A Monte Carlo Strategy to Integrate Detection and Model-Based Face Analysis

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

Department for Mathematics and Computer Science
University of Basel, Switzerland
{sandro.schoenborn,andreas.forster,bernhard.egger,thomas.vetter}@unibas.ch

Abstract. We present a novel probabilistic approach for fitting a statistical model to an image. A 3D Morphable Model (3DMM) of faces is interpreted as a generative (Top-Down) Bayesian model. Random Forests are used as noisy detectors (Bottom-Up) for the face and facial landmark positions. The Top-Down and Bottom-Up parts are then combined using a Data-Driven Markov Chain Monte Carlo Method (DDMCMC). As core of the integration, we use the Metropolis-Hastings algorithm which has two main advantages. First, the algorithm can handle unreliable detections and therefore does not need the detectors to take an early and possible wrong hard decision before fitting. Second, it is open for integration of various cues to guide the fitting process. Based on the proposed approach, we implemented a completely automatic, pose and illumination invariant face recognition application. We are able to train and test the building blocks of our application on different databases. The system is evaluated on the Multi-PIE database and reaches state of the art performance.

1 Introduction

Face image understanding is a very important problem in computer vision. We propose a method to extend the model-based image explanation concept combining Top-Down and Bottom-Up knowledge. Generative models are a wide-spread Top-Down method to interpret images. An image is explained by model parameters obtained with an Analysis-by-Synthesis approach [11]. Given a target image, the model's parameters are adapted (fitting) until the generated image is most similar to the input image and the corresponding parameter values (fit) are taken as image description.

We apply a 3DMM [16] to explain images of human faces. Traditionally, the fitting of a 3DMM to an image requires a good initialization of the applied optimization algorithm to find the best set of parameters. As automatic detection (Bottom-Up) of facial feature points is unreliable regarding strong pose and illumination variation, a new concept is needed to properly integrate such information.

We present a method to reinterpret the model fitting process, which opens doors to the integration of various sources of information. As a concrete example, we show how to integrate unreliable face and facial feature point detectors

J. Weickert, M. Hein, and B. Schiele (Eds.): GCPR 2013, LNCS 8142, pp. 101–110, 2013.

without forcing an early detection decision, which might be based on too little information. Not only the single best detection but the detector response over an area is fused with the prior knowledge of the 3DMM in the fitting process.

We formulate the 3DMM as a probabilistic model and the fitting problem as an inference problem. The fitting process is generalized to a sampling process, drawing samples from the posterior distribution over the model's parameters given an input image. The Metropolis-Hastings algorithm, which is the core of the integration, allows us to include different sources of information as proposal distributions. The used method is an example of a DDMCMC method, as proposed by Tu [21] for image parsing. Combined with Random Forests for feature point detection, this leads to a fully automatically initialized fitter which can deal with unreliable information of different origins in the form of proposal distributions.

To successfully integrate the detection into the model fitting of a 3DMM, we first need to detect face candidates. For each possible face box, the detection maps of facial features need then to be interpreted using model knowledge, this is stated as a sampling process. The samples from all the face boxes need to be combined into a single distribution ("detection posterior") which is then, in the last step, conditioned on the image to obtain the posterior distribution ("image posterior"). The samples from this final distribution represent the model-based image explanation.

To demonstrate the use of the proposed approach, we solve a face recognition task on the Multi-PIE database [12] with state of the art results. The recognition system is built as a direct application on-top of this general purpose face image understanding method. The result is a database-independent recognition system. As a big advantage, the concept of our approach would remain valid and applicable, if the model will be extended to incorporate additional information, such as expressions, ethnic variability or masks to cope with outliers like hair, glasses or beards.

2 Prior Work

2.1 Morphable Model

A 3DMM has been proposed by Blanz and Vetter [3,4] to generatively explain and analyze images of faces. The 3DMM consists of a parametric statistical model of the 3D shapes and textures of faces obtained from a 3D scanner. The faces are brought into dense correspondence before building the statistical model. The model [16] has been successfully used to solve a wide range of problems.

An image is generatively explained or interpreted by the 3DMM by adapting the set of parameters to the image. In a Analysis-by-Synthesis setting, a cost function, measuring the degree of fit of a rendered image, is optimized by standard procedures, such as LBFGS or conjugate gradient methods leading to a fit, often ending in a local optimum. The optimization process needs to be initialized since an exhaustive search is not feasible. This initialization is usually provided by manually annotated landmarks.

2.2 Detection

A good overview of approaches for face detection is given by [23]. The most influential work during the last decade was by Viola and Jones [13]. We use similar features but combine these with a derived Random Forest algorithm based on [5]. Many elaborate approaches tackle facial landmark localization, such as [2] or [6]. We use the same Random Forest algorithm as for face detection.

The idea of local detection is limited by principle, the different local parts have no global consistency. The global consistency additionally needs a model coupling the individual responses. We use the 3DMM to provide the prior knowledge needed. Other important approaches include the pictorial structures models operating in the image plane [9,1].

2.3 Markov Chain Monte Carlo

Markov Chain Monte Carlo methods proved to be a useful tool to handle probability distributions which are more complicated than the most simple analytically tractable ones. In our case, we apply the Metropolis-Hastings algorithm [20]. As the algorithm is very general in nature, it is applied to a variety of different computer vision problems [10,21,22,19].

In computer vision, a useful parametric model consists of many parameters of different scale and meaning to the image forming process. It is not straightforward to design a sampler which can efficiently deal with this.

Most MCMC methods rely completely on designed and fixed proposal distributions, mostly random walks in parameter space. A newer development in the sampling literature are data-driven proposal distributions which make use of the input data to form probably useful proposals (heuristics). DDMCMC methods have been used to segment images [21], do inference about a complex 3D scene using only monocular input [22], to infer the pose of a human body model [19] or to localize faces [15].

Compared to other approaches, our model is not of a composite form. We need to adapt a complex parametric model having many continuous parameters with different interactions and will not use detection to propose additional object hypotheses. We will focus not on model selection but on the adaption step of models with continuous parameters, the fitting. The model we use parameterizes the geometry and appearance of a surface rendered to a 2D view and has thus more parameters with very different roles in the image formation process.

3 Methods

3.1 Bayesian Face Model

The generative 3DMM is a parametric face model which is able to render an image $I_{\mathrm{M}}(\theta)$ given some parameter values θ. The parameters include camera settings, the illumination and the PCA face description split into a shape and a texture part. Using an additional noisy observation model of the generated

image $P(I|I_{\mathrm{M}}(\theta))$, this model can be interpreted in a probabilistic framework and used in a Bayesian setting. The Bayesian posterior distribution then consists of the image formation part and the prior on the model parameters $P(\theta)$:

$$P(\theta|I) \propto P(I|\theta)P(\theta). \tag{1}$$

The traditional fitting approach then corresponds to a maximum-a-posteriori (MAP) inference, finding the parameters with the highest posterior probability. In practice, this yields only local optima of the cost function.

As a noise model, we use the probabilistic interpretation of the traditional least squares cost function which is the isotropic Gaussian distribution, treating each pixel independently

$$P(I|\theta) = \prod_{p\in\mathrm{FG}} \mathcal{N}(I(p)|I_{\mathrm{M}}(p;\theta),\Sigma) \prod_{p\in\mathrm{BG}} \mathcal{N}(I(p)|\mu_{\mathrm{BG}},\Sigma_{\mathrm{BG}}), \tag{2}$$

where $\Sigma = \sigma^2 I_3$ is the covariance matrix. The pixels lying outside the generated face are considered background (BG) and their likelihood is evaluated using a multivariate Gaussian $\mathcal{N}(\mu_{\mathrm{BG}},\Sigma_{\mathrm{BG}})$ trained on all pixels in the observed image. A background model is needed to fully explain the observed image preventing partial explanation effects, such as "shrinking" of the face in the image. We evaluate all values related to pixels in the RGB color space.

We adopted the rendering process of the original 3DMM (see [3]) but changed the illumination model from a Phong model to a spherical harmonics-based global illumination model with two bands [18]. Such an illumination model allows us to obtain the optimal illumination coefficients by solving a linear system, for a fixed geometric setting.

For the 3DMM, we use a slightly modified Basel Face Model (BFM) [16] without ears and throat. The model comes with a statistical prior on the face shape and face texture. We use a broad multivariate Gaussian prior for the camera and illumination models, obtained by analyzing 20k face images in the AFLW database [14].

The 3DMM can also render the position of the facial feature points in the image plane $\hat{\boldsymbol{x}}_i(\theta)$. The observation model of these points is again an independent isotropic Gaussian distribution with standard deviation σ_{LM}. It provides the likelihood of F observed landmark positions $\{\boldsymbol{x}_i\}_{i=1}^{F}$

$$P(\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_F|\theta) = \prod_{i=1}^{F} \mathcal{N}(\boldsymbol{x}_i|\hat{\boldsymbol{x}}_i(\theta),\sigma_{\mathrm{LM}}^2). \tag{3}$$

3.2 Fitting by Sampling from Posterior

The probabilistic interpretation of the 3DMM allows us to deal with uncertainty and thus also to integrate unreliable hints properly. The fitting process changes from an optimization problem to a process inferring the posterior distribution (1).

The rather complicated image generation setting leads to a posterior distribution without a simple representation. We resort to the Metropolis-Hastings algorithm to simulate samples from the posterior. The algorithm transforms samples θ' from a proposal distribution $Q(\theta'|\theta)$ into samples distributed according to the target posterior distribution $P(\theta|I)$ by stochastically accepting or rejecting samples. The specific choice of this algorithm additionally allows us to work with an unnormalized posterior distribution.

Using the simple Propose-and-Verify architecture of the algorithm, we combine many different proposal distributions in a mixture distribution and thus integrate information from many sources, including our detections, directly into the posterior inference process. As only point-wise evaluation of (1) is necessary, we can also include proposals without a simple analytic representation. Traditional gradient moves and optimization steps can be integrated by restating them as additional proposals.

As basic proposals, we make use of Gaussian diffusion moves which lead to a random walk in parameter space. A random walk is not very efficient but can prevent the method from being stuck in local optima. Since the nature and scaling of the different parameters in the model (light, shape, texture and camera) varies drastically, we designed the random walk to be a mixture of block-wise alternating form, with different model parts as blocks. For each block we mix three different parameter scales, leading to a mixture of Gaussians distribution for the random walk stepping. Where appropriate, we included prior world knowledge to decorrelate the proposal distributions, such as compensating for scaling in distance modification proposals thus separating scaling and perspective effects. From time to time, the illumination is explicitly optimized, as the strongest part of mismatch between the rendered and the observed image is usually due to non-adapted light and dominates all other sources of misfit.

3.3 Detections

To include the face and landmarks detection results into the inference process we need a probabilistic output from the detectors in the form of a detection map, assigning each point in the image a likelihood of seeing a specific facial feature at that location.

The face detector and 9 facial feature point detectors (see Figure 1b) were trained using a standard Random Forest algorithm closely related to [5]. For each Random Forest detector we trained 256 trees. Each of these trees is learned using 30% of the training data, randomly selected. A node is split if it is not at a maximal depth of 30 and the data reaching that node is not pure. For a node to learn a random candidate set of 500 features is generated. Based on the information gain criterion the best threshold for each feature is calculated and the best split is selected. A leaf stores the percentage of positives in the data reaching that leaf. This is the certainty of a classification given the patch reaches the leaf. The response of the forest is then the mean of all responses of the trees

The training patches were gathered in a manner proposed in [7]. From the AFLW database [14], we selected for each detector approximately 25k positives

and 100k negatives. For the face detector, we selected additional 400k negative patches randomly sampled from the PASCAL-VOC 2012 [8] marked as not containing any person. To have more positive face samples we mirrored the face patches horizontally. As features we used Haar-like features also used in [13].

To detect faces we use a standard sliding window approach over all possible scales. The image is scaled by a factor of 0.9 between subsequent scales. We select the 10 best candidates not having a higher overlap than 60%. The feature response maps are then computed in a 40% enlarged area around each selected face box. The maps are averaged from three neighboring scales around the one the face is detected in.

The detection map $D_i(\boldsymbol{x})$ of landmark i needs to be combined with the observation noise model for landmarks (3). This is accomplished by performing a maximum convolution with the distance term on the response map as proposed in [9]. At each location $\boldsymbol{x}$ we then get

$$\log P(\boldsymbol{x}|D_i) = \max_{\boldsymbol{t}} \left\{ -\frac{\|\boldsymbol{x}-\boldsymbol{t}\|^2}{2\sigma_{\mathrm{LM}}^2} + \frac{1}{2}\log P(D_i(\boldsymbol{t})) \right\}. \tag{4}$$

3.4 Data-Driven Proposals

To properly integrate the information provided by the feature point detectors, this information must be stated as a proposal distribution which is used to generate samples in the parameter space of the model. As we have no explicit parameters encoding directly the position of the feature points, we resort to a generative type of inclusion.

The proposal distribution is created in an iterative Bayesian manner. For each possible face box, we build a proposal distribution by filtering unbiased proposals from the prior through a Metropolis acceptance-step, thus biasing the proposals with the ith face box's position and size and all the landmarks detection likelihoods $\mathcal{D}_i$ of the respective face box:

$$P(\theta) \to P(\theta|\mathrm{box}_i, \mathcal{D}_i). \tag{5}$$

The proposals from all the possible face boxes are combined in a mixture proposal and put through a Metropolis filter step which evaluates a proposed sample using the likelihood of the best face box available for each proposal, corresponding to an OR/union of the different boxes. This step thus mixes the different face box-conditionals (5) according to their consistency with the model:

$$\frac{1}{10}\sum_{i}^{10} P(\theta|\mathrm{box}_i, \mathcal{D}_i) \to P(\theta|\mathrm{allboxes}, \mathcal{D}). \tag{6}$$

The distribution $P(\theta|\mathrm{allboxes}, \mathcal{D})$ includes knowledge about all the possible detections but never forces us to take an explicit decision on the detection results.

Samples from the distribution (6) prefer other face boxes than the strongest detection in roughly one third of the cases. This implements an implicit model-based verification step without an explicit choice of a face box. A few samples

from (6) for a single image are visualized as a video and available as supplementary material.

3.5 Integration

The samples of the landmarks posterior (6) can now be used in the next step. By conditioning additionally on the actual image using the observation model (2) leads to the desired posterior distribution, summarizing all information:

$$P(\theta|\text{allboxes}, \mathcal{D}) \rightarrow P(\theta|I, \text{allboxes}, \mathcal{D}). \tag{7}$$

This rather wasteful generative approach to generate samples is feasible, as the evaluation of the landmark distribution is precomputed in (4) and is very cheap compared to the rendering needed to evaluate the image likelihood (2) at the end. The landmarks detection maps and the pinhole camera model used do not allow for a fast analytical calculation of the landmarks detection posterior anyway.

The system as a whole is able to integrate knowledge from different parts and allows uncertainty by its probabilistic nature. To gain full benefit of this integrative system, one is encouraged to include many different Bottom-Up heuristics increasing the probability that at least one makes a good guess.

4 Experiments and Results

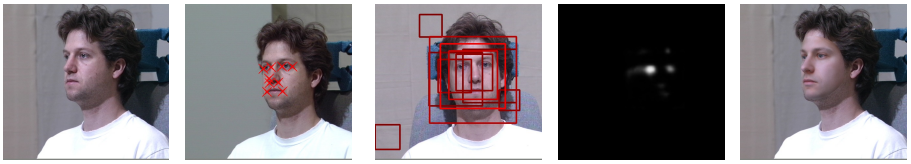
We evaluated our method on an unconstrained face recognition task on the Multi-PIE database [12]. Multi-PIE consists of 755k images including pose, illumination, expression and time (sessions). For our experiments we used the neutral photographs of 249 individuals in the first session in 3 poses (0°, 30°, 45°) cut to 512 x 512 pixels (see Figure 1). The exact setting can be easily reproduced by the pose and illumination indication in Table 1.

Contrary to most other approaches, we do not adapt any part of our recognition system to the Multi-PIE database.

The standard deviation $\sigma = 0.05$ of the image color[1] noise model has been obtained empirically by analyzing roughly 200 acceptably explained face images of an internal database. The standard deviation of the landmarks position is in the range of a few pixels. We use a value of $\sigma_{\text{LM}} = 4$ pixels. Our system does not use any given knowledge about landmarks, pose or illumination present in the image. The only assumption we take, is that there is exactly one face in every image.

We use the detection maps of 9 fiducial points (mouth corners, eye corners, nose tip, nose wingtips). By drawing 5000 samples from the Markov chain, we adapt the pose, the illumination, 50 texture and 50 shape parameters. For the recognition experiment, we use the best sample (maximum posterior probability) given the image and detection maps obtained during the sampling run. The overall runtime per image is under 10 minutes on current consumer hardware.

[1] Color values are ranged $c \in [0, 1]$.



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

Fig. 1. Sample images for the ID 1 in the Multi-PIE database. The poses used for our recognition experiment are a) 45°, b) 30° and c) 0°. Feature points we detect are shown in b). In c), the ten best face candidates are drawn. For the best face box (brightest red) we show the detection map for the right inner eye corner in d). Our fully automatic fitting result is shown in e).

Table 1. Rank-1 Identification rates (percent) and Rank-3 Identification rates (percent, in brackets) across pose, obtained by frontal 0° (`051_16`) images of all 249 first session individuals as gallery and the respective pose views as probes.

	30° (`130_16`)	45° (`080_16`)
our method	90.36 (96.39)	74.70 (86.75)
manual landmarks	93.57 (97.99)	81.93 (90.76)
3DGEM [17]	86.70	65.00

To measure the similarity between two faces f_1 and f_2, the cosine angle between the concatenation of our shape and color model coefficients is used, as suggested by Blanz and Vetter [4]: $d = \langle f_1, f_2 \rangle / (\|f_1\| \cdot \|f_2\|)$.

The Rank-1 identification rate refers to the proportion of probe images where the closest face in the gallery is of the same individual as the probe image. The Rank-3 allows the correct face to appear within the 3 closest faces.

Fig. 2. Original image with unreliable face detections (boxes), on the right: face region selected by the algorithm as original (top) and final face reconstruction result (lower) (Image: Keystone/epa/Jason Szenes)

Although our method is much more general (see Figure 2), we outperform state of the art methods for face recognition. Prabhu et al. reached the results closest to ours across pose variation using 3D Generic Elastic Models [17].

Previous approaches to fit a 3DMM [4,16] relied on manually annotated landmarks. If we use our system with these user-provided landmarks, implementing a perfectly reliable feature detection, we can slightly improve the recognition performance.

5 Conclusion

We presented a novel general concept to integrate unreliable information of various sources into the fitting process of a Morphable Model. In contrast to other fitting methods our proposed stochastic approach is not susceptible to local minima. Additionally, the DDMCMC integration concept, based on the Metropolis algorithm, is open to integrate further sources of information like an outlier model for glasses or segmentation of the face into different classes, such as skin, hair and eyes. Regression on the facial pose or expression would add further hints. More information can be used to explore probable hypotheses directly and should therefore improve the final result. All these noisy proposals can be integrated in the proposed approach in contrast to traditional fitters.

Using this concept, we demonstrated a straightforward application integrating unreliable face and landmark detection into model fitting without commitment to a single detection hypothesis. The developed method solves a face recognition task with state of the art performance, without any user input or database adaption.

References

1. Andriluka, M., Roth, S., Schiele, B.: Pictorial structures revisited: People detection and articulated pose estimation. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1014–1021. IEEE (2009)
2. Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: Localizing parts of faces using a consensus of exemplars. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011), pp. 545–552. IEEE (2011)
3. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: SIGGRAPH 1999 Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, pp. 187–194. ACM Press (1999)
4. Blanz, V., Vetter, T.: Face recognition based on fitting a 3d morphable model. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(9), 1063–1074 (2003)
5. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
6. Dantone, M., Gall, J., Fanelli, G., Van Gool, L.: Real-time facial feature detection using conditional regression forests. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012), pp. 2578–2585. IEEE (2012)
7. Eckhardt, M., Fasel, I., Movellan, J.: Towards practical facial feature detection. International Journal of Pattern Recognition and Artificial Intelligence 23(03), 379–400 (2009)

8. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. International Journal of Computer Vision 88(2), 303–338 (2010)
9. Felzenszwalb, P.F., Huttenlocher, D.P.: Pictorial structures for object recognition. International Journal of Computer Vision 61(1), 55–79 (2005)
10. Forsyth, D.A., Haddon, J., Ioffe, S.: The joy of sampling. International Journal of Computer Vision 41(1-2), 109–134 (2001)
11. Grenander, U.: Lectures in pattern theory. Applied Mathematical Sciences (1976)
12. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. Image and Vision Computing 28(5), 807–813 (2010)
13. Jones, M., Viola, P.: Fast multi-view face detection. Mitsubishi Electric Research Lab TR-20003-96 3 (2003)
14. Köstinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In: IEEE International Conference on Computer Vision Workshops (ICCV Workshops 2011), pp. 2144–2151. IEEE (2011)
15. Liu, C., Shum, H.-Y., Zhang, C.: Hierarchical shape modeling for automatic face localization. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002, Part II. LNCS, vol. 2351, pp. 687–703. Springer, Heidelberg (2002)
16. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: 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), pp. 296–301. IEEE (2009)
17. Prabhu, U., Heo, J., Savvides, M.: Unconstrained pose-invariant face recognition using 3d generic elastic models. IEEE Transactions on Pattern Analysis and Machine Intelligence 33(10), 1952–1961 (2011)
18. Ramamoorthi, R., Hanrahan, P.: An efficient representation for irradiance environment maps. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, pp. 497–500. ACM Press (2001)
19. Rauschert, I., Collins, R.T.: A generative model for simultaneous estimation of human body shape and pixel-level segmentation. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part V. LNCS, vol. 7576, pp. 704–717. Springer, Heidelberg (2012)
20. Robert, C.P., Casella, G.: Monte Carlo statistical methods, vol. 319. Citeseer (2004)
21. Tu, Z., Chen, X., Yuille, A.L., Zhu, S.C.: Image parsing: Unifying segmentation, detection, and recognition. International Journal of Computer Vision 63(2), 113–140 (2005)
22. Wojek, C., Roth, S., Schindler, K., Schiele, B.: Monocular 3D scene modeling and inference: Understanding multi-object traffic scenes. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 467–481. Springer, Heidelberg (2010)
23. Zhang, C., Zhang, Z.: A survey of recent advances in face detection. Tech. rep., Microsoft Research (2010)