3D FACE RECONSTRUCTION FROM STEREO: A MODEL BASED APPROACH

Ying Zheng, Jianglong Chang, Zhigang Zheng, Zengfu Wang

Department of Automation, University of Science and Technology of China
Contact: zfwang@ustc.edu.cn

ABSTRACT

Since the human faces are lowly textured, the conventional stereo methods based on intensity correlation can not give satisfying 3D face reconstruction results. In this paper, a model based stereo matching method is proposed. A reference 3D face is used as an intermedium for correspondence calculation. The virtual face images with known correspondences are first synthesized from the reference face. Then the known correspondences are extended to the incoming stereo face images, using face alignment and warping. The complete 3D face can thus be reconstructed from stereo images reliably.

Index Terms— Stereo Vision, 3D Face Reconstruction, Face Alignment

1. INTRODUCTION

3D face reconstruction has become a thriving research field with various applications for the past decade. Although some commercial laser scanners have been employed to directly capture 3D face data, the high expense of these equipments makes them difficult to be popularized. As a result, some researchers resort to active vision systems such as the structured light based ones [1, 2]. The drawbacks of these systems are obvious. The data in improperly reflected regions may be lost, and the dazzle caused by the projector illumination may displease the testees.

Stereo is a widely used passive vision technique which recovers the surface depth from geometric relation over calibrated stereo pairs. In stereo systems, the calculation of the disparity (or correspondence) between rectified stereo pairs is a crucial step. Once this step is done reliably and precisely, the depth information can be conveniently reconstructed. But the computation of disparity is not trivial, especially for face images. The human faces are lowly textured because of the smooth albedo across the facial skin. Consequently, the conventional stereo matching methods based on intensity correlation could probably fail since there would be significant ambiguity in the correspondence results. Moreover, the performances of these methods will deteriorate when several factors present, such as bad lighting or occlusions. So far, several attempts have been made to deal with the 3D face reconstruction from stereo images. Devernay and Faugeras [3] extended the classical correlation method to estimate both the disparity and the differential properties of the shape directly from image data. Lengagne *et al*. [4, 5] proposed to incorporate priori information in 3D stereo reconstruction process, including the differential information about the object shape and the geometric constrains such as curvature values and crest lines. Their reconstruction process is computationally expensive since each vertex has 6 parameters. Another kind of knowledge intergraded in stereo is shape from shading (SFS). SFS works well on textureless objects, which is considered to be complementary with stereo. Cryer *et al*. [6] integrated the high frequency information from the SFS and the low frequency information from stereo. In [7], Pua *et al*. used an objective function which is a weighted sum of stereo, SFS and smoothness, while in [8] they used the stereo to initialize information about the illumination, surface reflectance and shape, and then refined the results by using SFS.

From the related work mentioned above, we notice that most of the efforts were put on the post-processing steps while few were on stereo matching itself. The initial results derived from traditional stereo methods may be very far from the true surface. As a result, the refinement procedure is indispensable. However it is not only time consuming but also computationally very expensive, and can not obtain satisfying results in most cases. Actually, for a specific class of objects, such as faces, there are always some similar characteristics or priori knowledge between different individuals. We can use such similarity to get better results. Instead of reasonless post-processing, we propose a model based approach which is capable of recovering qualified 3D faces directly from stereo images. Our objective is to design a stereo method which is appropriate for a class of objects with similar characteristics, such as the human faces. There are two basic ideas lying in our method. One is that the pose of arbitrary face can be approximately estimated by fitting a general 3D face to it. The other one is that the alignment between images of different persons at the same pose is much easier than that between images of the same person at different poses.

The diagram depicting our method is shown in Figure 1. In this system, the poses of the incoming person in stereo im-

This work is supported by the National Natural Science Foundation of China under Grant No.60455001, and the Graduate Innovation Fund of University of Science and Technology of China under Grant No.KD2006042.

ages are first estimated. Then the virtual face images of a general 3D face (called the *reference*) at the same poses are synthesized. The correspondences between the two virtual face images are known precisely since they are both synthesized from the reference. The correspondences derived from the reference are extended to the incoming person, by face alignment and image warping. Once the correspondences are established, the disparity map can be calculated and the 3D model of target face can be generated from an organized cloud of recovered 3D points.

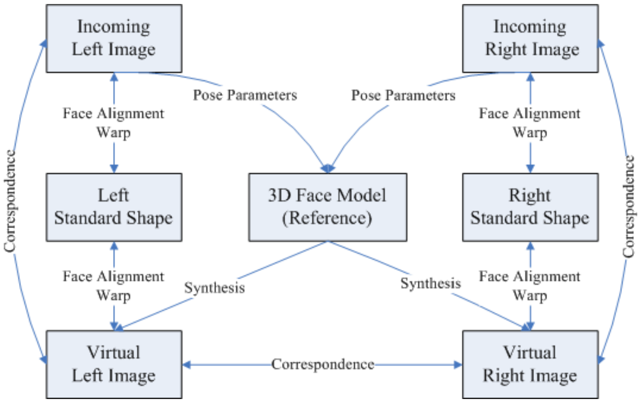

Fig. 1. The diagram of proposed method.

2. FITTING THE REFERENCE TO STEREO IMAGES

Our stereo vision system is composed of two cameras which are placed horizontally and calibrated carefully so that their intrinsic and extrinsic parameters are precisely known. The captured images are first rectified to make vertical disparities be zero. Formally stated, we denote the projected points of a scene point (x, y, z) as (u, v) and (u', v) respectively, then the disparity is defined as the difference $d = u' - u$ in rectified images. All of the successive discussions are based on the rectified images.

Given stereo images of an incoming person, the computation of the correspondences is not carried out directly between his stereo images. Alternatively, we first fit the reference 3D model to the face stereo images and synthesize virtual face images at the same poses. The correspondences between the virtual face images are ease to calculate. The process is described below.

To simplify the problem, we set the origin of the reference to the center of it. Then if we employ orthographic projection, a vertex $\mathrm{x} = (x, y, z)^T$ on the reference is mapped into a point $p = (p_x, p_y)^T$ on image plane according to the following expression:

$$p = f \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} R_\gamma R_\theta R_\phi \mathrm{x} + t_{2d} \qquad (1)$$

where $R = R_\gamma R_\theta R_\phi$ is the rotation matrix, γ, ϕ, θ are the corresponding rotation angles around the X, Y, Z axes, which control the pose of rendered face. $t_{2d} = (P_x, P_y)^T$ is the 2D translation vector, and f is the re-scaling factor.

The above six parameters $\rho = (\gamma, \phi, \theta, f, P_x, P_y)^T$ are used to control the projection from the 3D model to the image plane. The estimation of these parameters is obtained by minimizing the following cost function:

$$E = \sum_j \parallel q_j - p_j \parallel^2 \qquad (2)$$

where $q_j = (q_{x,j}, q_{y,j})^T$ are the feature points located on the inputted left or right images, and $p_j = (p_{x,j}, p_{y,j})^T$ are the projections onto the image plane from the corresponding vertices extracted from the 3D face model. It is obvious that the cost function will achieve minimum if the face poses in synthesized image and inputted images are similar. The Levenberg Marquardt (LM) algorithm [9] is applied for the nonlinear optimization. At least more than 6 points are requested to obtain the over-constrained solution of the problem. Here we choose 12 feature points which are easy to locate without ambiguity. Figure 2 shows a fitting example. The estimated pose, although not exactly the same as the incoming face, is usually very close to the true one. It should be pointed out that the estimated pose may be quite different from the true one in some cases. In that situation, our system permits users to correct the pose manually in an interactive manner.

The fitting is processed on left and right images respectively. It should be noticed that the estimated parameters f, P_y, γ, θ of left and right images which relate to the calculation of y coordinates should be very close since the images are rectified. This is supported by the quantitive fitting result shown in Table 1. To be more precise, we set the four parameters for left and right images to be the same.

Table 1. Estimated parameters.

	f	P_x	P_y	γ	ϕ	θ
left	**1.74**	528.5	**273.9**	**-0.107**	0.215	**0.0010**
right	**1.72**	194.7	**274.6**	**-0.091**	-0.151	**0.0013**

Once the parameters controlling the projection are recovered, the virtual "stereo" face images of the reference are synthesized. The computation of correspondences between the virtual images is trivial: the pixels on left and right virtual images both projected from the same vertex on the reference should correspond to each other.

3. EXTENSION OF CORRESPONDENCES TO STEREO IMAGES

Since the reference 3D face model used is not the true 3D model of the incoming person, the direct correspondences

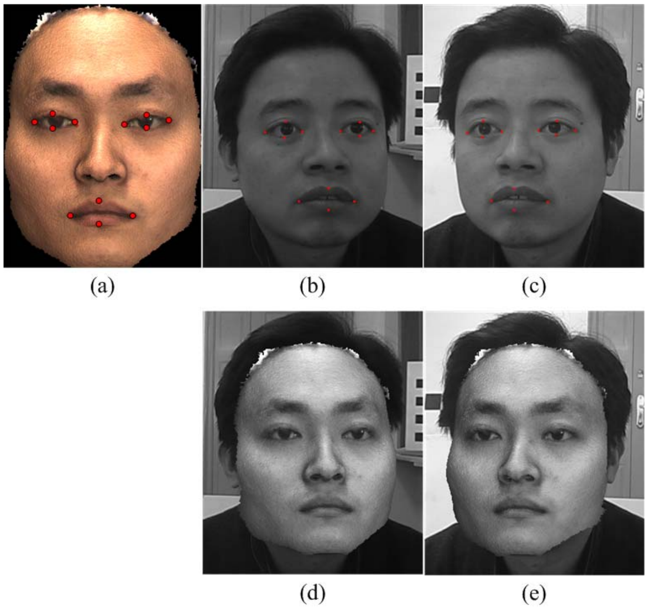

Fig. 2. Pose estimation. (a) The reference 3D face model with 12 feature points selected (red points). (b)-(c) Stereo images of an incoming person. The corresponding feature points are labeled on the images (red points). (d)-(e) The virtual face images synthesized from the reference redrawn into image planes of (b) and (c), using the estimated parameters.

between the reference and the inputted stereo images cannot help us obtain correct disparities of the person. To establish the correspondences between the inputted stereo images, some registration measures and nonlinear deformations should be employed. In this paper, face alignment and warping are used.

Face alignment is an important pre-precessing method in face recognition. In [10], the shapes of different people are aligned and the face images are warped to shape-free images to perform face recognition. Each pixel in the warped image could be considered have the same definition on the face (at the similar position on the original face). We utilize this scheme. We locate several feature points on the stereo images of the incoming person. The vertices corresponding to these features are labeled on reference previously so that their positions on the virtual images are known. The virtual and real stereo images of the same view (left or right) are warped to the shape-free images, so that the feature points on the original images are moved to overlap a standard shape. Generally, the standard shape is defined as the mean of the shapes in virtual and real images.

Thereafter, the correspondences of the stereo images are established according to the correspondences of the virtual images, by the following steps:

1. The pixels, which are on the left (right) images of the reference and the incoming person, and both are warped to the same point in the standard shape, are considered to be a matching pair. This builds the correspondences between virtual and inputted images of the same view.

2. The pixels, which are on the left and right images of the incoming person, and respectively correspond to two pixels that are matched on virtual images, are considered to correspond to each other. This builds the correspondences between inputted stereo images.

3. If a pixel corresponds to more than one pixels after applying step 2, the epipolar constraint (y coordinates should be the same) and the threshold of acceptable disparity are imposed.

It should be noticed that our method can only achieve pixel-to-pixel (not sub-pixel precision) correspondences right now. This may cause a problem that the reconstructed surface may not be smooth enough. If we extract a scan line across the reconstructed surface, we can see the roughness. We employ a simple one-dimension smoothing to solve the problem. Each scan line is extracted and smoothed using the moving average method (Figure 3). The smoothing is processed horizontally and then vertically which will bring satisfying face surface.

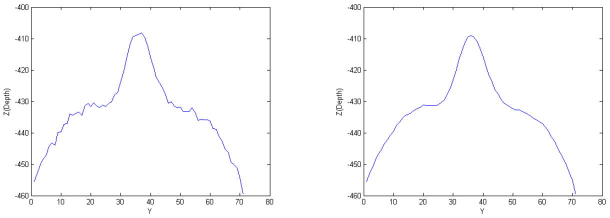

Fig. 3. Surface smoothing. Left: A horizontal scan line across the reconstructed surface which is rough. Right: Smoothing result using moving average method.

4. RESULTS

We test our method with a data set of 10 people. Figure 4 shows some qualitative results of reconstructed surfaces. Totally 68 feature points are located automatically on the faces using ASM [11] and then are warped to a standard shape. The number of the vertices in reconstructed face can be controlled by the size of warped images. We see in the figure that the stereo reconstructed 3D faces reliably reflect the characteristics of the incoming faces. The whole reconstruction process is fast. Much of the time is consumed in the face alignment while the process of stereo matching and reconstruction is fast within 4s on a P4 2.6GHz PC. The cost of post-processing is trivial since it is essentially a one-dimension smoothing.

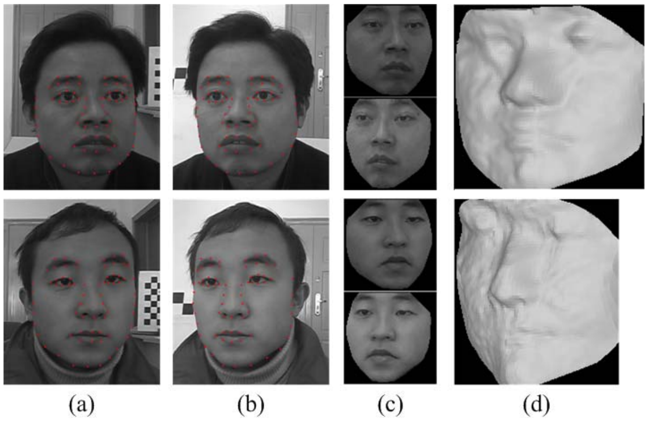

Fig. 4. Stereo reconstruction results. (a) The left images labeled with 68 feature points. (b) The right images with the same feature point definition. (c) The warped shape-free images of the left and right views. (d) The reconstructed surfaces from stereo.

5. CONCLUSION AND FUTURE WORK

In this paper, a stereo 3D reconstruction algorithm adapting to face images is proposed. This method does not follow the traditional stereo methods based on intensity correlation. It is able to recover satisfying face models directly from stereo pairs without intricacy post-processing. Our method is to use a reference 3D face model as an intermedium to facilitate the computation of stereo correspondence. The correspondences are first established between the virtual face images, which are synthesized from the reference 3D model at the same poses of inputted stereo images. Then the virtual correspondences are extended to stereo images using face alignment and image warping.

Our method gives good reconstruction results without any singularity point in a low computation cost. The method is automatic, stable and fast. Since the calculation is not based on intensity, the method will work well even bad lighting or occlusions are present (in these situations, the facial feature points can still be located by hand).

This work is about to be improved in the following ways. The method can be implemented in a multi-resolution manner which can bring sub-pixel correspondence precision. Although the robustness of our method against bad lighting and occlusions is predictable, it should be evaluated by experiments. More sophisticated post-processing methods and texture mapping should be employed to get realistic face models qualified for applications in virtual reality and computer animation.

6. REFERENCES

[1] B. Achermann, X. Jiang, and H. Bunke, "Face recognition using range images," in *International Conference on Virtual Systems and MultiMedia*, 1997, pp. 129–136.

[2] C. Beumier and M. Acheroy, "Automatic 3d face authentication," *Image and Vision Computing*, vol. 18, pp. 315–321, 2000.

[3] F. Devernay and O. Faugeras, "Computing differential properties of 3-d shapes from stereoscopic images without 3-d models," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, 1994, pp. 208–23.

[4] R. Lengagne, J. P. Tarel, and O. Monga, "From 2d images to 3d face geometry," in *Second IEEE International Conference on Automatic Face and Gesture Recognition*, Oct. 1996, pp. 301–306.

[5] R. Lengagne, P. Fua, and O. Monga, "3d face modeling from stereo and differential constraints," in *Third IEEE International Conference on Automatic Face and Gesture Recognition*, Nara, Japan, April 1998, pp. 148–153.

[6] J. E. Cryer, P. S. Tsai, and M. Shah, "Integration of shape from shading and stereo," *Pattern Recognition*, vol. 28, no. 7, pp. 1033–1043, 1995.

[7] P. Fua and Y. G. Leclerc, "Object-centered surface reconstruction: Combining multi-image stereo and shading," *International Journal of Computer Vision*, vol. 16, no. 1, pp. 35–56, 9 1995.

[8] D. Samaras, D. Metaxas, P. Fua, and Y.G. Leclerc, "Variable albedo surface reconstruction from stereo and shape from shading," in *IEEE Conference on Computer Vision and Pattern Recognition*, Hilton Head Island, SC, USA, 6 2000, pp. 480–487.

[9] W Press, S Teukolsky, W Vetterling, and B Flannery, *Numerical Recipes in C++: The Art of Scientific Computing*, Cambridge: Cambridge University Press, second edition, 2002.

[10] A Lanitis, CJ Taylor, and TF Cootes, "Automatic face identification system using flexible appearance models," *Image and Vision Computing*, vol. 13, no. 5, pp. 393–401, 1995.

[11] T.F. Cootes, C.J. Taylor, D.H. Cooper, and J. Graham, "Active shape models: Their training and application.," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.