Pose Normalization for Eye Gaze Estimation and Facial Attribute Description from Still Images

Bernhard Egger(✉), Sandro Schönborn,
Andreas Forster, and Thomas Vetter

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

Abstract. Our goal is to obtain an eye gaze estimation and a face description based on attributes (e.g. glasses, beard or thick lips) from still images. An attribute-based face description reflects human vocabulary and is therefore adequate as face description. Head pose and eye gaze play an important role in human interaction and are a key element to extract interaction information from still images. Pose variation is a major challenge when analyzing them. Most current approaches for facial image analysis are not explicitly pose-invariant. To obtain a pose-invariant representation, we have to account the three dimensional nature of a face. A 3D Morphable Model (3DMM) of faces is used to obtain a dense 3D reconstruction of the face in the image. This Analysis-by-Synthesis approach provides model parameters which contain an explicit face description and a dense model to image correspondence. However, the fit is restricted to the model space and cannot explain all variations. Our model only contains straight gaze directions and lacks high detail textural features. To overcome this limitations, we use the obtained correspondence in a discriminative approach. The dense correspondence is used to extract a pose-normalized version of the input image. The warped image contains all information from the original image and preserves gaze and detailed textural information. On the pose-normalized representation we train a regression function to obtain gaze estimation and attribute description. We provide results for pose-invariant gaze estimation on still images on the UUlm Head Pose and Gaze Database and attribute description on the Multi-PIE database. To the best of our knowledge, this is the first pose-invariant approach to estimate gaze from unconstrained still images.

1 Introduction

Faces play a fundamental role in human interaction. Facial attributes and gaze direction are very important for understanding the plot of a scene. The field of face analysis in still images evolved in the last years and a lot of very powerful

X. Jiang et al. (Eds.): GCPR 2014, LNCS 8753, pp. 317–327, 2014.
DOI: 10.1007/978-3-319-11752-2_25

methods have been developed. However, most of the research is put on the interpretation of a face regardless of its context and ignoring effects of pose variation. We take a step in the direction of facial interaction analysis by estimating the eye gaze. Points of attention can be estimated and faces can be described in their context (e.g. "Person A, male, is looking at Person B, female"). We show an overview of the presented method in Fig. 1.

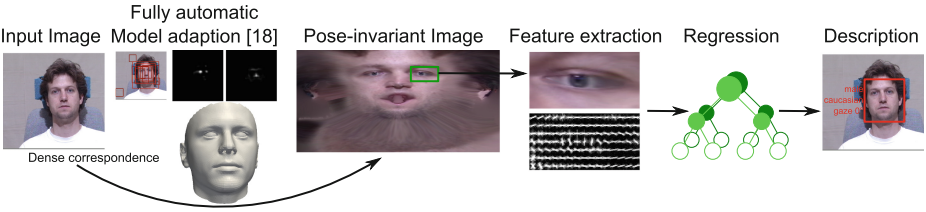


Fig. 1. System overview: The fully automatic 3DMM adaption method of Schönborn [17] is used to obtain a dense correspondence from the input image to the model reference. We extract a pose-invariant face representation preserving the texture from the original image. HOG features and image intensities are used as features for a Random Forest Regression. The output of the system is a gaze estimation and attribute-based image description.

In the broad research field of gaze estimation, most methods focus on tracking. For a single still image there is no pose-invariant method to automatically estimate eye gaze. We propose to use a pose-normalized version of the image and apply simple methods on this pose-invariant representation. Since a face is a three-dimensional object, a 3D Model is the natural way to obtain a pose-normalized representation. We use a generative 3D Morphable Model (3DMM) [3] of faces to solve the pose estimation and registration problem. A facial image is interpreted in an Analysis-by-Synthesis approach. The model is adapted to the face in the image as closely as possible (fitting).

The parameters (Shape, Color, Camera and Light) of the final representation in the model space (fit) contain information on the face and the scene. The eye gaze in the 3DMM is fixed and therefore the gaze estimation cannot be performed on the model parameters directly. The description by the model parameters is limited to what the model is able to reconstruct.

We overcome the model limitations with a discriminative approach. The normalization is based on full and perfect correspondence. We warp the image into a pose-normalized representation by the dense registration of the fit. The warped texture can be seen in Fig. 2. The remaining challenge for features and the classifier are the small correspondence inaccuracies in our fit.

For gaze estimation, we rely on the good registration and work on the histogram-normalized image intensities. For the attribute prediction, we use HOG features [7]. High frequency texture details cannot be encoded in our PPCA model parameters, and therefore the information captured by HOG features is valuable. The HOG features used for attribute estimation are invariant to small

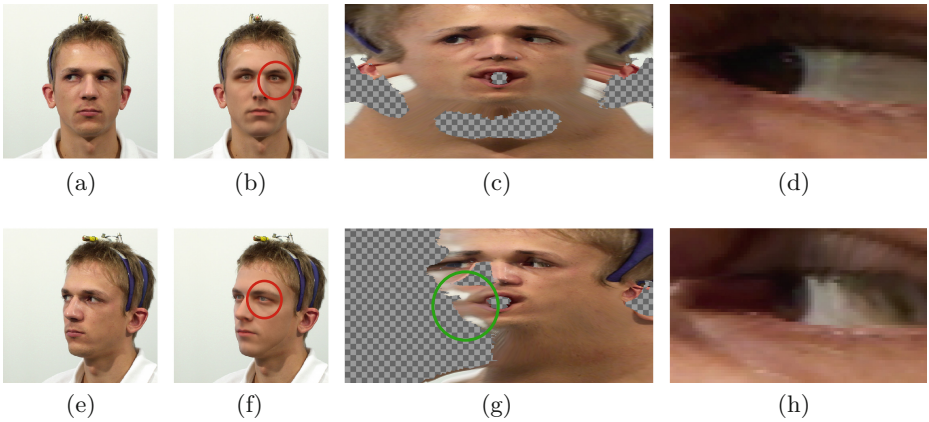


Fig. 2. We present a frontal 0° (a) and side view 40° (e) image of the UUlm HPG database. Both contain a relative eye gaze of 40°. The model fits (b) and (f) do not reflect the eye gaze (red circle). As one can see in (c) and (g), the pose-invariant representation obtained by the correspondence of the fit preserves the eye gaze. One can see small misalignments to the background at the border of the face model (green circle). (d) and (h) show the cropped regions we use to estimate the eye gaze (Color figure online).

misalignments due to the binning property. We use Random Forest Regression [6] for the Gaze estimation task and Random Forest Classification for the attribute estimation. To train our forests, we can use all image data which we can fit with the 3DMM.

In a gaze estimation experiment we show that the pose-normalized representation is suitable for gaze analysis. To the best of our knowledge, we present the first results for fully automatic gaze estimation on the UUlm Head Pose Gaze datasets [18] up to yaw angles of 40°. The database was chosen because of their wide range of gaze. The attribute-description experiment and the recognition experiment are based on the Multi-PIE database [9].

1.1 Prior Work

Most works on eye gaze estimation focus on tracking. Hansen et al. [10] give a nice overview of current methods. On single still images there are fewer works and all are limited to frontal pose or need-calibrated settings [8,13].

Kumar et al. [12] give a nice overview on prior work on facial attribute classification and demonstrate the power of attributes for face description and recognition. They classify attributes on affine-aligned face regions from near-frontal images. We extend this idea through a 3D model which adds full pose-invariance. Instead of a single global transformation we use dense local mappings incorporating the full 3D knowledge obtained by the 3DMM.

The power of pose-normalization using a 3D model was demonstrated by Blanz et al. [2]. The 3D Morphable Model is used for preprocessing for various

face recognition methods to produce frontal renderings. The viewport transformation improves the performance of 9 out of 10 systems on the Face Recognition Vendor Test FRVT2002.

Even though some facial attributes are encoded in the model parameters [1], the analysis of them has have not yet been explored. We show the limitations using the model parameters for attribute classification.

To obtain a pose-normalized face representation, we use the approach of Schönborn [17] for a fully automatic fitting of the 3DMM to an input face image. We work with a slightly modified version (without ears and throat) of the publicly available Basel Face Model [15].

2 Methods

2.1 Face Model

In this work, we use the 3DMM to extract the scene and face description. Both are obtained through a full adaption of the model to the input image. To achieve full automation, we make use of the probabilistic Data-Driven Markov Chain Monte Carlo integrative fitting algorithm of Schönborn [17], which can handle unreliable detection input. The fitting algorithm recovers the best face description, camera setup and illumination to reconstruct the image. The result of the adaption contains the image location of any point on the face through the correspondence with the model and the obtained camera setup. It also delivers a continuous face representation in terms of the PCA coefficients of the 3D face model.

In contrast to other automatic methods for extracting facial feature points, the 3DMM also results in a fully abstracted face representation which is invariant with respect to pose and illumination. We directly use this representation in terms of model coefficients for face recognition and attribute classification.

We adapted the model likelihood slightly to our needs by using a more general background model. We replace the restrictive original Gaussian background likelihood [17] by an empirical histogram model. Thus, we exchange

$$\mathcal{N}\left(I(p) \mid \mu_{\mathrm{BG}}, \Sigma_{\mathrm{BG}}\right) \quad \text{by} \quad \frac{1}{\delta}\mathrm{h}\left(I(p)\right), \tag{1}$$

where δ is the bin volume and $\mathrm{h}(I(p))$ is the relative frequency of the color value $I(p)$ at location p in input image I. Our histogram consists of 25 bins per RGB color channel.

2.2 Pose Normalization

Our pose-normalized representation is using the full correspondence of the fit for extraction of the image information. The 3D face is textured by the pixel information extracted from the image. The obtained representation corresponds to a texture map known from computer graphics. We use the texture representation proposed by Paysan [14] which builds on a quasi conformal mapping by

Kharevych et al. [11]. This texture map is a warp of the original image and still looks natural. We show examples for the pose-invariant representation in Fig. 2.

2.3 Gaze Estimation

We assume perfect registration and use the histogram-normalized image intensities of the pose-normalized representation. The gaze direction is parametrized relative to the head pose. A Random Forest regression is learned on a training set and used to predict the gaze direction.

2.4 Attribute Classifiers

The attribute classifiers are obtained similarly to the gaze estimation. Histograms of Oriented Gradients (HOG) [7] are used to represent textural details. The edge responses are binned into small spatial regions and can therefore cope with small misalignments.

We train a Random Forest Classifier to predict attributes. The output of the classifier is a certainty of the input image belonging to a class, respectively the face containing a specific attribute. The certainty gives us a more accurate description of the face than a (possibly wrong) binary output (e.g., if the classifier is "0.51 sure to see a male" than just "male").

A classifier is calculated per attribute. The eye, nose, mouth and eyebrow regions are used to predict the attributes. We combine different classifiers for the same attribute by the average prediction of all classifiers to obtain a single global attribute for face description.

2.5 Similarity Measure

A similarity measure in a face space is useful for all applications concerning identity. Different appearances (e.g., through pose or illumination) of the same face should always be similar.

The cosine angle between two face representations f_1 and f_2 is often used as similarity measure for face recognition based on the 3DMM [4]:

$$d = \frac{\langle f_1, f_2 \rangle}{(\| f_1 \| \cdot \| f_2 \|)} \tag{2}$$

In the classical setting, the vectors f_1 and f_2 are a concatenation of shape and texture parameters. To integrate our attribute predictions into the similarity measure, we concatenate them as a third component into the description vector.

3 Experiments and Results

To evaluate the pose-normalized face representation we performed two different experiments. First, we predict the eye gaze from the eye region cropped out of the normalized face texture. Second, we predict facial attributes from different

regions of the texture and evaluate their performance for face description on a face recognition task.

In all experiments we obtain a fit of the 3DMM to the image by a fully automatic DDMCMC method [17]. We draw 10 000 samples and take the best one (maximum posterior probability).

We use the OpenCV 2.4.4 [5] implementation of Random Forests and HOG features. We choose a tree depth of 10, select 10 features per split and trained with 2000 trees for all experiments. For the HOG features we took the preset parameters.

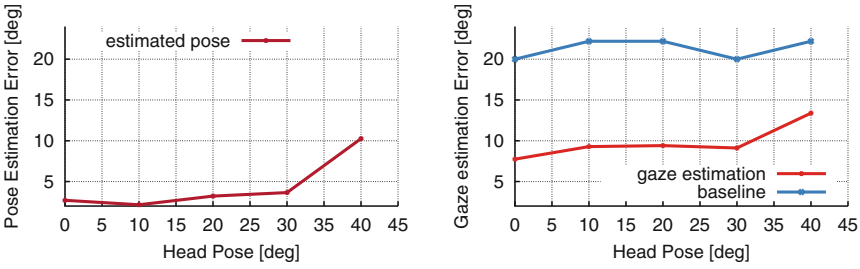
3.1 Gaze Estimation

The gaze estimation experiment was performed on the UUlm Head Pose and Gaze Database. It contains 20 subjects and 111 images per individual. We used the horizontal poses between 0° and 40° and relative gaze direction from −40° to 40°. The fitting is not reliable enough for gaze estimation for yaw angles above 45°. This selection leads to 940 images for our evaluation. We performed leave-one-out cross validation, always excluding all images of one subject. We show variations of the database in Fig. 3. Our gaze estimation is trained on the relative gaze (pose-corrected). The relative gaze is dependent on the estimated head pose and therefore the pose estimation error is propagated to the gaze estimation. The pose estimation error obtained by the model adaption is shown in Fig. 4a. For our gaze estimation experiment we reach a total Mean Approximation Error (MAE) of 9.74°. In Fig. 4b we show the estimation error itemized on each pose seperately. Both plots are reflected in Table 1. Note that due to pose normalization we are able to train a single regression for all poses. The proposed gaze estimator trained on the UUlm HPG database delivers reasonable results on real world images, see Fig. 5.

Fig. 3. These are 5 of 20 subjects of the UUlm HPG database. The variation used in our experiments is shown from left to right. We use yaw angles from 40° to 0° and relative gaze direction from 40° to −40°. The database contains different lighting conditions, glasses and occlusion through hair.

3.2 Attributes

We use 16 attributes to describe a face, see Table 2. For each attribute we learn a regressor on the eyes, nose, mouth or eyebrows. We compare the prediction



(a) (b)

Fig. 4. (a) shows the pose estimation performance of the fully automatic fitting method on UUlm HPG database. (b) shows gaze estimation performance per head pose on UUlm HPG database. The baseline results are obtained by always predicting a gaze of 0°.

Table 1. head pose and gaze estimation error (MAE) in degree. The baseline results are obtained by always predicting a gaze of 0°.

	0°	10°	20°	30°	40°
Pose estimation error	2.73	2.16	3.22	3.65	10.26
Gaze estimation error	7.75	9.29	9.41	9.13	13.40
Baseline error	20.00	22.22	22.22	20.00	22.22

based on HOG features and color intensities by a prediction obtained on model parameters. The performances of the particular attribute classifiers are shown in Table 2. The attributes were learned and evaluated on separated subsets of the Multi-PIE database [9]. The database contains over 750 000 images of 337 different persons. We used the 249 identities from the first session for evaluation and the first appearance in session two to four of the other 88 identities for training. The experiment was performed on five poses (15°, 30°, 45°, 60° resp. in Multi-PIE camera names `051`, `140`, `130`, `080`, `090`) and illumination `16`. This leads to 440 images for the training of each classifier. For one attribute we train a single classifier over all poses. There is no tuning to a specific pose.

3.3 Recognition

We use the similarity measure (2) for a face recognition experiment. The performance of attributes estimated on the texture is compared with the recognition rate obtained by the model parameters. We used the output of the classifiers for all 16 attributes on all 4 selected regions (64 attribute estimations). We evaluate our recognition method on the Multi-PIE database. We use the exact same setting for the recognition as Schönborn et al. [17]. The 249 individuals from the first session are used for the recognition task. The results are listed in Table 3. The attribute classifiers are the same as in the attributes section and

Table 2. Prediction performance in % of binarized attribute classifiers on pca coefficients and on the pose-normalized representation using HOG features and color intensities. The region selected for the image-based classifier is shown in the fourth column. Attributes indicated by a * are underrepresented in the test set ($\leq 20\,\%$). The pose-normalized image is especially useful for attributes like glasses which are not contained in the 3DMM.

Attribute	PCA	HOG	Region
African American*	97.6	98.2	Mouth
Asian	83.4	85.1	Eye
Beard*	96.9	95.5	Eyebrow
Black hair	78.2	75.2	Nose
Blond hair*	87.9	87.4	Eyebrow
Bue eyes	70.9	77.7	Eye
brown eyes	67.8	80.6	Eye
Caucasian	85.4	82.9	Eye
Glasses	71.1	88.5	Nose
Hair on forehead	73.4	67.8	Eyebrow
Indian*	90.3	93.3	Eye
Male	76.2	76.0	Mouth
Mustache	95.3	94.8	Mouth
Nasolabial fold	74.8	75.7	Nose
Thick lips	66.0	69.0	Mouth
Wide nose	59.9	63.9	Nose

Table 3. Rank-1 Identification rates (percent) across pose, obtained by frontal 0° (`051_16`) images as gallery and the respective pose views as probes.

	15° (`140_16`)	30° (`130_16`)	45° (`080_16`)	60° (`090_16`)
3DMM shape, texture and attributes	97.6	95.2	80.7	50.6
Attributes only	93.2	82.3	65.5	30.1
3DMM shape only	86.4	63.9	44.2	11.2
3DMM texture only	98.4	94.0	77.5	43.0
3DMM shape and texture	97.6	94.8	79.5	49.0
3DMM shape and texture [17]	-	90.4	74.7	-
3DGEM [16]	97.6	86.7	65.0	44.9

were trained on images from subjects not occurring in the recognition experiment. We compare our results to other fully automatic approaches based on 3D Generic Elastic Models [16] and previous results obtained with a 3DMM [17]. The effect of the added attribute detections is shown in Table 3. As we use

an empirical histogram background model, we obtain better recognition results using shape and texture coefficients than previous results by Schönborn et al. [17]. The experiment shows that the description by attributes is powerful and slightly improves the recognition performance.

(a) (b)

(c) (d)

Fig. 5. Our gaze estimation approach also works on unconstrained real world images. The automatically extracted gazes relative to head pose are (a): 15°, (b): 27°, (c): 20°, (d): −2°. The images where cropped to the face region after processing. Images: (a) KEYSTONE/AP Photo/Richard Drew, (b) KEYSTONE/EPA/Justin Lane, (c) KEYSTONE/EPA/Dennis M. Sabangan, (d) KEYSTONE/AP Photo/Alastair Grant

4 Conclusion

We proposed to use the registration obtained from the 3DMM fit for gaze estimation and attribute description. A pose-normalized face representation arises through the dense image correspondence. A regression and classification function is learned on the region of interest and profits from the pose normalization. The pose-normalized input image conserves the textural information from the input image. The information can be extracted by classical image features and lead to a description not contained in the model parameters. In contrast to the 3DMM, which needs high resolution 3D scans, our predictors can be learned directly on image data. By this we overcome model limitations. This approach is fully automatic, using a fully automatic 3DMM adaption method. In the experiments we present the first fully automatic and pose-invariant gaze estimation results on the UUlm HPG database. The gaze estimation is not limited to the database. The learned regression can be applied on real world images, see Fig. 5. In addition we show the limitation of the model parameters describing attributes not contained in the model (e.g. glasses, see Table 2). Attribute-based description combined with the 3DMM parameters achieves higher face recognition rates than other automatic approaches, especially for yaw angles larger than 30°.

Acknowledgment. This work has been partially founded by the Swiss National Science Foundation.

References

1. Amberg, B., Paysan, P., Vetter, T.: Weight, sex, and facial expressions: on the manipulation of attributes in generative 3D face models. In: Bebis, G. (ed.) ISVC 2009, Part I. LNCS, vol. 5875, pp. 875–885. Springer, Heidelberg (2009)
2. Blanz, V., Grother, P., Phillips, P.J., Vetter, T.: Face recognition based on frontal views generated from non-frontal images. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 2, pp. 454–461. IEEE (2005)
3. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3D faces. In: SIGGRAPH'99 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 Trans. Pattern Anal. Mach. Intell. **25**(9), 1063–1074 (2003)
5. Bradski, G.: The opencv library. Dr. Dobb's J. Softw. Tools **25**, 120–126 (2000)
6. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
7. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 886–893. IEEE (2005)
8. Florea, L., Florea, C., Vrânceanu, R., Vertan, C.: Can your eyes tell me how you think? a gaze directed estimation of the mental activity (2013)
9. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. Image Vis. Comput. **28**(5), 807–813 (2010)

10. Hansen, D.W., Ji, Q.: In the eye of the beholder: a survey of models for eyes and gaze. IEEE Trans. Pattern Anal. Mach. Intell. **32**(3), 478–500 (2010)
11. Kharevych, L., Springborn, B., Schröder, P.: Discrete conformal mappings via circle patterns. ACM Trans. Graph. (TOG) **25**(2), 412–438 (2006)
12. Kumar, N., Berg, A., Belhumeur, P., Nayar, S.: Describable visual attributes for face verification and image search. IEEE Trans. Pattern Anal. Mach. Intell. **33**(10), 1962–1977 (2011)
13. Marku, N., Frljak, M., Pandi, I.S., Ahlberg, J., Forchheimer, R.: Eye pupil localization with an ensemble of randomized trees. Pattern Recogn. **47**(2), 578–587 (2014)
14. Paysan, P.: Statistical modeling of facial aging based on 3D scans. Ph.D. thesis, University of Basel, Switzerland (2010)
15. 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)
16. Prabhu, U., Heo, J., Savvides, M.: Unconstrained pose-invariant face recognition using 3D generic elastic models. IEEE Trans. Pattern Anal. Mach. Intell. **33**(10), 1952–1961 (2011)
17. Schönborn, S., Forster, A., Egger, B., Vetter, T.: A monte carlo strategy to integrate detection and model-based face analysis. In: Weickert, J., Hein, M., Schiele, B. (eds.) GCPR 2013. LNCS, vol. 8142, pp. 101–110. Springer, Heidelberg (2013)
18. Weidenbacher, U., Layher, G., Strauss, P.M., Neumann, H.: A comprehensive head pose and gaze database (2007)