Estimation of Face Parameters using Correlation Analysis and a Topology Preserving Prior

Stella Grasshof, Hanno Ackermann and Jörn Ostermann
Institut für Informationsverarbeitung, Leibniz Universität Hannover, Germany

Abstract

Candide-3 is a well-known model, used to represent triangular meshes of human faces. It is common to only estimate 17 to 21 of the 79 model parameters. We show that these are insufficient to fit model vertices to facial feature points with low error and if more parameters are estimated, the model mesh deforms to unnatural configurations. To overcome this problem, we propose a novel solution: Given facial feature points, we propose to estimate the model parameters in subsets in which they are uncorrelated. Additionally we present a term to penalize topologically incorrect triangular mesh configurations. As a result the average mean squared error between facial feature points and model vertices is reduced by 90%, while face topology is preserved.

1 Introduction

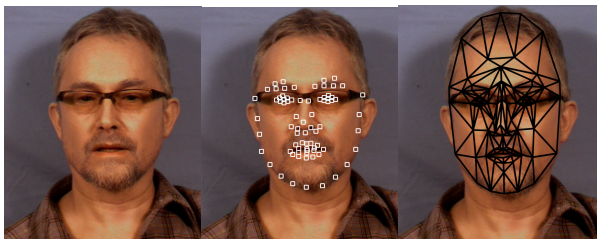
Candide [1, 2] is a parametrized model consisting of 3D points, arranged in a triangular mesh, representing a human face. The mesh configuration is altered by the model parameters, which are divided in shape and action parameters. Later the MPEG-4 facial animation standard [4] introduced a standard nomenclature for facial points (FP) and facial animation parameters (FAP). The definition of the updated Candide Model, Candide-3 [3] includes references to corresponding parameters of the MPEG-4 facial animation standard.

This model is often used for face tracking [5, 6, 7], where usually only 6 to 9 of the 65 action parameters are used. But especially for 3D face reconstructions, using a few parameters often causes large errors between projected mesh vertices and provided facial feature points. In this case the mesh does not approximate the shape of the face well, because it violates constraints resulting from topology of human faces.

Using the existing algorithms, which neglect proper 3D perspective projection, with a larger set of parameters causes numerical instabilities, which in turn cause even more topological errors. To prevent this, we propose to divide the parameters in uncorrelated subsets and estimate them separately and consecutively. Furthermore, we propose an algorithm based on nonlinear parameter estimation using a prior to prevent mesh configurations which violate the topology of human faces. One result of our algorithm is displayed in Figure 1. To summarize, our contributions are:

- Using subsets of uncorrelated parameters in the estimation procedure
- Nonlinear optimization including a topology preserving penalty term
- A full perspective camera model instead of the weak perspective model

The paper is organized as follows: In Section 2 previous work is described, including model definition and two algorithms for parameter estimation. In Section 3 we present our algorithms. Experimental evaluation of all approaches follows in Section 4. Discussions can be found in Section 5.



(a) (b) (c)

Figure 1: (a) Original Image of MUCT Face Database. (b) Image with corresponding Facial Feature Points. (c) Resulting model mesh retrieved with our algorithm.

2 Previous Work

In this paper we refer to Candide-3 introduced by Ahlberg [3], but as there are slight differences in literature, we point out that the version used here is Candide-3.1.6.

2.1 Candide-3 - Model Definition

Candide-3 [3] is defined by a set of 113 vertices, connected by 184 triangles; 14 shape parameters and 65 action units enable to change the model mesh configuration. As there are some vertices not included in any triangle, we end up with a total of 104 vertices.

We define the initial position of stacked 3D model vertices as $\bar{v} \in \mathbb{R}^{3N}$. This standard configuration can be changed by adding a matrix-vector product of sparse shape and action matrices $S \in \mathbb{R}^{3N\times 14}$, $A \in \mathbb{R}^{3N\times 65}$, weighted by their parameter vectors $s \in \mathbb{R}^{14}$ and $a \in \mathbb{R}^{65}$. As a result any 3D facial configuration $\widehat{v} \in \mathbb{R}^{3N}$ is defined as $\widehat{v} = (\widehat{v}_{1,x}, \cdots, \widehat{v}_{N,x}, \widehat{v}_{1,y}, \cdots, \widehat{v}_{N,y}, \widehat{v}_{1,z}, \cdots, \widehat{v}_{N,z})^T$.

$$\widehat{v} = \bar{v} + Ss + Aa \tag{1}$$

Let one of these vertices be $\widehat{v}_k \in \mathbb{R}^3$, $k = 1, \ldots, N$, then v_k defines the globally transformed $\widehat{v}_k$ as

$$v_k = R\, C\, \widehat{v}_k \; + t \tag{2}$$

where R defines a 3D rotation matrix (with rotation angles $\theta_x, \theta_y, \theta_z$), C a scaling matrix (with scaling parameters c_x, c_y, c_z) and $t = (t_x, t_y, t_z)^T \in \mathbb{R}^3$ a translation vector. As the global motion parameters are assumed to be equal for all vertices, Eq. (2) is extended

to a formulation for all model vertices:

$$v = \underbrace{(R \otimes I_N)}_{=Q} \underbrace{(C \otimes I_N)}_{=D} \widehat{v} + \underbrace{t \otimes 1_N}_{=d}$$
$$v = Q\ D\ \Big(\bar{v} + Ss + Aa\Big) + d \tag{3}$$

where $I_N \in \mathbb{R}^{N \times N}$ is the unit matrix, $1_N \in R^N$ only contains ones and $\otimes$ defines the Kronecker product.

To fit the model to any individual face, the introduced 79 local (s, a) parameters of Eq. (1) and 9 global parameters $(\theta_x, \theta_y, \theta_z,\ c_x, c_y, c_z,\ t_x, t_y, t_z)$ of Eq. (2) have to be estimated.

2.2 Estimation of Model Parameters

Given a set of N_f facial feature points in 3D $f \in \mathbb{R}^{3N_f}$ or in 2D $f' \in \mathbb{R}^{2N_f}$, we assume correspondences to a subset $\mathcal{I}$ of Candide-3 model vertices, which are known as $v_\mathcal{I} \in \mathbb{R}^{3N_f}$ and $v'_\mathcal{I} \in \mathbb{R}^{2N_f}$. Please note that the definition of $v'_\mathcal{I}$ differs between algorithms. To estimate the parameters of the model, the square of the sum of the euclidean distances between model vertices and feature points is minimized:

$$\min_p \|v_\mathcal{I} - f\|^2 \ \text{ or } \ \min_{p'} \|v'_\mathcal{I} - f'\|^2 \tag{4}$$

with respect to the global motion parameters defined in Section 2.1 and the shape and action parameters s and a, combined in parameter vector p or p'.

2.2.1 One-step solution

Ahlberg [8, 9] presents a one-step solution to estimate all necessary parameters. Under the assumptions that (1.) shape and action parameters are applied to the scaled version of the initial model vertices $\bar{v}$, (2.) only small rotations occur and (3.) shape and action units (S and A) are invariant to rotation, the original model Eq. (3) can be approximated as

$$v \approx \widetilde{v} = \widetilde{Q}\bar{v} + D\bar{v} + Ss + Aa + u \tag{5}$$
$$\text{with} \quad \widetilde{Q} = \widetilde{R} \otimes I_N \tag{6}$$
$$\text{and} \quad \widetilde{R} = r_x R_1 + r_y R_2 + r_z R_3 + I_3 \tag{7}$$

where R_1, R_2, R_3 are infinitesimal rotation matrices.

In case facial feature points are provided in 3D, the euclidean distance between model and feature points can be easily calculated, however if provided in 2D, the last coordinate of the model vertices $v_\mathcal{I}$ is dropped to obtain $v'_\mathcal{I}$, which can be regarded as orthographic projection.

2.2.2 Weak Perspective Projection

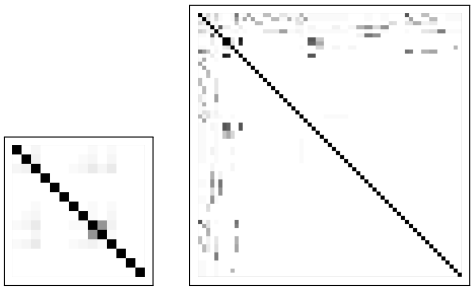
A weak perspective projection model is used in [9, 5].

Let z be a scaling factor, $d_2 = (t_x, t_y)^T \otimes 1_N \in \mathbb{R}^{2N}$, $K = [1, 0, 0;\ 0, 1, 0] \otimes I_N$ and Q be as in Eq. (3). We can then rewrite Eq. (3) as

$$v' = (z + 1)\ K\ Q\,(\bar{v} + Ss + Aa) + d_2 \tag{8}$$

Whereas the method is used to minimize a distance between input texture and a synthesized model texture [9, 5], we apply this approach to minimize the distance introduced in Eq. (4) with 2D facial feature points.

To estimate the parameters of the models defined by Eq. (5) and Eq. (8) the authors [8, 9, 5] suggest to first estimate all parameters jointly, then refine the result by assuming shape parameters as fixed, only adapting global and action parameters. However usually only 6 out of 65 action parameters are estimated. This restrictions are made to overcome problems caused by correlations in parameter space. As suggested in [5] correlation can be removed by application of PCA. Causing a new parameter space, where standard facial animation parameters (FAPs) are not represented anymore and leading to global instead of local deformations, we avoid usage of PCA here.



(a) (b)

Figure 2: Figures (a) and (b) show the correlation matrices for shape and action matrices S and A. Dark values indicate high absolute values of the corresponding correlation coefficient.

3 Improved Model Estimation

3.1 Projective Camera

To improve 3D reconstruction, we replace the weak perspective projection of Eq. (8) by a projective camera model. By estimation of shape and action parameters, we obtain an estimate for the 3D model vertices $\widehat{v}$ of Eq. (1). Using the 3D model vertices and the corresponding 2D facial feature points, camera parameters can be estimated by Direct Linear Transformation (DLT) [10]. We alternate estimations of model and camera parameters in our optimization procedure.

3.2 Parameter Subsets

Analyzing the matrices S and A defined in Eq. (1), we found that some columns are correlated. Figure 2 shows the correlation matrices for S and A, i.e. each entry of the two matrices shown in Figure 2(a) and 2(b) is computed by the correlation between two columns of S and A. As can be seen in Figure 2(b), high correlation values occur especially for action parameters. We therefore propose to divide the shape and action parameters into subsets, in which they are pairwise uncorrelated. We define $C_A \in \mathbb{R}^{N_a \times N_a}$ as correlation matrix of A, where the element $C_A(i, j)$ contains the pairwise correlation coefficient of parameters i and j. Parameters a_i and a_j, $i \neq j$ are assumed to be uncorrelated if $|C_A(i, j)| < \lambda_c$. We collect uncorrelated parameters in sets U_k, which are disjoint, but if united,

contain all parameters. If $|C_A(i,j)| \geq \lambda_c$, a_i and a_j will not be elements of the same U_k. If one parameter is correlated with all the others, it defines its own set. The sets U_k for uncorrelated shape parameters are computed analogously.

For Optimization, the camera parameters are estimated first, then the parameters in the sets U_k are estimated independently and consecutively. The estimation of camera parameters is alternated with the ones of U_k.

3.3 Topological Constraint

Topologically incorrect mesh model configurations occur, if more than six action parameters are estimated with the algorithms described in section 2.2.1 or 2.2.2. Figure 3(f) shows one example of such a mesh. As a result of unconstrained optimization, model vertices without corresponding facial feature points move unrestricted and cause large topologically unsuitable triangles in the model mesh.

To avoid large differences between original and adapted model topology, we introduce a constraint which penalizes deviations of the direction of surface normal vectors compared to triangles of the original model configuration.

The initial model vertex configuration is known as $\bar{v}$ from Eq. (1). We define the corresponding set of triangles as $\overline{T} = \{\overline{T}_1, \ldots, \overline{T}_{N_t}\}$, analogously $\widehat{T}$ is the set of triangles for $\widehat{v}$ of Eq. (1). As for each triangle T_k a surface normal vector $n(T_k)$ with $\|n(T_k)\| = 1$ can be computed, we define our topology preserving penalty

$$g\left(\widehat{T}\right) = \sum_{k=1}^{N_t} \left\| n\left(\widehat{T}_k\right) - n\left(\overline{T}_k\right) \right\|^2 \qquad (9)$$

Adding this to the minimization problem defined in Eq. (4), we obtain

$$\min_p \quad \left\{ \|v'_{\mathcal{I}} - f'\|^2 \; + \; \lambda_t \cdot g\left(\widehat{T}\right) \right\} \qquad (10)$$

where p contains parameters for shape, action and global motion as in Eq. (4), the parameter λ_t controls how strictly the estimated mesh should adhere to the topology of the template mesh model.

4 Experiments

4.1 MUCT Face Database

The MUCT database [11] provides a total of 3755 face images of 276 individuals shown in five camera views and ten light settings. Providing the complete set of 76 facial landmarks, we chose frontal view images and one specific light setting, leading to 104 test sets. From the provided landmarks, we picked 49 corresponding to Candide-3 model vertices, which are used to estimate local and global parameters for model fits by algorithms described in sections 2.2 and 3. The chosen facial feature points with corresponding matched model vertices are displayed in Figure 3(a) to (d).

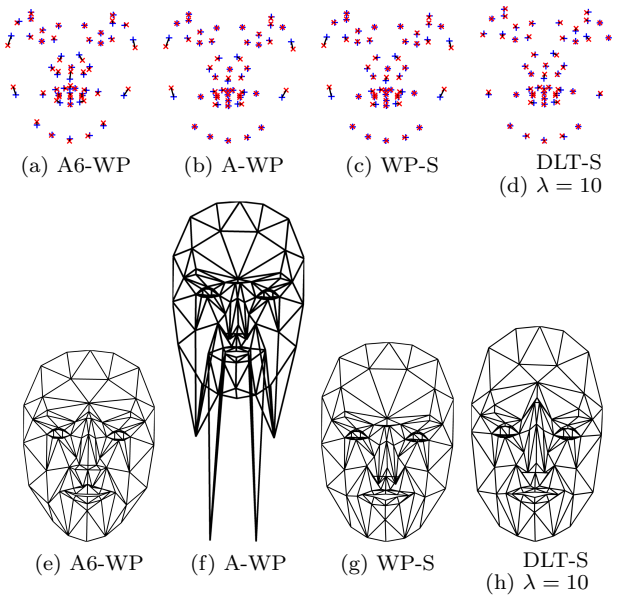


Figure 3: Different face model configurations, as a result of application of different algorithms: The first row shows the distance of provided facial feature points (blue plus-sign) compared to the matching model vertices (red cross) obtained by the corresponding algorithms. The second row shows the corresponding model mesh configuration.

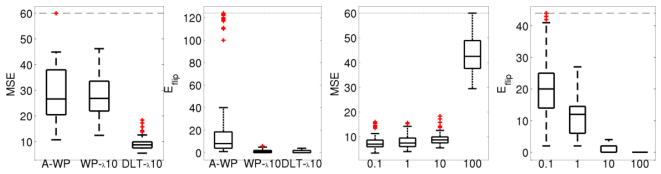
4.2 Quantitative Quality Measure

For some algorithms the resulting mean squared error (MSE) indicates an excellent fit, however it only takes the distance between corresponding facial feature points and model vertices into account, while model vertices without correspondence are not included. As can be seen in Figure 3(b) there is a relatively small error between facial feature points and model vertices, however this result corresponds to an unfavorable mesh configuration shown in Figure 3(f). We observed that, compared to the initial model configuration $\bar{v}$, some triangles of $\widehat{v}$ change the direction of their surface normals drastically, i.e. by more than 90 degrees. This effect will further be called a "flip" of a triangle, the total amount of *flip* events E_{flip}. Additionally the total amount can be replaced by the sum of triangle areas E_{flipw}, to which a flip occurred. To summarize: the lower the values of MSE, E_{flip} and E_{flipw}, at the same time, the better the model fit.

4.3 Evaluation

Some shape or action parameters only influence model vertices for which no corresponding facial feature point is available. Therefore we restrict the parameters to those, which alter at least one component of the model vertices $v_{\mathcal{I}}$, leading to a total of 38 of 65 action and 12 of 14 shape parameters, so 50 of 79 model parameters are estimable.

In the following we denote the algorithm defined by Eq. (5) as A, and Eq. (8) as A-WP. If only 6 action parameters are estimated: A6 and A6-WP, respectively. The abbreviation DLT refers to algorithm of section 3.1, adding "S" refers to the usage of subsets described



(a) MSE (b) E_{flip} (c) MSE (d) E_{flip}

Figure 4: Illustration of performance of algorithms. (a) and (b) show MSE and E_{flip} for previous and improved algorithms. Effects of the topological penalty term are shown in (c) and (d): Higher penalty weights λ_t increase the MSE values (slightly), while E_{flip} decreases considerably. Red crosses indicate outliers.

in section 3.2, while an additional λ denotes application of the constraint defined in Eq. (9).

We found that estimation of six action parameters with algorithms A6 and A6-WP leads to a large MSE. These are lowered by incorporating more action parameters with A and A-WP, but at cost of larger E_{flip} values, including unfavorable mesh configurations as can be seen in Figure 3(f). These high E_{flip} values are lowered by using uncorrelated parameter subsets in the estimation procedure, i.e. WP-S and DLT-S, both lead to better results than WP or DLT, alone; one example is illustrated in Figure 3(f) and (g). Furthermore we found that DLT induces lower MSE than using WP for all experiments, i.e. DLT outperforms WP, which is illustrated in 4(a).

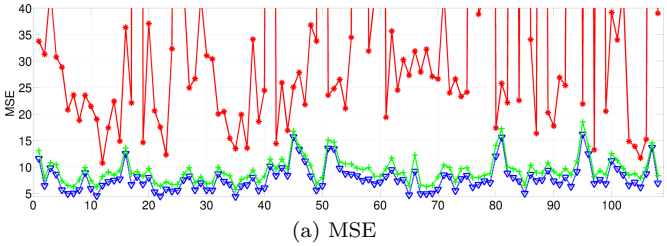
We tested different values for λ_t introduced in Eq. (9). Figure 4(d) shows the expected effect that increasing λ_t decreases the number of flips E_{flip}. As expected the MSE increased slightly for $\lambda_t = 0.1, 1, 10$, which can be seen in 4(d). $\lambda_t = 100$ caused a high increase of MSE, as the model mesh is stressed to stay in original configuration. We received best results for $\lambda_t = 10$. For $\lambda_t = 1, 10$ algorithms DLT-S and DLT-S-λ lead to comparable results of good quality, in terms of low MSE and E_{flip}-values. However DLT-S-λ is superior to DLT-S in terms of lower E_{flipw} values, which is illustrated in Figure 5.

5 Discussion

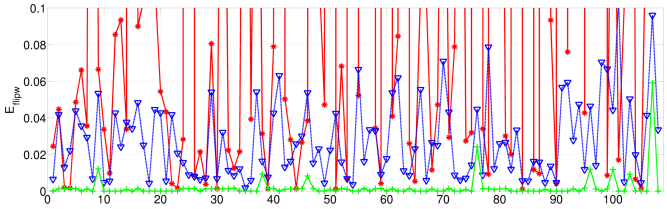
Previous algorithms only estimated a part of the parameters provided by the Candide-3 face model, leading to large errors and unsatisfactory results for 3D-face reconstruction. We show that increasing the number of parameters leads to unfavorable model mesh configurations, which we avoid by estimating the shape and action parameters in subsets of uncorrelated parameters. Furthermore we introduce a topological penalty which favors mesh configurations, implying surface normals close to the initial model, which improves the results even more. An additional decrease of MSE was reached by replacing the weak perspective projection by a perspective camera model.

Compared to the original algorithm A6-WP, our approach DLT-S-λ, reduces the mean MSE from 98.95 to 9.42 and mean E_{flip} from 19.68 to 1.05, which is a decrease of over 90% for both criteria.

Though we demonstrated our algorithm on facial feature points only, the adaptations are applicable to



(a) MSE



(b) E_{flipw}

Figure 5: Results obtained by application of algorithms A-WP (red stars), DLT-S (blue triangles) and DLT-S-λ (green plus signs) ($\lambda_t = 10$) on 108 individuals (x-axis) of MUCT face database. (a) MSE-values of A-WP lie always over the ones obtained by algorithms DLT-S and DLT-S-λ, while DLT-S outperforms DLT-S-λ, slightly. (b) E_{flipw}-values are considerably lower for DLT-S-λ, compared to DLT-S. (Figure is best viewed in color)

textures. This will be done in future work.

References

[1] M. Rydfalk, "CANDIDE, a parameterized face", *techn. report*, Report No. LiTH-ISY-I-866, Dept. of Electrical Engineering, Linköping University, Sweden, 1987

[2] B. Welsh, "Model-Based Coding of Images", PhD dissertation, British Telecom Research Lab, Jan. 1991.

[3] Ahlberg, Jörgen, "CANDIDE-3 - An Updated Parameterised Face", *technical report*, Report No. LiTH-ISY-R-2326, Linköping University, 2001, `http://www.icg.isy.liu.se/candide/main.html`

[4] MPEG Working Group on Visual, International Standard on Coding of Audio-Visual Objects, Part 2 (Visual), ISO-14496-2, 1999

[5] Ahlberg, Jörgen and Forchheimer, Robert, "Face Tracking for Model-based Coding and Face Animation", *International Journal of Imaging Systems and Technology*, Vol. 13, Issue 1, pp. 8-22, 2003

[6] Windows Kinect Face Tracking SDK, `http://msdn.microsoft.com/en-us/library/jj130970`

[7] Zepeda, J.A.Y. and Davoine, F. and Charbit, M., "A linear estimation method for 3D pose and facial animation tracking", *IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR '07*,

[8] Ahlberg, Jörgen, "Extraction and Coding of Face Model Parameters", *techn. report*, Linköping University, 1999

[9] Ahlberg, Jörgen, "Model-baed Coding - Extraction, Coding and Evaluation of Face Model Parameters" *phd-thesis*,SE-581 83, Linköping University, 2002

[10] Hartley, Richard and Zissermann, Andrew, "Multiple View Geometry", 2003

[11] Milborrow, S. and Morkel, J. and Nicolls, F., "The MUCT Landmarked Face Database" *Pattern Recognition Association of South Africa*, 2010, `http://www.milbo.org/muct`