A METHOD OF MODELING DEFORMATION OF AN OBJECT EMPLOYING SURROUNDING VIDEO CAMERAS

Joo Kooi TAN, Seiji ISHIKAWA
Department of Mechanical and Control Engineering
Kyushu Institute of Technology, Japan
etheltan@is.cntl.kyutech.ac.jp, ishikawa@is.cntl.kyutech.ac.jp

KEY WORDS: Modeling, 3-D modeling, Shape recovery, Deformation, Motion, Non-rigid objects, Factorization.

ABSTRACT

A novel technique is presented for modeling deformation of an object, not only its front side but also the rear side, employing surrounding multiple video cameras. In order to make a 3-D model of an object from its actual images, 3-D shape recovering techniques based on image capturing are usually employed. Unlike existent techniques which make a model by mutually connecting partially recovered shapes, the proposed technique provides a model of an object by recovering its entire shape simultaneously employing multiple fixed video cameras surrounding the object. In the technique, $F(\geqq 3)$ pairs of, therefore $2F$, video cameras are placed around the object concerned and take the images of the object's deformation. A single measurement matrix is defined by extracting feature points locations from the $2F$ video image streams and factorization is applied to the matrix once to derive the 3-D coordinates of all the feature points on the object during observation. Thus the entire model of the object is obtained.
Factorization based on orthographic projection is employed for 3-D shape recovery. This causes approximately 3 to 4% of recovery errors in the technique. Instead, the present technique necessitates neither camera calibration except for making the light axes of a facing pair of video cameras parallel, nor registration of recovered shape. The technique is presented and experimental results are shown.

1 INTRODUCTION

Three-dimensional object modeling has drawn much attention mainly in computer vision community. Various 3-D shape recovering techniques for rigid objects have been studied (Horn,1986) and some of them are now commercially available such as a stereo vision system and a structured lights projection system. When deformable objects are taken into account, optical methods represented by stereo vision seem promising compared with non-optical methods such as goniometers and magnetic sensors which one has to put on his body. Existent optical methods including various motion capturing methods, however, necessitate strict camera calibration, which is not very convenient especially for outdoor use. A technique called factorization has been proposed by Tomasi & Kanade(1992) for recovering 3-D shape of a rigid object and camera motion from a video image stream without using camera parameters. This technique has been extended to a deformable object's shape recovery by Tan *et al.*(1998) and Tan & Ishikawa(1999) employing uncalibrated multiple video cameras. The technique is, however, insufficient for making an entire 3-D model of an object (irrespective of rigid or non-rigid), since only the object part commonly visible from the multiple video cameras is recovered. Normally an entire model of an object is, if it cannot be modeled by a CAD system because of its complicated shape, made by mutually connecting partially recovered shape of the object employing existent techniques.

In the present paper, a technique is described for modeling deformation of an object, not only its front side but also the rear side, employing surrounding multiple video cameras. In the technique, F pairs of, therefore $2F$, video cameras are placed around the object concerned and take the images of the object's deformation. In these $2F$ video image streams, feature points on the object are tracked during observation time and their image coordinates are all stored into a single measurement matrix. This is the novelty of the present paper. Once the single measurement matrix is obtained, factorization is applied to the matrix, resulting in the recovery of the 3-D coordinates of all the chosen feature points. Thus the entire model of the object is obtained. If the object is rigid, the entire shape of the object is modeled. If the object is deformable, the entire shape at each sampled time and the entire deformation during observation time is modeled. The technique is presented and experimental results are shown. Since factorization based on orthographic projection is employed in the technique, some recovery errors are inevitable. This is also discussed.

2 SHAPE RECOVERY OF A DEFORMABLE OBJECT

An outline of 3-D shape recovery for deformable objects based on factorization is presented according to Tan & Ishikawa(1999).

$F(\geqq 3)$ video cameras are fixed around an object. During discrete observation time $t(t=1,2,\ldots,T)$, video camera f takes images of the object, producing image streams $I_f(t)(t=1,2,\ldots,T;\ f=1,2,\ldots,F)$. The feature points defined on the object observed by all the video cameras at time t are denoted by $\boldsymbol{s}_p(t)(p=1,2,\ldots,P_t)$. Feature point $\boldsymbol{s}_p(t)$ is a 3×1 column vector and is projected at $(x_{fp}(t),y_{fp}(t))$ on the image plane of video camera f at time t.

A measurement matrix at time t denoted by $W(t)$ is defined as a $2F\times P_t$ matrix whose fth component of the pth column contains $x_{fp}(t)$, whereas $(f+F)$th component of the pth column contains $y_{fp}(t)$. The measurement matrices $W(t)(t=1,2,\ldots,T)$ are unified into a single matrix called an extended measurement matrix in the following form;

$$W = \left(W(1)\ \mid\ W(2)\ \mid\ {}_{3}\ \mid\ W(T)\right). \tag{1}$$

The mean value of each row of matrix W is subtracted from the components of the corresponding row to yield a matrix $\widetilde{W}$ as

$$\widetilde{W} = W - \frac{1}{Q} W \cdot E\ , \tag{2}$$

where $Q = \sum_{t=1}^{T} P_t$, *i.e.*, the number of all the feature points, and E is a $Q\times Q$ matrix whose entries are all unity. Let the world origin be specified at the centroid of all the Q feature points and let the feature points be newly denoted by $\boldsymbol{s}_p(t)(p=1,2,\ldots,P_t)$ with respect to the world origin. Then, since orthographic projection is assumed in the present technique, the fp component of matrix $\widetilde{W}$ can be described by $(\boldsymbol{i}_f, \boldsymbol{s}_p(t))$ and the $(f+F)p$ component by $(\boldsymbol{j}_f, \boldsymbol{s}_p(t))$, where $\boldsymbol{i}_f$ and $\boldsymbol{j}_f$ are 3×1 column vectors and the set of vectors $\boldsymbol{i}_f$, $\boldsymbol{j}_f$, and $\boldsymbol{i}_f \times \boldsymbol{j}_f$ defines the orthonormal system of the lens coordinate system of video camera f. This leads to the following decomposition of the matrix $\widetilde{W}$ (Tomasi & Kanade, 1992);

$$\widetilde{W} = M \cdot S\ , \tag{3}$$

where matrices M and S have the following form;

$$M = \left(\boldsymbol{i}_1,\ \ \boldsymbol{i}_2,\ \ {}_{3}\ ,\ \ \boldsymbol{i}_F\ \ \mid\ \ \boldsymbol{j}_1,\ \ \boldsymbol{j}_2,\ \ {}_{3}\ ,\ \ \boldsymbol{j}_F\right)^{\mathrm{T}}, \tag{4}$$

$$S = \left(\boldsymbol{s}_1(1),\ \ \boldsymbol{s}_2(1),\ \ {}_{3}\ ,\ \ \boldsymbol{s}_{P_1}(1)\ \mid\ \boldsymbol{s}_1(2),\ \ \boldsymbol{s}_2(2),\ \ {}_{3}\ ,\ \ \boldsymbol{s}_{P_2}(2)\ \mid\ {}_{3}\ \mid\ \boldsymbol{s}_1(T),\ \ \boldsymbol{s}_2(T),\ \ {}_{3}\ ,\ \ \boldsymbol{s}_{P_T}(T)\right). \tag{5}$$

Here 'T' signifies transpose of a matrix or a vector. Equations (3), (4), and (5) indicate that 3-D locations of all the Q feature points during the observation time $t=1,2,\ldots,T$ are calculated simultaneously from matrix $\widetilde{W}$ along with the orientations of the fixed F video cameras. It should be noted that the feature points commonly visible from the F video cameras at each observation time are recovered their 3-D locations. Those which are visible from less than F video cameras must therefore be eliminated from matrix W.

3 TRANSFORMATION BETWEEN REAR AND FRONT IMAGE PLANES

The deformation recovery technique described in the former section recovers partial 3-D shape of an object commonly visible from all the F video cameras. In order to recover the entire shape, F pairs of cameras are placed around the object in the proposed technique. Each pair of video cameras (called a front video camera and a rear video camera) face with each other beyond the object and their light axes are mutually set parallel. The feature points on the object projected onto the image plane of the rear video camera are transformed into the projected feature points on the image plane of the front video camera in order to make a single measurement matrix.

Let us take the fth pair of video cameras (See **Figure 1**). Video camera C_f takes the image of the front side of object O, whereas video camera D_f takes the image of the rear side of object O. Let us denote image planes of video cameras C_f and D_f by I_f and J_f, respectively. Suppose m feature points $\boldsymbol{s}_i(i=a_1,a_2,\ldots,a_m)$ are projected onto image plane I_f and n feature points $\boldsymbol{s}_j(j=b_1,b_2,\ldots,b_n)$ are projected onto image plane J_f. Locations of the projected feature points are denoted by $\boldsymbol{x}_i=(x_i,y_i)^{\mathrm{T}}$ and $\boldsymbol{\xi}_j=(\xi_j, \eta_j)^{\mathrm{T}}$, respectively. Their sets are denoted by $S_{If}=\{\boldsymbol{x}_i \mid i=1,2,\ldots,m\}$ and $S_{Jf}=\{\boldsymbol{\xi}_j \mid j= 1,2,\ldots,n\}$, respectively. Generally, $S_{If}\cap S_{Jf}\neq \phi$.

Assuming that at least three feature points are commonly observable on image planes I_f and J_f (f=1,2,…,F), affine transform A is defined employing these feature points between the two image planes. The feature point $\boldsymbol{\xi}_j$ in S_{Jf} are then transformed into the points on image plane I_f by

$$\hat{\boldsymbol{x}}_j = A\xi_j\,. \tag{6}$$

Thus the projected front feature points in S_{If} and the projected rear feature points in S_{Jf} are unified on the front image plane newly denoted by $\hat{I}_f$.

4 PRODUCING A SINGLE MEASUREMENT MATRIX

In the technique(Tan & Ishikawa, 1999), the images of a deformable object obtained from F video cameras define a single measurement matrix, which receives singular value decomposition to yield recovery of 3-D deformation process. In this particular technique, F successive video cameras, $f_k, f_{k1}, f_{k2},\ldots, f_{k,F\text{-}1}$ (k=1,2,…,2F) are chosen among 2F video cameras surrounding the object. Here kl=k+l (mod 2F) (l=1,2,…,F-1). With respect to each set of the F video cameras, the feature points visible from all the F video cameras at time t (t=1,2,…,T) are taken correspondence to yield a matrix $W(t)$ of the form given by Eq.(7) (Case of F=3).

$$W(t)=\begin{pmatrix}
x^1_{11}(t) & \cdots & x^1_{1P_1}(t) & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & x^5_{11}(t) & \cdots & x^5_{1P_5}(t) & x^6_{11}(t) & \cdots & x^6_{1P_6}(t) \\
x^1_{21}(t) & \cdots & x^1_{2P_1}(t) & x^2_{21}(t) & \cdots & x^2_{2P_2}(t) & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & x^6_{21}(t) & \cdots & x^6_{2P_6}(t) \\
x^1_{31}(t) & \cdots & x^1_{3P_1}(t) & x^2_{31}(t) & \cdots & x^2_{3P_2}(t) & x^3_{31}(t) & \cdots & x^3_{3P_3}(t) & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & x^2_{41}(t) & \cdots & x^2_{4P_2}(t) & x^3_{41}(t) & \cdots & x^3_{4P_3}(t) & x^4_{41}(t) & \cdots & x^4_{4P_4}(t) & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & x^3_{51}(t) & \cdots & x^3_{5P_3}(t) & x^4_{51}(t) & \cdots & x^4_{5P_4}(t) & x^5_{51}(t) & \cdots & x^5_{5P_5}(t) & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & x^4_{61}(t) & \cdots & x^4_{6P_4}(t) & x^5_{61}(t) & \cdots & x^5_{6P_5}(t) & x^6_{61}(t) & \cdots & x^6_{6P_6}(t) \\
y^1_{11}(t) & \cdots & y^1_{1P_1}(t) & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & y^5_{11}(t) & \cdots & y^5_{1P_5}(t) & y^6_{11}(t) & \cdots & y^6_{1P_6}(t) \\
y^1_{21}(t) & \cdots & y^1_{2P_1}(t) & y^2_{21}(t) & \cdots & y^2_{2P_2}(t) & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & y^6_{21}(t) & \cdots & y^6_{2P_6}(t) \\
y^1_{31}(t) & \cdots & y^1_{3P_1}(t) & y^2_{31}(t) & \cdots & y^2_{3P_2}(t) & y^3_{31}(t) & \cdots & y^3_{3P_3}(t) & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & y^2_{41}(t) & \cdots & y^2_{4P_2}(t) & y^3_{41}(t) & \cdots & y^3_{4P_3}(t) & y^4_{41}(t) & \cdots & y^4_{4P_4}(t) & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & y^3_{51}(t) & \cdots & y^3_{5P_3}(t) & y^4_{51}(t) & \cdots & y^4_{5P_4}(t) & y^5_{51}(t) & \cdots & y^5_{5P_5}(t) & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & y^4_{61}(t) & \cdots & y^4_{6P_4}(t) & y^5_{61}(t) & \cdots & y^5_{6P_5}(t) & y^6_{61}(t) & \cdots & y^6_{6P_6}(t)
\end{pmatrix} \tag{7}$$

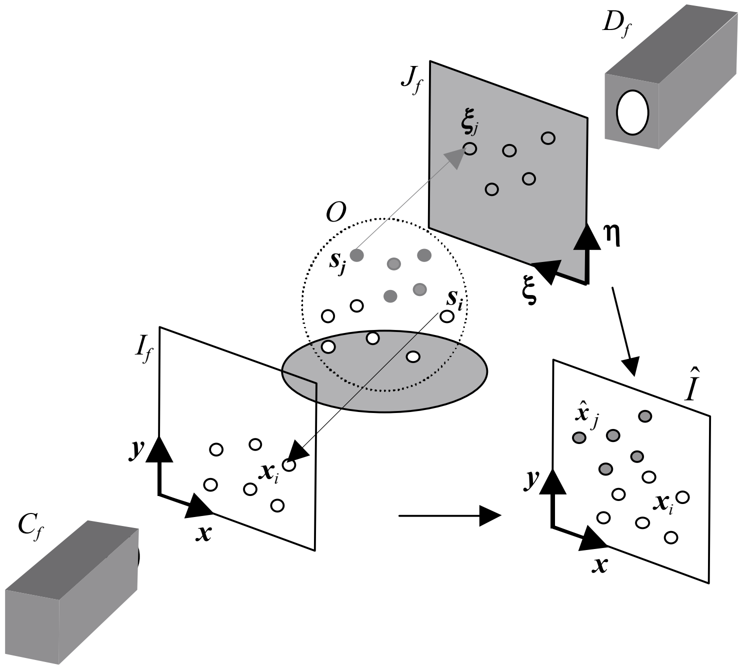

Figure 1. Integrating two images acquired by mutually facing cameras

Then, by the procedure stated in **3**, the projected feature points on the image of video camera *F*+*m* (*m*=1,2,…,*F*) are transformed into the points on the image of video camera *m* to yield the matrix *V*(*t*) of the form given by Eq.(8) (Case of *F*=3).

$$V(t)=\begin{pmatrix} x^1_{11}(t) & \cdots & x^1_{1P_1}(t) & \hat{x}^2_{41}(t) & \cdots & \hat{x}^2_{4P_2}(t) & \hat{x}^3_{41}(t) & \cdots & \hat{x}^3_{4P_3}(t) & \hat{x}^4_{41}(t) & \cdots & \hat{x}^4_{4P_4}(t) & x^5_{11}(t) & \cdots & x^5_{1P_5}(t) & x^6_{11}(t) & \cdots & x^6_{1P_6}(t) \\ x^1_{21}(t) & \cdots & x^1_{2P_1}(t) & x^2_{21}(t) & \cdots & x^2_{2P_2}(t) & \hat{x}^3_{51}(t) & \cdots & \hat{x}^3_{5P_3}(t) & \hat{x}^4_{51}(t) & \cdots & \hat{x}^4_{5P_4}(t) & \hat{x}^5_{51}(t) & \cdots & \hat{x}^5_{5P_5}(t) & x^6_{21}(t) & \cdots & x^6_{2P_6}(t) \\ x^1_{31}(t) & \cdots & x^1_{3P_1}(t) & x^2_{31}(t) & \cdots & x^2_{3P_2}(t) & x^3_{31}(t) & \cdots & x^3_{3P_3}(t) & \hat{x}^4_{61}(t) & \cdots & \hat{x}^4_{6P_4}(t) & \hat{x}^5_{61}(t) & \cdots & \hat{x}^5_{6P_5}(t) & \hat{x}^6_{61}(t) & \cdots & \hat{x}^6_{6P_6}(t) \\ y^1_{11}(t) & \cdots & y^1_{1P_1}(t) & \hat{y}^2_{41}(t) & \cdots & \hat{y}^2_{4P_2}(t) & \hat{y}^3_{41}(t) & \cdots & \hat{y}^3_{4P_3}(t) & \hat{y}^4_{41}(t) & \cdots & \hat{y}^4_{4P_4}(t) & y^5_{11}(t) & \cdots & y^5_{1P_5}(t) & y^6_{11}(t) & \cdots & y^6_{1P_6}(t) \\ y^1_{21}(t) & \cdots & y^1_{2P_1}(t) & y^2_{21}(t) & \cdots & y^2_{2P_2}(t) & \hat{y}^3_{51}(t) & \cdots & \hat{y}^3_{5P_3}(t) & \hat{y}^4_{51}(t) & \cdots & \hat{y}^4_{5P_4}(t) & \hat{y}^5_{51}(t) & \cdots & \hat{y}^5_{5P_5}(t) & y^6_{21}(t) & \cdots & y^6_{2P_6}(t) \\ y^1_{31}(t) & \cdots & y^1_{3P_1}(t) & y^2_{31}(t) & \cdots & y^2_{3P_2}(t) & y^3_{31}(t) & \cdots & y^3_{3P_3}(t) & \hat{y}^4_{61}(t) & \cdots & \hat{y}^4_{6P_4}(t) & \hat{y}^5_{61}(t) & \cdots & \hat{y}^5_{6P_5}(t) & \hat{y}^6_{61}(t) & \cdots & \hat{y}^6_{6P_6}(t) \end{pmatrix} \tag{8}$$

The above matrix *V*(*t*) therefore contains the information on the entire shape of the object at time *t*.

By merging the matrices *V*(*t*) (*t*=1,2,…,*T*), we have

$$V=\big(V(1)\ \mid\ V(2)\ \mid\ \cdots\ \mid\ V(T)\big)\ . \tag{9}$$

This matrix contains the information on deformation process over the entire shape of the object during observation time *T*. Matrix *V* is transformed into matrix $\widetilde{V}$ in the same way as shown by Eqs.(1) and (2). Singular value decomposition is applied to matrix $\widetilde{V}$ yielding decomposition of matrix $\widetilde{V}$ into two matrices, *i.e.*,

$$\widetilde{V} = M \cdot S \ . \tag{10}$$

Shape matrix *S* gives the model of the entire 3-D shape of the object at each time *t* and the entire 3-D deformation of the object during observation time *T*.

5 EXPERIMENTAL RESULTS

A toy balloon is recovered its inflation and deflation process three-dimensionally employing the proposed technique. In this experiment, *F*=3, *i.e.*, 6 video cameras surround the balloon, the light axes of each pair of video cameras being parallel with each other. Observation time is 27seconds: Sampling is done every 0.2 second: Hence *T*=135. As feature points, small colored papers are affixed on the surface of the balloon.

Correspondence of the feature points is found among every three images at each time *t* semi-automatically by calculating normalized correlation to yield matrix *W*(*t*) given by Eq.(7). With respect to every video camera pairs, affine transform is defined and the projected feature points on the rear image are transformed into the points on the front image by Eq.(6) to obtain matrix *V*(*t*) of Eq.(8). This procedure is iterated for *t*=1,2,…,*T* and matrix *V* of Eq.(9) is finally obtained. The total number of feature points chosen during the observation is 4,995(=*Q*).

Views of the deformation of the toy balloon from one of the 6 video cameras are shown in **Figure 3** at some sampled times. Results of the 3-D recovery are illustrated in **Figure 4**. Time flows in the direction of the arrows. Computation time from feeding *V* of Eq.(9) into the program to obtaining the shape matrix *S* in Eq.(10) is 33.2 seconds by PC (PentiumII:330MHz).

6 DISCUSSION

The proposed technique is an extension of the factorization method (Tomasi & Kanade, 1992). The idea of storing feature points at different times into a single measurement matrix has extended factorization to shape recovery of deformable objects (Tan *et al.*, 1998, Tan & Ishikawa, 1999). Further extension has been done in the present paper by integrating the front and the rear feature points on the respective video images and yielding a single measurement matrix. This measurement matrix, also called an extended measurement matrix, is a novel idea contained in the present technique.

The recovery error defined by Tan *et al.*(1998) was approximately 3.1% in average in the performed experiment. The errors mainly result from the assumption of orthographic projection and positional displacements when finding correspondence of feature points among the video images. A number of performed experiments employing various rigid objects and human motions always provide us with 3 to 4% of recovery errors. In this sense, the experiment gives a satisfactory result. Although the recovery errors are larger in the proposed technique than in existent techniques, one is almost able to escape from the laborious camera calibration in use of the present technique. It is another advantage that, unlike existent techniques, the proposed technique is a registration-free technique.

The technique may have applications to modeling human realistic motions for VR systems, TV games, or for man-machine interfaces. It may as well be applied to analyzing not only athletic or dancing motions but also analyzing those motions of the aged or the handicapped.

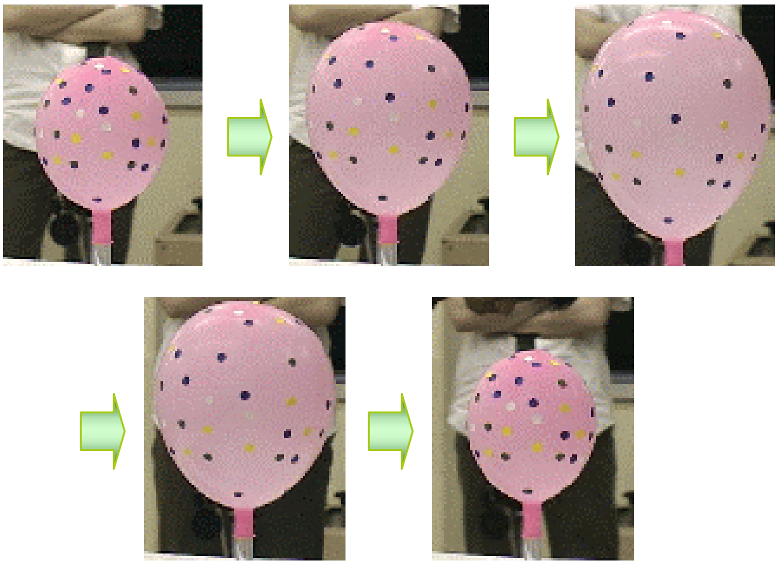

Figure 3. Views of a toy balloon becoming inflated and deflated as time flows in the direction of arrows

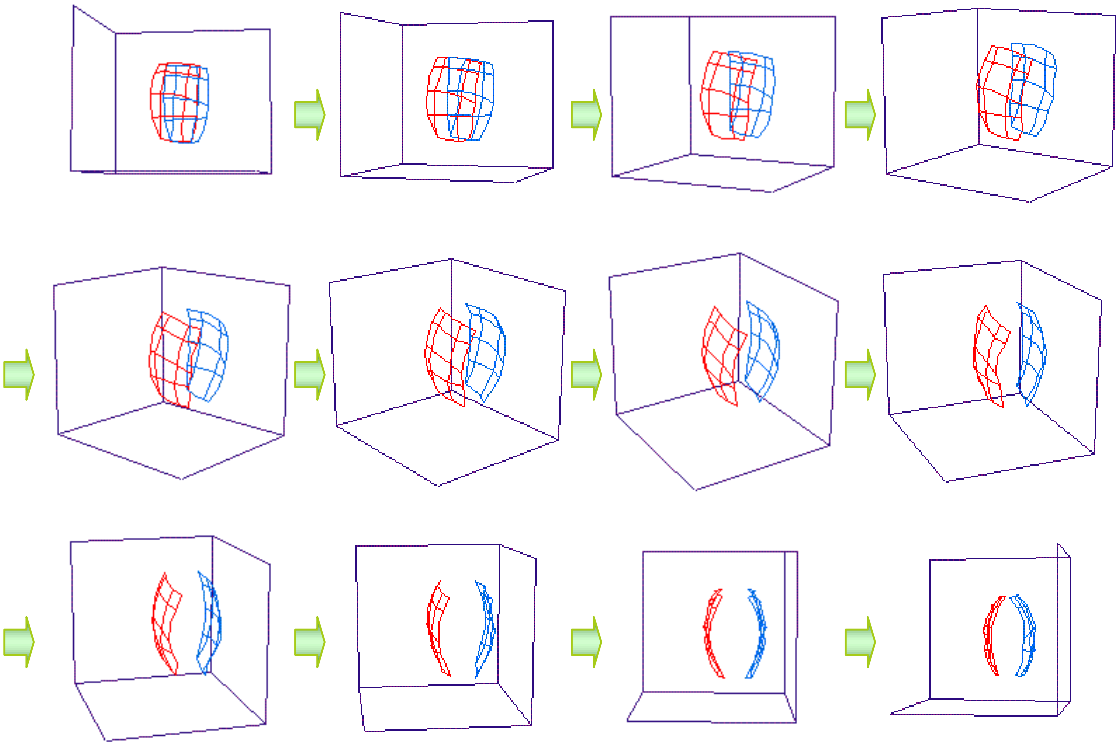

Figure 4. Recovered deformation of the balloon: Time proceeds as indicated by arrows

7 CONCLUSIONS

A technique was proposed for recovering 3-D entire shape and deformation of an object employing surrounding video cameras. More than three pairs of video cameras take video images of the object concerned. The feature points in the front image stream and those in the rear image stream are integrated to yield a single measurement matrix to which factorization is applied. Advantages of the technique include that it is a calibration-free technique except for aligning the light axes of the facing front and rear video cameras and that it is a registration-free technique. The proposed technique may find various application fields such as modeling human realistic motions or analyzing motions of the aged or the handicapped as well as athletic or dancing motions.

REFERENCES

Horn, B. K. P., 1986. Robot vision, MIT Press, Cambridge, MA.

Tan, J. K., Kawabata, S., Ishikawa, S., 1998. An efficient technique for motion recovery based on multiple views, Proc. IAPR Workshop on Machine Vision Applications, pp.270-273.

Tan, J. K., Ishikawa, S., 1999. Extracting 3-D motions of individuals at work by uncalibrated multiple video cameras, Proc. 1999 IEEE Int. Conf. on Systems, Man and Cybernetics, pp.III-487-490.

Tomasi, C., Kanade, T., 1992. Shape and motion from image streams under orthography: A factorization method, International Journal of Computer Vision, 9(2), pp.137-154.