DENSE POINT-TO-POINT CORRESPONDENCES BETWEEN 3D FACES WITH LARGE VARIATIONS FOR CONSTRUCTING 3D MORPHABLE MODELS

Moritz Kaiser, Nicolas Lehment, and Gerhard Rigoll

Institute for Human-Machine Communication
Technische Universität München, Germany
moritz.kaiser@ tum.de

ABSTRACT

In this contribution a novel method to compute dense point-to-point correspondences between 3D faces is presented. The faces are aligned in 3D space with a Generalized Procrustes Analysis and subsequently mapped into 2D space. To compute a correspondence flow between two faces an energy function is minimized which is based on the following assumptions: smoothness of the flow, mapping of landmarks on their counterparts, and texture and depth consistency. Based on these correspondences, the 3D faces are resampled, so that each face is represented by the same amount of 3D points and for any point there is a corresponding point in all other faces. The accuracy of the point-to-point correspondences is demonstrated on the basis of two applications, namely facial texture mapping and the construction of 3D Morphable Models.

Index Terms— 3D surface registration, correspondence estimation, face image processing, face synthesis

1. INTRODUCTION

The 3D Morphable Model (3DMM), presented by Blanz and Vetter [1], is a powerful tool applicable for many computer vision tasks, such as face recognition [1], expression transfer between individuals [2], and face tracking [3]. The crucial step in constructing a 3DMM is to find dense point-to-point correspondences between 3D faces of a database, so that Principal Component Analysis (PCA) can be applied. The accuracy of the correspondences determines the quality of the 3D face model and the performance of the whole application. Establishing point-to-point correspondences between faces is particularly challenging if the faces have different facial expressions. Thus, in [2] spots had to be painted on the skin of the faces in the expression database. Most recent 3DMMs [4, 5] do not model any facial expressions.

Methods to determine dense point-to-point correspondences can be classified into two categories. The first approach is to estimate correspondences based on texture and depth information. In the original 3DMM paper, Blanz and Vetter applied a modified Lucas-Kanade optical flow to determine correspondences. Similar approaches have been presented in [6, 7, 8]. Amberg et al. [9] proposed a nonrigid Iterative Closest Point Algorithm where dense correspondences are computed based on only depth information. Correspondence estimation based on texture and depth has several drawbacks. The estimation might be corrupted if two individuals have different skin color, facial hair or different facial expressions. Note that in the works [1, 4, 9] the faces have no facial hair and neutral facial expression. However, employing a database with a variety of expressions, different amounts of facial hair, or different skin color is desirable, since it makes the model more powerful. We will show in this work that a multi-expression 3DMM built with correspondences based on only texture and depth is not satisfying. Most recently published 3D face databases come already with hand-labeled landmarks (for example [10] or [11]). This information is completely ignored by correspondences based on only texture and depth.

Thus, a second approach is to base the correspondence estimation only on landmarks. The correspondences for all other points are subsequently interpolated by, for example, thin-plate spline functions, as proposed in [5], radial basis functions, as in [12], or triangle quadrisection, as in [13]. However, the landmarks are a quite sparse set of points compared to the overall number of points of a 3D scan. It is obvious that in between landmarks correspondences might not be interpolated correctly.

In this work we propose a novel method that integrates both texture/depth consistency and information from landmarks into one energy function. Based on the landmarks of each face we perform a Generalized Procrustes Analysis so that all faces are aligned in 3D. Subsequently, the 3D surfaces are mapped into the 2D plane and missing pixels are interpolated (Section 2). One reference face is selected and correspondences in 2D are estimated based on texture/depth consistency, landmark consistency, and smoothness (Section 3). The 3D faces are resampled, so that all faces are represented by the same amount of points and corresponding points appear at the same position in the point list (Section 4). In Section 5 we demonstrate the accuracy of our point-to-point correspondences on the bases of two applications, namely facial texture transfer and the construction of a 3DMM. Section 6 concludes the paper.

2. ALIGNMENT AND MAPPING INTO 2D

The output of a 3D scanner is usually a list of 3D points. In addition to that, certain landmarks on the 3D facial surface are given.

Generalized Procrustes Analysis. First, all faces need to be aligned in 3D. For this purpose the Generalized Procrustes Analysis based on the 3D landmarks is applied to all faces. The output of the Generalized Procrustes Analysis is a translation, rotation, and scaling factor per face. For each face, we apply those three transformations to all points and also to the 3D landmarks. For a detailed description of the Generalized Procrustes Analysis refer to [14].

Mapping 3D Surfaces Into 2D. After the faces have been aligned, the 3D points are mapped into the 2D plane. We use parallel projection, i.e., the x- and y-coordinates of a 3D point (if necessary scaled) are directly employed to determine its position in the 2D image. The x- and y-coordinates are usually not whole numbers and the 3D points are not equidistant from each other, which causes holes in

the image. Barycentric interpolation is used to assign a well-defined value to each pixel. Based on the 2D coordinates the points are Delaunay triangulated. Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3$ be the 2D coordinates of the three vertices of a triangle and $\boldsymbol{x} = (x, y)^T$ a pixel position (whole numbers) inside this triangle. The barycentric coordinates of $\boldsymbol{x}$ are:

$$\begin{aligned} b_1(\boldsymbol{x}) &= A(\boldsymbol{x}, \boldsymbol{x}_2, \boldsymbol{x}_3)/A(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3), \\ b_2(\boldsymbol{x}) &= A(\boldsymbol{x}, \boldsymbol{x}_3, \boldsymbol{x}_1)/A(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3), \\ b_3(\boldsymbol{x}) &= A(\boldsymbol{x}, \boldsymbol{x}_1, \boldsymbol{x}_2)/A(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3), \end{aligned} \tag{1}$$

where $A(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3)$ denotes the area of the triangle spanned by $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, and $\boldsymbol{x}_3$. For example, the red component of pixel $\boldsymbol{x}$ is

$$R(\boldsymbol{x}) = \sum_{k=1}^{3} b_k(\boldsymbol{x}) \cdot R(\boldsymbol{x}_k). \tag{2}$$

Green, blue, and depth channels are computed alike. The result is an image where each pixel has an explicit red, green, blue and depth value. Any image resolution can be chosen by the user without causing holes in the image. Note that we also tried cylindrical projection. However, the results with parallel projection were better than with cylindrical projection for the two databases employed in the results in Section 5. A reason for this is that the longitudinal axis of the cylindrical coordinate system has to be determined automatically (e.g. by the center of mass) and may vary considerably between faces. Furthermore, faces are usually quite flat.

Pixels with assigned values are foreground pixels and we store these pixel positions in form of a foreground mask for the correspondence estimation.

3. CORRESPONDENCE ESTIMATION

One face from the database is chosen as *reference face*. For each foreground pixel of the reference face image a corresponding subpixel position in all other face images, here called *new faces*, is searched. The correspondence is encoded by a flow field $\boldsymbol{u} = (u, v)^T$, where u and v are the shifts from a foreground pixel in the reference face to the corresponding subpixel position in a new face in x-direction and y-direction, respectively. The correspondence estimation is based on three assumptions: smoothness, landmark consistency, and texture/depth consistency.

Smoothness. The first requirement for the correspondence estimation is smoothness. The facial skin is a flexible but connected surface and thus neighboring pixels should have similar $\boldsymbol{u}$ values. Only for the mouth and the eyes we need to break this constraint. This leads to the first term of the energy function

$$E_{\text{Smooth}} = \sum_{\boldsymbol{u} \in \Omega_{\text{Smooth}}} \sum_{\boldsymbol{u}_n \in \mathcal{N}} (u - u_n)^2 + (v - v_n)^2, \tag{3}$$

where $\mathcal{N}$ denotes a 3×3 neighborhood around the current pixel. The binary mask Ω_{Smooth} is depicted in Fig. 1 (a). It is set to 1 for all pixels except for three thin lines between mouth and eye corners. The lines are computed with the landmark information.

Landmark Consistency. In addition to smoothness, we require each landmark to be matched to its counterpart in the other face:

$$E_{\text{LM}} = \sum_{\boldsymbol{u} \in \Omega_{\text{LM}}} (u - u_{\text{LM}})^2 + (v - v_{\text{LM}})^2, \tag{4}$$

where u_{LM} and v_{LM} are the required shifts to guarantee landmark consistency. The binary mask Ω_{LM} is set to 1 only for the pixel

Fig. 1. Binary masks that are employed for the three components of the energy function. White means that the pixel is considered, black that it is not considered. (a) Mask for the smoothness constraint. The flow has to be smooth everywhere, even in the background, except for a thin line at eyes and mouth. (b) Mask for the landmark consistency term. Only landmark positions are considered. (c) Texture/depth consistency is required for all foreground pixels.

locations of the few landmarks, as depicted in Fig. 1 (b). Usually landmarks have subpixel positions, so all four affected pixels around the subpixel position are set to 1. In Fig. 2 (a) an exemplary flow field based on only smoothness and landmark consistency is shown. The flow field is evenly interpolated between the four landmarks and it adapts smoothly if one of the landmarks receives a stronger force, as shown in Fig. 2 (b).

This is beneficial for low textured regions, such as cheeks or forehead. However, for other regions a pure interpolation of the flow is not enough, especially if landmarks lie farther away from each other. A very fine and accurate alignment also between landmarks is desired.

Texture and Depth Consistency. Therefore, we introduce a third term, the texture and depth consistency term. We require the intensity, gradient, and depth of corresponding pixels to be equal:

$$I_{c,\text{ref}}(x, y) = I_{c,\text{new}}(x + u, y + v), \tag{5}$$

where index c stands for the three channels (intensity, intensity gradient, depth). The linearized version of this assumption leads to

$$I_{c,x} u + I_{c,y} v + \Delta I_c = 0, \tag{6}$$

where $\Delta I_c = I_{c,\text{new}} - I_{c,\text{ref}}$ and $I_{c,x}$, $I_{c,y}$ are the derivatives in x- and y-direction. Hence, the third term of the energy function is

$$E_{\text{T/D}} = \sum_{\boldsymbol{u} \in \Omega_{\text{T/D}}} \sum_{c} (I_{c,x} u + I_{c,y} v + \Delta I_c)^2. \tag{7}$$

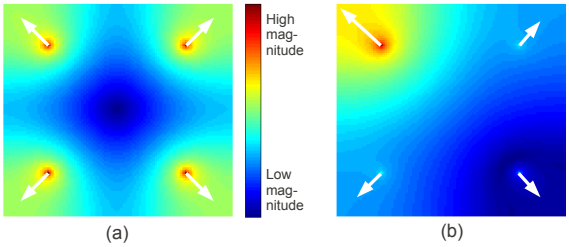

Fig. 2. The flow field is smoothly interpolated between the landmarks. Landmarks must move into a certain direction due to the landmark consistency term. (a) All four landmarks are affected by the same shift. (b) One landmark receives a stronger force and the flow field smoothly adapts.

The binary mask $\Omega_{T/D}$, which is depicted in Fig. 1 (c) is only set to 1 for the foreground pixel positions of the reference face image. Numerical derivatives are computed with a Sobel filter.

Energy Function. The energy function, which will be minimized with respect to $\boldsymbol{u}$, is then

$$E = E_{\text{Smooth}} + \alpha E_{\text{LM}} + \beta E_{\text{T/D}}, \tag{8}$$

where α and β are weighting factors for the landmark and texture/depth consistency term, respectively. In this work both weighting factors are set to 1. Tests confirmed this choice. In order to minimize E, it is derived with respect to u and v for all pixels and subsequently all derivatives are set to zero:

$$\frac{\partial}{\partial \boldsymbol{u}} E = \boldsymbol{0}. \tag{9}$$

Equation (9) is a linear system of equations that can be reformulated to $\boldsymbol{A}\boldsymbol{u} = \boldsymbol{b}$, where $\boldsymbol{A}$ is a highly sparse $2K \times 2K$ matrix, with K being the number of pixels per image. The sparse linear system of equations can be solved with iterative methods for sparse systems. We choose the biconjugate gradient method due to its computational efficiency.

Warping. To overcome convergence to local minima we apply the smart warping approach Brox et al. proposed in [15]. A multiresolution image pyramid is created and the energy function is minimized for the coarsest level. The solution is taken as starting point for the next finer level and so on. Instead of creating the image pyramid with the standard downsampling factor $\mu = 0.5$, a factor closer to 1, here $\mu = 0.8$ as suggested in [15], is employed. Note that before each downsampling step the image is smoothed with a Gaussian kernel with standard deviation $\sigma = 1/\mu$. Previous to the whole correspondence estimation, the required number of pyramid levels L is computed by $L = \lceil \log(\max(\{u_{\text{LM}}\}, \{v_{\text{LM}}\})) / \log(1/\mu) \rceil$, where $\{u_{\text{LM}}\}$ and $\{v_{\text{LM}}\}$ are the sets of the required shifts for all landmarks.

4. RESAMPLING

Subsequently, all faces are resampled. The process is illustrated in Fig. 3. It is important not to lose precision when going from the 3D space into the image domain or using subpixel correspondences. For each 3D point in the reference facial surface, the subpixel position $\boldsymbol{x}_{\text{ref}}$ in the reference face image is determined and the flow $\boldsymbol{u}$ is bilinearly interpolated at this position. The corresponding subpixel position in the new face image is $\boldsymbol{x}_{\text{new}} = \boldsymbol{x}_{\text{ref}} + \boldsymbol{u}$. For the four affected pixels, the corresponding 3D position and the color values have been determined with barycentric interpolation in the 3D-2D mapping step (Section 2).

For each new face the resampled 3D points with accompanying color information are stacked into a face vector with N resampled points. All face vectors have the same size and points with the same index correspond among faces.

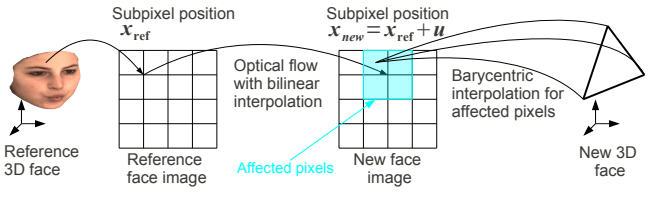

Fig. 3. Illustration of the resampling step. For each 3D point of the reference face a resampled point of the new face is computed. Bilinear and barycentric interpolation are employed to not lose precision.

5. RESULTS

In this section we show two applications for point-to-point correspondences between 3D faces. The applications also demonstrate the accuracy of our method.

5.1. Facial Texture Transfer

With the resampled 3D faces it is trivial to perform facial texture transfer. Assume two vectors with resampled face vectors $\boldsymbol{f}_1$ and $\boldsymbol{f}_2$. The facial texture of face 2 is transferred to the shape of face 1:

$$\boldsymbol{f}_{2\to1} = (x_{1,1}, y_{1,1}, z_{1,1}, r_{2,1}, \ldots, z_{1,N}, r_{2,N}, g_{2,N}, b_{2,N})^T, \tag{10}$$

where $x_{i,j}$ is the x-coordinate of point j of face i. Fig. 4 shows a 3D facial shape with neutral expression. The textures of other faces with strong expressions are mapped on the neutral shape as explained above. It can be seen that the lips, the eyes, and nose textures are mapped accurately on the right positions on the neutral shape indicating proper point-to-point correspondence estimation of our method. For the facial texture transfer, the Bosphorus Database [10] (22 landmarks per face) has been employed.

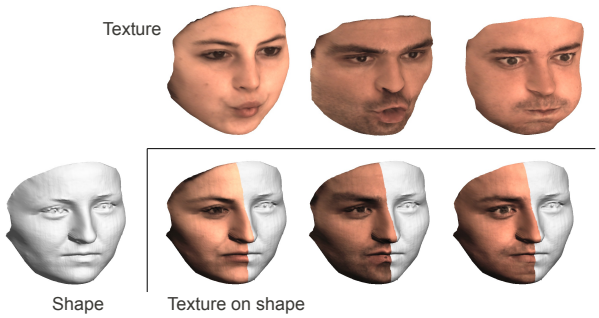

Fig. 4. With dense point-to-point correspondences between faces the texture of one face can be painted onto a shape of another face. Here textures of faces with expressions are mapped on a neutral shape.

5.2. 3D Morphable Models

A more powerful application is the construction of 3D Morphable Models [1]. Depending on the training set a multi-gender, multi-ethnic, or multi-expression 3DMM can be built with accurate point-to-point correspondences.

Construction. The mean face $\bar{\boldsymbol{f}}$ of all resampled faces $\boldsymbol{f}_m$, $1 \leq m \leq M$ is computed and subtracted from each face vector. Subsequently, the mean-subtracted face vectors are written column-wise into a matrix and Principle Component Analysis (PCA) is applied on this matrix. Synthetic faces can be obtained by adding a linear combination of the eigenfaces $\boldsymbol{a}_m$ to the mean face:

$$\boldsymbol{f}_{\text{synth}}(\boldsymbol{\alpha}) = \bar{\boldsymbol{f}} + \sum_m \alpha_m \cdot \boldsymbol{a}_m, \tag{11}$$

where α_m are the weights corresponding to each eigenface.

Visual Comparison. We compare the 3DMM with correspondences based on only texture and depth consistency, as proposed

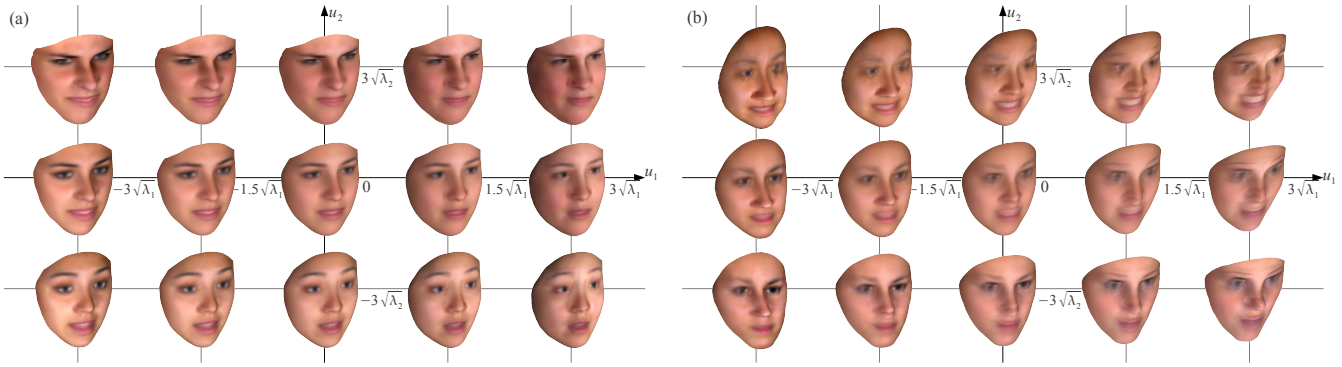

Fig. 5. Synthetic faces for $\alpha_1 \in [-3\sqrt{\lambda_1}, 3\sqrt{\lambda_1}]$ and $\alpha_2 \in [-3\sqrt{\lambda_2}, 3\sqrt{\lambda_2}]$, with λ_1 and λ_2 being the eigenvalues of the first two eigenfaces. The model was trained with a set of 500 faces with multiple expressions from the Binghamton Database. (a) Point-to-point correspondences have been computed with our method. Eyes, eyebrows, mouth, and teeth are clearly defined. (b) Point-to-point correspondences have been computed as proposed in [1]. Facial features, such as nose, eyes, eyebrows, and mouth are blurred, or even appear twice.

in [1], to a 3DMM with correspondences computed with our method. Synthetic faces are created for $\alpha_1 \in [-3\sqrt{\lambda_1}, 3\sqrt{\lambda_1}]$ and $\alpha_2 \in [-3\sqrt{\lambda_2}, 3\sqrt{\lambda_2}]$, where λ_1 and λ_2 are the eigenvalues corresponding to the first two eigenfaces. Fig. 5 shows two 3DMMs which were trained with data from 20 female individuals and 25 different facial expressions per individual. Fig. 5 (b) demonstrates that the strong expressions let the optical flow method [1] fail, so that eyes, eyebrows, and mouth are blurred or even appear twice. In Fig. 5 (a) it can be seen that our method shows promising accuracy for the dataset with multiple expressions. Eyes, eyebrows, mouth, and teeth are well-defined and change smoothly when λ_1 and λ_2 are varied.

6. CONCLUSIONS

In this work a method that computes dense point-to-point correspondences between 3D faces is presented. Correspondences are computed based on texture/depth consistency and landmark consistency assumptions. The accuracy of correspondences computed with our method is demonstrated via two applications. Facial texture mapping is possible with individuals of different ethnicity and different facial expressions. Furthermore, a 3D Morphable Model from a dataset with faces with different facial expressions is constructed. The model based on correspondences computed with our method is visually compared to a model based on correspondences computed with a previously proposed method. In our ongoing research we are trying to apply the 3D Morphable Model for tracking applications.

7. REFERENCES

[1] Volker Blanz and Thomas Vetter, "Face recognition based on fitting a 3D morphable model," *TPAMI*, vol. 25, no. 9, pp. 1063–1074, 2003.

[2] Volker Blanz, Curzio Basso, Tomaso Poggio, and Thomas Vetter, "Reanimating faces in images and video," *Comput. Graph. Forum*, vol. 22, no. 3, pp. 641–650, 2003.

[3] Enrique Muñoz, José M. Buenaposada, and Luis Baumela, "A direct approach for efficiently tracking with 3D morphable models," in *ICCV*, 2009, pp. 1–8.

[4] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter, "A 3D face model for pose and illumination invariant face recognition," in *AVSS*, 2009, pp. 296–301.

[5] Ankur Patel and William A. P. Smith, "3D morphable face models revisited," in *CVPR*, 2009, pp. 1327–1334.

[6] N. Litke, M. Droske, M. Rumpf, and P. Schröder, "An image processing approach to surface matching," in *Symposium on Geometry Processing*, 2005, pp. 207–216.

[7] A. Savran and B. Sankur, "Non-rigid registration of 3D surfaces by deformable 2D triangular meshes," in *CVPR Workshops*, June 2008, pp. 1–6.

[8] Moritz Kaiser, Andre Störmer, Dejan Arsić, and Gerhard Rigoll, "Non-rigid registration of 3D facial surfaces with robust outlier detection," in *WACV*, 2009, pp. 430–435.

[9] Brian Amberg, Sami Romdhani, and Thomas Vetter, "Optimal step nonrigid ICP algorithms for surface registration," in *CVPR*, 2007.

[10] A. Savran, N. Alyüz, H. Dibeklioglu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun, "Bosphorus database for 3D face analysis," in *BIOID*, 2008, pp. 47–56.

[11] Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J. Rosato, "A 3D facial expression database for facial behavior research," in *FG*, 2006, pp. 211–216.

[12] Brian Amberg, Andrew Blake, Andrew W. Fitzgibbon, Sami Romdhani, and Thomas Vetter, "Reconstructing high quality face-surfaces using model based stereo," in *ICCV*, 2007, pp. 1–8.

[13] Moritz Kaiser, Gernot Heym, Nicolas Lehment, Dejan Arsić, and Gerhard Rigoll, "Non-rigid registration of 3D facial surfaces with robust outlier detection," in *WACV*, 2011, pp. 39–44.

[14] J. Gower, "Generalized procrustes analysis," *Psychometrika*, vol. 40, pp. 33–51, 1975.

[15] Thomas Brox, Andrés Bruhn, Nils Papenberg, and Joachim Weickert, "High accuracy optical flow estimation based on a theory for warping," in *ECCV*, 2004, pp. 25–36.