3D FACE RECONSTRUCTION FROM ROTATIONAL MOTION

Yoshiko Sugaya, Shingo Ando, Akira Suzuki and Hideki Koike

NTT Cyber Space Laboratories
NTT Corporation 1-1 Hikari-no-oka, Yokosuka, Kanagawa, 239-0847 Japan
E-mail: sugaya.yoshiko@lab.ntt.co.jp

ABSTRACT

3D reconstruction of a human face from an image sequence remains an important problem in computer vision. We propose a method, based on a factorization algorithm, that reconstructs a 3D face model from short image sequences exhibiting rotational motion. Factorization algorithms can recover structure and motion simultaneously from one image sequence, but they usually require that all feature points be well tracked. Under rotational motion, however, feature tracking often fails due to occlusion and frame out of features. Additionally, the paucity of images may make feature tracking more difficult or decrease reconstruction accuracy. The proposed 3D reconstruction approach can handle short image sequences exhibiting rotational motion wherein feature points are likely to be missing. We implement the proposal as a reconstruction method; it employs image sequence division and a feature tracking method that uses Active Appearance Models to avoid the failure of feature tracking. Experiments conducted on an image sequence of a human face demonstrate the effectiveness of the proposed method.

Keywords: 3D Face Reconstruction, Rotational Motion, Factorization, Active Appearance Models

1. INTRODUCTION

Recovering the 3D shape of an object and the camera motion from an image stream is one of the important research topics in computer vision. One popular solution is to employ factorization by Tomasi and Kanade[1] which can recover structure and motion simultaneously from one image sequence. This algorithm is robust and efficient to use under the assumption that all features are well-tracked. Feature tracking, however, remains a fundamental problem, and no best solution has been found. A major cause of this problem is missing features.

The approaches used to deal with missing data can be categorized as two types. The first type estimates or interpolates the feature tracking data[2]. Estimation and interpolation of the feature trajectory, however, is only possible over a few frames. Tracking fails when the adjacent frames are very different or the camera moves rapidly. In addition, it is impossible to recover the complete 3D shape of the object using data captured by one or more cameras set on a circle centered on the object.

The second approach, on the other hand, deals with missing data by constructing a 3D shape from some short image sequences extracted from a long input image sequence. Some researchers recover structure and motion by applying sequential factorization[3][4] which overcomes the problem by regarding features as a vector time series; others try to divide a long input image sequence into some subsequences and then integrate the results from the shorter sequences[5]. The main problem with this technique is that reconstruction accuracy of the recovered structure tends to be low because the camera motion becomes very small due to the division. For the structure and motion problem, in general, high accuracy is achieved when the camera motion is enough large.

We propose a 3D face reconstruction method that can handle an image stream exhibiting rotational motion but likely to be missing some data. In order to overcome the difficulties mentioned above, we introduce a feature tracking method based on Active Appearance Models(AAM) to deal with the missing features, and a reconstruction method that recovers camera motion from the whole input sequence first by factorization, and then reconstructs the 3D shape from several short sequences to prevent a drop in reconstruction accuracy. All derived 3D shapes are finally integrated into one final shape.

We begin with a brief review of AAM[6] and Factorization [8] in Section 2. In Section3 we describe the proposed method. In Section 4, experimental conditions and reconstruction results are presented, followed by the conclusion in Section 5.

2. BACKGROUND

This section presents two short overviews of AAM by Cootes, Edwards and Taylor and the paraperspective factorization method by Poelman and Kanade. More detailed descriptions of these methods can be found in [6][7][8].

2.1 Active Appearance Models

AAM[6][7] is a fitting algorithm commonly used to handle deformable objects in single images. This algorithm is based on the statistical models generated by combining a model of shape variation with a model of texture variation.

The statistical models are built by using a set of labeled training images. Corresponding points have been marked on each training image. Let shape vector $\mathbf{x}$ be a column vector

containing point locations in the image, and $\overline{\mathbf{x}}$ be the mean shape calculated as an average of the shape vectors in the training set. After each training image is warped by mean shape to match the points to those of mean shape, we obtain texture vector $\mathbf{g}$, which has the components of texture in the warped shape and the average of texture vectors $\overline{\mathbf{g}}$ as a mean texture. Using Principal Component Analysis (PCA), shape vector $\mathbf{x}$ and texture vector $\mathbf{g}$ in the training image can be approximated as:

$$\mathbf{x} = \overline{\mathbf{x}} + \quad {}_s\mathbf{c} \tag{1}$$

$$\mathbf{g} = \overline{\mathbf{g}} + \quad {}_g\mathbf{c} \tag{2}$$

where $\quad {}_s$, $\quad {}_g$ are matrices describing the mode of variation derived from the training set, and $\mathbf{c}$ are the appearance parameters that control shape and texture.

Fitting the model to a new image, called AAM search, involves finding the optimal appearance parameters $\mathbf{c}$ and the pose parameter $\mathbf{t}$ which contains the in-plane rotation, translation, and scale in the new image, by minimizing the residual error between the geometrically normalized input image and the current model texture.

2.2 Factorization

When P features are being tracked in F image frames, let $(u_f \ , v_f \ )$ be the projection point observed in the ith frame. The $2F \quad P$ measurement matrix, $\mathbf{W}$, is then defined as follows:

$$\mathbf{W} = \begin{bmatrix} u_{11} \cdots u_{1P} \\ \vdots \ddots \vdots \\ u_{F1} \cdots u_{FP} \\ v_{11} \cdots v_{1P} \\ \vdots \ddots \vdots \\ v_{F1} \cdots v_{FP} \end{bmatrix} \quad \begin{bmatrix} \overline{u}_1 \\ \vdots \\ \overline{u}_F \\ \overline{v}_1 \\ \vdots \\ \overline{v}_F \end{bmatrix} \begin{bmatrix} 1 \cdots 1 \end{bmatrix} \tag{3}$$

where $\overline{u}_f, \overline{v}_f$ is the average of feature points in frame f. The measurement matrix, $\mathbf{W}$, is thus composed of the elements based on the coordinate system whose origin is the centroid of the feature points.

$$\acute{u}_f \ = u_f \quad \overline{u}_f, \quad \acute{v}_f \ = v_f \quad \overline{v}_f \tag{4}$$

Let us now consider $\mathbf{s}$ as the 3D feature in world coordinates; each image, f, was taken at some camera orientation, which we describe by the orthogonal unit vector $\mathbf{i}_f, \mathbf{j}_f$ and $\mathbf{k}_f$ where $\mathbf{i}_f$ and $\mathbf{j}_f$ correspond to the x and y axes of the image plane, respectively, with $\mathbf{k}_f$ points along the camera's line of sight. Under paraperspective projection, a feature point on frame f can be expressed as follows:

$$u_f \ = \mathbf{m}_f^T \mathbf{s} \ + x_f, \quad v_f \ = \mathbf{n}_f^T \mathbf{s} \ + y_f \tag{5}$$

where

$$z_f = \quad \mathbf{t}_f \mathbf{k}_f, \ \ x_f = \frac{\mathbf{i}_f \mathbf{t}_f}{z_f}, \ \ y_f = \frac{\mathbf{j}_f \mathbf{t}_f}{z_f} \tag{6}$$

$$\mathbf{m}_f = \frac{\mathbf{i}_f \quad x_f \mathbf{k}_f}{z_f}, \qquad \mathbf{n}_f = \frac{\mathbf{j}_f \quad y_f \mathbf{k}_f}{z_f} \tag{7}$$

where z_f is the depth to the object's center of mass, and (x_f, y_f) is the center of mass projected onto frame f. Note that these equations are simplified by assuming unit focal length.

Without loss of generality, we can place the world origin at the center of mass of the object. Because of this, $\overline{u}_f, \overline{v}_f$ equal x_f, y_f, respectively.

$$\overline{u}_f \quad = \quad \frac{1}{\mathrm{P}} \sum_{=0}^{\mathrm{P}} (\mathbf{m}_f^T \mathbf{s} \ + x_f) = x_f \tag{8}$$

$$\overline{v}_f \quad = \quad \frac{1}{\mathrm{P}} \sum_{=0}^{\mathrm{P}} (\mathbf{n}_f^T \mathbf{s} \ + y_f) = y_f \tag{9}$$

Thus the elements of $\mathbf{W}$,$\acute{u}_f \ , \acute{v}_f \ $, can be written as:

$$\acute{u}_f \quad = \quad u_f \quad x_f = \mathbf{m}^T \mathbf{s} \tag{10}$$

$$\acute{v}_f \quad = \quad v_f \quad y_f = \mathbf{n}^T \mathbf{s} \tag{11}$$

These equations show that $\mathbf{W}$ can be decomposed into motion matrix $\mathbf{M}$ (which represents the camera motion) and shape matrix $\mathbf{S}$ (which represents the 3D shape):

$$\mathbf{W} = \begin{bmatrix} \mathbf{m}_1^T \\ \vdots \\ \mathbf{m}_F^T \\ \mathbf{n}_1^T \\ \vdots \\ \mathbf{n}_F^T \end{bmatrix} \begin{bmatrix} \mathbf{s}_1 \cdots \mathbf{s}_P \end{bmatrix} = \mathbf{MS} \tag{12}$$

In order to ensure that this decomposition of (12) is unique, it is required to find appropriate linear transformation $3 \quad 3$ matrix $\mathbf{A}$ that transforms $\hat{\mathbf{M}}$ and $\hat{\mathbf{S}}$ into the true solutions $\mathbf{M}$ and $\mathbf{S}$ as follows:

$$\mathbf{M} = \hat{\mathbf{M}}\mathbf{A}, \qquad \mathbf{S} = \mathbf{A}^{-1}\hat{\mathbf{S}} \tag{13}$$

An appropriate $\mathbf{A}$ can be determined by using the following equations suggested by the constraints of paraperspective projection.

$$\frac{\|\mathbf{m}_f\|^2}{1 + x_f^2} = \frac{\|\mathbf{n}_f\|^2}{1 + y_f^2} \tag{14}$$

$$\mathbf{m}_f \mathbf{n}_f = \frac{x_f y_f}{2} \left(\frac{\|\mathbf{m}_f\|^2}{1 + x_f^2} + \frac{\|\mathbf{n}_f\|^2}{1 + y_f^2} \right) \tag{15}$$

3. OVERVIEW OF THE PROPOSED METHOD

The proposed 3D reconstruction approach can handle an image sequence containing rotational motion in which feature tracking is likely to fail. Fig. 1 overviews the proposed method. First, the input image sequence is divided into some subsequences to overcome the dropping of feature points. In each short subsequence and the whole input sequence, the feature points that are visible in all views are

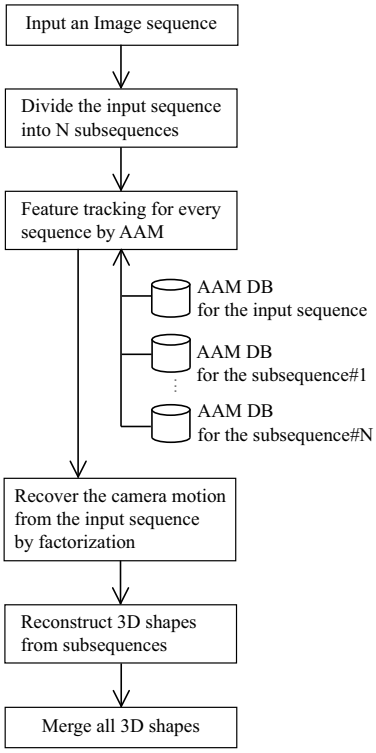


Fig. 1: Overview of the proposed method

tracked by using AAM. A set of tracked features from the whole sequence is then input to the factorization algorithm to recover camera motion. Using the recovered camera motion, local 3D shapes are reconstructed from the short subsequences. By utilizing the relationship between the local coordinates of short parts and that between the overlapping features of adjacent subsequences, we can reconstruct one large 3D face model by merging the 3D face parts obtained from the subsequences.

3.1 Sequence Division

The original input sequence is divided into N subsequences, see Fig 2. This enables us to deal with many more features in the subsequences than is possible in the whole sequence.

3.2 Feature Tracking

AAM is well-suited for fitting deformable objects like human faces. Its fitting performance strongly depends on the training set used to build the model. AAM can fit any directions of face if the training set has a variety of directions of face images. In the proposed method, N+1 AAMs for every sequence are created in consideration of this characteristic.

As a feature tracker, AAM is applied to all frames in each image sequence. We can then obtain the appearance parameters $\mathbf{c}$ of every single image. Using the equations mentioned in Section 2.1, shape vector $\mathbf{x}$ can be derived from $\mathbf{c}$.

AAM fitting has another advantage over other feature trackers, it doesn't require any time series information. Thus

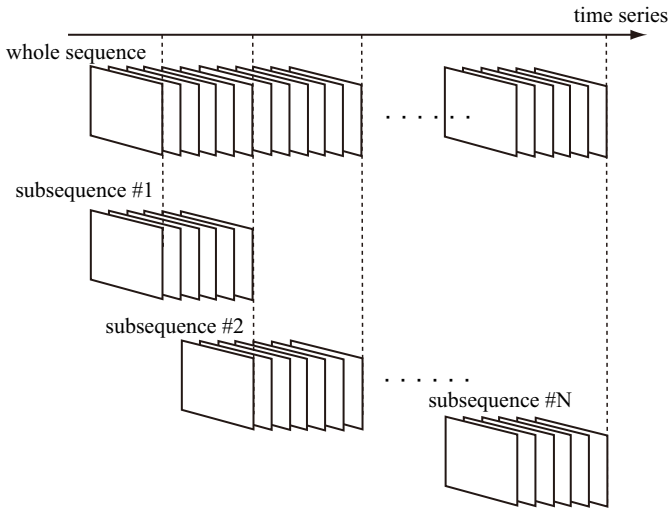


Fig. 2: Division of the input image sequence

an AAM base feature tracker can extract features even if adjacent frames have quite different appearances.

3.3 Camera Motion and Shape Recovery

The camera motion recovered by applying the factorization algorithm to the entire sequence is used in the shape recovery of subsequences. Of course we can obtain both the camera motion and the shape simultaneously from the factorization results of short subsequences, but those derived from subsequences tend to be less accurate than that from the entire sequence. There is thus a need to replace the camera motion when recovering the shape from subsequences.

When the camera motion is known, the shape matrix can be obtained from (12) as:

$$\mathbf{S} = \mathbf{M}^{+}\mathbf{W} \tag{16}$$

where $\mathbf{M}^{+}$ is the pseudo-inverse matrix of $\mathbf{M}$. Using this equation and the recovered camera motion, the 3D shapes of the subsequences can be obtained.

3.4 Merging All Shapes

N+1 shapes reconstructed from (16) have different local coordinates. In order to acquire one final shape by merging the partial shapes, we employ Horn's method, which uses unit quaternion, to find the conversion matrices that unify the different local coordinates into one common coordinate [10]. After conversion of all 3D shapes in each local coordinate, the final 3D face shape is created by smoothing overlapping points. The number of vertices in the merged final shape may not be enough to express a human face because the AAMs used in feature tracking have sparse feature points as shape vectors. To make the 3D face dense, a subdivision mesh[11] is applied after merging the 3D shapes.

4. EXPERIMENTAL RESULTS

In order to confirm the proposed method's ability to reconstruct a 3D shape from an image stream exhibiting rotational motion, we challenged it with a real image sequence.

In this experiment, 13 images of a human face from different orientations in the HOIP Database[9] were used as the input image sequence. Fig.3 shows some of the input images. These face images represent 5 degree increments in horizontal rotation with no vertical movement. The input sequence was divided into 2 subsequences where the number of frames in both subsequences was set to 7 with 1 image overlap. That is, the first subsequence consisted of frame #1 to frame #7, and the second subsequence consisted of frame #7 to frame #13.

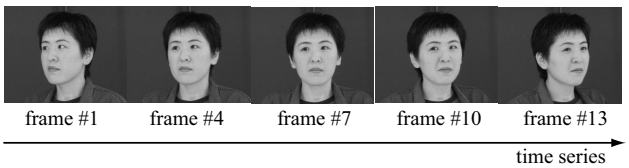


Fig. 3: Input image sequence

AAMs for each subsequence were built using 210 face images taken from 7 directions as a training set. Fig. 4 shows examples of the AAM training images; the shape vectors were created by aligning the white points in the picture. Fig. 4 (a) is one of the training samples for the entire sequence; (b) and (c) are for subsequence #1 and subsequence #2, respectively. Each AAM consisted of 69 features. The AAM for the entire sequence is narrower than that for the subsequences, but this is enough to recover camera motion.

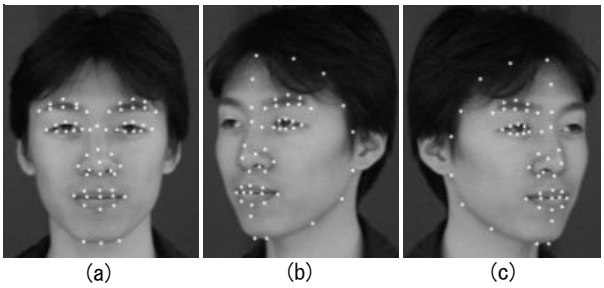


Fig. 4: AAM training samples

Fig.5 compares the factorization result of the camera motions from the sequences where and are the azimuth angle and the polar angle, respectively. It is readily understand from Fig.5 that the subsequences, especially , yield less accurate results. Note that a consideration of the coordinates defined in factorization sets the vertical axis as the y-axis.

Fig. 6 shows the merging result and the results after making the subdivision mesh. Three views of the 3D face with texture mapping on the subdivided shape are shown in Fig. 7.

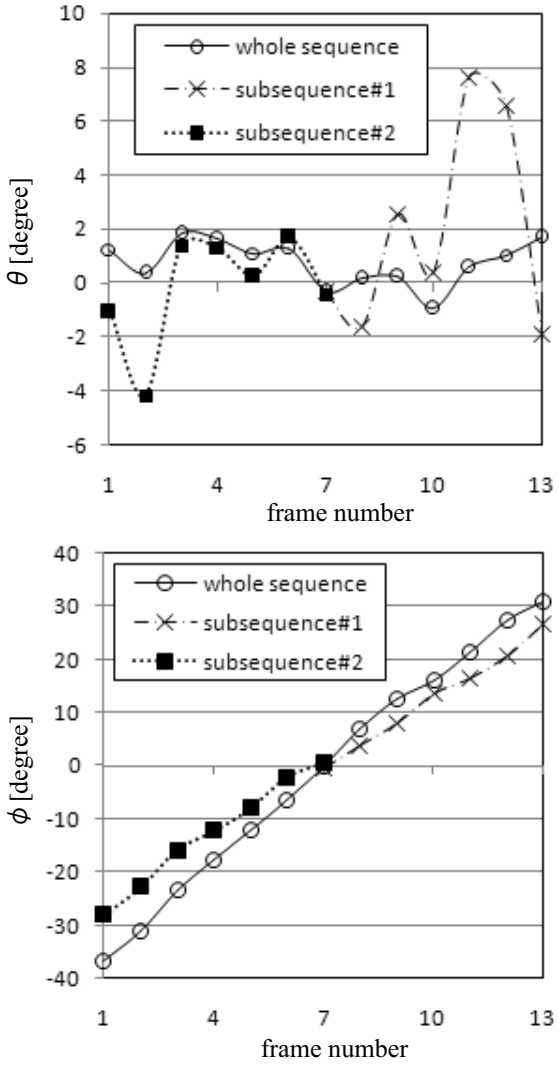


Fig. 5: Recovered Camera Motions

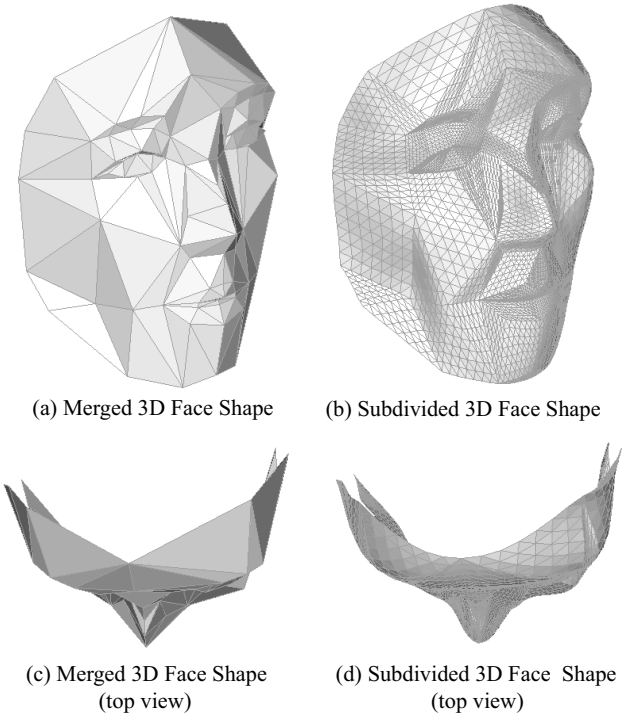
5. CONCLUSIONS AND FUTURE WORKS

An method of 3D face reconstruction from an image sequence exhibiting rotational motion has been presented. This method reconstructs a 3D face by dividing the input image sequence into several short subsequences and merging all 3D points derived from the subsequences. This allows a human face to be reconstructed from the image sequence in which feature tracking is likely to fail and to use short image sequences, especially when the images were captured under large camera motion. To further investigate the possibilities of the proposed approach, we will conduct a quantitative evaluation.

6. ACKNOWLEDGEMENTS

The authors would like to thank Isao Miyagawa of the Image Media Processing Group at NTT Cyber Space Laboratories for his constructive comments and discussions.

Fig. 7: Resulting 3D Face



(a) Merged 3D Face Shape (b) Subdivided 3D Face Shape

(c) Merged 3D Face Shape (top view) (d) Subdivided 3D Face Shape (top view)

Fig. 6: Reconstruction Results of 3D Shape

7. REFERENCES

[1] C. Tomasi and T. Kanade: ”Shape and motion from image streams – a factorization method” International Journal of Computer Vision, 9(2), pp. 137-154 (1992)

[2] D. Jacobs :”Linear Fitting with Missing Data: Applications to Structure-from-Motion and to Characterizing Intensity Images,” IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’97), pp.206-212 (1997)

[3] T. Morita and T. Kanade: ”A Sequential Factorization Method for Recovering Shape and Motion from Image Streams”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.19, No.8, pp.858-867 (1997)

[4] E. Shibusawa, W. Mitsuhashi: ”A Sequential Factorization Method for Euclidean Reconstruction from Image Sequence with Missing Data” IEICE technical report. Vol.106, No.449, pp. 189-192 (2007)

[5] T. Li, W. Chengke, L. Shigang and Y. Yaoping: ”Complete structure recovery from long image sequence with occlusions”, Third International Symposium on Multispectral Image Processing and Pattern Recognition. Proceeding of the SPIE, Vol. 5286, pp. 529-534 (2003)

[6] T. F. Cootes, G. J. Edwards and C. J. Taylor:”Active appearance models”, Proceedings of the 5th European Conference on Computer Vision, Vol. 2, pp. 484-498 (1998)

[7] T. F. Cootes and C. J. Taylor: ”Statistical models of appearance for computer vision” Technical report, Wolfson Image Analysis Unit, University of Manchester (1999)

[8] C.J. Poelman and T. Kanade: ”A Paraperspective Factorization Method for Shape and Recovery” Proceedings of the 3rd European Conference on Computer Vision, pp. 97-108 (1994)

[9] HOIP Face Database
http://www.softopia.or.jp/rd/facedb.html
The facial data in this paper are used by permission of the Softopia Japan Foundation. It is strictly prohibited to copy, re-use, or distribute the facial data without permission.

[10] B. K. P. Horn: ”Closed-form solution of absolute orientation using unit quaternions” Journal of the Optical Society of America A, vol.4, pp.629-642, 1987

[11] C. Loop: ”Smooth Subdivision Surfaces Based on Triangles” master’s thesis. Dept. of Math., Univ. of Utah (1987)