SYNTHESIS OF 3D FACES

R. Enciso, J. Li, D.A. Fidaleo, T-Y Kim, J-Y Noh and U. Neumann

Integrated Media Systems Center

University of Southern California

Los Angeles, CA 90089, U.S.A.

Abstract

In this paper, we present a new technique for creating photorealistic textured 3D facial models from photographs. Starting with several pairs of calibrated views of a human subject, we first automatically recover a 3D dense reconstruction for each pair. We employ a user-assisted technique to align these 3D meshes with a generic face mesh. A scattered data interpolation technique is then used to deform the generic mesh to fit the particular geometry of the subject's face. Having recovered the geometry we automatically blend the different texture maps to obtain a final 3D face with a high degree of realism. The rich texture map is 512x1024 pixels and captures the complete head. Faces produced by our method are only 1700 polygons and suited to realistic animations of specific individuals in ***applications*** *like special effects, games, and 3D teleconferencing.*

Keywords: 3D modeling, stereo vision, 3D reconstruction, volume morphing

I. INTRODUCTION

Recent interest in facial modeling and animation is spurred by the increasing appearance of virtual characters in film and video and computer games. Another possible applications are 3D teleconferencing, 3D email or chatting over the Internet. Ever since the pioneering work of Parke [13], researches attempted to generate realistic facial models and animation. For an excellent survey the reader is referred to [14].

The complexity of human facial anatomy and our natural sensitivity to facial appearance increase the difficulty of modeling human facial appearance and subtle expressions. Although some recent work [3, 10, 15] produces realistic results with relatively fast performance, the process of generating a specific person's head model suitable for facial animation often entails extensive human intervention [15], physical markers on the face [10], or the need for a huge database of people's faces [3].

Our modeling approach is based on computer vision techniques in which different pairs of stereo images are used to create precise geometry. To obtain a complete model of the head, we acquire frontal and side views of the person's face, a scalp view and a back view, each producing separate 3D reconstructions of the visible head regions. Our interactive system allows the user to specify a set of correspondences between each stereo mesh and the prototype model vertices. These one-to-one correspondences are used to automatically align the stereo meshes with the prototype model and to deform the last one to interpolate the geometry of the new subject.

Fig. 1 shows a set of input images and the 3D face synthesized in a similar pose. Fig. 2 illustrates the generic model before and after the morphing. Our system enables the user to refine the initial fit as needed.

As our prototype mesh contains only 1700 polygons (for speed purposes) compared to about 20-40K triangles per reconstruction, therefore the fitting process is implicitly also a simplification or compression process. The prototype mesh contains the animation mechanism, (in our case we have tried two methods: volume morphing anchore vertices or muscle-based information per vertex [19]), so after the fitting is complete, we can animate the personalized complete head animation model with predictability.

The main advantage of our method is that we retrieve the complete head geometry compared to [10], [3] where only the frontal part is modeled. We achieve this without tuning any parameter or putting markers on the face [10]. The use of the stereo algorithm to capture the geometry avoids the use of very expensive devices like 3D scanners [3], [10] or 3D digitizers like Inspeck's©. Another advantage is that a typical video camera is able of extracting texture maps of higher resolution than scanners (Cyberware© scans typically produce a 512x256 grid).

The main differences between our approach and [15] is the use of different views of the person's head allowing us to recover the whole head's geometry including the back and the scalp, and the decrease in the usual number of correspondences needed to build a specific face model from 99 [15] or 182 [10] to less than 50. Note also that our solution is cheaper because we only use a pair of cameras instead of five as in [10].

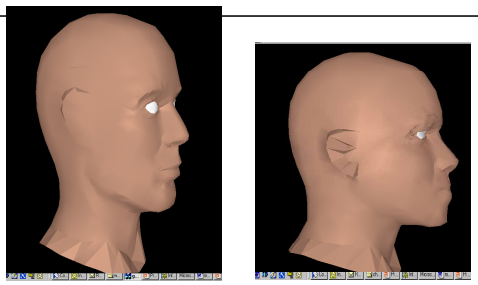
The remainder of this paper is organized as follows. Section II describes the stereo algorithm and section III describes the volume morphing and refinement method to make the final head. The section IV contains the conclusion of the paper and future research.

II. 3D MODELS FROM STEREO RECONSTRUCTION

Stereo reconstruction has been an active area of research in computer vision for more than two decades. The idea is to use two cameras looking at the scene from a slightly different point of view, and based in projective geometry to find the depth information. Fig. 3 shows a flow diagram of our method.

Calibration: Off-line calibration of the cameras is necessary in this approach. If calibration is not available, one can compute the epipolar geometry [21] and then obtain a Euclidean reconstruction assuming some of the internal parameters are known or constant [16] or introducing some human interaction to adjust the focal length [5]. As our goal is to obtain an accurate Euclidean reconstruction with as little manual input as possible we have chosen to calibrate the cameras based on [12]. The calibration step computes the perspective projection matrix relating a 3D point in space to its projection on the image plane. To compute this projection we need two types of information: a) the 3D coordinates of at least 11 points $\mathbf{M} = (x, y, z, 1)$ known in a world coordinate frame with high accuracy, and b) the 2D projections of these points detected on the image with subpixel accuracy: $\mathbf{m} = (u, v, 1)$. A 10^{th} of a pixel accuracy is currently used.

Fig. 1 –Top row: a set of input images. Bottom row: the 3D personalized head.



(a) **(b)**

Fig. 2 - Our approach morphs a prototype face model (a) to fit one or more stereo reconstructed meshes to produce a personalized complete 3D head (b).

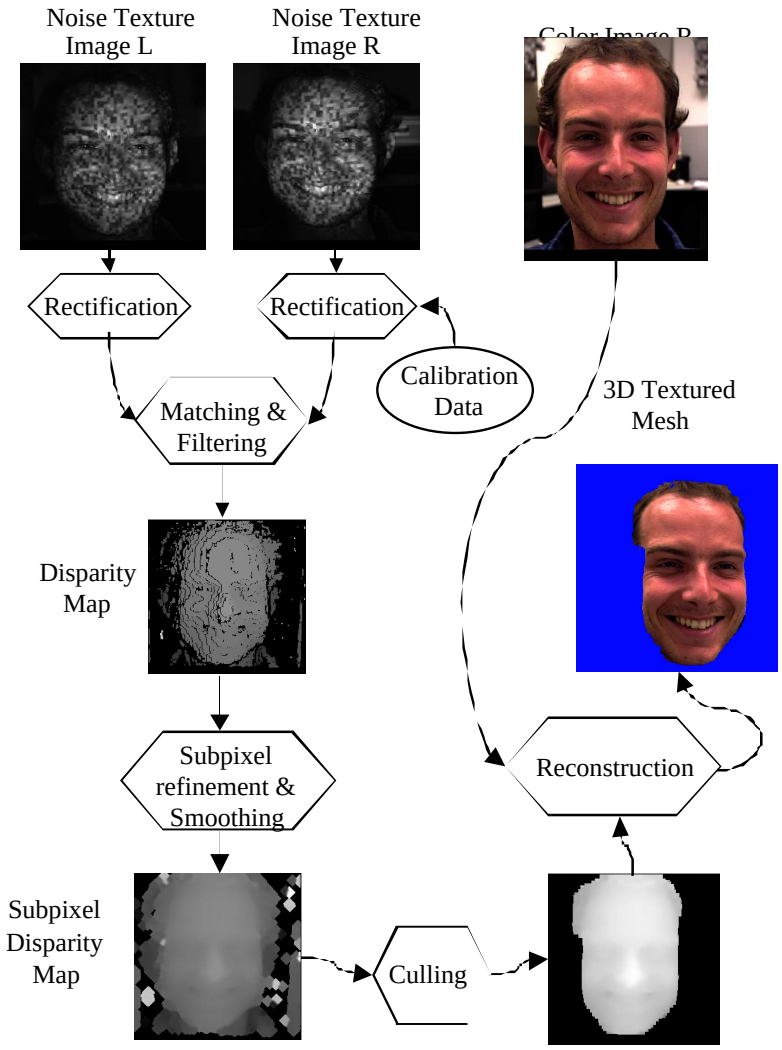


Fig 3 – Stereo Reconstruction Algorithm Flow Chart.

Rectification and matching: Once calibrated, the first step is to rectify the images to reduce the search of correspondences in the two images to a one-dimensional search [1]. The rectification step provides the 3x3 matrices $\mathbf{R}_i$ for each camera i. After rectification, for a fixed point in the left rectified image (u_r, v_r), its correspondence in the right image is in the same row, so it can be represented as (u_r, v_r+d). We call the distance d, the *integer disparity*.

To find correspondences we slide a correlation window along the corresponding row. Our choice is an area-based cross-correlation

method [7], i.e. the metric used is the gray-value similarity over a window of the image. The stereo algorithm we use is based on [17, 2]. To limit the searching space only disparities d in a predefined range are considered (this range depends on the camera's configuration and the distance to the subject). The hypotheses can finally be pruned by a selection procedure based on the visibility constraint [9], the ordering constraint [20], and the gradient limit constraint [4]. A subpixel correction of the integer disparity map can be obtained by a simple linear interpolation of the correlation scores, or a more sophisticated approach described in [6], which takes into account the perspective distortion between the left and right correlation windows for a planar region of surface. Even the more sophisticated stereo algorithm will produce a disparity map with holes, outliers and false matches. So, once the disparity map is computed, we first interpolate the subpixel disparity then filter outliers and smooth the resulting disparity map.

3D Reconstruction and mesh generation: From the subpixel disparity map, 3D points are computed by triangulation (this is a sub-product of the calibration). The 3D reconstruction finished, we tessellate to define a quadrangular facet's mesh (it decreases the number of polygons compared to a triangular one). With 256x256 image's size, the mesh's generated size is approximately 20-40K polygons. Fig. 5 shows some 3D stereo meshes.

Texture map generation: To ensure accurate geometry over the entire face and head, we momentarily project a salt and pepper noise texture onto the face with a slide projector when acquiring a stereo image pair (Noise Texture Images L and R in Fig. 3). A high-resolution color image for texturing the final model is acquired immediately after disabling the noise texture (Color Image R in Fig. 3).

To ensure alignment between the Noise Texture Images and the Color Image we have on-line control of the slide projector through an electronic switcher connected to the serial port of the computer. This equipment allows us to switch from one configuration to the other in less than a second. To segment the face from the background we do manually or we use a panel on the background with a distinctive color and disregard any region with this color.

For each 3D point **M** we know the projection on the rectified image (u_r, v_r, *1*) because we are calibrated and the texture mapping coordinates $(s\,u, s\,v, s) = \mathbf{R}^{-1}.\,(u_r, v_r, 1)^T$, where **R** is the rectification matrix and *s* is a scalar factor.

Execution times: In a 450MHz PC we produce a stereo mesh in two seconds for original image's size 512x512 pixels and final disparity map of 256x256.

III. MESH FITTING BY VOLUME MORPHING

Stereo models have variable complexity and mesh arrangements. Although these capture the geometry and coloring of a person, they are not suitable for direct animation. The prototype mesh, on the other hand, has the parameters suited for animation, but its geometry and texture does not match a specific individual. By fitting the prototype mesh to the reconstruction model(s), the animation mesh takes on the shape and coloring of a specific individual.

There are several steps in a fitting process. The first step is *a landmark-based volume morphing* where the transformation and deformation of the prototype mesh is guided by the interpolation of a set of landmark points with a radial basis function [18-8]. The landmark volume morphing only guarantees that the morphing of the prototype mesh is accurate near the landmark points, and since these are sparse and scattered, a further optimization is needed to ensure the overall quality of the morph.

The second step of *surface optimization* minimizes a cost function to further improve the overall similarity between the reconstructed model and the prototype mesh based on the Euclidean distance between vertices. To this effect, we define a cost function as follows:

$E = E_{dis} + E_{spring}$

where E_{dis} is the sum of the minimum distances from each point in the reconstructed model to the surface of the prototype mesh, and E_{spring} is a term that measures the energy of the prototype by placing on each edge of the mesh a spring of rest length zero with a fixed spring constant. The positions of all the vertices of the prototype are iteratively adjusted to minimize E. The inclusion of the E_{spring} term guarantees the existence of a minimum and regularizes the optimization to a desirable local minimum [11].

A third step is the *texture extraction*. Fig. 4 shows the spherical texture map coordinates used.

When multiple reconstructed models are available, a fourth step of *blending* merges the corresponding spherical texture maps into one aggregate final texture map. We consider the following rules: The texture inside the rectangle of Fig. 4 comes only from the frontal view, since it is the most important region. The texture outside the rectangle is extracted by a weighted average, where the weighting coefficient is taken as the cosine of the angle between the normal of the corresponding point in the generic model and the viewing direction from which the stereo images are taken.

Fig. 6 shows the results of the fitting procedure to three members of our research group.

IV. CONCLUSIONS

We introduced a method to create realistic head models with high resolution texture maps and realistic geometry. Our method modifies a prototype animation mesh to fit a specific person with minimal interaction. The manual interactions consist only of clicking a few correspondences on both the prototype mesh and each stereo model. In general, about 12 correspondences distributed among the eyes, nose, ears, and mouth are sufficient to fit the prototype mesh to a new stereo model set. These correspondences could be automatically initialized using the information provided by a vision face recognition/tracking system.

The problems inherent to the stereo system are the same as the scanner: semi-transparent objects like glasses and specially hair do not work very well. We are working on modeling and rendering of hair so it will be possible to model it separately.

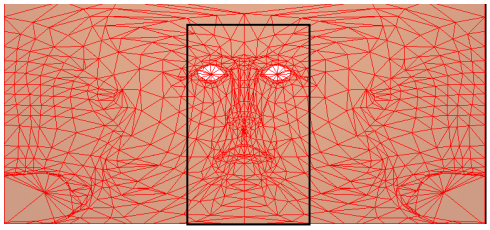


Fig. 4 - Texture mapping for the prototype mesh based on a spherical projection.

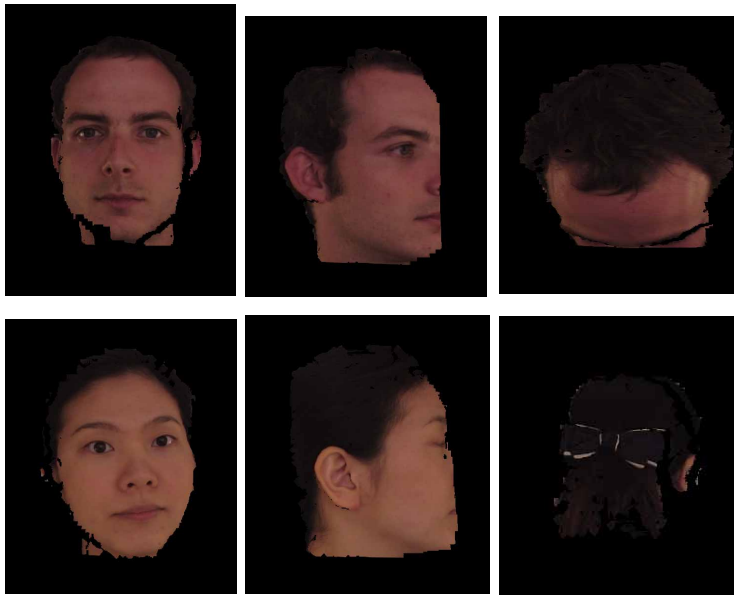


Fig. 5: Some 3D Stereo reconstructed views for three subjects: frontal, side, back or scalp.

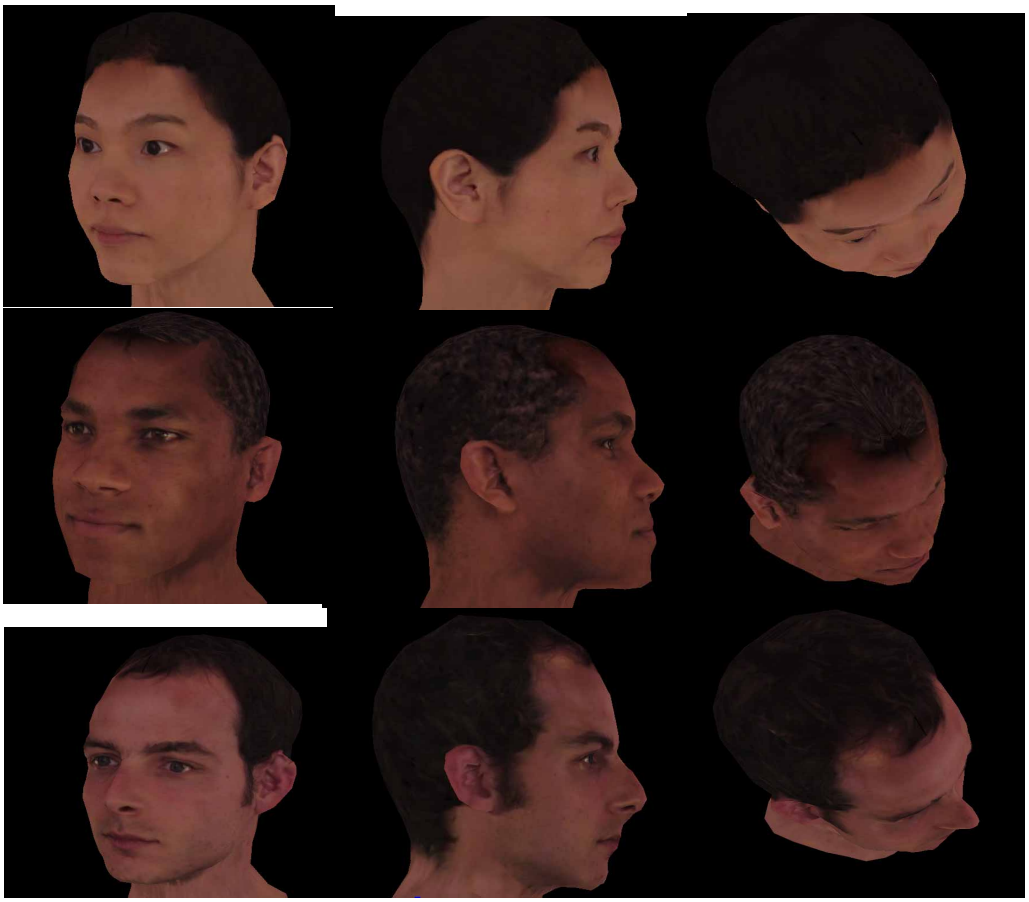


Fig. 6 – Different views of the personalized 3D models.

ACKNOWLEDGEMENTS

This work has been partially funded by *The Annenberg Center for Communication*, the *Integrated Media Systems Center* at the University of Southern California and *Advanced Network & Services, Inc.* as part of the National Tele-Immersion Initiative (NTII).

REFERENCES

[1] N. Ayache and C. Hansen. "Rectification of images for binocular and trinocular stereovision". In *Int'l Conf. on Pattern Recognition*, Oct. 1988.

[2] R. Bajcsy, R. Enciso, G. Kamberova, L. Nocera, and R. Sara. "3D Reconstruction of Environments for Virtual Collaboration". *In 4th IEEE Workshop on Applications of Computer Vision*, WACV'98, Princeton, USA, 1998.

[3] V. Blanz, T. Vetter. "A Morphable Model for the Synthesis of 3D Faces". *SIGGRAPH'99 Conf. Proc.*, Los Angeles, USA, 1999.

[4] P. Burt and B. Julesz. "A disparity gradient limit for binocular fusion". *Perception*, vol. 9, pp 671-682, 1980.

[5] Q. Chen and G. Medioni. "A Semi-Automatic System to Infer Complex 3-D Shapes from Photographs". *IEEE Multimedia Systems*'99, Florence, Italy.

[6] F. Devernay and O. Faugeras. "Computing Differential Properties of 3D Shapes from Stereoscopic Images Without 3D Models". In Proc. *of Int'l Conf. on Computer Vision and Pattern Recognition*, pp 208-213, Seattle, WA, June 1994.

[7] U. R. Dhond and J. Aggarwal, "Structure from Stereo - A Review". IEEE *Transactions on Systems, Man, and Cybernetics*, 19(6), pp 1489-1510, 1989.

[8] S. Fang, R. Raghavan, J. T. Richtsmeier. "Volume Morphing Methods for Landmark Based 3D Image Deformation". *SPIE Int'l. Symposium on Medical Imaging*, 1996.

[9] O. Faugeras. "Three-dimensional Computer Vision: A Geometric Viewpoint". *MIT Press*, 1993.

[10] B. Guenter, C. Grimm, D. Wood, H, Malvar, F. Pighin. "Making Faces". *SIGGRAPH'98 Conf. Proc.*, pp. 55-66.

[11] H. Hoppe, T. DeRose, T. Duchamp, J. McDonald, W. Stuetzle, "Mesh Optimization", *Computer Graphics Proceedings*, 1993.

[12] R.K. Lenz and RY. Tsai. "Techniques for Calibration of the Scale Factor and Image Center for High Accuracy 3D Machine Vision Metrology". *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 10:713-720, 1988.

[13] F. I. Parke. "Computer Generated Animation of Faces". Proc. *ACM Annual Conference*, August 1972

[14] F.I. Parke and K. Waters. “Computer Facial Animation”. A K Peters, Wellesley, Massachusetts, 1996.

[15] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, D. H. Salesin. “Synthesizing Realistic Facial Expressions from Photographs”. *SIGGRAPH Conf. Proc*., 1998, pp. 75-84.

[16] M. Pollefeys, R. Koch and L.V. Gool. “Self-calibration and Metric Reconstruction in spite of Varying and Unknown Internal Camera Parameters”. In *Proc. Int’l Conf. Computer Vision* 98, pp 90-95, 1998.

[17] R. Sara, R. Bajcsy, G. Kamberova and R.A. McKendall. “3-D Data Acquisition and Interpretation for Virtual Reality and Telepresence”..In Proc. *IEEE Workshop Computer Vision for Virtual Reality Based Human Communications* (ICCV’98), pp 88-93, India, January 1998.

[18] F. Ulgen. “A Step Toward Universal Facial Animation via Volume Morphing”. In 6th IEEE *Int’l Workshop on Robot and Human Communication*, 1997, pp. 358-363.

[19] K. Waters. “A muscle model for animating three-dimensional facial expression”. In *SIGGRAPH 87 Conf. Proc*., vol. 21 pp. 17-24, July 1987.

[20] A. Yuille and T. Poggio. “A Generalized Ordering Constraint for Stereo Correspondence”. *Technical Report 777*, MIT, 1984.

[21] Z. Zhang, R. Deriche, O. Faugeras, Q.-T. Luong, “A Robust Technique for Matching Two Uncalibrated Images Through the Recovery of the Unknown Epipolar Geometry”, *Artificial Intelligence Journal*, Vol.78, pp 87-119, Oct. 1995.