

Person-specific Face Shape Estimation under Varying Head Pose from Single Snapshots

F. Dornaika^{1,2}

¹*University of the Basque Country
San Sebastian, Spain*

²*IKERBASQUE, Basque Foundation for Science
Bilbao, Spain*

Email: fadi_dornaika@ehu.es

B. Raducanu

*Computer Vision Center
Bellaterra, Barcelona, Spain
Email: bogdan@cvc.uab.es*

Abstract—This paper presents a new method for person-specific face shape estimation under varying head pose of a previously unseen person from a single image. We describe a featureless approach based on a deformable 3D model and a learned face subspace. The proposed approach is based on maximizing a likelihood measure associated with a learned face subspace, which is carried out by a stochastic and genetic optimizer. We conducted the experiments on a subset of Honda Video Database showing the feasibility and robustness of the proposed approach. For this reason, our approach could lend itself nicely to complex frameworks involving 3D face tracking and face gesture recognition in monocular videos.

Keywords—Person-specific face shape; 3D head pose; face subspace; holistic approaches;

I. INTRODUCTION

Offline or online computed 3D face shapes can be used in many applications such as face recognition [1], [2], 3D face pose tracking [3], [4], and facial expression recognition [5], [6]. Model-based applications exploiting monocular vision systems (the face model is given by a 3D mesh or a range model) need to personalize the face model of the person utilizing the system in order to achieve an accurate estimation. This holds true even with simple 3D models such as cylinders and ellipsoids. Recently some researchers proposed the use of special sensors such as a travelling camera or a 3-D scanner in order to build personalized facial shapes [7]. These shape models are then used for art production or for 3D face detection and recognition. Such systems suffer from several shortcomings. Some of the shortcomings can be alleviated by using stereo vision sensors. In [8], the authors propose to infer side-view shape parameters from one single frontal image using learned statistical correlation between the frontal-view parameters and the side-view parameters. The facial points (MPEG-4 points) and the frontal view parameters (relative distances) are extracted from the frontal image using some heuristics and prior knowledge.

The mainstream for shape estimation relies on extracting and matching some salient facial features such as the

locations and local statistics of the eyes, nose, and mouth in one or more views. Thus, feature-based shape estimation not only require the extraction of the facial features but also requires a frontal view of the face. Feature-based approaches suffer from self-occlusions and drifting. A solution to overcome the drawbacks of feature-based approaches is given by holistic approaches (appearance-based approaches), which try to analyze the whole facial appearance [9], [10]. For example, Active Appearance Models (AAMs) were mainly used for 2D model fitting and tracking.

In this paper, we present a new method for specific face shape estimation under varying head pose of a previously unseen person from a single image. Although the shape estimation is the main focus, our approach is intrinsically related with 3D head pose estimation, since they are both included in the proposed mathematical framework.

The proposed holistic approach estimates both the face shape control parameters as well as the 3D pose parameters by registering the input texture (warped region of the image) to a statistical face texture. Compared to AAMs methods our proposal has two advantages. First, there is no need to compute a Jacobian matrix neither offline nor online. Second, while AAMs merge both the inter and intra-person shape variabilities, our method separates these variabilities, and therefore the proposed method can be easily and efficiently used for initializing a real time 3D face tracker and facial expression recognizer in videos (both the personalized 3D model and its 3D pose are computed for the first frame in the video sequence). However, it is not clear how these tasks can be performed with AAMs.

Our approach does not use neither 2D AAM nor 3D AAM. The only similarity with AAMs is the use of a statistical facial texture model based on Principal Component Analysis (PCA). The remainder of the paper is organized as follows. Section II describes the face modelling aspects. Section III presents the proposed holistic approach for the simultaneous estimation of the 3D pose and shape. Section IV presents some experimental results. Section V concludes the paper.

II. MODELLING FACES

A. A deformable 3D mesh

In our study, we use *Candide* 3D face model [11]. This common 3D deformable wireframe model accounts for person specific shape variation as well as for facial animation. The 3D shape of this wireframe model (triangular mesh) is directly recorded in coordinate form. It is given by the coordinates of its n 3D vertices. Thus, the shape up to a global scale can be fully described by the $3n$ -vector \mathbf{g} ; the concatenation of the 3D coordinates of all vertices. The vector \mathbf{g} is written as:

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{S} \boldsymbol{\tau}_s + \mathbf{A} \boldsymbol{\tau}_a \quad (1)$$

where $\bar{\mathbf{g}}$ is the standard shape of the model, $\boldsymbol{\tau}_s$ and $\boldsymbol{\tau}_a$ are shape and animation control vectors, respectively, and the columns of \mathbf{S} and \mathbf{A} are the Shape and Animation Units. Both matrices are provided by *Candide* model package. A Shape Unit provides a means of deforming the 3D wireframe so as to be able to adapt eye width, head width, eye separation distance, etc (see Figure 1). Thus, the term $\mathbf{S} \boldsymbol{\tau}_s$ accounts for shape variability (inter-person variability) while the term $\mathbf{A} \boldsymbol{\tau}_a$ accounts for the facial animation (intra-person variability). With this model, the ideal neutral face configuration is represented by $\boldsymbol{\tau}_a = \mathbf{0}$. In this study, we assume that the images are depicting quasi-neutral faces. Thus, the expression for the deformable mesh becomes:

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{S} \boldsymbol{\tau}_s \quad (2)$$

The shape modes were created manually to accommodate the subjectively most important changes in facial shape. In the model package, the number of modes associated with facial Shape Units matrix \mathbf{S} (inter-person variability) is twelve. However, for the purpose of our study which deals with the automatic image-based extraction of the control vector $\boldsymbol{\tau}_s$ only six components are considered as the most significant indicators of the perceived person-dependent facial shape in a given near frontal facial image. These components are: Head height, vertical position of the eye brows, vertical position of the eye, eyes separation distance, vertical position of the nose, vertical position of the mouth. The remaining components are set to nominal values.

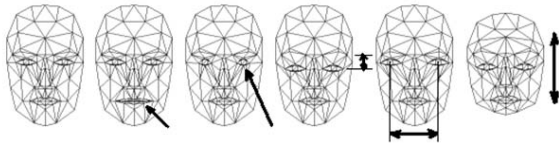


Figure 1. Effects of some facial shape control parameters on the deformable 3D model (standard shape, mouth width, eyes width, eyes vertical position, eye separation distance, head height).

In equation (1), the 3D shape is expressed in a local coordinate system. However, one should relate the 3D coordinates to the image coordinate system. To this end, we should add the six degrees of freedom associated with the 3D face pose. The mapping between the 3D face model and the image adopts the weak perspective projection model.

Thus, the state of the 3D wireframe model is given by the 3D face pose parameters (three rotations and three translations) and the shape control vector $\boldsymbol{\tau}_s$. This is given by the 12-dimensional vector \mathbf{b} :

$$\mathbf{b} = [\theta_x, \theta_y, \theta_z, t_x, t_y, t_z, \boldsymbol{\tau}_s^T]^T \quad (3)$$

Note that if only the aspect ratio of the camera is known, then the component t_z is replaced by a scale factor having the same mapping role between 3D and 2D. Estimating the camera intrinsic parameters can be carried out using offline and online calibration techniques [12].

B. Shape-free facial patches

A facial patch is represented as a shape-free image (geometrically normalized rawbrightness image). The geometry of this image is obtained by projecting the standard shape $\bar{\mathbf{g}}$ using a centered frontal 3D pose onto an image with a given resolution. The texture of this geometrically normalized image is obtained by texture mapping from the triangular 2D mesh in the input image (see figure 2) using a piecewise affine transform.

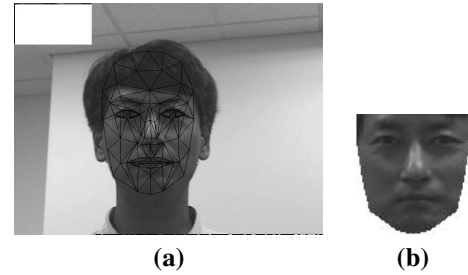


Figure 2. (a) an input image with correct fitting. (b) the corresponding shape-free facial patch.

III. SHAPE AND 3D HEAD POSE PARAMETER ESTIMATION

A. Face subspace

The statistical facial texture model describes the appearance variation of the shape-free facial patches \mathbf{x} (see figure 2.(b)). These patches are obtained from the training images (individual snapshots or video sequences) by fitting the 3D deformable model to the face. This fitting can be manual or automatic [11]. Using these training patches one can easily build the face subspace. For this purpose we use the Principal Component Analysis (PCA)—a well-known technique used for modeling face subspaces. We assume that we have K shape-free patches. Applying a PCA on the

training patches we can compute the mean and the principal modes of variation. The use of linear subspace (PCA) can be justified by 1) the facial images are geometrically normalized, and 2) the proposed method will be carried out in a relatively constrained environment. However, currently we are investigating the use of non-linear manifold learning techniques.

B. Optimization

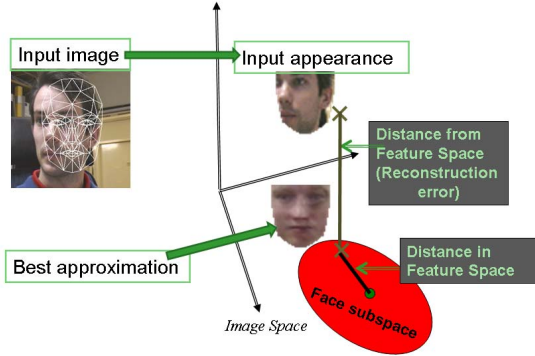


Figure 3. The unknown parameters are estimated by maximizing a likelihood measure taking into account the reconstruction error and the distance in feature space. The face subspace is linear.

The basic idea is to estimate the 3D face pose and shape parameters, i.e. the vector \mathbf{b} , such that the associated shape-free patch will be as close as possible to the facial sub-space. This can be carried out by maximizing a certain likelihood measure. For this purpose, we use the likelihood measure proposed in [13]:

$$p(\mathbf{x}|\mathbf{b}) \propto \exp\left(-\frac{1}{2} \sum_{i=1}^M \frac{\xi_i^2}{\lambda_i}\right) \exp\left(-\frac{e}{2\rho^*}\right) \quad (4)$$

where e is the reconstruction error, λ_i s are the M largest eigenvalues given by the PCA, ξ_i s represent the texture projection onto the corresponding M eigenvectors, and ρ^* is the arithmetic average of the remaining eigenvalues (in the complementary subspace). The reconstruction error is the distance between the original shape-free texture \mathbf{x} and its projection onto the PCA subspace.

The above likelihood measure takes into account two distances (i) the distance-from-feature-space, and (ii) the distance-in-feature-space. These two distances are illustrated in Figure 3. Maximizing this likelihood is equivalent to minimizing the *Mahalanobis* distance over the original textures. The unknown 3D face pose and shape parameters (the vector \mathbf{b}) can be estimated by seeking the maximum of the

likelihood (4):

$$\mathbf{b} = \arg \max_{\mathbf{b}} p(\mathbf{x}|\mathbf{b}) \quad (5)$$

To this end, we use the Differential Evolution (DE) algorithm [14] in order to maximize (4) with respect to the 3D face pose and shape parameters. The DE algorithm is a practical approach to global numerical optimization that is easy to implement, reliable and fast. The crucial idea behind DE is a scheme for generating trial parameter vectors. Basically, DE adds the weighted difference between two population vectors to a third vector.

In our case, the initial population is randomly selected between the lower and upper bounds defined for each variable using uniform distributions. The distributions associated with the translational part of the 3D face pose are centered on the output of the 2D face detector [15].

IV. EXPERIMENTAL RESULTS

Experiments were conducted to evaluate the performance of the proposed fitting algorithm in image snapshots extracted from several video sequences of the Honda video database [16]. The video sequences were recorded in realistic conditions, with persons featuring unconstrained in-plane and out-of-plane head movements. The sequences are at least 15 seconds long and are recorded at 15 frames per second. A subset of 20 video sequences (corresponding to 20 different persons) has been retrieved for our experiments.

However, since our 3D model is based on Candide model not all snapshots extracted from the videos database can be used. Only those depicting 3D head poses belonging to the interval $[-40^\circ, +40^\circ]$ for the pitch and yaw angles were considered. In total, we selected for our experiments about 900 images. A PCA model is built from a set of 500 shape-free templates, belonging to 5 persons. The remaining persons were used for test. We have found that PCA models with 20 principal components are usually enough for representing the face sub-space. More precisely, we found that the retained variance is above 95% of the total variance.

Figure 4 illustrates the fitting results obtained with four unseen persons. As can be appreciated, the face pose and shape parameters (relative positions of eyebrows, eyes, nose, and mouth) are correctly fitted on the face.

Quantitatively speaking, we performed an evaluation process taking into account both intra-person and inter-person estimation accuracy. Ground-truth parameters have been obtained manually. Table I depicts the average error of shape parameters in the first case. Due to lack of space, we present the results of only 5 persons from the total 15. The shape parameters are normalized, i.e., each parameter belongs to the interval $[-1, 1]$.

Table II depicts the average error between the manually obtained parameters and the automatically estimated ones for all the persons considered.



Figure 4. 3D face pose and person-specific shape estimation associated with four unseen persons.

	eyebrow	eye	eyes separa.	nose	mouth
#1	0.7%	6.2%	3.2%	6.4%	1.4%
#7	2.8%	0.4%	0.3%	2.8%	1.1%
#10	3.1%	4.3%	1.1%	5.9%	0.7%
#13	1.9%	1.6%	2.3%	3.6%	1.6%
#15	2.5%	3.5%	2.1%	5.3%	1.7%

Table I
AVERAGE ERROR FOR FIVE INDIVIDUALS.

V. CONCLUSION

This paper presented a featureless method that fits a generic deformable 3D face model to a single facial image where the face is not required to be frontal. The fitted parameters are some salient shape control parameters as well as the 3D face pose parameters. The proposed method has several advantages that make it attractive, being useful for 3D face pose tracking and facial expression recognition in real-time.

	eyebrow	eye	eyes separa.	nose	mouth
All	2.4%	3.6%	1.6%	4.2%	1.4%

Table II
GLOBAL AVERAGE ERROR.

ACKNOWLEDGMENTS

This work has been supported in part by the projects TIN2009-14404-C02-00 and CONSOLIDER-INGENIO CSD 2007-00018, Ministerio de Educación y Ciencia, Spain.

REFERENCES

[1] A. Bronstein, M. Bronstein, and R. Kimmel, "Three-dimensional face recognition," *International Journal of Computer Vision*, vol. 64, pp. 5–30, 2005.

[2] F. Haar and R. Veltkamp, "3D face model fitting for recognition," in *European Conference on Computer Vision*, 2008.

[3] F. Dornaika and F. Davoine, "On appearance based face and facial action tracking," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 16, no. 9, pp. 1107–1124, 2006.

[4] E. Murphy-Chutorian and M. Trivedi, "Head pose estimation in computer vision: a survey," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 4, pp. 607–626, 2009.

[5] F. Dornaika and F. Davoine, "Simultaneous facial action tracking and expression recognition in the presence of head motion," *International Journal of Computer Vision*, vol. 76, no. 3, pp. 257–281, 2008.

[6] Y. Sun and L. Yin, "Facial expression recognition based on 3D dynamic range model sequences," in *Computer Vision – ECCV 2008*, ser. LNCS, D. Forsyth, P. Torr, and A. Zisserman, Eds., vol. 5303, 2008.

[7] M. D. Breitenstein, D. Kuettel, T. Weise, L. Gool, and H. Pfister, "Real-time face pose estimation from single range images," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2008.

[8] C. J. Kuo, R. Huang, and T. Lin, "3D facial model estimation from single front-view facial image," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 12, no. 3, pp. 183–192, 2002.

[9] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–684, 2001.

[10] I. Matthews and S. Baker, "Active appearance models revisited," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 135–164, 2004.

[11] J. Ahlberg, "An active model for facial feature tracking," *EURASIP Journal on Applied Signal Processing*, vol. 2002, no. 6, pp. 566–571, June 2002.

[12] F. Dornaika and R. Chung, "An algebraic approach to camera self-calibration," *Computer Vision and Image Understanding*, vol. 83, no. 3, pp. 195–215, 2001.

[13] B. Moghaddam and A. Pentland, "Probabilistic visual learning for object representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 696–710, 1997.

[14] K. V. Price, J. A. Lampinen, and R. M. Storn, *Differential Evolution: A Practical Approach To Global Optimization*. Springer, 2005.

[15] P. Viola and M. Jones, "Robust real-time object detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[16] K. Lee, J. Ho, M. Yang, and D. Kriegman, "Video-based face recognition using probabilistic appearance manifolds," *IEEE Conf. On Computer Vision and Pattern Recognition*, vol. 1, pp. 313–320, 2003.