Automatic Animated Face Modeling Using Multiview Video

Fu-Che Wu, Murphy Chien-Chang Ho, Ming Ouhyoung

Department of Computer Science and Information Engineering, National Taiwan University

Abstract

An image-based 3-D modeling system is presented in this paper. Our modeling system consists of three main stages: camera calibration, depth estimation and 3-D geometry reconstruction. All of these steps are executed automatically. In the camera calibration stage, some patterns are used that help to determine the camera's position in an environment. The camera's intrinsic and external parameters are determined using epipolar geometry. After the camera parameters are determined, the camera's location in each projected frame is determined. The depth for each pixel in a base image is estimated from the camera's focus to the object's surface by measuring the similarity between the base image and the neighboring images. The object's 3-D geometry is reconstructed with texture from the base image using the depth information.

1. Introduction

In this paper, we will present an automatic procedure to construct the 3-D structure of an object from an image sequence. Our discussion will focus on head modeling. However, this methodology is not limited to head modeling and can be applied to other objects. Head modeling is such an interesting topic because it is a very complex structure. Head modeling has received much attention in recent years. Two research clusters are involved in generating a 3-D head model (in wireframe plus texture) from multi-angle images. The first approach involves using structured light or a regular pattern to project onto the target object, a head, to solve the problem of feature point correspondence. The major advantage of this approach is that it facilitates very precise modeling. However, because a head model consists of hundreds to thousands of points, and since we don't know which points belong to the lips or eyes, the model thus obtained can not be used in facial animation unless vertices are tagged manually to provide required semantics. This method requires some special peripherals and also has some constraints on capturing a head model. For example, the environmental lighting can not be too bright. The eyes must be closed. It can not manage a black area. It is also difficult for this approach to capture a deforming process.

The second approach involves image analysis. In this approach, a generic model is used. A small number of feature points are designated on this model, such as the corner points of the eyes, lips, or nose, to help construct the 3-D wireframe model. Since these feature points are supplied manually, the model generated can be readily used in facial animation. However, this approach lacks the precision of the first approach, since fewer points are used in constructing the model. In addition, the captured head image must be similar to the generic model. If the head looks very different, for instance, the hairstyle, with eyeglasses or a difference in age, all of these changes will make the result less accurate.

Imagine that you are standing in front of a camera and shaking your head while a computer constructs a 3-D model of your head. If you smile, the virtual head smiles. This is the final vision for this project. To construct the 3-D structure of a head model from an image sequence, the camera parameters in each frame must be determined. The location of each point in each frame in the image model can be reconstructed in Euclidean space.

The outline of this project is as follows. We will briefly describe the related fundamental theory. Works by other researchers will be discussed with some pitfalls in this topic that we have faced. Following this, we will introduce our work and results.

2. Fundamental Theory

2.1. Camera model

To estimate the inverse structure from the projected image, a camera model must be determined[1]. The coordinates of 3-D point $M = [X,Y,Z]^T$ in the world coordinate system and its image coordinates $m = [u, v]^T$ are related by

$$s \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = P \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix},$$

where s is an arbitrary scale, and P is a 3x4 matrix.

This is projective geometry. A full perspective model is rather similar to a real camera model. A line in Euclidean space will become a point in projection space, and a plane will become a line. In the homogeneous transformation form, this kind of transformation can be represented using a 4x3 matrix P. The matrix P can be decomposed into $P = A[Rt]$ where A is a 3x3 matrix, mapping the normalized image coordinates to the retinal image coordinates. $[Rt]$ is the 3-D displacement from the world coordinate system to the camera coordinate system. A is the intrinsic parameters and $[Rt]$ is the external parameters. In general, A can be written as

$$A = \begin{bmatrix} \alpha & \gamma & u \\ 0 & \beta & v \\ 0 & 0 & 1 \end{bmatrix},$$

Where α, β are the scale factors in the image u and v axes. γ describes the skewness of the two image axes. (u, v) are the coordinates of the principal point.

2.2. Epipolar geometry

The case of the two cameras is shown in fig. 1.

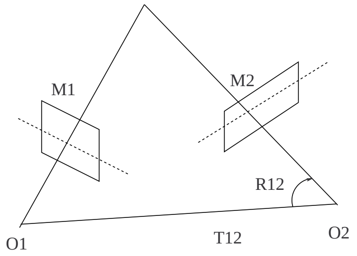

Figure 1: *Epipolar geometry is co-planarity constraint.*

The relationship between the two cameras is T_{12} and R_{12}. First, we observe in Euclidean space.

$$\begin{aligned} M_1 &= R_{12} \times M_2 + T_{12} && (1) \\ T_{12} \times M_1 &= T_{12} \times R_{12} \cdot M_2 && (2) \\ 0 &= M_1 \cdot T_{12} \times R_{12} \cdot M_2 && (3) \end{aligned}$$

$T_{12} \times R_{12}$ can be presented as $E = \lfloor T_{12} \rfloor_\times \cdot R_{12}$ where E is called the *essential matrix*. $\lfloor T_{12} \rfloor_\times$ is a mapping from a 3-D vector into a 3x3 matrix. Using this mapping, we can express the cross product of the two vectors using the matrix multiplication of a 3x3 matrix and a column matrix. Considering the projection plane, let $sm_1 = A_1 [I0] M_1$, $sm_2 = A_2 [I0] M_2$, then,

$$m_2 A_2^{-T} [T_{12}] \times R_{12} A_1^{-1} m_1 = 0$$

Define the *fundamental matrix F* to be

$$F = A_2^{-T} [T_{12}] \times R_{12} A_1^{-1}, then$$

$$m_2 F m_1 = 0$$

From previous observations, given a point m_1 in the first image, its corresponding point in the second image is constrained on a line called the m_1 epipolar line. Geometrically, Fm_1 defines the epipolar line of point m_1 in the second image.

3. Previous Work

Many research groups have focused on head modeling and developed many methods to construct a realistic model of the human head. The data input methods include the Cyberware laser scanner, active light striper, stereo images and video sequencing. In this paper, we will focus on the image-based approaches.

Most of the researchers in image-based approaches have used generic head models as predefine models. This generic model is then adjusted to approximate a new model that corresponds to the image sequence. Certain manual interactions are required to specify certain feature points such as the corners of the eyes , nose top, or mouth to fit the generic model.

Pighin et al.[2] developed a system that employs a user-assisted technique to recover the camera poses corresponding to the views as well as the 3-D coordinates for a sparse set of locations on the subject's face. This method can produce very realistic facial expressions. However, numerous manual adjustments are required in this method.

Blanz and Vetter[3] built a morphable face model by exploiting large 3-D face scan statistics and recovered domain knowledge about the facial variations by applying a pattern classification method. Thus, a 3-D face can be generated automatically from one or more photographs or modeled directly through an intuitive user interface.

Zhang[4] presented a procedure that can automatically match most facial feature points and rapidly produce a head model . Image matching is a heuristic approach that can involve many errors. In this method, the false matched points are filtered out and the correctly matched points are reconstructed in 3-D space. The reconstructed 3-D points are then adapted into a face model.

Lee and Thalmann[5] presented a method to reconstruct a head model from two orthogonal pictures. They provided a semi-automatic feature point extraction method with a user interface for interactive correction if required. They acquired (x,y) from the front view and (y,z) from a side view. A generic model was then deformed with the detected feature points. Their deformation approach was based on the Dirichlet Free-Form Deformations(DFFD) used to get new geometrical coordinates for the generic head modification . In this approach, because the feature points are given by manual interaction, the reconstructed 3-D face can be animated immediately with the given expression parameters.

Fua's[6] work fit a complex model into uncalibrated image sequences. Initially, they manually supplied the approximate 2-D locations for five feature points in one reference image. Their system then automatically determined the position and orientation of a central camera that brings the key-point projections as close as possible to those positions. The generic model contains many bundle-adjustment triangulations. This initialization guarantees that the bundle-adjustment triangulation vertice projections roughly fall onto the face. A least squares adjustment of this control mesh is then performed so that the model projection matches the corresponding point relationship in the image sequence. In this approach, the model details are more consistent with actual head model than Lee and Thalmann's work.

Wu's[7] work falls between the previous two approaches. Thirty-five feature points are manually chosen on a face from at least two frames. The 3-D positions for the coarse model control points are then chosen. A generic model is then deformed to match this coarse model.

TCT international (TCTi) is a company[8]. The company has develop a 3-D surface capture technology that can acquiring the human shape, texture and surface color to digitally display a person in three dimensions. They employ projecting a random light pattern on the subject and capturing him/her with precisely synchronized digital cameras set at various angles. By filtering different wavelengths of light, 3-D surface geometry and surface texture are acquired simultaneously, resulting in a very accurate texture map when applied to the surface data.

In these previous researches, a generic model is deformed to match the corresponding points in the image sequence. Our approach is similar to Fua's work, but we ignore the generic head model. The camera's intrinsic and external parameters are calculated first. Next the depth from the object surface to the camera focus in each pixel is estimated onto a base image. Using this result, the 3-D point positions in the facial image sequence are estimated. The major idea in processing this problem is similar to the other researches. In our previous experiments, we have faced some problems. We will briefly introduce these problems in the following.

4. Preliminary Testing

The major problem in this area is how to determine the corresponding points from different frame viewpoints. The corresponding point in each pixel is always located on the epipolar line of another frame. As in Fua's work[9], this constraint can be used to search for the corresponding point. With this approach, corresponding point pairs can be used to construct a head model. Using two stereo images to recover 3-D information has some basic problems. It is easy to find correct matching points if the two views are close. However, estimating a precise depth is a limitation in this situation. A large baseline is more robust for estimating depth. However, It is hard to compare when the appearance in the two images is different.

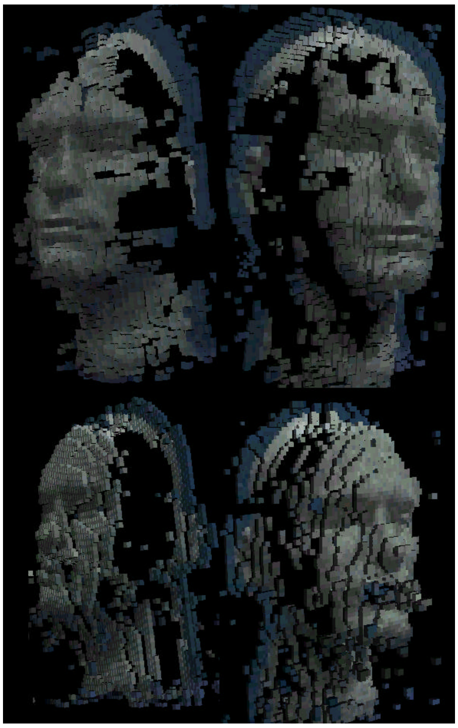

Figure 2: *Our previous results were constructed using different baseline lengths. The above baseline length is large. Note the holes after the initial processing.*

To solve this situation, we used multi-frame data to involve more information into the feature point transformation process. Feature points are tracked on each frame using an image sequence. To estimate the 3-D location from a set of tracked points, estimated location is projected onto each frame. A nonlinear minimization method is then used to minimize the errors on each frame to determine the sum of the distance between the projected point and the tracked point. This method did not produce a good result. The major problem was that the tracking mechanism did not work.

Some false matched points affected the precision in estimating the 3-D position. More information could not prevent bad results and also generated turbulence.

Tracking feature points on an image sequence is difficult. There are two clusters of approaches that can solve this problem. One way is to use the template matching technique to track feature points frame by frame. Another way is based on an optical flow technique to estimate the variations on each frame. Both methods have their limitations. In addition to the aperture and occlusion problems, the results may suffer from accumulated errors in the frame tracking process. The later approach is based on the assumption that the intensity is conserved. That is not the case in actual situations. The optical flow approach is well known as an ill-pose problem. In general cases, optical flow can only work under certain constraints. This approach is suitable for estimating global information, for instance, the motion of the camera or object. DeCarlo et al.[10] and Essa et al.[11] used this approach to estimate facial expressions. To estimate the precise 3-D structure of an object, the precise relationship among the feature points is required. Under general conditions, the optical flow technique is used to estimate the possible velocity at some position in a frame. This technique is not adaptable for acquiring the specific location correlated with a given point.

After several experiments, we found that the major problem in recovering the object's 3-D location from an image sequence is to find a robust corresponding matching method. We attempted to filter out these false matching points. We collected the silhouette and epipolar constraints from each frame to detect false matching points. Afterwards we still got a model that contained many holes. We were not able to solve this problem.

The silhouette constraint and the epipolar constraint are very strong limitations in solving the 3-D structure recovery problem. If the camera's intrinsic and external parameters have been determined, the silhouette constraint can construct a model such as Matusik et al.[12]'s visual hull. If the base frame depth from the objects' surface to the camera's focus can be estimated successfully, it is enough to reconstruct the 3-D surface from this viewpoint. Okutomi and Kanade[13] developed a stereo matching method that estimates the disparity from a multiple-baseline. Their approach inspired us to estimate the depth by comparing correlation sum from multi-angle images. We will briefly introduce our work in the next section.

5. Our work

There are three main steps in the current 3-D modeling process: camera calibration, depth estimation and 3-D geometric reconstruction. All of these steps can be executed automatically. In the first step, a special pattern is used to calibrate the camera's intrinsic and external parameters. The object's silhouette and a front view are chosen as the base image. From the silhouette information a visual hull is constructed. This visual hull is used to initialize the depth from the object's surface to the camera's focus in each pixel of the base image. Similar to the Marching Cube approach, each voxel's correlation sum is calculated from multi-angle images. The maximum correlation location is chosen as the best candidate for object's surface. After determining the depth of each pixel, a median filter is applied to smooth this estimation and produce its' 3-D position. The base mesh is then connected to produce the texture from the base image. The following is a detailed description of this process.

5.1. Camera calibration

This step is very important. It is the key point for success in the following stages. In order to certify that the camera's parameters can be estimated correctly, an 8x8 block pattern is placed onto the background. The corners of this block are easier to detect. In order to avoid a mismatch situation, each frame must contain the entire pattern. The corners of this pattern can be checked to separate data from noise.

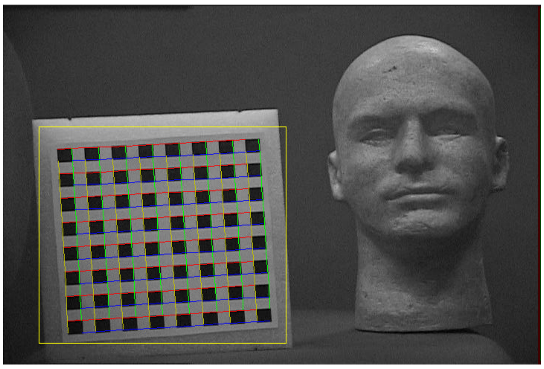

Figure 3: *The camera calibration pattern*

Four types of kernel filters are used to detect out four types of corners. In order to evaluate if a pixel whether belongs to a given corner or not, the correlation function is used to measure their similarities. The threshold that discriminates between pixels is dynamically updated to exactly 256 dots. Each row and column of dots is kept in line. If the filtering result for a frame passes these conditions, the examined corner corresponds to the other frame's corner. After some iterations,if the result still does not meet the requirement, the frame is discarded. The epiploar constraint is thus used to get the camera's intrinsic and external parameters.

We assume that the camera's intrinsic parameters are fixed. A model plane with m points was used to test this system. The maximum likelihood estimate can be obtained by minimizing the following function[14]:

$$\sum_{i-1}^{n}\sum_{j-1}^{m}\left\|m_{ij}-\hat{m}(A,R_i,t_i,M_i)\right\|^2$$

5.2. depth estimation

The depth of one pixel in the base image is the distance between the object's surface and the camera's focus. Initially, a silhouette can be used to construct a coarse model to reduce the search space. In order to extract the silhouette, the color of a pixel determines whether or not it belongs to the object using a probability distribution model.

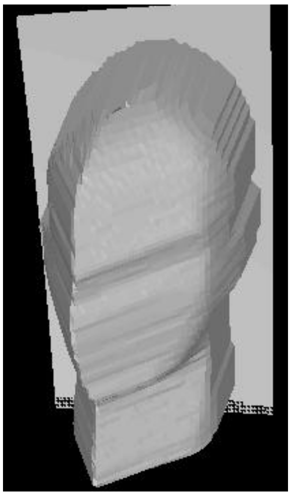

Figure 4: *A coarse model construct from the silhouette*

A point in 3-D space can be projected onto each frame. The appearance in each frame will be similar. In order to estimate the depth of one pixel in a base image, the correlation from the base image to its' neighboring image is correlated. In order to determine whether two blocks correspond, a template matching technique is used to evaluate the block pair correspondence. A *coarse to fine* algorithm is used to reduce the probability of a mismatch. Each frame is sub-sampled into a decomposed pyramid using a Gaussian operator. We compare The template block is compared using coarse to fine scales. The template matching function is based on the following correlation value operator[15]:

$$C(y)=\frac{<IT>-<I><T>}{\sigma(I)\sigma(T)}$$

where I is the image patch that must be matched to T. T is the template, $<>$ is the average operator and σ is the standard deviation over the area being matched. The image features that are compared include the intensity, the gradient and the color component.

Based on our observation, the distribution of the correlation sum can be divided into three main classes: The first class is the best candidate in normal distribution peak. That is the best case. The second class also has a normal distribution, but the maximum peak does not occur in the expected position. The last class does not have a discriminative peak in its' distribution, having only has some little turbulences. In the last two classes the base image does not present a discernable texture or contrast intensity. To filter out these points, the peak value of these points is compared with the mean average. If the value is under a predefined threshold, we assume that the correct depth for this pixel cannot be produced using its' neighboring pixels.

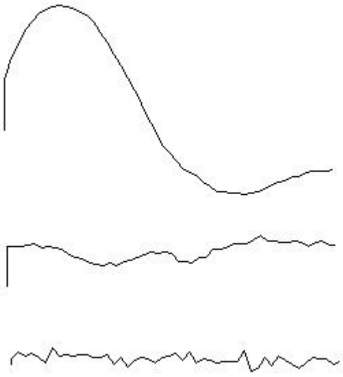

Figure 5: *The correlation distribution sum*

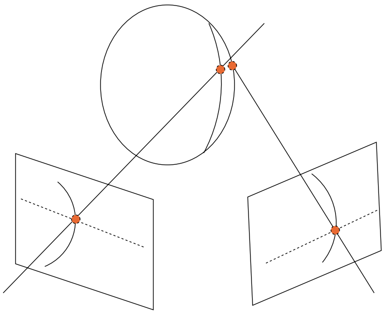

Figure 6: *The silhouette observed from different viewpoints*

There are two cases in which a precise depth estimation can be corrupted. The first is the edge effect that will dominate the template matching result. As Figure 6 shows, point A and point B are not the some point in Euclidean space. The image shown in the projection view is similar. The same fault will appear in the specular light. When the view is changed, the position of the specular light also varies. This false match will effect the depth estimation.

5.3. 3-D geometry reconstruction

If the camera is calibrated, a pixel (x,y) in the projected plane will shoot a line from the camera focus across this pixel in Euclidean space. After the distance from the focus to the 3-D point in Euclidean space is determined, the position of this point in Euclidean space can be calculated as follows.

$$s\begin{bmatrix} m_x \\ m_y \\ 1 \end{bmatrix} = \begin{bmatrix} \alpha & \gamma & u \\ 0 & \beta & v \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \Rightarrow s\cdot m = AM$$

Let the distance from focus to point in 3-D is z. Then, $s = \frac{z}{\|A^{-1}m\|}$, thus,

$$X = \frac{\left(m_x - u - \gamma\cdot\frac{(m_y - v)}{\beta}\right)\cdot s}{\alpha} \quad (4)$$

$$Y = \frac{(m_y - v)\cdot s}{\beta} \quad (5)$$

$$Z = s \quad (6)$$

In our processing method, the depth of each point is estimated one by one. Connecting the neighboring points can generate a regular mesh and this base image can be used as the texture. This is only two and half D solution. To generate a full 3-D model, additional images from different angles are needed to complete the entire model.

6. Experiment Results

Two desktop video cameras were used to capture our image sequences. The results were not very good because the physical characteristics of the two cameras were very different. It is not easy to determine the corresponding pairs when the average luminance and the distortion are both different. A Nikon Coolpix digital camera was used to capture this data. About 70 percent of the corresponding points could be determined. A Sony digital video camera was used as the input device. The resolution of the image was 720x480.

As mentioned before, the key to estimating accurate 3-D geometry is the calibration of the camera. In the calibration stage, if not enough accurate information is acquired about the camera, a correct epipolar constraint cannot be estimated. The calibration pattern must be held on the screen to ensure that accurate and complete camera calibration information is acquired. This limitation limits the viewing angle for the object to a small range. In our experimental sequence, only choose 21 frames were chosen for the test sequence. A frontal image was chosen as the base. Another ten images were added before and after the base image. The captured frames for the input sequence are as follows.

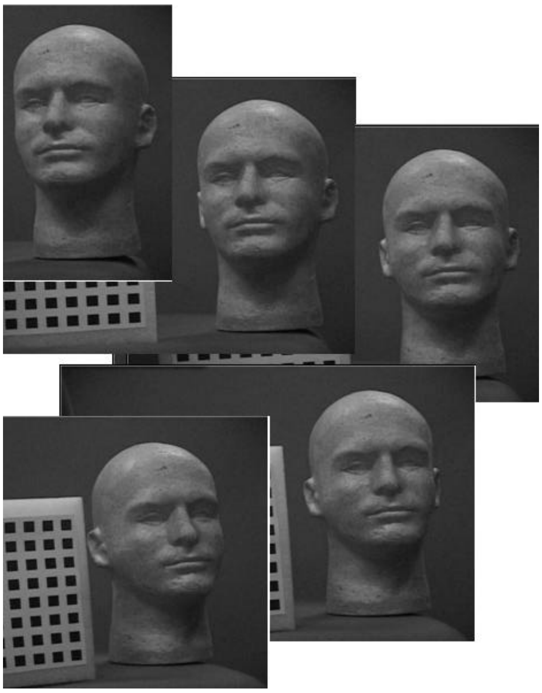

Figure 7: *The experimental sequence*

Table 1: *The results of different reference frames*

	Root mean square error(mm)	*Max error distance(mm)*	*Execution Time(min)*
3 frames	5.9	52.9	15
5 frames	4.5	35.5	33
9 frames	4.3	39.6	60
17 frames	4.2	39.6	115
17 frames†	3.5	18.9	116

† :with smooth operator.

To evaluate the result's quality, we also scan the model using a 3-D laser scanner. We manually adjust two models to locate at the same place roughly. Then, the Iterative Closest Point (ICP) algorithm is used to minimize the distance of the two models. The results of different reference frames are compared as above.

No computation speed optimization was used in this experimental series. The computation load was therefore very heavy. Most of this time was spent on estimating the depth. The result is as follows.

7. Conclusion and future work

We have presented an image-based modeling system. This system can execute and produce a precise head model auto-

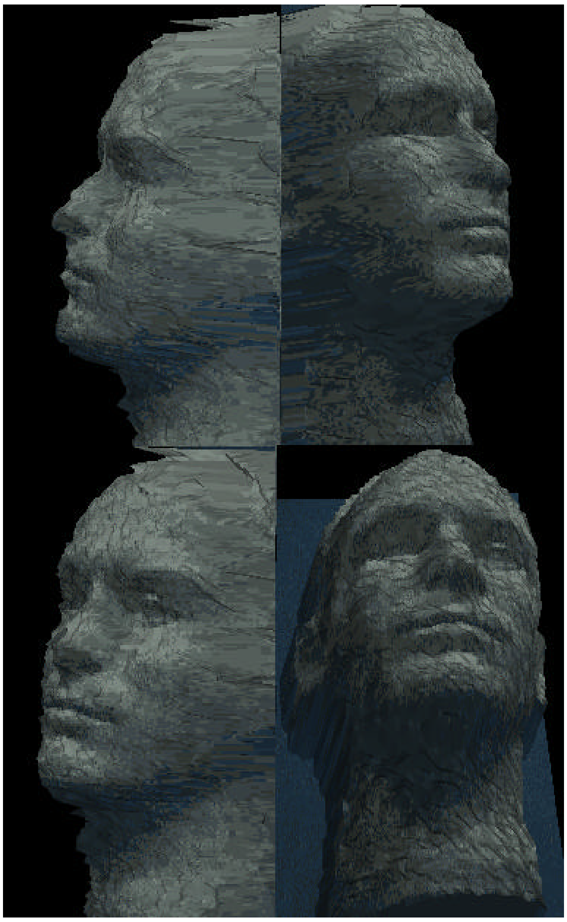

Figure 8: *The results shown in different viewpoints*

matically. In the future, we will improve the execution time for this model. We will collect several different surface viewpoints to complete a full model. This approach can also be applied to other objects with fewer constraints to get better results.

References

1. Z. Zhang, R. Deriche, O. Faugeras, and Q.-T. Luong. "A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry". *Artificial Intelligence Journal*, 78:87-119, Oct. 1995.

2. F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D. H. Salesin, "Synthesizing realistic facial expressions from pho-tographs," in *Computer Graphics, Annual Conference Series*, pp. 75-84, Siggraph, July 1998.

3. V. Blanz and T. Vetter, "A morphable model for the synthesis of 3d faces," in *Computer Graphics, Annual Conference Series*, pp. 187-194, Siggraph, August 1999.

4. Z. Liu, Z. Zhang, C. Jacobs, and M. Cohen, "Rapid modeling of animated faces from video," in *Proc. 3rd Inter-national Conference on Visual Computing*, (Mexico City), Sept. 2000. Also available as MSR technical report from http://research.microsoft.com/∼zhang/Papers/TR00-11.pdf.

5. W.S. Lee, N.M. Thalmann, "Generating a Population of Animated Faces from Pictures", *IEEE International Workshop on Modelling People (ICCV'99 Workshop mPeople)*, Corfu Holiday Palace, Kerkyra (Corfu), Greece, September, 20, 1999

6. P. Fua, "Using model-driven bundle-adjustment to model heads from raw video sequences," *International Conference on Computer Vision* , pp. 46-53, Sept. 1999. .

7. Wu, Tzer-Yih, "Face Reconstruction Using Computer Vision Techniques", *Master Thesis*, Department of Computer Science and Information Engineering, National Taiwan University, 2000

8. http://www.tcti.com

9. P. Fua. "A parallel stereo algorithm that produces dense depth maps and preserve iamge features". *Machine Vision and Applications*, 6:35-49, 1993

10. D. Decarlo, D. Metaxas, "Deformable Model-Based Shape and Motion Analysis from Images using Residual Error". *Proceeding ICCV '98*,pp.113-119.

11. I. Essa, S. Basu, T. Darrell, A. Pentland. "Modeling, Tracking and Interactive Animation of Faces and Heads Using Input from Video". *Proccedings of Computer Animation '96 Conference*, June 1996.

12. W. Matusik,C. Buehler,R. Raskar,S. J. Gortler,Leonard McMillan. "Image-Based Visual Hulls" *Computer Graphics, Annual Conference Series*, pp. 75-84, Siggraph, July 1998.

13. M. Okutomi, T. Kanade. "A Multiple-baseline Stereo", *IEEE Transactions on Pattern Analysis and Machine Intelligence* **15**(4), pp. 371–372 (April 1993).

14. Z. Zhang, "Flexible camera calibration by viewing a plane from unknown orientations," in *Proceedings of the 7th Interna-tional Conference on Computer Vision*, pp. 666-673, IEEE Computer Society Press, (Corfu, Greece), Sept. 1999. software is available at http://research.microsoft.com/∼zhang/Calib/.

15. R. Brunelli, T. Poggio. "Face Recognition: Features versus Templates", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1042-1052, vol. 15, no 10, Oct. 1993.