model (BFM) presented by Paysan *et al.* [10]. The model was interpreted as a Bayesian face model using PPCA by Schönborn *et al.* [13].

The aim of face model adaptation (fitting) is to synthesise a face image that appears as similar as possible to the target image. A likelihood model is used to rate parameters given a target image. The likelihood model consists of a product over the pixels i of the target image $\tilde{I}_i$. Pixels are covered by the face model ($\mathcal{F}$) or by the background model ($\mathcal{B}$). The foreground and background likelihoods (ℓ_{face}, b) compete to explain pixels in the image. The full likelihood model covering all pixels i in the image is

$$\ell\left(\theta;\tilde{I}\right) = \prod_{i\in\mathcal{F}} \ell_{\text{face}}\left(\theta;\tilde{I}_i\right) \prod_{i'\in\mathcal{B}} b\left(\tilde{I}_{i'}\right). \tag{1}$$

2.2 Occlusion Extension of the Face Model

We extend (1) to be able to handle occlusion. Therefore, we introduce a label z to distinguish between pixels which contain information about the face ($z=1$) and pixels which depict non-face regions ($z=0$). The generative face model will be adapted to the face pixels only, non-face pixels should be excluded from the face model adaptation. Non-face pixels are only characterised by not being covered by the explanation of the face model, they can be outliers, occlusions or background pixels. The likelihood (1) is therefore rewritten as

$$\ell\left(\theta;\tilde{I},z\right) = \prod_{i} \ell_{\text{face}}\left(\theta;\tilde{I}_i\right)^{z} \cdot \ell_{\text{non-face}}\left(\theta;\tilde{I}_i\right)^{1-z}. \tag{2}$$

The main difference to the formulation by [13] is that the face model does not have to fit to all pixels in the face region. Those pixels can alternatively be explained as occlusion via the non-face model. In Figure 3 we give an overview over the different labels and regions.

The label z is a random variable. We use an extension of the variational Chan-Vese segmentation technique [3] for a MAP-estimation of the label z given θ and our appearance prior. The classical simple colour models from the Chan-Vese segmentation are replaced by our face and non-face models. We reformulate the energy term E to fit in our probabilistic framework:

$$-\log \mathrm{p}\left(z|\tilde{I},\theta\right) = E = \Psi + \int_{\Omega} z(x)\log \ell_{\text{face}}(\theta;\tilde{I}(x)) + (1-z(x))\log \ell_{\text{non-face}}(\theta;\tilde{I}(x))\,\mathrm{d}x \tag{3}$$

Ψ is the length term of the classical Chan-Vese formulation regularising the boundary of the level set. The aim of the segmentation is to find our binary label z which is deduced from the level-set formulation ϕ using the Heavyside function: $z(x) = H(\phi(x))$, where x is the continuous position in the whole image domain Ω. We use the original discretisation scheme proposed by [3] on our pixel grid.

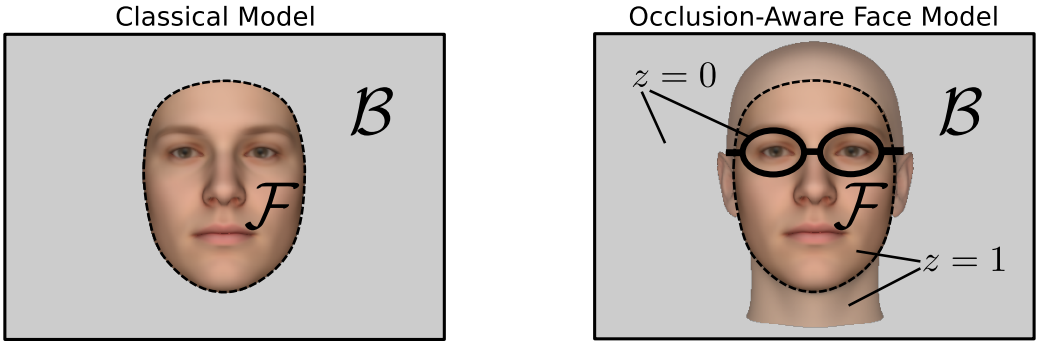

Figure 3: On the left we present the regions used by the likelihood model by [13] presented in (1). Each pixel belongs to the face model region $\mathcal{F}$ or the background model region $\mathcal{B}$. The assignment to foreground or background is based on the face model visibility only. To integrate occlusions we add a label z (see (2)) as shown on the right. We assign a label z indicating if the pixel belongs to face or non-face. Occlusions in the face model region $\mathcal{F}$ (in this case glasses) can hereby be excluded from the face model adaptation.

Depending on the label z we choose a likelihood model for each pixel. The likelihood of pixels labelled as face ($z = 1$) is the following:

$$\ell_{\text{face}}\left(\theta;\tilde{I}_i\right) = \begin{cases} \frac{1}{N}\exp\left(-\frac{1}{2\sigma^2}\left\|\tilde{I}_i - I_i(\theta)\right\|^2\right) & \text{if } i \in \mathcal{F} \\ \frac{1}{\delta}h(\tilde{I}_i) & \text{if } i \in \mathcal{B}. \end{cases} \tag{4}$$

Pixels are evaluated by the face model if they are labelled as face ($z = 1$) and are located in face region $\mathcal{F}$. The rendering function I generates a face for given parameters. This synthesised image is compared to the observed image $\tilde{I}$. The likelihood model for pixels covered by the face model is assuming per-pixel Gaussian noise in the face region. The likelihood ℓ_{face} is defined over the whole image and therefore also in the non-face region $\mathcal{B}$. Those pixels can not be explained by the generative face model, therefore we use a simple colour model for their likelihood. We use a colour histogram h with δ bins estimated on all pixels in $\mathcal{F}$ labeled as face ($z = 1$).

Pixels i which are labelled as non-face ($z = 0$) are evaluated by the following likelihood:

$$\ell_{\text{non-face}}\left(\theta,\tilde{I}_i\right) = b\left(\tilde{I}_i\right). \tag{5}$$

This likelihood allows us to integrate simple colour models based on the observed image $\tilde{I}$. The paper on background modelling by Schönborn *et al.* [14] gives an overview over possible models; we use the simplest one, a constant likelihood.

2.3 Inference

The full model consists of the likelihoods for face model adaptation shown in (2) and segmentation from (3). Those equations depend either on a given parameter set θ or a given label z. Both are not known in advance and are adapted during the inference process to get a MAP-estimate of the face model parameters and the segmentation. We therefore use an EM-like algorithm [5] for alternating inference of the full model. In the expectation step, we update the label assignment z of the segmentation. In the maximisation step, we adapt the face model parameters θ. An overview of those alternating steps can be found in Figure 1.

Algorithm 1: Robust Illumination Estimation.

Input: Pose and camera parameters, mean shape and colour
Output: Illumination parameters, initial segmentation label z
for *200 iterations* **do**
1. Select 30 random points on face surface
2. Estimate illumination on those points
3. Evaluate prediction for other points (consensus set)
4. Save consensus set if better than previous

Face model adaptation is performed by a Markov Chain Monte Carlo strategy [13] with our extended likelihood from (2). Segmentation is performed using a extended version of the Chan-Vese segmentation, where the simple colour models are replaced by our face model and a background model (3).

During segmentation we assume fixed parameters θ and during fitting we assume given labels z. Since those fixed values are only an approximation during the optimisation process, we account for small inaccuracies. This enables to also include interesting regions like the eye, nose and mouth regions which are often ignored due to their high variability in appearance when using robust error measures. The likelihood of the face model used for segmentation is adapted to consider a small misalignment given by the face model parameters (compare to (4)). Some facial features in the synthesised image are not yet aligned perfectly to the location in the target image:

$$\ell'_{\text{face}}(\theta;\tilde{I}_i) = \frac{1}{N}\exp\left(-\frac{1}{2\sigma^2}\min_{n\in N(i)}\left\|\tilde{I}_i - I_{i,n}(\theta)\right\|^2\right) \text{if } i \in \mathcal{F}. \tag{6}$$

The small misalignment of the current status of the fit is taken into account by the neighbouring pixels N in the target image. In our case we take the minimum over a patch of the $9x9$ neighbouring pixels direction (interpupillary distance is ~120 pixels).

The likelihood of the non-face model during the face model adaptation is also extended (compare to (5)). Pixels which are masked as non-face can be explained not only by the background model, but also by the face model if it can do better:

$$\ell'_{\text{non-face}}\left(\theta,\tilde{I}_i\right) = \max\left(\ell_{\text{face}}\left(\theta,\tilde{I}_i\right), b\left(\tilde{I}_i\right)\right) \text{if } i \in \mathcal{F}. \tag{7}$$

Both extensions of the likelihoods lead to a convergence behaviour which tends to label more pixels as face and consider them during face model adaptation.

2.4 Initialisation and Robust Illumination Estimation

In the early steps of the face model adaptation under occlusion we need an initial label z. Occlusions are however hard to determine in the beginning of the face model adaptation due to the strong influence of illumination on facial appearance (see Figure 2). We therefore first adapt a simplified face model using a RANSAC strategy [6].

We adapt only the illumination of the simplified face model and use the reflectance and shape of the mean face. Given shape, reflectance and the target image, the spherical harmonics-based illumination parameters can be estimated by solving a linear system [13].

Method	neutral	glasses	scarf
Trimmed Estimator 70%	0.73 (0.69|0.27)	0.76 (0.68|0.54)	0.64 (0.53|0.44)
Trimmed Estimator 80%	0.77 (0.75|0.30)	0.73 (0.65|0.45)	0.66 (0.55|0.44)
Trimmed Estimator 90%	0.82 (0.80|0.33)	0.66 (0.60|0.31)	0.68 (0.57|0.45)
Robust illumination estimation	0.79 (0.77|0.28)	0.81 (0.72|0.64)	0.75 (0.61|0.59)
Full model	**0.90 (0.89|0.42)**	**0.88 (0.83|0.73)**	**0.84 (0.77|0.66)**

Table 1: Comparison of segmentation performance in SMC and in brackets the JSC (face|non-face) on the AR face database [9]. We compare our results to a robust fitting method that uses n percent of pixels matching the target image best (lines 1-3). We present seperate results for our initalisation using robust illumination estimation (line 4) and the full model including segmentation (line 5).

Since we do not know in advance which pixels can be used for this estimation, we perform 200 RANSAC iterations with 30 points on the surface. Pseudo-Code of the RANSAC-like procedure can be found in Algorithm 1. The algorithm was in our experiments not sensitive to the exact choice of those parameters. The points are sampled in regions where the colour variance of the face model is smaller than half the maximal variance. Points with higher variance are not suited for illumination estimation using the reflectance of the mean face. This procedure leads to a good initialisation of the illumination setting - and is robust to occlusions and outliers.

Our robust illumination estimation gives a rough estimate of the illumination and segmentation. However the obtained mask is underestimating the face region. Especially the eye, eyebrow and mouth regions are not included in this first estimate. Those regions differ from the skin regions of the face by their higher variance in appearance. In the inference process using the full face model, those regions are incorporated gradually by (6) and (7).

3 Experiments and Results

We present results investigating the performance of our segmentation and quality of the fit. For the 3DMM adaptation, the 3D pose has to be initialised. In the literature, this is performed manually [1, 2, 12] or by using fiducial point detections [13]. For all our experiments, we use automatic fiducial point detection results from the CLandmark Library [15]. Our method is therefore fully automatic and does not need manual input.

For all experiments, we perform alternatingly 2,000 Metropolis-Hastings sampling steps (best sample is taken to proceed) followed by a segmentation step with 500 iterations and repeat this procedure five times. This amounts to a total of 10,000 samples and 2,500 segmentation iterations.

3.1 Segmentation

We evaluate the segmentation on the AR face database [9]. We use a subset that contains the first 10 male and female participants appearing in both sessions. In contrast to other evaluations, we include images under illuminations from the front and the side. We selected the neutral (containing glasses and beards) images from the first session and images with scarves and sunglasses from the second session. The total set consists of 120 images. We had to exclude 4 images because the fiducial point detection failed for them. For

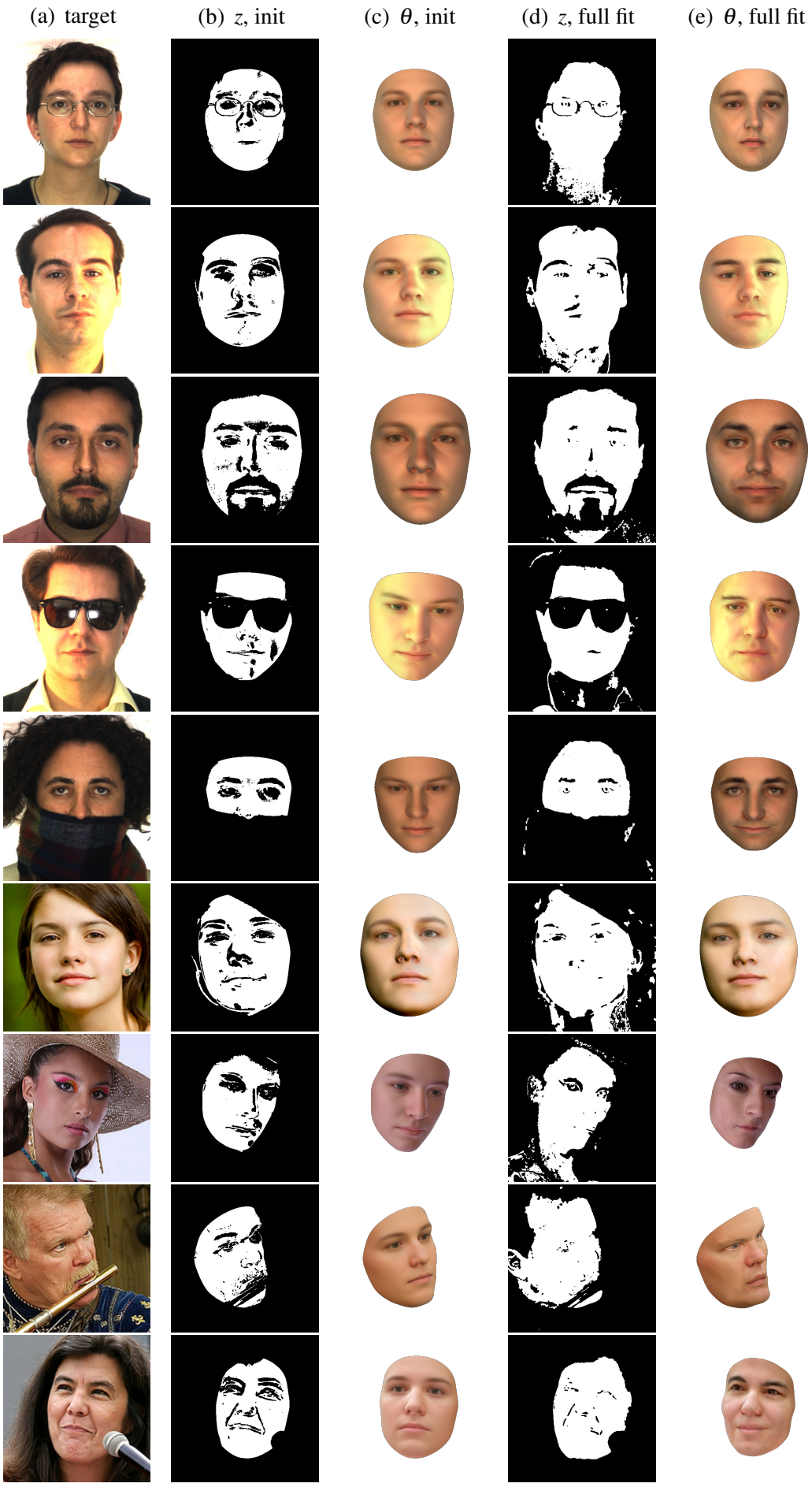

Figure 4: (a) shows the target images from the AR face database (first five) [9] and the AFLW database [8]. (b) and (c) depict our initialisation arising from the robust illumination estimation, (d) and (e) present the final results. Our final segmentation and synthesized face includes much more information of the eye, eyebrow, nose and mouth regions than the initalisation.

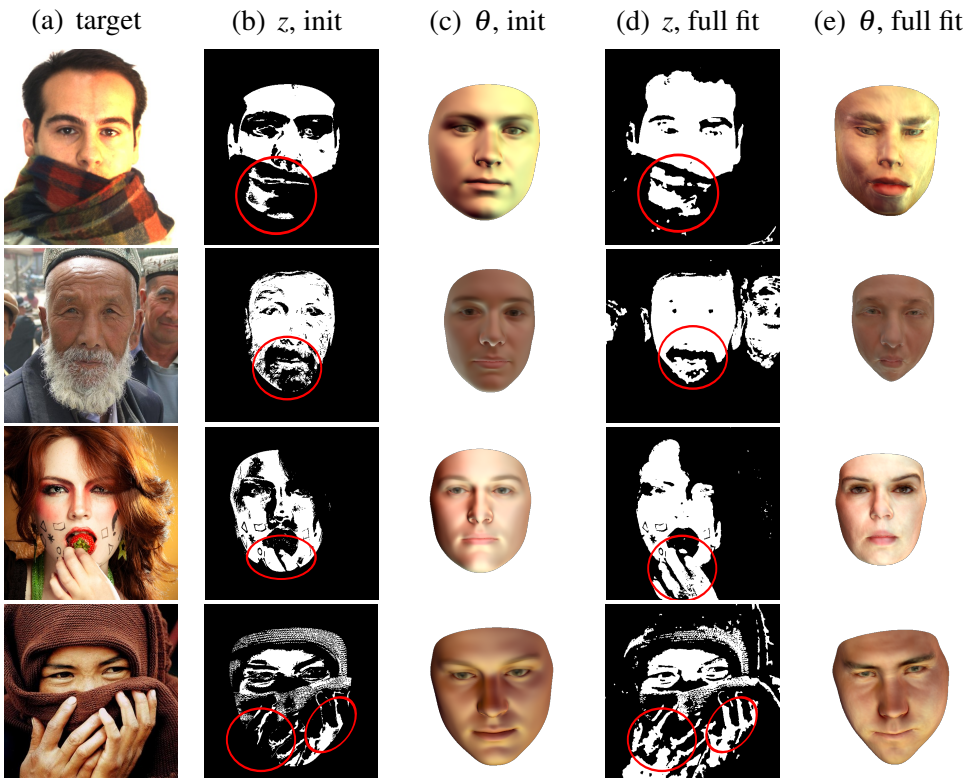

Figure 5: We present usual cases of failure (errors in circles): beards and scarves can be explained by the light and colour model and are therefore mislabelled. Hands have similar color appearance and do not distort the face model adaptation but lead to a wrong segmentation. Note that our method is not adapted to a specific kind of outlier.

evaluation we labeled an elliptical face region and occlusions manually. The segmentation error was measured in the elliptical face region. Our manual annotations, used for evaluation only, are available under http://gravis.cs.unibas.ch/publications/2016/2016_Occlusion-aware_3D_Morphable_Models.zip.

We compare our method to a standard technique, namely a trimmed estimator including only n % of the pixels which are best explained by the face model. In Table 1 we present the simple matching coefficient (SMC) and the Jaccard similarity coefficient (JSC) for the three image settings: neutral, scarf and sunglasses. We include the result of the initial illumination estimation to depict its contribution and show that the fitting improves the segmentation even more. Our approach reaches the best result on all settings. In Figure 2 we present an example where the trimmed estimator fails whilst our approach succeeds.

3.2 Quality of Fit

We present qualitative results of our fitting quality on the AR face database [9] and the Annotated Facial Landmarks in the Wild (AFLW) database [8]. In our results in Figure 4, we include various sources of occlusions like glasses, beards, facial hair, make-up and others. In Figure 5 we also include results where our method fails. Our method detects occlusions by an appearance prior from the face model. If occlusions can be explained by the colour or illumination model, the segmentation will be wrong. The fitting quality results on the AFLW database show almost the same performance as we obtain on data without occlusions. Interesting parts of the face are included gradually during the fitting process.

4 Conclusion

We proposed a novel approach for combined segmentation and 3D Morphable Face Model adaptation to integrate occlusion. The result of our method is both a set of Morphable Face Model parameters and a segmentation of the image into face and non-face regions. Our method succeeds in integrating regions like the eye, eyebrow, nose and mouth regions. Those regions are harder to fit by the face model and are therefore often excluded by other robust methods. For initialisation, we use a RANSAC-like robust illumination estimation. The resulting occlusion-aware face analysis framework does not require any manual input or database adaptation and is therefore fully automatic. This is the first fully automatic approach to occlusion-aware 3DMM adaptation. We are also first to present results under complex illumination settings and on an "in the wild" database for various kinds of occlusions.

References

[1] 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. ISSN 0162-8828. doi: 10.1109/TPAMI.2012.206.

[2] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *SIGGRAPH'99 Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194. ACM Press, 1999.

[3] Tony F Chan and Luminita A Vese. Active contours without edges. *Image processing, IEEE transactions on*, 10(2):266–277, 2001.

[4] 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.

[5] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, pages 1–38, 1977.

[6] 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.

[7] 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.

[8] Martin Koestinger, 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 *First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.

[9] Aleix M Martinez and Robert Benavente. The ar face database. *CVC Technical Report*, 24, 1998.

[10] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *Proceedings of the 6th IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS)*, pages 296–301. IEEE, 2009.

[11] 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.

[12] 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. IEEE, 2003.

[13] Sandro Schönborn, Andreas Forster, Bernhard Egger, and Thomas Vetter. A monte carlo strategy to integrate detection and model-based face analysis. In *Pattern Recognition*, pages 101–110. Springer, 2013.

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

[15] Michal Uřičář, Vojtěch Franc, Diego Thomas, Sugimoto Akihiro, and Václav Hlaváč. Real-time Multi-view Facial Landmark Detector Learned by the Structured Output SVM. In *11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2015*, volume 02, pages 1–8, May 2015. doi: 10.1109/FG.2015.7284810.

[16] Xiangyu Zhu, Junjie Yan, Dong Yi, Zhen Lei, and Stan Li. Discriminative 3d morphable model fitting. In *Proceedings of 11th IEEE International Conference on Automatic Face and Gesture Recognition FG2015*, Ljubljana, Slovenia, 2015.