A Facial Expression Model with Generative Albedo Texture

Songnan Li, Fanzi Wu, Tianhao Zhao, Ran Shi, and King Ngi Ngan
The Chinese University of Hong Kong
E-mails: {snli,fzwu,thzhao,rshi,knngan}@ee.cuhk.edu.hk

***Abstract*—A facial expression model (FEM) is developed which can synthesize various face shapes and albedo textures. The face shape varies with individuals and expressions. FEM synthesizes these shape variations by using a bilinear face model built from the Face Warehouse Database. On the other hand, the generative albedo texture is directly extracted from a neutral face model - the Basel Face Model. In this paper, we elaborate the model construction process and demonstrate its application in face reconstruction and expression tracking.**

I. Introduction

3D morphable face model (3DMM) was proposed in 1999 by Blanz and Vetter [1]. Its role is found to be more and more important in face-related computer vision tasks due to at least the following reasons. Firstly, morphable face model can provide prior information that is essential to ill-posed problems, such as 3D face reconstruction from a single color image. Secondly, the analysis-by-synthesis method enabled by this model can extract rich semantic information from the image, such as the head pose, identity, expression, facial landmark positions, the lighting condition, and so on. These automatically-extracted semantic information can benefit machine learning algorithms, most of which nowadays still learn from the limited manually-labelled data. Thirdly, as depth sensors become cheap, compact and portable, 3D data will be easily accessible, making the fitting process of the 3D morphable model more computationally practical.

Generally, 3DMM is a linear model trained on registered samples of human neural face scans. Each face sample is taken as a vector consisting of x, y, z coordinates (shape) or RGB values (albedo texture) of all vertices. Principal component analysis is performed on all face samples to derive the mean and the principal components, whose linear combination can generate an arbitrary new face. The Basel Face Model (BFM) [2] improved 3DMM by offering higher accuracy of shape, texture and registration of the face samples. The Global-to-Local Model (GLM) [3] made further progress by incorporating multi-resolution analysis and local support on high frequency components, which facilitates the model fitting process. However, the three face models mentioned-above can synthesize identity with neural expression only. To reconstruct both the identity and various facial expressions, blendshape model was used in [14]. To construct the user-specific blendshapes, recent works applied deformation transfer on the user's neural expression mesh and then performed refinement usually frame-by-frame on-the-fly [15]. A more integral approach for identity and expression modelling was the multilinear face model proposed in [4], which is adopted in this paper and elaborated in Section II-A. Cao et al. presented the FaceWarehouse Database (FWD) in [7]. Following [4] they developed a bilinear face model using FWD's around 7000 face samples generated from 150 individuals, and demonstrated its usage in several intriguing applications, such as facial image manipulation and face component transfer. However, since FWD dose not provide the albedo texture of each face mesh, the bilinear face model in [4] can synthesize the face shape only.

In this paper, we develop to our best knowledge the first facial expression model that is able to not only synthesize face shape variations due to different identities and expressions but also generate different albedo textures. Specifically, we use the Face Warehouse Database to construct the bilinear model as in [4] for the face shape modelling. On the other hand, the generative albedo texture is extracted from BFM, which provides the albedo values for each of its vertices. By mesh registration, correspondence to BFM is determined for each vertex of the bilinear model, so that the albedo texture can be transferred. Meanwhile, the region segmentation of BFM (eyes, nose, mouth, and the rest of the face) can be transferred similarly. We also manually segment the upper face region, 49 inner facial landmarks and 17 face boundary lines to facilitate the model fitting in applications like rigid head pose tracking [5] and face reconstruction. Furthermore, vertex decimation is performed using the QSLIM method [6] so that the model fitting can be adapted to different resolutions of the input data, or implemented hierarchically to improve the fitting efficiency.

The rest of the paper is structured as follows. In Section II, we present the model development methods, especially the albedo texture transfer from BFM to the new model. Section III illustrates the use of the developed model in face reconstruction and face tracking. Conclusion is drawn in Section IV.

II. The Face Model

A. Bilinear Shape Model

We use the FaceWarehouse Database [7] and adopt the bilinear method of [4,7] to model the face shape deformation due to identity and expression variations. In general, all face meshes of FWD are assembled into a third-order (3-mode)

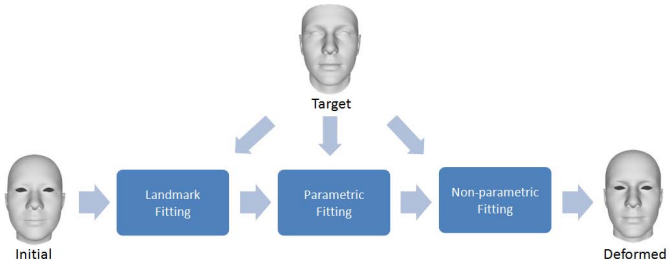


Fig. 1. Flow chart of the registration to Basel Face Model.

data tensor T[1]. The first mode contains vertex positions, while the second and third modes correspond to the identity and expression, respectively. The Higher-Order Singular Value Decomposition (HOSVD) is applied to the data tensor T as

$$T = C \times_2 U_{id} \times_3 U_{exp} \tag{1}$$

where C is the core tensor, operation $\times_n$ is the mode-n multiplication, U_{id} and U_{exp} are orthonormal matrices consisting of the model-2 and mode-3 left singular vectors. As analyzed in [17], most properties of the matrix SVD have a clear higher-order counterpart in HOSVD, especially the approximation property, i.e., the original tensor can be well approximated by discarding the smallest singular values. Therefore, we have

$$\check{T} = C_r \times_2 \check{U}_{id} \times_3 \check{U}_{exp} \tag{2}$$

where C_r, $\check{U}_{id}$ and $\check{U}_{exp}$ are the truncated versions of C, U_{id} and U_{exp}, respectively. $\check{T}$ is a well approximation (though not best) of the original tensor T in a least-square sense. C_r is called the reduced core tensor, and can be used to generate a new face shape as given below

$$S = C_r \times_2 \alpha_{id} \times_3 \beta_{exp} \tag{3}$$

where α_{id} and β_{exp} are the identity and expression parameters, whose dimensions are chosen to be 50 and 25, respectively, leading to an average reconstruction error of about 0.25 mm on FWD. The identity and expression parameters are assumed to have a multivariate Gaussian distribution, of which the mean (μ_α and μ_β) and the standard deviation (σ_α and σ_β) are derived from matrices $\check{U}_{id}$ and $\check{U}_{exp}$. The reduced core tensor C_r together with the multivariate Gaussian distributions of the parameters constitute our bilinear face shape model.

B. Generative Albedo Texture

The Basel Face Model (BFM) can synthesize a neutral face with various albedo textures. As introduced in Section I, the albedo texture is modelled using Principle Component Analysis (PCA), i.e., each vertex of BFM is associated with a 3×1 vector of mean RGB values and a $3 \times M$ matrix of principle components (M is the dimension of the texture parameters), from which new albedo values can be generated. In order to transfer this generative albedo texture to the bilinear face model, we need to find the vertex correspondences between these two models. Therefore, mesh registration is performed in three steps, as shown in Fig. 1. The first step is the landmark fitting. A rough registration is obtained by tuning the identity parameters of the bilinear face model to fit the 26 3-D landmarks (8, 4, 6, 8 points around eyes, nose, mouth, and ears, respectively, which are manually chosen on both models) as follows

$$\min_{R,t,\alpha} \sum_{i \in N_l} ||f_i(R,t,\alpha,\mu_\beta) - v_i||^2 + \lambda(\alpha - \mu_\alpha)^\top Q_\alpha (\alpha - \mu_\alpha) \tag{4}$$

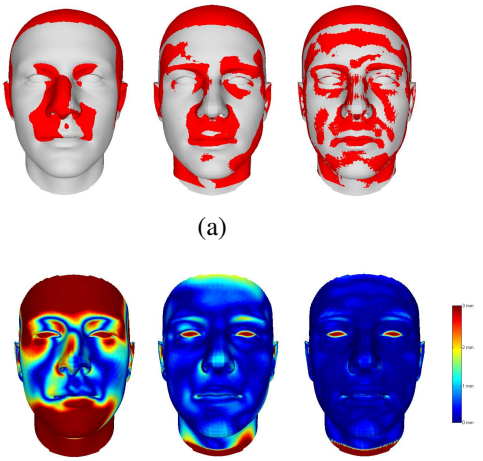
and

$$f_i(R,t,\alpha,\beta) = R \times (C_{r,i} \times_2 \alpha \times_3 \beta) + t \tag{5}$$

where N_l is the set of landmark indices, R and t represent the rigid rotation and translation, α denotes the identity parameters, μ_α and μ_β denote the mean parameters for identity and expression, respectively, v_i is the 3-D position of the i^{th} landmark on BFM, $C_{r,i}$ is a portion of the reduced core tensor associated with the i^{th} vertex, Q_α is a diagonal matrix which contains the reciprocal of the variance of each identity parameter. The regularization term in (4) penalizes deviation from the mean, with λ controlling the regularization intensity. The optimization is implemented using MATLAB with gradient-descent method.

The second step is the parametric fitting using dense corresponding vertex pairs. The objective function is similar to (4), with N_l replaced by a set of denser vertices, which are located by the nearest neighbour method. A corresponding vertex pair is rejected as outlier if the distance between the two vertices or their difference in normal directions is larger than a pre-defined threshold (3 mm). Finding correspondences and minimizing the objective function are performed iteratively until the fitting result converges. Finally, non-parametric fitting is performed in the last step, where the Laplacian mesh deformation method in [8] is adopted to further enhance the registration accuracy.

Fig. 2(a) shows the overlay of the deformed mesh (red) and the target mesh (white) after step one (left), two (middle) and three (right), respectively. It can be observed subjectively

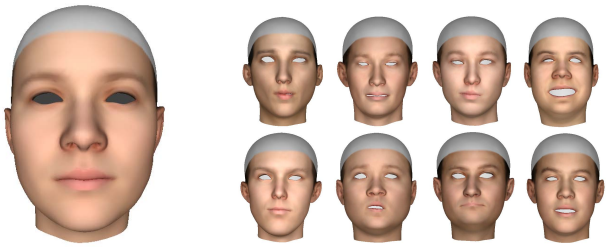


(a)

(b)

Fig. 2. From left to right (a) the overlay of the deformed mesh (red) and the target mesh (white) in steps 1-3, and (b) the hot maps of fitting error generated from steps 1-3.

[1]Tensors are higher-order equivalents of vectors (first-order) and matrices (second-order). Refer to [17] for details.



(a) (b)

Fig. 3. (a) Mean texture. (b) Random identities, expressions and textures.

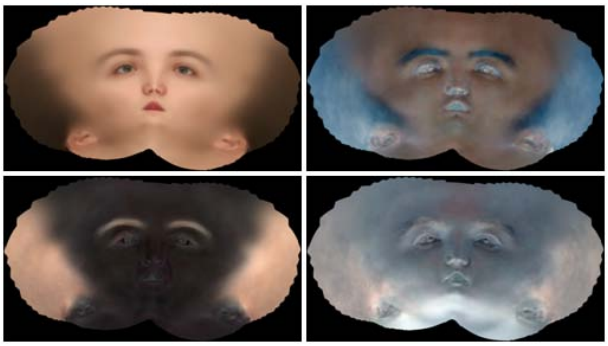


Fig. 4. For top to bottom and left to right: texture maps generated for the mean, 1st, 2nd, and 3rd principle components, respectively.

that the registration accuracy improves from left to right. Fig. 2(b) uses the hot maps to visualize the fitting error. From left to right, the hot maps are generated from step one, two, and three, respectively. Notice that after the non-parametric fitting, the fitting error for most vertices is close to 0 mm. The high fitting errors such as those around the eyes are caused by the topology differences between BFM and the bilinear face model, specifically, there are two holes in the eyes region of the bilinear face model while BFM is a disc-like mesh.

After the mesh registration, vertex correspondences are determined by the nearest neighbour method, and the PCA-based generative albedo of BFM is transferred to the bilinear face model vertex-by-vertex. Fig. 3(a) shows the bilinear face model with the mean albedo texture transferred from BFM. Fig. 3(b) presents eight exemplar faces with random identities, expressions and textures.

It should be noted that BFM has a much higher vertex density, which is the reason we directly find correspondence for each vertex of the bilinear model without interpolation. But in other words, only part of the BFM texture is transferred to the bilinear model. To preserve the BFM texture more completely, UV unwrapping [11] is performed on BFM to bijectively project each 3D vertex onto an 2D texture map. A dense texture map is then generated by interpolation. For the mean texture and each of its principle components, a texture map can be produced following this process. Fig. 4 shows the texture maps generated for the mean and the first three principle components. Using these texture maps and UV coordinates which can be derived from the above mesh registration procedure, faces with more detailed textures can be synthesized.

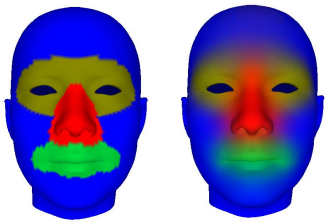


Fig. 5. Binary and weighted segmentations transferred from BFM.

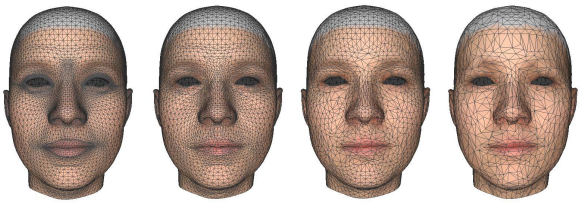


Fig. 6. From left to right: the original model with 11K vertices, down-sampled models with 8K, 4K, and 2K vertices, respectively.

C. Semantic Segmentation and Vertex Decimation

Besides the texture, BFM's segmentation information can also be transferred to the bilinear model after the mesh registration, as shown in Fig. 5, which is indispensable for segment-based face model fitting. We also manually mark the upper face region (excluding the mouth and jaw) for rigid head pose tracking [5], 49 inner landmarks and 17 boundary lines [9] to facilitate the landmark-based face reconstruction. Furthermore, vertex decimation is performed using the QSLIM method [6], as shown in Fig. 6. These down-sampled models can be used to improve the model fitting efficiency, by using the sparser model to suit the low-resolution input or provide the rough initialization result for the further refinement.

III. Applications

We developed two applications, i.e., face reconstruction (Section III-A) and expression tracking (Section III-B), to demonstrate the usability of the textured bilinear model. Limited by the paper length, a brief description on the implementation is given blow.

A. Face Reconstruction

In face reconstruction, the model parameters for identity, expression, and texture are determined by fitting an input frame captured by Kinect v1. Specifically, we use 2-D facial landmarks ([12,13]) and the depth map to initialize the identity and expression parameters by minimizing an objective function similar to (4), which consists of two data terms measuring total distances of corresponding 2-D landmarks and 3-D vertices, respectively, and two regularization terms for identity and expression parameters. Then, the lighting condition is estimated using 1st order spherical harmonics [16], and the texture parameters are determined under this lighting condition by fitting the input color image pixel-wise. Finally, identity and expression parameters are further refined by fitting the color image with the fixed albedo texture. Fig. 7 shows

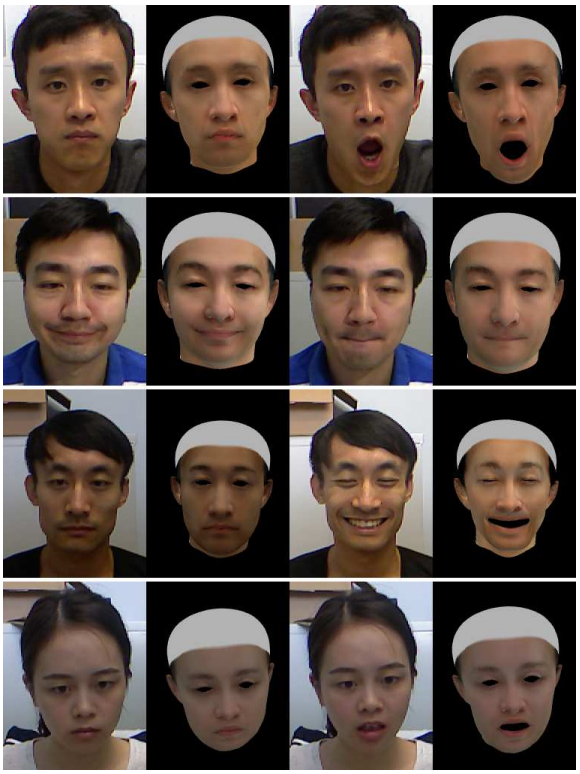


Fig. 7. Face reconstruction examples.

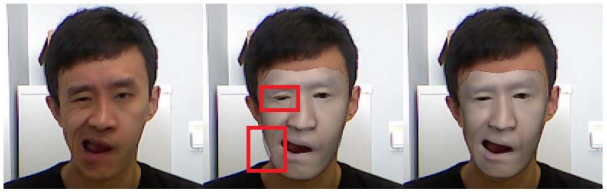


Fig. 8. Comparison of face reconstruction results between using generative albedo texture (right) and not (middle).

several examples of the reconstruction results. Notice that using the textured bilinear model we can fit not only the neutral face but also various expressions with personalized textures. Fig. 8 illustrates the effectiveness of using the generative albedo texture in face reconstruction. Compared with using the landmarks and depth map only (Fig. 8 middle), integrating the albedo texture can improve the fitting accuracy around the eyes and face boundary (Fig. 8 right). Similar improvements can be commonly observed in other images.

B. Expression Tracking

In expression tracking, a similar procedure is performed on each input frame, i.e., which fits landmarks and depth values firstly and the color values next, by tuning the expression parameters with the fixed identity, albedo texture, and frame-wisely-estimated lighting condition. (The identity and texture parameters are determined in the first input frame using the face reconstruction algorithm introduced above.) A exemplar tracking result can be seen from the project website [10]. Notice that the synthesized faces can well represent the input. Further improvement can be made by using multiple frames for the identity and texture reconstruction, and considering temporal smoothness of the expression change.

IV. Conclusion

By incorporating merits from the bilinear face model [4,7] and the Basel Face Model [2], a facial expression model is developed which can synthesize face variations caused by three factors: identity, expression, and albedo texture. To our best knowledge, it is the first face model that combines all three factors together. In the future work, we plan to make further improvements to the current model by incorporating multi-resolution analysis, local-supports [3], details synthesis (e.g., wrinkles, expression lines) ability, and eyeball/inner-mouth synthesis ability.

Acknowledgment

We thank Prof. Thomas Vetter and Prof. Kun Zhou for sharing the Basel Face Model and the FaceWarehouse Database which motivates this work.

References

[1] V. Blanz and T. Vetter, "A Morphable Model For the Synthesis of 3D Faces," Proc. of SIGGRAPH, pp. 187-194, 1999.
[2] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, T. Vetter, "A 3D face model for pose and illumination invariant face recognition," Proc. of AVSS, pp. 296-301, 2009.
[3] R. Knothe, "A global-to-local model for the representation of human faces," PhD Thesis, University of Basel, 2009.
[4] D. Vlasic, M. Brand, H. Pfister, J. Popovic, "Face Transfer with Multilinear Models," ACM TOG, 24(3), pp. 426-433, 2005.
[5] S. Li, K. N. Ngan, R. Paramesran, L. Sheng, "Real-time Head Pose Tracking with Online Face Template Reconstruction," IEEE TPAMI, accepted, 2016.
[6] M. Garland, P. S. Heckbert, "Surface simplification using quadric error metrics," Proc. of SIGGRAPH, pp. 209-216, 1997.
[7] C. Cao, Y. Weng, S. Zhou, Y. Tong, K. Zhou, "FaceWarehouse: a 3D Facial Expression Database for Visual Computing," IEEE Transactions on Visualization and Computer Graphics, 20(3): 413-425, 2014.
[8] O. Sorkine, D. Cohen-Or, Y. Lipman, M. Alexa, C. Rossl, H. P. Seidel, "Laplacian surface editing," Proc. of SGP, pp. 175-184, 2004.
[9] C. Cao, Q. Hou, K. Zhou, "Displaced Dynamic Expression Regression for Real-time Facial Tracking and Animation," ACM TOG, 33(4), 2014.
[10] Project Website: http://www.ee.cuhk.edu.hk/~snli/face_model.htm/~snli/face_model.htm
[11] B. Levy, S. Petitjean, N. Ray, J. Maillot, "Least squares conformal maps for automatic texture atlas generation," Proc. of SIGGRAPH, pp. 362-371, 2002.
[12] T. Baltrusaitis, P. Robinson, L. P. Morency, "Constrained Local Neural Fields for Robust Facial Landmark Detection in the Wild," ICCVW, pp. 354-361, 2013.
[13] X. Xiong, F. Torre, "Supervised Descent Method and its Applications to Face Alignment," CVPR, pp. 532-539, 2013.
[14] T. Weise, S. Bouaziz, H. Li, M. Pauly, "Realtime performance-based facial animation," ACM TOG, 30(4), 2011.
[15] P. L. Hsieh, C. Ma, J. Yu, H. Li, "Unconstrained Realtime Facial Performance Capture," CVPR, 2015.
[16] S. Suwajanakorn, I. Kemelmacher, S. M. Seitz, "Total Moving Face Reconstruction," ECCV, pp. 796-812, 2014.
[17] L. D. Lathauwer, B. D. Moor, J. Vandewalle, "A Multilinear Singular Value Decomposition," SIAM Journal on Matrix Analysis and Applications, 21(4):1253 - 1278, 2000.