3-D SHAPE RECONSTRUCTION FROM ENDOSCOPE IMAGE SEQUENCES BY THE FACTORIZATION METHOD

Koichiro DEGUCHI, Tsuyoshi SASANO,
Himiko ARAI, Yutaka YOSHIKAWA
Faculty of Engineering
University of Tokyo

Hongo 7-3-1, Bunkyo-ku, Tokyo 113 Japan

Abstract

A new application of the factorization method is reported for 3-D shape reconstruction from endoscope image sequences. The feasibility of the method is verified with some theoretical considerations and results of many kinds of extensive experiments.

INTRODUCTION

In this paper, we present some theoretical considerations and results of several kind extensive experiments on the factorization method to verify the feasibility of the method to reconstruct 3-D shapes from endoscope image sequences.

The purpose is to reconstruct inner wall shapes from images observed by an endoscope moving within human stomach(**Fig.1(a)**). For this case, it should be noted that the movement of the camera head cannot be controlled or accurately measured. So that, we must also estimate its movement from the image sequence to reconstruct the object shape.

The factorization method was developed by Tomasi and Kanade, and improved by Poelman and Kanade [1, 2], which intends to achieve high accuracy of shape reconstruction for such cases where the observing camera movement was unknown. This method uses a large number of points and image frames, and robustly applies a well-understood matrix computations.

There have been proposed several techniques for such a problem. These existing solutions for the structure from motion problem (Fig.1(b)) work well for perfect images without reading noise, but it is common knowledge that they are very sensitive to the noise. There reported some techniques for noisy data based on the Kalman filter etc.[3], but they were too complicated to understand what is the essential to achieve their performances.

Among them, the factorization method provided a possibility to achieve the robust accuracy by using a large number of points and image frames. However, the latter half process of the method, named *normalization*, was not so well-understandable as the use of singular value decomposition in its first half. Actually, as shown in this paper, many choices are possible for this normalization and a variety of results have been obtained according to the choice.

We admit that this method is easy to understand, easy to implement, and providing enough accuracy for the case where the approximation of the optical system holds well. But, the detail theoretical basis has been open to study.

In this paper, we focus on the normalization process of the factorization method, and present some kinds of strategies for the *fitting* with theoretical considerations. Then propose some formulations of the metric constraints of camera movements for the normalization, and also propose criteria on the constraints for more sophisticated projection models, the Scaled Orthographic Projection and Paraperspective Projection as well as for the originally proposed Orthgraphic Projection.

Then, we apply the method for the 3D shape reconstruction of inner walls of human stomach from endoscope images.

FACTORIZATION METHOD

Given an image sequence for a target object, as shown in Fig.1(b), we are supposing to have tracked P feature points over F frames. We then obtain trajectories of image coordinates $\{(x_{fp}, y_{fp}) | f = 1, \ldots, F,\ p = 1, \ldots, P\}$. We organize all the feature coordinates (x_{fp}, y_{fp}) into a $2F \times P$ *measurement matrix* W' as

$$W' = \begin{array}{c} \begin{matrix} \longleftarrow & P & \longrightarrow \end{matrix} \\ \begin{bmatrix} x_{11} - cx_1 & \cdots & x_{1P} - cx_1 \\ \vdots & \ddots & \vdots \\ x_{F1} - cx_F & \cdots & x_{FP} - cx_F \\ y_{11} - cy_1 & \cdots & y_{1P} - cy_1 \\ \vdots & \ddots & \vdots \\ y_{F1} - cy_F & \cdots & y_{FP} - cy_F \end{bmatrix} \end{array} \begin{matrix} \uparrow \\ \\ 2F \\ \\ \downarrow \end{matrix} \qquad (1)$$

where (cx_f, cy_f) is the center-of-mass of the image coordinates of points in the f-th frame, and

$$cx_f = \overline{x}_f = \frac{1}{P}\sum_{p=1}^{P} x_{fp}, \qquad cy_f = \overline{y}_f = \frac{1}{P}\sum_{p=1}^{P} y_{fp} \qquad (2)$$

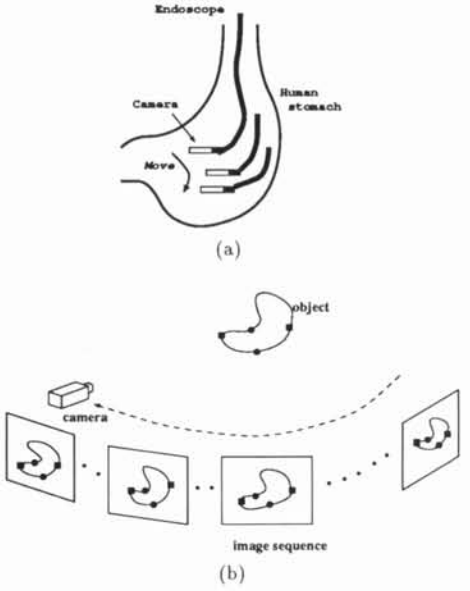


Figure 1: (a)Observation by endoscope moving within human stomach. (b)Image sequence for a rigid object obtained by a moving endoscope camera. P feature point on the object are tracked over F frames.

The essential of the factorization method is that these image coordinates can be rewritten in forms as

$$x_{fp} = \boldsymbol{m}_f^T \boldsymbol{s}_p + cx_f, \qquad y_{fp} = \boldsymbol{n}_f^T \boldsymbol{s}_p + cy_f \tag{3}$$

under some approximations of the perspective imaging system described below.

Approximation of the imaging system

Let us define the coordinate systems in the imaging system as shown in **Fig.2**. We call $S = [\boldsymbol{s}_1, \cdots, \boldsymbol{s}_P]$ the *shape matrix*, $\boldsymbol{t}_f$ the *camera location vector* (of the f-th frame), and $C_f = [\boldsymbol{i}_f, \boldsymbol{j}_f, \boldsymbol{k}_f]^T$ the *camera pose matrix* (of the f-th frame). For the simplicity, we set the world origin at the center-of-mass of the object feature points, i.e.,

$$\sum_{p=1}^{P} \boldsymbol{s}_p = \boldsymbol{0} \tag{4}$$

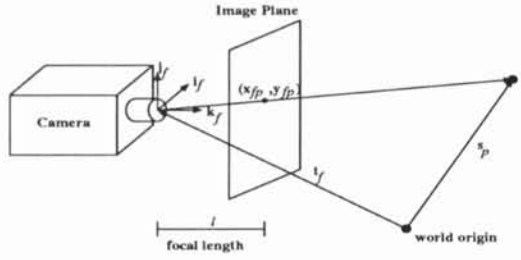


Figure 2: Coordinate system

The main idea of the factorization method is to employ approximation models with linear formulations for this perspective imaging to avoid the computational complexity and un-stability at the small cost of the approximation errors. We employ next three approximation models, where the respective relations hold.

For the *Orthographic Projection Model* which is shown in **Fig.3**(a),

$$\begin{aligned} cx_f &= -\boldsymbol{i}_f^T \boldsymbol{t}_f, \qquad & cy_f &= -\boldsymbol{j}_f^T \boldsymbol{t}_f \\ \boldsymbol{m}_f &= \boldsymbol{i}_f, & \boldsymbol{n}_f &= \boldsymbol{j}_f \end{aligned} \tag{5}$$

For the second *Scaled Orthographic Projection Model* shown in Fig.3(b), letting $z_f = -\boldsymbol{k}_f^T \boldsymbol{t}_f$,

$$\begin{aligned} cx_f &= -(l/z_f)\boldsymbol{i}_f^T \boldsymbol{t}_f, \qquad & cy_f &= -(l/z_f)\boldsymbol{j}_f^T \boldsymbol{t}_f \\ \boldsymbol{m}_f &= (l/z_f)\boldsymbol{i}_f, & \boldsymbol{n}_f &= (l/z_f)\boldsymbol{j}_f \end{aligned} \tag{6}$$

For the third *Paraperspective Projection Model* shown in Fig.3(c), letting $z_f = -\boldsymbol{k}_f^T \boldsymbol{t}_f$,

$$\begin{aligned} cx_f &= -(l/z_f)\boldsymbol{i}_f^T \boldsymbol{t}_f, \qquad & cy_f &= -(l/z_f)\boldsymbol{j}_f^T \boldsymbol{t}_f \\ \boldsymbol{m}_f &= (l\boldsymbol{i}_f - cx_f \boldsymbol{k}_f)/z_f, & \boldsymbol{n}_f &= (l\boldsymbol{j}_f - cy_f \boldsymbol{k}_f)/z_f \end{aligned} \tag{7}$$

This implies that the measurement matrix W' can be decomposed into a product of two matrices, as

$$W' = MS = \begin{matrix} \uparrow \\ 2F \\ \downarrow \end{matrix} \overset{\leftarrow 3 \rightarrow}{\begin{bmatrix} \boldsymbol{m}_1^T \\ \vdots \\ \boldsymbol{m}_F^T \\ \boldsymbol{n}_1^T \\ \vdots \\ \boldsymbol{n}_F^T \end{bmatrix}} \overset{\leftarrow P \rightarrow}{\begin{bmatrix} \boldsymbol{s}_1 & \cdots & \boldsymbol{s}_P \end{bmatrix}} \begin{matrix} \uparrow \\ 3 \\ \downarrow \end{matrix} \tag{8}$$

This means next two facts: Firstly, W' is a product of $2F \times 3$ matrix M and $3 \times P$ matrix S, so that $\mathrm{rank}(W') = 3$. Secondly, S is the shape matrix itself, and M is the motion matrix because it contains only informations on the camera motion.

The camera pose matrix $C_f = [\boldsymbol{i}_f, \boldsymbol{j}_f, \boldsymbol{k}_f]$ and its position $\boldsymbol{t}_f$ can be reconstructed from the matrix M ($\boldsymbol{t}_f$ for the orthographic projection model cannot be reconstructed)[1, 2, 4].

Decomposition by SVD Technique

If the approximations hold, based on the rank theorem, W' can be decomposed into $2F \times 3$ and $3 \times P$ matrices by using the singular value decomposition (SVD) technique.

From $\mathrm{rank}(W') \leq 3$, (8) can be rewritten into

$$\begin{matrix} \uparrow \\ 2F \\ \downarrow \end{matrix} \overset{\leftarrow P \rightarrow}{\begin{bmatrix} \\ W' \\ \\ \end{bmatrix}} = \begin{matrix} \uparrow \\ 2F \\ \downarrow \end{matrix} \overset{\leftarrow 3 \rightarrow}{\begin{bmatrix} \\ U' \\ \\ \end{bmatrix}} \Sigma \overset{\leftarrow P \rightarrow}{\begin{bmatrix} \\ V'^T \\ \\ \end{bmatrix}} \begin{matrix} \uparrow \\ 3 \\ \downarrow \end{matrix} \tag{9}$$

where Σ is a 3×3 diagonal matrix whose diagonal values are non-zero singular values of W'.

Then, denoting

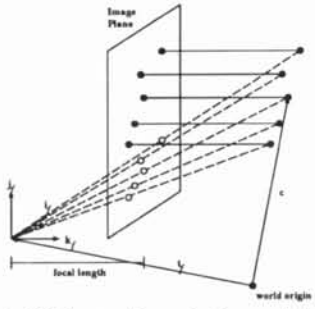
$$\hat{M} = U', \qquad \hat{S} = \Sigma\, V'^T \tag{10}$$

we have a decomposition of

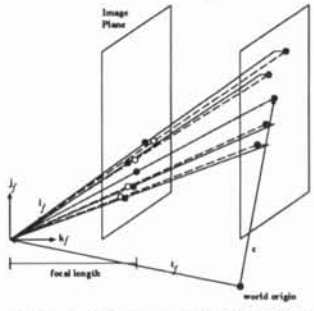
$$W' = \hat{M}\hat{S} \tag{11}$$

Constraints on unknown parameters

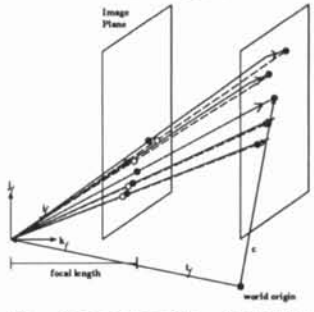
We just introduced a decomposition (11) of W', and this has the same form as (8). But this decomposition does not necessarily result in that $M = \hat{M}$ and $S = \hat{S}$, because, for any regular matrix A, $M' = \hat{M}A$ and $S' = A^{-1}\hat{S}$ satisfy $M'S' = W'$. Therefore, the next problem is to find the matrix which implies



(a) Orthographic projection model.



(b) Scaled Orthographic Projection Model.



(c) Paraperspective Projection Model.

Figure 3: Three approximation models for imaging system. (Broken lines indicate true perspective projection, and they are substituted with respective solid lines.)

$$M = \hat{M}A, \qquad S = A^{-1}\hat{S} \tag{12}$$

This process was called *normalization* in [1, 2]. It is claimed that this A can be determined from physical constraints on $\boldsymbol{m}_f$ and $\boldsymbol{n}_f$, which are led from that $\boldsymbol{i}_f$, $\boldsymbol{j}_f$, and $\boldsymbol{k}_f$ compose an orthonormal system.

This constraints are different for each respective models, and according to (5), (6), and (7) they are given as followings, respectively.

For Orthographic Projection Model

$$\begin{array}{l} \boldsymbol{m}_f^T\boldsymbol{m}_f = \boldsymbol{n}_f^T\boldsymbol{n}_f = 1 \\ \boldsymbol{m}_f^T\boldsymbol{n}_f = 0 \end{array} \qquad f = 1, \cdots, F \tag{13}$$

For Scaled Orthographic Projection Model

$$\begin{array}{l} \boldsymbol{m}_f^T\boldsymbol{m}_f = \boldsymbol{n}_f^T\boldsymbol{n}_f \left(= l^2/z_f^2\right) \\ \boldsymbol{m}_f^T\boldsymbol{n}_f = 0 \end{array} \qquad f = 1, \cdots, F \tag{14}$$

For Paraperspective Projection Model

$$\frac{\boldsymbol{m}_f^T\boldsymbol{m}_f}{l^2 + cx_f^2} = \frac{\boldsymbol{n}_f^T\boldsymbol{n}_f}{l^2 + cy_f^2} = \frac{\boldsymbol{m}_f^T\boldsymbol{n}_f}{cx_f\,cy_f}\left(= \frac{1}{z_f^2}\right) \qquad f = 1, \cdots, F \tag{15}$$

METRIC CONSTRAINTS FOR NORMALIZATION

Formulation of the constraints

Firstly, we formulate the metric constraints for respective models in common forms, and introduce criteria for the respective normalizations for the models of Scaled Orthographic Projection and Paraperspective Projection , besides Orthographic Projection model.

The problem here is to determine unknown matrix A introduced in the previous section. We described the constraints on A as (13), (14), and (15) for respective models. But they are not sufficient conditions to determine A uniquely, because, if $\hat{A}$ satisfies one of them, then $\hat{A}U$ satisfies it with any orthonormal matrix U. This means that the recovered camera pose is not absolute and open to rotation and mirror symmetry.

On the other hand, for any matrix $\hat{A}$, there exists an orthonormal matrix U and $\hat{A}U$ is symmetry. Therefore, for the solution of the constraints (13), (14), and (15), we can limit A to be symmetry.

Then we denote

$$A = \begin{bmatrix} x_1 & x_6 & x_5 \\ x_6 & x_2 & x_4 \\ x_5 & x_4 & x_3 \end{bmatrix} \tag{16}$$

$$\boldsymbol{x} = (x_1, \cdots, x_6) \tag{17}$$

and find $\boldsymbol{x}$, for A, which minimize a criterion on the metric constrains.

Criterion on the constraints

In our experiments, the criterion on the metric constraints for the respective models are given as followings.

Orthographic Projection Model The solution A or $\boldsymbol{x}$ given as (16) or (17) is uniquely determined for the condition of constraint (13) if noise free. Then we introduce two types of criterion on the constraint of (13) as

$$g_1(\boldsymbol{x}) = \sum_{f=1}^{F}\left[\left(\boldsymbol{m}_f^T\boldsymbol{m}_f - 1\right)^2 + \left(\boldsymbol{n}_f^T\boldsymbol{n}_f - 1\right)^2 + \left(\boldsymbol{m}_f^T\boldsymbol{n}_f\right)^2\right] \tag{18}$$

or

$$g_2(\boldsymbol{x}) = \sum_{f=1}^{F}\left[\left(\|\boldsymbol{m}_f\| - 1\right)^2 + \left(\|\boldsymbol{n}_f\| - 1\right)^2 + \left(\frac{\boldsymbol{m}_f^T\boldsymbol{n}_f}{\|\boldsymbol{m}_f\|\|\boldsymbol{n}_f\|}\right)^2\right] \tag{19}$$

Scaled Orthographic Projection Model For the Scaled Orthographic Projection model, as is the same case for Paraperspective Projection model, if $\hat{A}$ satisfies its constraints of (14) (or constraints of (15) for Paraperspective Projection), $k\hat{A}$ also satisfies the condition with any real number k. This means there remains an ambiguity of absolute size of the object or absolute distance to the object, that is, if we limit A to be symmetry, it is still up to a scale.

Then we minimize one of the next two criteria under the condition of $\det(A) = 1$.

$$g_3(\boldsymbol{x}) = \sum_{f=1}^{F}\left[\left(\boldsymbol{m}_f^T\boldsymbol{m}_f - \boldsymbol{n}_f^T\boldsymbol{n}_f\right)^2 + \left(\boldsymbol{m}_f^T\boldsymbol{n}_f\right)^2\right] \tag{20}$$

or

$$g_4(\boldsymbol{x}) = \sum_{f=1}^{F}\left[\left(\|\boldsymbol{m}_f\| - \|\boldsymbol{n}_f\|\right)^2 + \left(\frac{\boldsymbol{m}_f^T\boldsymbol{n}_f}{\|\boldsymbol{m}_f\|\|\boldsymbol{n}_f\|}\right)^2\right] \tag{21}$$

Paraperspective Projection Model As is the same case of the Scaled Orthographic Projection model, we minimize the next criterion under the condition of $\det(A) = 1$.

$$\begin{aligned} g_5(\boldsymbol{x}) = \sum_{f=1}^{F}\Bigg[& \left(\frac{cx_f\,cy_f}{l^2 + cx_f^2}\boldsymbol{m}_f^T\boldsymbol{m}_f - \boldsymbol{m}_f^T\boldsymbol{n}_f\right)^2 \\ & + \left(\frac{cx_f\,cy_f}{l^2 + cy_f^2}\boldsymbol{n}_f^T\boldsymbol{n}_f - \boldsymbol{m}_f^T\boldsymbol{n}_f\right)^2 \\ & + \left(\frac{cx_f\,cy_f}{l^2 + cx_f^2}\boldsymbol{m}_f^T\boldsymbol{m}_f - \frac{cx_f\,cy_f}{l^2 + cy_f^2}\boldsymbol{n}_f^T\boldsymbol{n}_f\right)^2\Bigg] \end{aligned} \tag{22}$$

Minimization of the criteria

To minimize g_1 and g_2, next three algorithms were used.

- Powell's method (*Conjugate Direction Method*)
- FRPR algorithm (*Conjugate Gradient Method*)
- DFP algorithm (*quasi-Newton Method* or *Variable Metric Method*)

which are offered in many numerical software packages.

For the conditional ($\det(A) - 1 = 0$) minimization of g_3, g_4, and g_5, the algorithm was the Lagrange's method of indeterminate multiplier, where the substantial minimization were carried out with above mentioned FRPR algorithm and DFP algorithm.

EXPERIMENTAL RESULTS

Experimental Simulation

The experiments here is to evaluate the performance improvements by introducing more sophisticated projection models, the Scaled Orthographic Projection and the Paraperspective Projection .

The first simulation was for the object of ten random points distributed within a rectangular parallelepiped. The readings of the image coordinates of these points were added with some Gaussian noise.

These situations are listed in **Table 1** with other assumed parameters. The details and the results for other set-up conditions are reported in [4].

The accuracies of the object shape reconstruction were evaluated using next error amounts. We denote the original true length between a pair of points p, q ($p < q$) with d_{pq} , and its reconstructed length with $\tilde{d}_{pq}$. That is, denoting the true shape matrix with $S = [\boldsymbol{s}_1, \cdots, \boldsymbol{s}_P]$,and its reconstruction with $\tilde{S} = [\tilde{\boldsymbol{s}}_1, \cdots, \tilde{\boldsymbol{s}}_P]$, we have

Table 1: Basic set-up of the simulations.

item	values
Object	Randomly distributed points within $50cm \times 50cm \times 5cm$
Number of points	$P = 10$
Projection	Perspective
Image error	Gaussian, with ave. 0, s.d. 0.25pixels
Number of frames	$F = 5, 6, 10, 20, 40$
Camera distance	$3.0m$ (constant)
Focal length	$l_{focal} = 20mm$
Pixel size	$\rho = 9.28\mu m$ (both in x and y)

Table 2: The reconstruction results of point object distributed within thick depth. The camera was controlled with fixed-eye movement.

Solution			Frame	Shape
Model	Criterion	Minimization	# F	Error e
Orthographic Projection	g_1	FRPR	5	0.0359
	g_1	FRPR	6	0.0361
	g_1	Powell	10	0.0366
	g_1	FRPR	20	0.0370
	g_1	Powell	40	0.0372
Scaled Orthographic Projection	g_3	FRPR	5	0.0379
	g_3	FRPR	6	0.0382
	g_3	FRPR	10	0.0387
	g_3	FRPR	20	0.0390
	g_3	FRPR	40	0.0392
Paraperspective Projection	g_5	FRPR	5	0.0878
	g_5	FRPR	6	0.0941
	g_5	FRPR	10	0.0971
	g_5	DFP	20	0.2783
	g_5	FRPR	40	0.1595

$$
\begin{aligned}
d_{pq} &= \|\boldsymbol{s}_p - \boldsymbol{s}_q\| \\
\widetilde{d}_{pq} &= \|\widetilde{\boldsymbol{s}}_p - \widetilde{\boldsymbol{s}}_q\| \\
& \quad p, q \in \{1, \cdots, P\}, \quad p < q
\end{aligned}
$$

We consider the ratio of them

$$
r_{pq} = \frac{\widetilde{d}_{pq}}{d_{pq}}, \qquad p, q \in \{1, \cdots, P\}, \quad p < q \tag{23}
$$

Its average $\overline{r}$ and the variance σ_r^2 can be expressed as

$$
\overline{r} = \frac{1}{P(P-1)/2} \sum_{p=1}^{P-1} \sum_{q=p+1}^{P} r_{pq} \tag{24}
$$

$$
\sigma_r^2 = \frac{1}{P(P-1)/2} \sum_{p=1}^{P-1} \sum_{q=p+1}^{P} (r_{pq} - \overline{r})^2 \tag{25}
$$

If this variance σ_r^2 is small, the reconstruction can be said to be well performed. To compare the results which have different average $\overline{r}$, we define the *shape reconstruction error* as

$$
e = \frac{\sigma_r}{\overline{r}} \tag{26}
$$

and we conclude that, the smaller this value was, the better accuracy we obtained.

The experimental results are shown in tables of **Table 2**.

Other than this, we have obtained many results of experimental simulations. Here, we summarize on on the reconstruction accuracies with respect to the conditions.

Accuracies with respect to the projection models The model which achieved the highest accuracy was the Orthographic Projection model. The next was by Scaled Orthographic Projection model, then Paraperspective Projection model. This order is thought to be caused by the weakness of the metric constraints on the respective models.

Accuracies with respect to the projection models The model which achieved the highest accuracy was the Orthographic Projection model. The next was by Scaled Orthographic Projection model, then Paraperspective Projection model. This order is thought to be caused by the weakness of the metric constraints on the respective models.

This is a kind of irony! ; Paraperspective Projection model has the highest fidelity to the true perspective projection among the models, next is Scaled Orthographic Projection model, and then Orthographic Projection model.

Accuracies with respect to the number of frames Generally, the minimum sufficient number of image frames were 5. In some case the accuracy was improved by using a few more frames. However, the model approximation error have not be overcome by increasing the number of frames.

Accuracies with respect to the camera movement In the simulations, we employed fixed-eye control and random movements. The accuracy of the camera pose recovery was higher for the fixed-eye control than the random movement. But, the accuracy of the object shape reconstruction was almost same among both movements.

Accuracies with respect to the depth of object When object feature points were distributed within thin range along the eye-line of camera, higher accuracy of the shape reconstruction was achieved. This can be concluded intuitively from the properties of the approximation models of the projection.

APPLICATION FOR REAL ENDOSCOPE IMAGES

We show results of shape reconstruction from real image sequence of endoscope in human stomach. For this case, its actual shape and movement of the endoscope camera could not be measured and were unknown. 10 frames in the sequence were used and 9 feature points were traced in the every image frames. A part of inner wall shape was reconstructed.

Fig.4 is a part of the image sequence.

The reconstruction was done with Orthographic Projection model, the criterion $g_1(\boldsymbol{x})$, and the conjugate gradient method for minimization. The obtained 3-D positions of these feature points are shown in **Fig.5**. **Fig.6** shows interpolated surface of these 3D positions of the reconstructed 9 feature points with 8th degree polynomial surface. The true shape could not known, but a shape of concave inner wall of stomach was observed.

CONCLUSIONS

Many kinds of extensive experiments to evaluate the performance improvements by introducing these models and criteria for them are reported by using synthetic simulation images and real images. The summary of the experimental results includes : The model which achieve the highest accuracy was the Orhtographic Projection model. Generally, the minimum sufficient number of image frames were 5, and in some case the accuracy was improved by using a few more frames. The accuracy of the camera pose recovery was higher for the fixed-eye control than the random movement, but the accuracy of the object shape reconstruction was almost same among both movements. When object feature points were distributed within thin range along the eye-line of camera, higher accuracy of the shape reconstruction was achieved.

We conclude these features promise feasibility to reconstruct 3-D shape from endoscope image sequences.

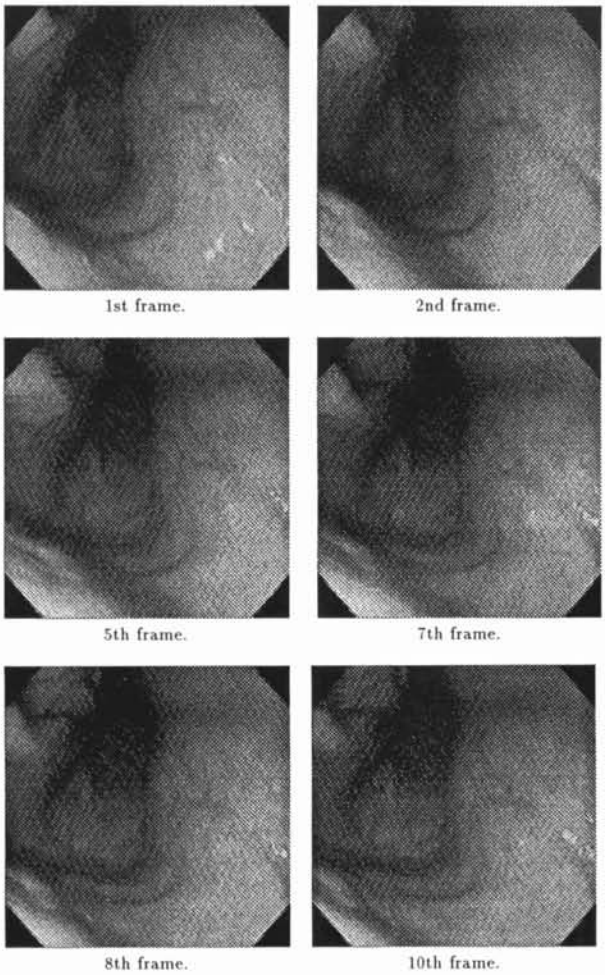


Figure 4: A part of image sequence taken with an endoscope in human stomach.

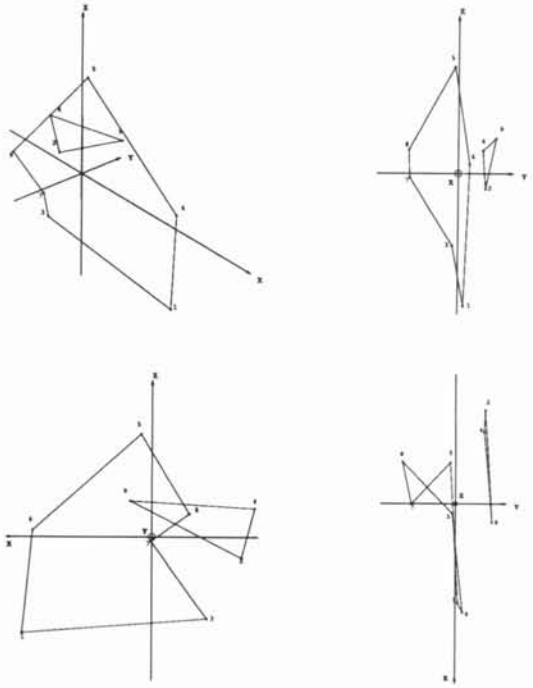


Figure 5: Reconstructed shape of inner wall of human stomach.

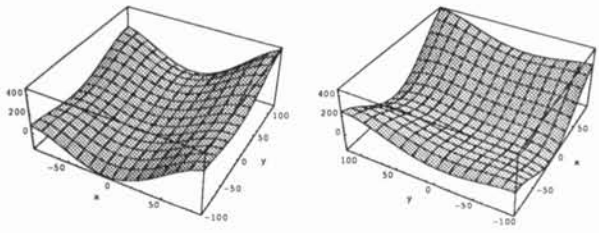


Figure 6: Reconstructed shape of inner wall of human stomach.

References

[1] C.Tomasi, and T.Kanade, "The Factorization Method for the Recovery of Shape and Motion from Image Streams," *Proceedings: Image Understanding Workshop*, pp.459-472, January 1992.

[2] C.J.Poelman, and T.Kanade, "A Paraperspective Factorization Method for Shape and Motion Recovery," *Technical Report*, CMU-CS-92-208, October 1992.

[3] N.Cui, *et al.*, "Extended Structure and Motion Analysis from Monocular Image Sequences," *Proc. 3rd ICCV*, pp.222-229, 1990.

[4] K.Deguchi, *et al.*, "Feasibility Study of The Factorization Method to Reconstruct Shape from Image Sequences," *Technical Report, Meip7-93017*, Dept. of MEIP, Univ. of Tokyo, 1994.

This work was supported in part by Nakatani Electronic Measuring Thechnology Association of Japan. The authors are grateful to the Association.