PCA-based 3D Face Photography

Jesus P. Mena-Chalco
IME - Universidade de Sao Paulo
jmena@ime.usp.br

Ives Macedo
Instituto Nacional de Matematica Pura e Aplicada
ijamj@impa.br

Luiz Velho
Instituto Nacional de Matematica Pura e Aplicada
lvelho@impa.br

Roberto M. Cesar-Jr
IME - Universidade de Sao Paulo
cesar@ime.usp.br

Abstract

This paper presents a 3D face photography system based on a small set of training facial range images. The training set is composed by 2D texture and 3D range images (i.e. geometry) of a single subject with different facial expressions. The basic idea behind the method is to create texture and geometry spaces based on the training set and transformations to go from one space to the other. The main goal of proposed approach is to obtain a geometry representation of a given face provided as a texture image, which undergoes a series of transformations through the texture and geometry spaces. Facial feature points are obtained by an active shape model (ASM) extracted from the 2D gray-level images. PCA then is used to represent the face dataset, thus defining an orthonormal basis of texture and range data. An input face is given by a gray-level face image to which the ASM is matched. The extracted ASM is fed to the PCA basis representation and a 3D version of the 2D input image is built. The experimental results on static images and video sequences using seven samples as training dataset show rapid reconstructed 3D faces which maintain spatial coherence similar to the human perception, thus corroborating the efficiency of our approach.

1. Introduction

The process of construction of 3D facial models is an important topic in computer vision which has recently received attention within the research community. This is an example of the so called computational photography where computer vision and graphics methods are used to solve a given problem. Modelling facial data depends on the nature of the considered problem. Usually, models with accurate geometry are preferred for face recognition, and more simpler models are preferred in applications where the speed of the process is a critical factor [5], e.g. face transmission or augmented reality.

Face images play a central role in different applications of computer vision and graphics. Different methods for 3D face detection, tracking and representation have been developed to address applications such as face recognition [1, 7, 12, 15], facial expression analysis [13, 16], face synthesis [10, 17] and video puppeteering [2, 6]. As far as face synthesis is concerned, most 3D face reconstruction methods proposed so far are based on artificial 3D face models such as public available avatars [2]. Jiang *et al.* [7] explored an alignment and facial feature extraction algorithm for automatic 3D face reconstruction. Their algorithm subsequently applies principal component analysis on the shape to compute 3D shape coefficients. However, a frontal face image of a subject with normal illumination and neutral expression is required. Despite more than 3 decades of research [8, 9], there are still some important 3D face photography open problems. This paper presents a new approach for 3D face computational photography using real-data based models. The main contributions of this paper rely on the new approach itself which, because of being based on real geometry data, produces more realistic 3D reconstructed faces. The system works with few training samples and relies on standard vision and graphics algorithms and representations, thus leaving space for different improvements in the future.

Starting from the work of Vlasic *et al.* [14] and Macedo *et al.* [11], an automatic system for 3D face reconstruction from 2D color images using a small training set of range images (registered texture and geometry data) has been created. It is worth noting that these previous works [11, 14] do not explore 3D data. The training set is composed by a small set of range images corresponding to some different facial expressions of a single subject. Our approach employs Principal Component Analysis to represent the face model (texture and geometry separately). In the system, the PCA face model is composed by two separate orthonormal

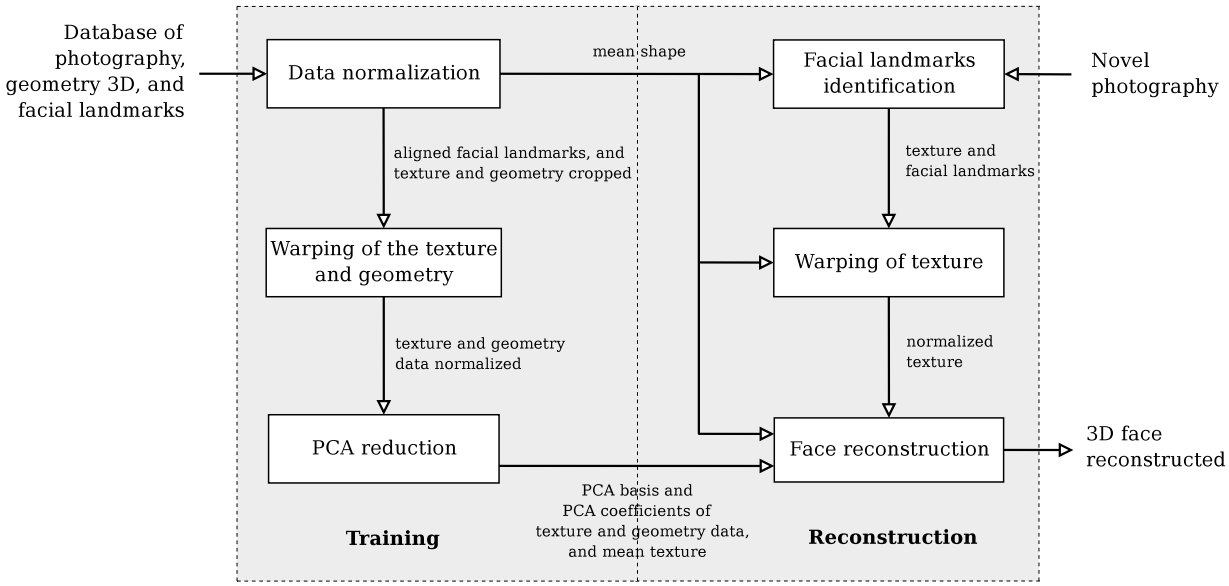


Figure 1. Schematic data flow diagram of the 3D facial reconstruction system (dotted lines). Each block (solid lines) represents a process while each arrow represents the information flow between processes.

basis which represent texture and geometry, respectively.

Given an input frontal face image to be 3D reconstructed, an Active Shape Model (ASM) is used to extract the 2D facial feature points on 2D gray-level images. The set of feature points is used to normalize the input texture. The 3D facial geometry is produced by projecting the normalized texture onto the geometry space (obtained in the training procedure). The projection is produced by using a PCA vectorial basis and a linear optimization function to relate 2D texture and 3D geometry information. Finally, the 3D reconstruction is obtained by directly mapping the normalized texture onto the geometry. Figure 1 summarizes the proposed system architecture.

This paper is organized as follows. An overview of the proposed mathematical scheme is presented in Section 2. The introduced method is described in Section 3. Experimental results are shown in Section 4. The paper is concluded with some comments on our ongoing work in Section 5.

2. Mathematical model overview

The list below summarizes the symbols used in the current paper, being presented to help the reader:

- l_i^t, l_i^g: i-th texture and geometry landmarks, respectively;
- L_i^t, L_i^g: i-th texture and geometry landmark matrices, respectively;
- x^t, x^g: input texture face and corresponding output reconstructed geometry, respectively;
- E^t, E^g: texture and geometry PCA basis;
- α^t, α^g: texture and geometry coefficients expressed in terms of E^t, E^g, respectively;
- s_x: weighting coefficients of x^t in terms of the training samples.

The proposed approach is based on learning a 3D face model using texture and geometry of a training face for some different facial expressions. An input 2D face image (i.e. only texture) is then reconstructed by projecting it on the trained 2D texture space, decomposing it as weights of the training samples. The obtained weighting coefficients are then used to build a 3D model from the 3D training geometry samples.

The training set is composed by pairs of texture and geometry data from a given subject with some different facial expressions. A set of landmarks $\{l_1^t, l_2^t, ..., l_K^t\}$ are placed on the texture image and used to represent the input texture information. Therefore, each facial texture is represented by a matrix L^t composed by the landmarks information. Because texture and geometry data are registered, the texture landmarks have corresponding geometry counterparts, which are used to define the geometry landmarks $\{l_1^g, l_2^g..., l_k^g\}$. Hence, each facial geometry is represented by a matrix L^g composed by the geometry landmarks informa-

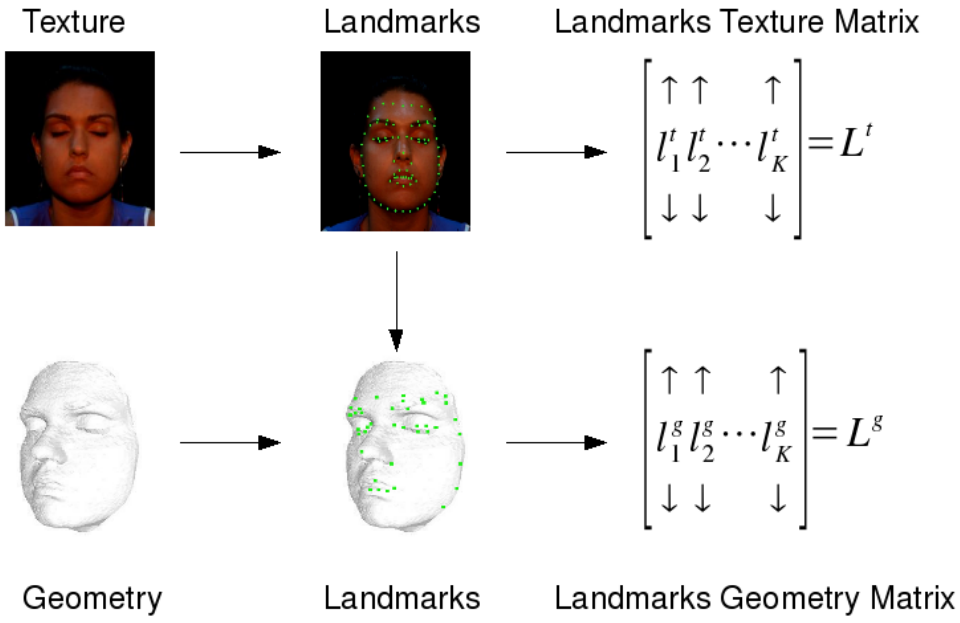


Figure 2. Training set formation: texture and geometry landmarks.

tion. This initial landmarks representation scheme is illustrated in Figure 2.

The training phase consists in defining good texture and geometry space representations based on a given of set facial expression samples of a given subject. Therefore, texture and geometry landmark matrices are obtained for N different facial expressions, being denoted as $\{L_1^t, L_2^t..., L_N^t\}$ and $\{L_1^g, L_2^g..., L_N^g\}$, respectively. These matrices help to define the initial texture and geometry spaces, as illustrated in Figures 3(a) and (b). In order to have a more efficient and statistically optimized representation, both texture and geometry spaces are PCA-transformed (Figures 3(c) and (d)). Each training sample represents a vector expressed in these spaces.

The main goal of the 3D photography system is to obtain a geometry representation of a given face provided as a texture image. A landmark representation x^t is automatically extracted from such input image and undergoes a series of transformations through the texture and geometry spaces, as illustrated in Figure 3. The final result is the reconstructed geometry of the input face, i.e. a point in the geometry space.

3. The 3D face reconstruction system

The proposed system is composed by three parts, the first two being executed off-line: data acquisition, system training and 3D face reconstruction.

3.1. Data acquisition and face model

3D face data is acquired using a non-contact 3D scanner KONICA MINOLTA VIVID 910. The scanner is composed by a laser distance sensor and a digital video camera. The scan volume specifications are: $111 \times 84 \times 40mm$ (min) to $1200 \times 903 \times 400mm$ (max) (width $\times$ depth $\times$ height, respectively). Texture images have been acquired with a 320×240 pixels resolution. The 3D geometry associated to each texture image contains appromimately 15000 points.

The data is hence composed by registered texture and geometry data. Images from a single subject have been acquired with 7 different facial expressions (one single image per facial expression). The training data has been obtained in a controlled environment (illumination and subject pose and position).

Once the training images have been acquired, facial landmarks are manually placed over the texture images and aligned to the corresponding range images. We adopted a face model with $K = 77$ landmark points, and one triangulation contains 120 elements (triangles).

Figure 4 shows an example of a face obtained.

3.2. Training

The training procedure is composed by three phases. Firstly, the input data is normalized by Procrustes analy-

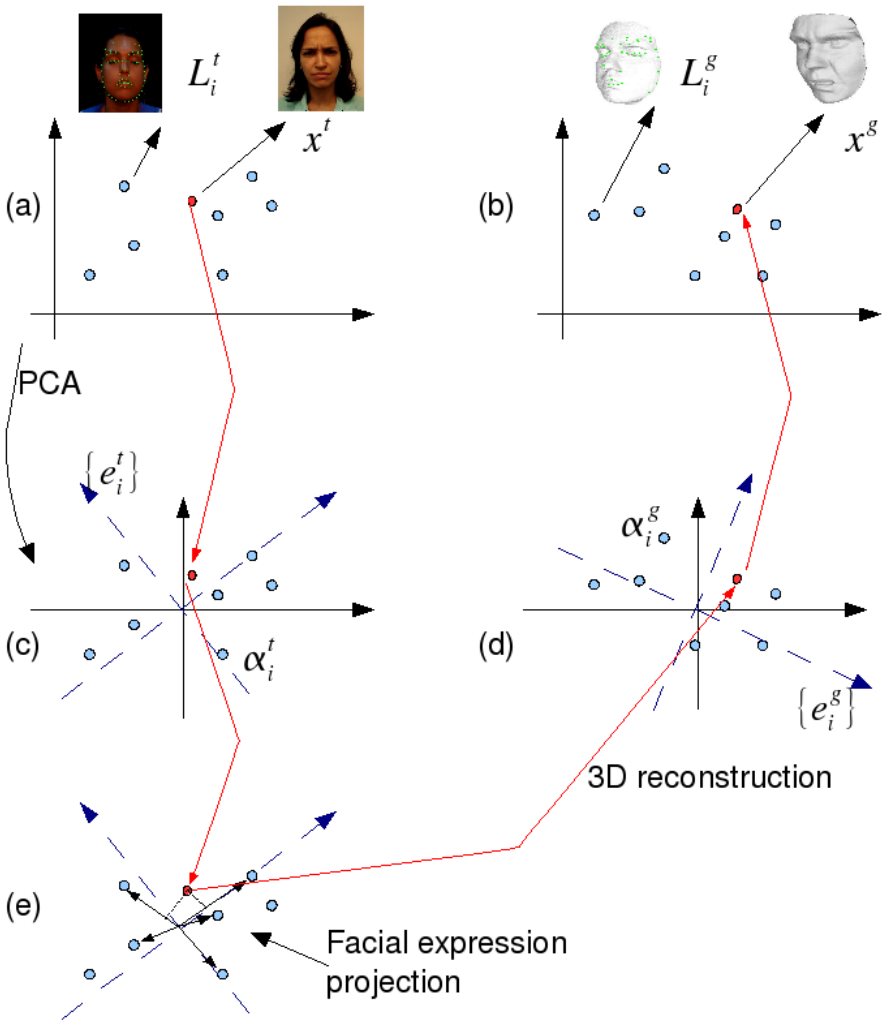


Figure 3. Texture and geometry spaces. x^t is an input texture face which undergoes a series of transformations through texture and geometry space until its geometry x^g is built.

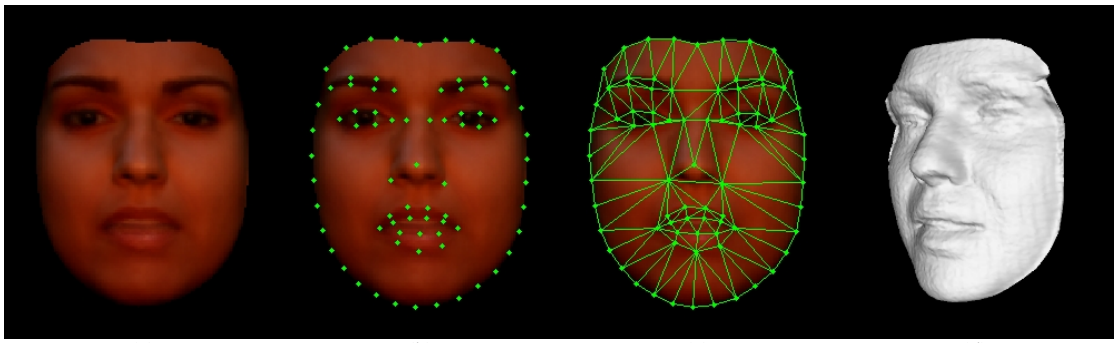
sis [4], thus resulting in a dataset with landmarks aligned in a common coordinate system. The facial landmarks may be aligned for the different input images because of homology among the individual representations. This fact allows to map each input image by warping onto the average face data. The geometry data is also mapped onto the average face produced by the Procrustes analysis.

Two PCA procedures are carried out separately for the geometry L_i^g and for the texture L_i^t data. Such analysis lead to:

- An average texture model (t_0), an orthonormal basis ($E^t = \{e_i^t\}$) for the facial texture space and the coefficients ($\{\alpha_i^t\}$) for each texture image in the training set expressed w.r.t. $\{e_i^t\}$;
- An average geometry model (g_0), an orthonormal basis ($E^g = \{e_i^g\}$) for the facial geometry space and the coefficients ($\{\alpha_i^g\}$) for each 3D geometry data in the training set expressed w.r.t. $\{e_i^g\}$.

In order to work with the same number of principal components in the aforementioned spaces, we use the minimum amount of components representing a pre-defined amount of total variance kept by both basis. The results shown in this paper were drawn from those in which PCs kept at least 95% of the total variance.

The training pipeline is summarized in the system overview for 2D-to-3D face reconstruction of Figure 1.



(a) (b) (c) (d)

Figure 4. Face data obtained in training phase. (a) texture; (b) facial landmarks; (c) triangulation; (d) geometry.

3.3. Face reconstruction

The input to the system is a frontal face image to which the ASM is applied in order to automatically detect the facial landmarks. The ASM landmarks are extracted from the gray scale input image. ASM has been proposed in [3] and allows the alignment and representation of image data using a statistical model of the target object obtained from training data. A face model is represented by a set of landmarks manually placed over training face images (not necessarily those obtained for the 3D face model). The sets of landmarks for the training images are aligned in order to minimize the distance between corresponding points (i.e. homologous points). A point distribution model (PDM) is obtained from the variance of the distances among the different points. The PDM is used to constraint the shape variation in the ASM matching process.

The facial landmarks are aligned to the mean face shape obtained in the training process. Thus, the texture is warped to the mean shape (similar to the process done in training). This process allows to normalize the texture of the input image.

Let x^t be the warped texture of the input image, and t_0 the normalized average texture obtained in training process. The texture coefficients, α_x^t, are calculated by projecting $(x^t - t_0)$ onto the respective orthonormal basis ($\{e_i^t\}$):

$$\alpha_x^t = E^t.(x - t_0) \tag{1}$$

where E^t is a transformation matrix defined by the orthonormal basis for the texture space learned in the training process.

Once the texture coefficients α_x^t are obtained, the texture coefficients α^t of all images considered in the training process are used to calculate the coefficients s_x, defined as:

$$\alpha^t.s_x = \alpha_x^t \tag{2}$$

where α^t is the matrix defined by the coefficients for each texture image in the training set. Intuitively, s_x represents weighting coefficients obtained by projecting α_x onto α (Figure 3(e)). It is important to recall that each sample represented in α is associated to a different facial expression. Therefore, s_x represents a decomposition of x^t in terms of the different facial expressions learnt by the system (e.g. as we would say that x^t is $a\%$ happy, $b\%$ angry, $c\%$ neutral, etc.).

The geometry coefficients α_x^g of x are then calculated using the geometry coefficients of all training geometry samples α^g:

$$\alpha_x^g = \alpha^g.s_x \tag{3}$$

The normalized geometry x^g of the test face image x is then reconstructed by:

$$x^g = (E^g.\alpha_x^g) + g_0 \tag{4}$$

where E^g is a transformation matrix defined by the orthonormal basis for the geometry space learned in the training process. Laplacian smoothing has been applied to reduce noise on the reconstructed facial geometry (surface). This smoothing technique was selected because of being robust and efficient to smooth a general mesh.

Finally, the input texture warped to the average shape face is directly mapped onto the 3D smooth geometry. It is important to note that missing blank areas are filled by interpolation of adjacent 3D points. The test process is summarized in the system overview of Figure 1.

4. Results

The proposed system has been tested using real data. All experiments were performed with only seven training images (texture and geometry). Figure 5 shows the different

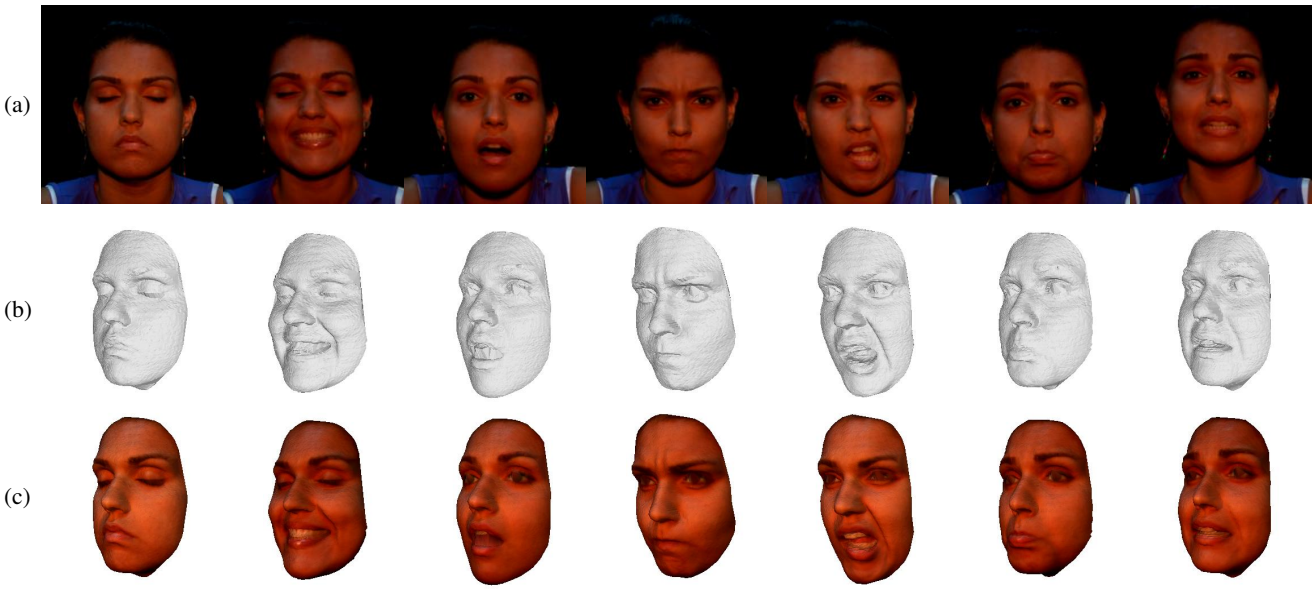


Figure 5. Training dataset. (a) frontal faces with different expressions of one subject; (b) face geometry; (c) 3D face with mapped texture.

training images used in this work. We used a 3D scanner KONICA MINOLTA VIVID 910 for the data acquisition as mentioned in Section 3.1. It is important to note that, for all experiments, a geometry of 14104 points have been obtained at most 4 seconds using a simple implementation on MATLAB.

Two different types of experiments have been performed to evaluate the 3D facial reconstruction quality of the system. One experiment was considered to examine the 3D reconstruction taking as input an image of the same person seen in the training process, but with different facial expression. In other words, the geometry of the same person used to train the system but with a different facial expression has been used. A second experiment was performed for 3D reconstruction using a face of a different person from that of the training phase. In this second case, a video sequence of the subject has been obtained using a standard webcam and the facial features have been tracked using the ASM approach described above. This input data has been fed to the system in order to reconstruct a 3D geometry video sequence.

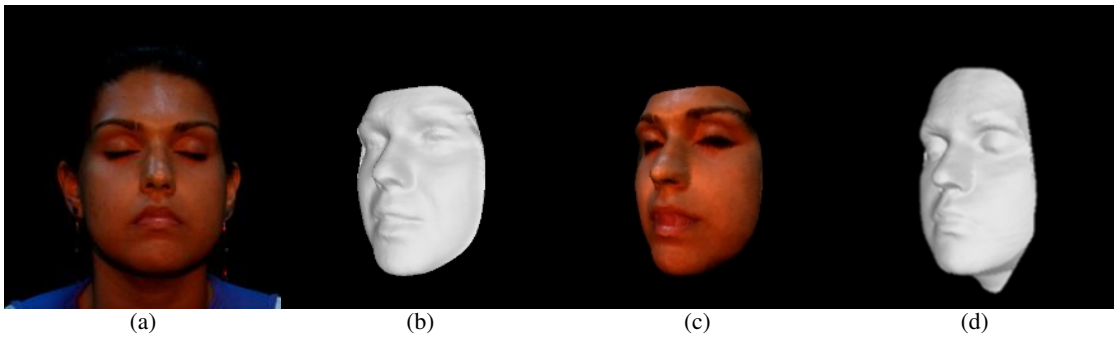
The texture shown in Figure 6(a) has been used as input to the system. This facial expression was not present in the training phase. Figure 6(b) shows the corresponding reconstructed 3D facial geometry. Figure 6(c) shows the facial geometry of the input image with the texture mapped onto it. As can be seen, the system was able to successfully reconstruct the 3D face structure of the subject. Figure 6(d) shows the face geometry acquired of subject in Figure 6(a) using a 3D scanner.

A different experiment has been devised to test the system using a subject that is not present in the training phase. A video sequence of a different subject has been obtained using a standard webcam in a non-controlled environment. It is worth noting that the so obtained texture information has different technical specifications from those used to train the system. ASM has been applied to track the video sequence and used as input to the face reconstruction system. Figure 7 shows the 3D face reconstruction for five frames of the video sequence. The calculated 3D reconstructions are show in Figures 7(b)-(c).

5. Conclusions

This paper describes a system for 3D reconstruction of faces from 2D photographs based on a small set of training samples. The mathematical model behind the system is based on building suitable texture and geometry spaces from the training set and transformations between such spaces. Face reconstruction from an input texture image is then carried out by transforming the input data through these spaces until a geometry model is created. The experimental results have shown that the proposed method may be applied to 3D reconstruction of faces from a video sequence.

Our experimentation with the system has shown that a key point for the performance of the system with faces not present in the training phase relies on the identification of



(a) (b) (c) (d)

Figure 6. 3D reconstruction of a facial expression not present in the training phase: (a) frontal face; (b) reconstructed 3D face geometry; (c) 3D geometry with the mapped texture; (d) face geometry acquired with the scanner;

the facial landmarks. The system is sensitive to the identification of the facial landmarks considered as the first step of the reconstruction, i.e. $\{l_1^t, l_2^t, ..., l_K^t\}$. Unfortunately, the ASM has not always performed as would be desirable, thus leading eventually to bad reconstruction results. Thus, a better method for automatic landmark identification would certainly help in improving the performance obtained. This is one of the main topics of our ongoing work.

Acknowledgments

Financial support for this research has been provided by CAPES, CNPq and FAPESP. The authors are grateful to anonymous reviewers for the critical comments and valuable suggestions.

References

[1] V. Blanz. Face recognition based on a 3D morphable model. In *International Conference on Automatic Face and Gesture Recognition*, pages 617–624, 2006.

[2] J.-X. Chai, J. Xiao, and J. Hodgins. Vision-based control of 3D facial animation. In D. Breen and M. Lin, editors, *Eurographics/SIGGRAPH Symposium on Computer Animation*, pages 193–206, San Diego, California, 2003. Eurographics Association.

[3] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models: Their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, January 1995.

[4] I. L. Dryden and K. V. Mardia, editors. *Statistical Shape Analysis*. John Wiley and Sons, Chichester, 1998.

[5] E. Elyan and H. Ugail. Reconstruction of 3D human facial images using partial differential equations. *Journal of Computers*, 2(8):1–8, 2007.

[6] P. Hong, Z. Wen, T. S. Huang, and H. Y. Shum. Real-time speech-driven 3D face animation. In *3D Data Processing Visualization and Transmission*, pages 713–716, 2002.

[7] D. L. Jiang, Y. X. Hu, S. C. Yan, L. Zhang, H. J. Zhang, and W. Gao. Efficient 3D reconstruction for face recognition. *Pattern Recognition*, 38(6):787–798, June 2005.

[8] T. Kanade. Picture processing system by computer complex and recognition of human faces. In *Doctoral dissertation, Kyoto University*. November 1973.

[9] J. V. Kittler, A. Hilton, M. Hamouz, and J. Illingworth. 3D assisted face recognition: A survey of 3D imaging, modelling and recognition approaches. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 3, pages 114–114, 2005.

[10] M. W. Lee and S. Ranganath. 3D deformable face model for pose determination and face synthesis. In *International Conference on Image Analysis and Processing*, pages 260–265, 1999.

[11] I. Macedo, E. V. Brazil, and L. Velho. Expression transfer between photographs through multilinear AAM's. In *SIBGRAPI*, pages 239–246. IEEE Computer Society, 2006.

[12] D. Onofrio and S. Tubaro. A model based energy minimization method for 3D face reconstruction. In *ICME*, pages 1274–1277. IEEE, 2005.

[13] H. Soyel and H. Demirel. Facial expression recognition using 3D facial feature distances. In *International Conference on Image Analysis and Recognition*, pages 831–838, 2007.

[14] D. Vlasic, M. Brand, H. Pfister, and J. Popović. Face transfer with multilinear models. *ACM Transactions on Graphics*, 24(3):426–433, 2005.

[15] Y. J. Wang and C. S. Chua. Face recognition from 2D and 3D images using 3D Gabor filters. *Image and Vision Computing*, 23(11):1018–1028, October 2005.

[16] T. Yabui, Y. Kenmochi, and K. Kotani. Facial expression analysis from 3D range images; comparison with the analysis from 2D images and their integration. In *International Conference on Image Processing*, pages 879–882, 2003.

[17] Y. Zhang and S. Xu. Data-driven feature-based 3D face synthesis. In *3-D Digital Imaging and Modeling*, pages 39–46, 2007.

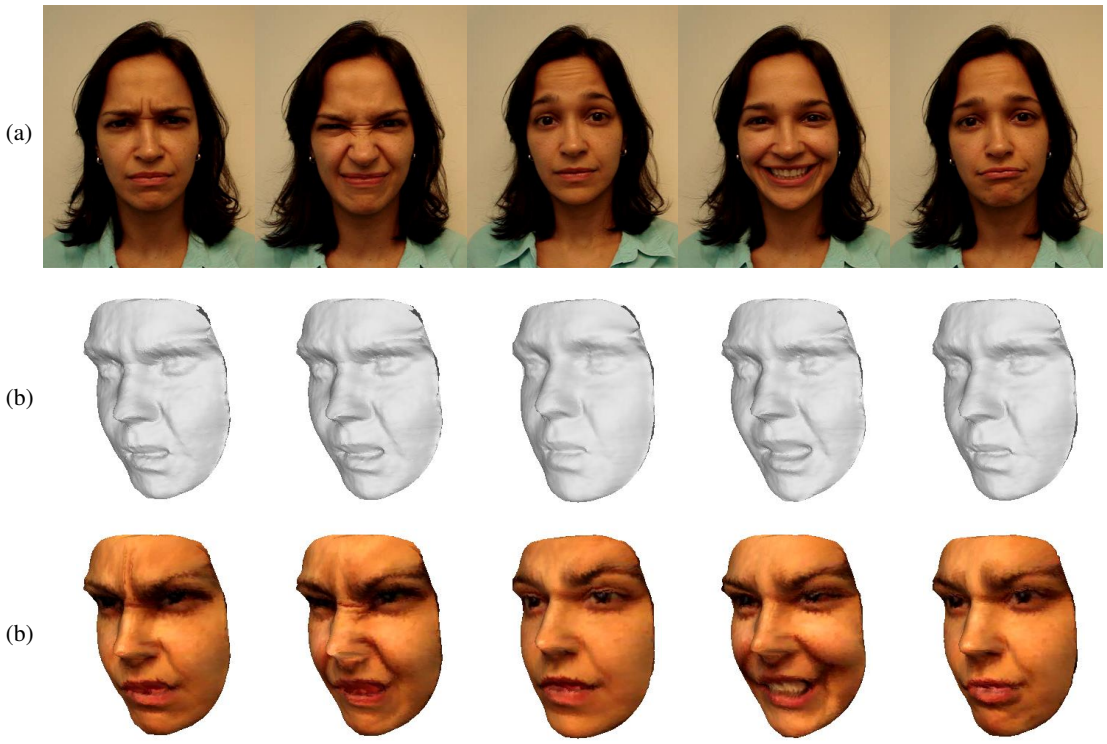


Figure 7. Results of 3D reconstruction of a video sequence using a standard webcam from a subject not present in the training database. (a) frontal faces with different expressions; (b) 3D reconstructed face geometry; (c) 3D geometry with the mapped texture.