Robust Template Matching Using Multiview Video for Head Modeling

Fu-Che Wu, Murphy Chien-Chang Ho, Ming Ouhyoung
Department of Computer Science and Information Engineering, National Taiwan University
E-mail: {joyce,murphyho}@cmlab.csie.ntu.edu.tw,ming@csie.ntu.edu.tw

Abstract

Searching corresponding points is an important task in "structure from motion" related applications. In traditional approaches the method used to search for a corresponding pair from two stereo images usually focuses on 2D space. In this paper, we propose a new approach to solve the corresponding pair problem in 3D space. If a point is a corresponding point of a certain point in another image, both points have a common location in Euclidean space. To robustly recover this relationship, more frame information is used in this method and the common position in Euclidean space is estimated using epipolar geometry. Epipolar geometry describes a camera as a pinhole model. First, the camera's intrinsic and extrinsic parameters are determined using a regular grid pattern. In this way an epipolar geometric model can be constructed. With this model, the depth of a target pixel from the base image can be estimated by searching along the epiolar line of its neighboring images to find the best matched position. In this paper, we present a procedure that uses this technique to reconstruct the 3D geometry of a "human head" successfully from multiview images. The root mean square errors of the resulting model ranges from 5.9 mm obtained from 3 views to 3.5 mm from 17 views.

1 Introduction

A 3D structure is useful in many areas, such as animation, interactive communication, compression, environment browsing and so on. There are many tools available for constructing a 3D model. The data input methods include the Cyberware laser scanner, active light striper, stereo, video sequencing and manual interaction. Those methods can be divided into three main approaches: hardware, image-based and manual interaction. The major advantage of the hardware solution is that it facilitates very precise modeling. However, the hardware is expensive and not popular. In general, the image-based approaches cannot easily achieve a precise measure, which is limited by the character of the image. The manual approach is time consuming and cannot easily construct a realistic model. Until now, a convenient, cheap and precise tool for constructing a 3D structure has not existed. A robust approach to estimate the corresponding points from the image sequence is needed.

To determine a corresponding pair is an important issue in constructing an object's geometry from stereo images. The image content contains only 2D information. To recover the 3D structure, more data frames are needed to determine the relationship among these frames. To determine the relationship among these frames in Euclidean space involves estimating the camera's intrinsic and extrinsic parameters. That is epipolar geometry. Recovering a corresponding pair involves finding a common position in Euclidean space that is consistent among each projected plane. This method can be divided into three main approaches: optical flow, template matching and feature extraction.

The optical flow method 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-posed 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.[1] and Essa et al.[2] used this approach to estimate facial expressions. To estimate the precise 3D 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.

The template matching method involves comparing the similarity between two blocks of an image. There are also many limitations in this method. For example, there are three situations that will affect the comparison result: no discernible texture, the edge effect and luminance variations. If the compared block does not present a discernible texture, the similarity between the two blocks cannot be discriminated. If the compared block contains some strong signals such as edges or specular light, that will dominate the template matching result. As Figure 1 shows, point A

and point B are not the some point in Euclidean space. The image shown in the projection view is similar. The luminance variation will make a region at the same position in Euclidean space, but the two points will look very different in the two images. In the above situations, the template matching approach will produce a greater number of unexpected results.

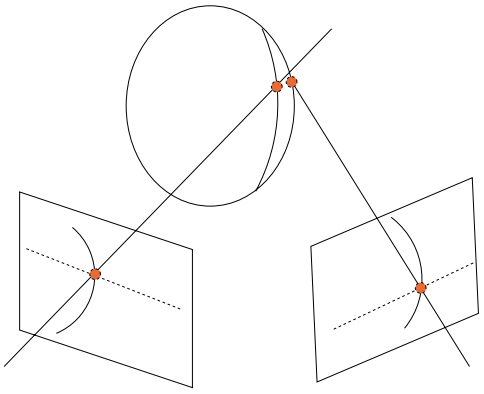

Figure 1. The silhouette observed from different viewpoints

The feature extraction method involves extracting some special feature points, such as a corner point or a line, and then finding a matching point among the feature points in the two images. This method can survive luminance variations. It still does not work under discernible texture or edge effect situations. This method can only find feature points' positions. In the general approach, most of the researchers used generic head models as predefined models. This generic model is then adjusted to approximate a new model that corresponds to these feature points.

Most of the researchers in images-based approaches only compare the data between two images. In this paper, we present a procedure to perform template matching on multi-view images. In this way, we can get more precise results. The outline of this paper is as follows. We will briefly describe the related works by other researches. Following this, we will introduce our work and some experiment results.

2 Related work

Pighin et al.[3] developed a system that employs a user-assisted technique to recover camera poses corresponding to the views as well as the 3D 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[4] built a morphable face model by exploiting large 3D face scan statistics and recovered domain knowledge about the facial variations by applying a pattern classification method. Thus, a 3D face can be generated automatically from one or more photographs or modeled directly through an intuitive user interface.

Liu et al.[5] 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 3D space. The reconstructed 3D points are then adapted into a face model.

Lee and Thalmann[6] 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 produce new geometrical coordinates for the generic head modification. In this approach, because the feature points are produced using manual interactions, the reconstructed 3D face can be animated immediately with the given expression parameters.

Fua et als.[7] 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[8] 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.

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.

3 Fundamental Theory

Our work is constructed under a calibrated environment. Each input frame must determine its intrinsic and extrinsic camera parameters. In this section, we will introduce the camera model and epipolar geometry first.

3.1 Camera model

To estimate the inverse structure from the projected image, a camera model must be determined[9]. 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 represents the intrinsic parameters and $[Rt]$ represents 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.

3.2 Epipolar geometry

The case of the two cameras is shown in fig. 2.

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 from 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,

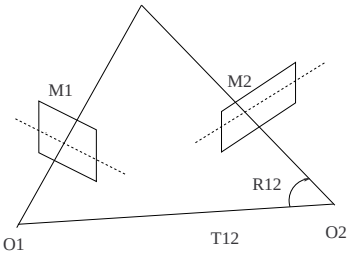

Figure 2. Epipolar geometry is a co-planarity constraint.

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

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.

4 Implementation

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.

4.1 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.

Four types of kernel filters are used to detect the four types of corners. In order to evaluate if a pixel belongs to a given corner or not, the correlation function is used to measure their similarities. The threshold that discriminates

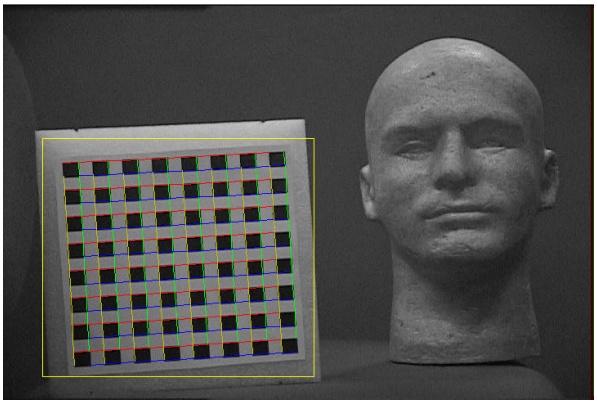

Figure 3. The camera calibration pattern

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 acquire 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[10]:

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

4.2 Template matching

After the camera parameters have been estimated, the corresponding points on each frame can be determined. Okutomi and Kanade[11] developed a stereo matching method that estimates the disparity from a multiple-baseline. Their approach inspired us to estimate the depth by comparing the correlation sum from multi-angle images. A point in 3D 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 subsampled into a decomposed pyramid using a Gaussian operator. The template block is compared using coarse to fine scales.

The template matching function is based on the following correlation value operator[12]:

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

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.

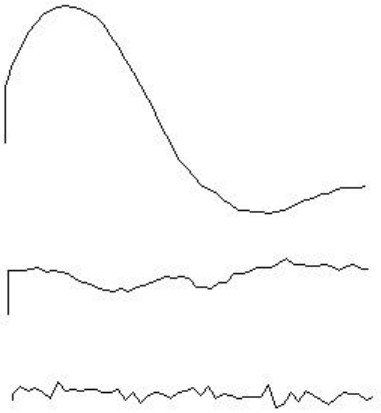

Figure 4. The correlation distribution sum

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 a 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 this value is under a predefined threshold, we assume that the correct depth for this pixel cannot be produced using its' neighboring pixels.

4.3 Coordinate Transformation

In this subsection, we will introduce how to locate the position at neighboring frames when a pixel of the base image is at a certain depth z.

As mentioned before, the coordinate of 3D point M_b in the base image's local coordinate system and its image co-

ordinates m_b are related by

$$M_b = s \cdot A^{-1} m_b$$

If the distance from the focus to the surface of the object is z. then

$$s = \frac{z}{\|A^{-1} m_b\|}$$

Then, we translate the local coordinate into a world coordinate using following relation

$$M_w = R_b^{-1}(M_b - T_b)$$

From above coordinate we can translate to each frame's project plane.

$$s \cdot m_i = P(R_i M_w + T_i)$$

After the camera parameter is estimated, the translation and rotation of each frame is known. Thus, we can iteratively increase the depth z and compare the similarity among these frames.

5 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 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 were as follows.

To evaluate the quality of the result, the model was scanned using a 3D laser scanner. We manually adjusted the two models so they would be scanned roughly at the same place. The Iterative Closest Point (ICP) algorithm was used to minimize the distance between the two models. The results from the different reference frames are compared as follows.

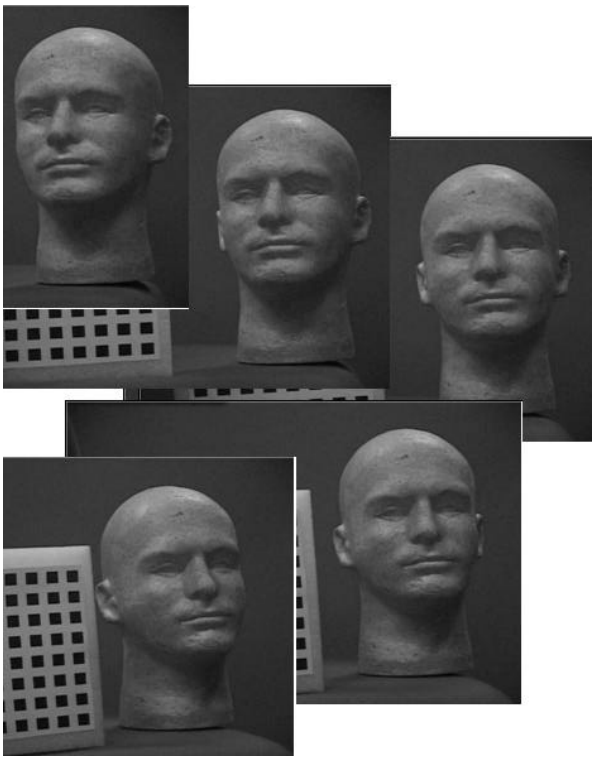

Figure 5. The experimental sequence

6 Conclusion

We have presented a new template matching approach for an image-based modeling system. This system can execute and produce a precise head model automatically. 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.

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.

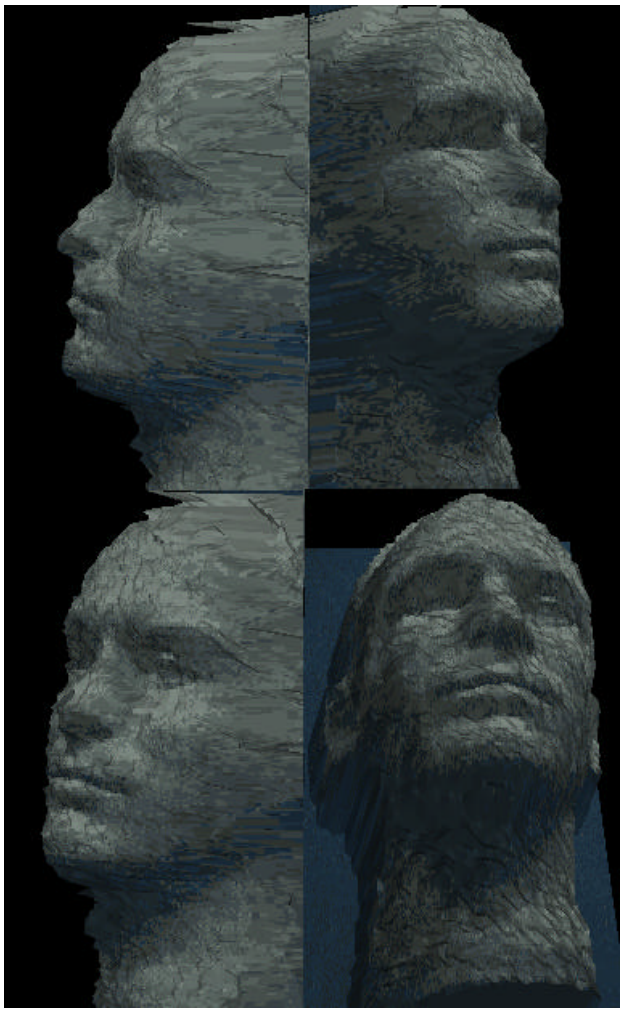

Figure 6. The results shown in different viewpoints

References

[1] D. Decarlo, D. Metaxas, ”Deformable Model-Based Shape and Motion Analysis from Images using Residual Error”. *Proceeding ICCV ’98*,pp.113-119.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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

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

[8] Wu, Tzer-Yih, ”Face Reconstruction Using Computer Vision Techniques”, *Master Thesis*, Department of Computer Science and Information Engineering, National Taiwan University, 2000

[9] 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.

[10] 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/.

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

[12] 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.