3-D Reconstruction from a Monocular Image Sequence by Tracking Markers and Natural Features

Tomokazu Sato, Masayuki Kanbara, Haruo Takemura* and Naokazu Yokoya

Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara 630-0101, Japan
E-mail: {tomoka-s, masay-ka, takemura, yokoya}@is.aist-nara.ac.jp

Abstract

Three-dimensional (3-D) reconstruction of a scene from an image sequence has been widely investigated for object recognition, robot navigation, mixed reality, and so on. However, there exist some problems concerning calculation cost and accuracy in the 3-D reconstruction from a long sequence of images that may contain features becoming invisible from time to time by occlusion. In this paper, we propose a new 3-D reconstruction method from a monocular image sequence, which uses a number of predefined markers whose positions in real world, color and shape are known, as well as unknown natural features. In this method, the extrinsic camera parameters and 3-D positions of natural features are estimated efficiently in every frame by tracking these markers and natural features automatically. Finally, the accumulation of estimation errors is minimized by optimization over the whole input. We demonstrate two experimental results of 3-D reconstruction from real image sequences to show the feasibility of the proposed method.

1 Introduction

Three-dimensional (3-D) reconstruction from an image sequence is widely used for object recognition, robot navigation, mixed reality, and so on. Shape-from-motion, which is a method based on features captured in a monocular image sequence, has attracted much attention. Many researchers have studied intensively 3-D reconstruction methods from monocular image sequences by tracking features.

One of the methods of shape-from-motion is the factorization algorithm [1] which is well known. In this method, using the linear approximation for camera model, a 3-D scene can be reconstructed stably and efficiently. However, when the 3-D scene is not suitable for affine camera model, reconstructed scene is distorted. Also the reconstruction is difficult when features are not observed in all frames of the input sequence.

Some of other shape-from-motion methods are based on non-linear optimization algorithms [2] [3]. They employ perspective projection as a camera model, and a 3-D scene is reconstructed by minimizing the error of points estimated in input images. Therefore, a 3-D scene can be reconstructed accurately even if the scene is not suitable for affine camera model, or if features are not observed in all frames of input sequence. However, a problem of calculation cost still exists.

Because of the difficulty in reconstructing 3-D scene and estimating camera motion at the same time from only a monocular image sequence, sensors such as gyro and knowledge such as camera path are often used [4] [5]. However, scenes which can be reconstructed are limited.

In order to solve these problems, we propose a new 3-D reconstruction method which uses a small number of predefined markers of known positions, colors, shapes and natural features in real world. In general, extrinsic camera parameters can be estimated by using the linear least square minimization method from at least six markers whose positions in real world and in the image are known. Therefore, the first frame of the input image sequence must contain six or more markers. It should be noted that intrinsic camera parameters must be estimated in advance.

In each frame, the extrinsic camera parameters and 3-D positions of natural features are estimated by tracking both markers and natural features automatically. The markers are tracked by using color and shape information, and natural features are tracked by using tentative camera parameters presumed by a

*Presently with Cybermedia Center, Osaka University.

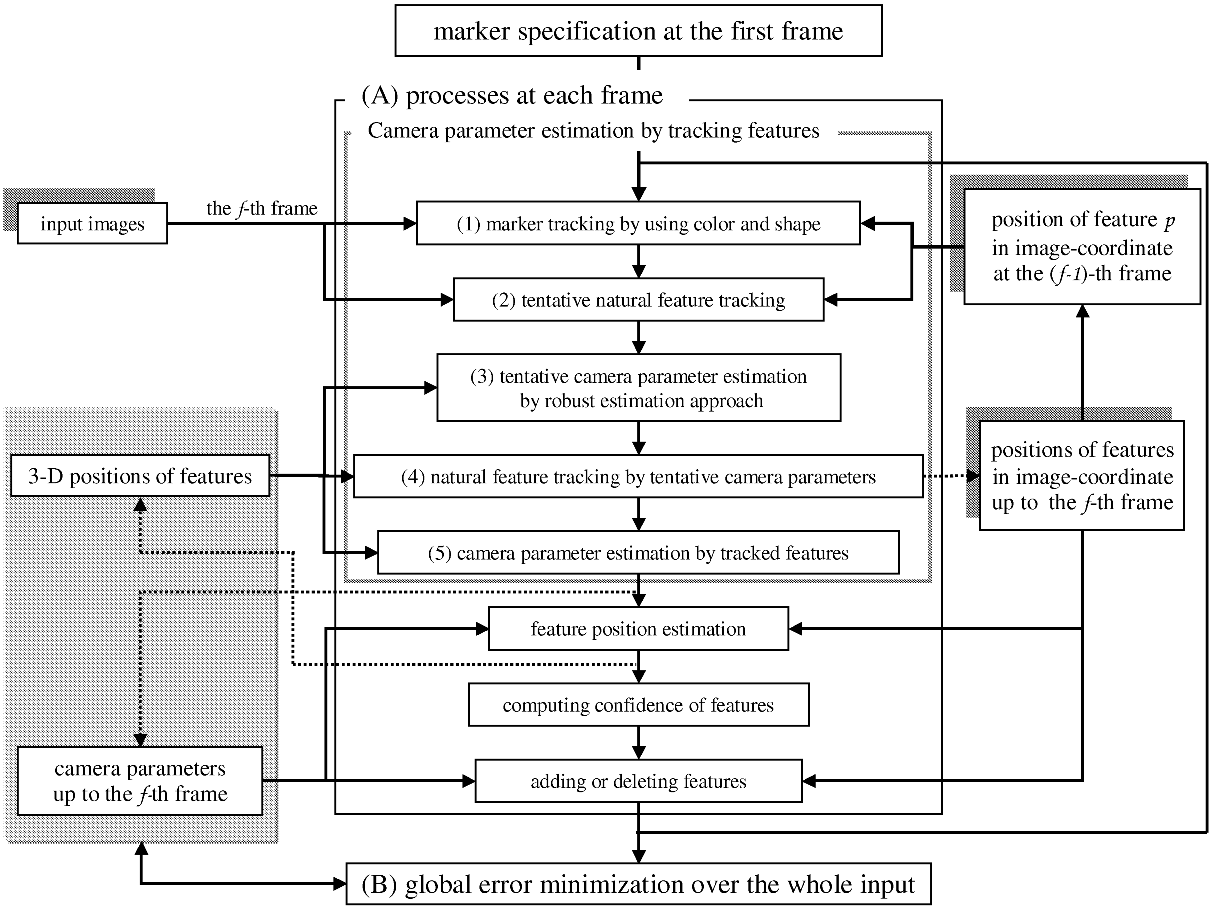

marker specification at the first frame
(A) processes at each frame
Camera parameter estimation by tracking features
input images
the f-th frame
(1) marker tracking by using color and shape
(2) tentative natural feature tracking
(3) tentative camera parameter estimation by robust estimation approach
3-D positions of features
(4) natural feature tracking by tentative camera parameters
(5) camera parameter estimation by tracked features
position of feature p in image-coordinate at the (f-1)-th frame
positions of features in image-coordinate up to the f-th frame
feature position estimation
computing confidence of features
camera parameters up to the f-th frame
adding or deleting features
(B) global error minimization over the whole input

Figure 1. Flow diagram of 3-D reconstruction.

robust estimation approach. Finally, the accumulation of estimation errors is minimized by optimization over the whole input. Using our approach, a 3-D scene can be reconstructed accurately by efficiently analyzing several hundreds images of a very long sequence, even if the sequence contains features becoming invisible from time to time.

This paper is structured as follows. Section 2 describes a method of tracking markers and natural features, and estimating camera parameters and natural feature positions in real world. In Section 3, we demonstrate two experimental results of 3-D reconstruction from real image sequences to show the feasibility of the proposed method. Finally, Section 4 gives conclusion and future work.

2 Reconstruction by tracking markers and natural features

This section describes a 3-D reconstruction method which is based on tracking features (markers and natural features). Figure 1 shows the flow diagram of our algorithm. First, we must specify the positions of six or more markers in the first frame of input sequence. Then the following processes are divided in two groups, one is camera parameter estimation and 3-D reconstruction at each frame (A) and the other is global error minimization over the whole input (B).

2.1 Feature tracking and camera parameter estimation

2.1.1 Tracking features

In this paper, we assume that the camera posture and position change greatly during input sequences, therefore how natural features look in the images may greatly change too. Tracking natural features only by image features usually suffers from two problems: One is that a center of tracked natural feature drifts because of accumulation of tracking error (a), the other is that a natural feature is tracked incorrectly when similar image pattern exists near by the feature (b).

To solve the problem (a), we employ Harris's interest operator [6] to detect corners or cross-points of edges in the input images. Local maxima of this operator are used as candidate positions of tracking features. For the problem (b), tentative camera parameters computed by robust estimation are used to limit the search region for natural feature tracking.

Processing steps for the f-th frame ($f \geq 2$) of input image sequence are described as follows.

(1) The markers used in the $(f-1)$-th frame are searched in the f-th frame by using color and shape information.

(2) A measure evaluating an interest point is computed by Harris's operator. Then the local maxima of the measure are selected as candidate positions of natural features. Every feature in the $(f-1)$-th frame is tentatively matched with candidate feature points which exist within a search window placed around the feature position in the $(f-1)$-th frame by using a standard template matching.

(3) Then the robust estimation is started. At the i-th iteration, first, k features $\mathbf{P}_i = \{p_{i1}, p_{i2}, \ldots, p_{ik}\}$ are randomly selected from the tentatively tracked natural features, and temporary camera parameter $\hat{\mathbf{M}}_i$ is estimated using $\mathbf{P}_i$. Next, the median RM_i of re-projection errors R_{ifp} is computed for estimated temporary camera parameter $\hat{\mathbf{M}}_i$. The re-projection error of feature p is defined as the square of distance between the tracked position $\mathbf{x}_{fp}$ and the position $\hat{\mathbf{x}}_{fp}$ that is the re-projected position of estimated feature position $\mathbf{S}_p$ onto the image at f-th frame using camera parameter $\hat{\mathbf{M}}_i$. The re-projection error R_{ifp} and the median RM_i of R_{ifp} are defined by the following equations.

$$R_{ifp} = |\mathbf{x}_{fp} - \hat{\mathbf{x}}_{fp}|^2. \tag{1}$$

$$RM_i = med(R_{if1}, R_{if2}, \ldots, R_{ifn}). \tag{2}$$

After g times iteration of these steps, the tentative camera parameter $\bar{\mathbf{M}}_f$ is selected from temporary camera parameters $\hat{\mathbf{M}}_i$ that minimizes the following LMedS criterion.

$$LMedS = min(RM_1, RM_2, \ldots, RM_g). \tag{3}$$

The algorithm for estimation of the camera parameter $\hat{\mathbf{M}}_i$ from tracked features $\mathbf{P}_i$ is described in Section 2.1.2.

(4) The features at the $(f-1)$-th frame are bound to the candidate positions in the f-th frame by searching the limited searching area. The center of the limited searching area is a position where S_p is projected onto the image by using camera parameter $\bar{\mathbf{M}}_f$.

(5) Finally, the camera parameter $\mathbf{M}_f$ at the f-th frame is determined by tracked features.

2.1.2 Extrinsic camera parameter estimation

In this section, a method of estimating extrinsic camera parameters from tracked features is described. In the proposed method, the re-projection error defined in Equation (1) is used as a measure for estimation error. The camera parameter $\mathbf{M}_f$ at the f-th frame is estimated by minimizing the estimation error E_f defined as follows:

$$E_f = \sum_p W_{fp} R_{fp}, \tag{4}$$

where W_{fp} is a weighting coefficient for the feature p at the f-th frame and is computed by considering the confidence that is described in Section 2.3. In this paper, we assume that a camera parameter has six degrees of freedom and its coordinate system is an orthogonal coordinate system.

Since estimating camera parameters is a non-linear minimization problem, there exist problems concerning local minima and calculation cost. In the first step, an initial camera parameter $\hat{\mathbf{M}}_f$ is estimated by a linear least square minimization method. However $\hat{\mathbf{M}}_f$ has twelve degrees of freedom. Next, the estimated camera parameter $\hat{\mathbf{M}}_f$ is linearly adjusted to reduce the degree of freedom to six by assuming that the direction of optical axis is correctly estimated. Finally, E_f is minimized by gradient descent from the adjusted camera parameter. Because the initial camera parameter is expected to be close to the true camera parameter, the estimation error E_f could be globally minimized.

2.2 Position estimation of natural features in real world

The position $\mathbf{S}_p$ of the natural feature p in real world is estimated from $\mathbf{x}_{fp}$ and $\mathbf{M}_f$ that has already been determined. The position $\mathbf{S}_p$ is computed by minimizing a sum of distances between $\mathbf{S}_p$ and straight lines that connect the center of projection in f-th frame and position $\mathbf{x}_{fp}$ of feature p in the image as shown in Figure 2.

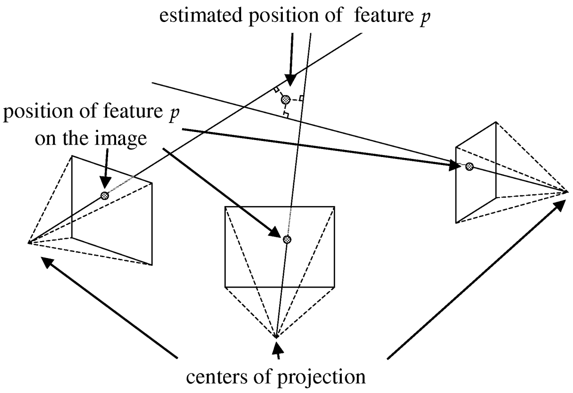

estimated position of feature p
position of feature p on the image
centers of projection

Figure 2. Estimating 3-D position of natural feature in real world.

2.3 Computing confidences of features

The confidence of features are used for the weighting coefficient of camrera parameter estimation. The position $\mathbf{x}_{(f+1)p}$ of feature p at the $(f+1)$-th frame does not usually correspond to the re-projected position $\hat{\mathbf{x}}_{(f+1)p}$ because of tracking error. We assume that the distribution of tracking error can be approximated by a Gaussian probability density function. The probability that $\mathbf{x}_{(f+1)p}$ corresponds to the true position is represented as follows:

$$p(\mathbf{x}_{(f+1)p}) = \frac{1}{2\pi\sigma_p^2} exp(-\frac{|\mathbf{x}_{(f+1)p} - \hat{\mathbf{x}}_{(f+1)p}|^2}{2\sigma_p^2}). \quad (5)$$

The total probability $P_{(f+1)}$ for all the features is given by the following equation.

$$P_{(f+1)} = \prod_p p(\mathbf{x}_{(f+1)p}). \quad (6)$$

The camera parameter $\mathbf{M}_{(f+1)}$ that maximizes above $P_{(f+1)}$ is obtained by minimizing

$$EM_{(f+1)} = \sum_p \frac{|\mathbf{x}_{(f+1)p} - \hat{\mathbf{x}}_{(f+1)p}|^2}{2\sigma_p^2}, \quad (7)$$

where σ_p^2 is computed by re-projection errors up to the f-th frame. The confidence W_{fp} of feature p that is tracked from $(f-k)$-th to f-th frame is defined by comparing Equations (4) and (7) as follows:

$$W_{(f+1)p} = \frac{1}{2\sigma_p^2} = \frac{k+1}{2} \left\{ \sum_{i=f-k}^{f} |\mathbf{x}_{ip} - \hat{\mathbf{x}}_{ip}|^2 \right\}^{-1} \quad (8)$$

2.4 Addition and deletion of natural features

Feature candidates that satisfy all the following conditions are added to the set of natural features at every frame.

- The confidence is over a given threshold.
- The matching error is less than a given threshold.
- The output value of Harris's operator is more than a given threshold.
- The maximum angle between lines that connect the center of projection and estimated 3-D position of the feature candidate is more than a given threshold.

On the other hand, natural features which satisfy at least one of the following conditions are deleted at every frame.

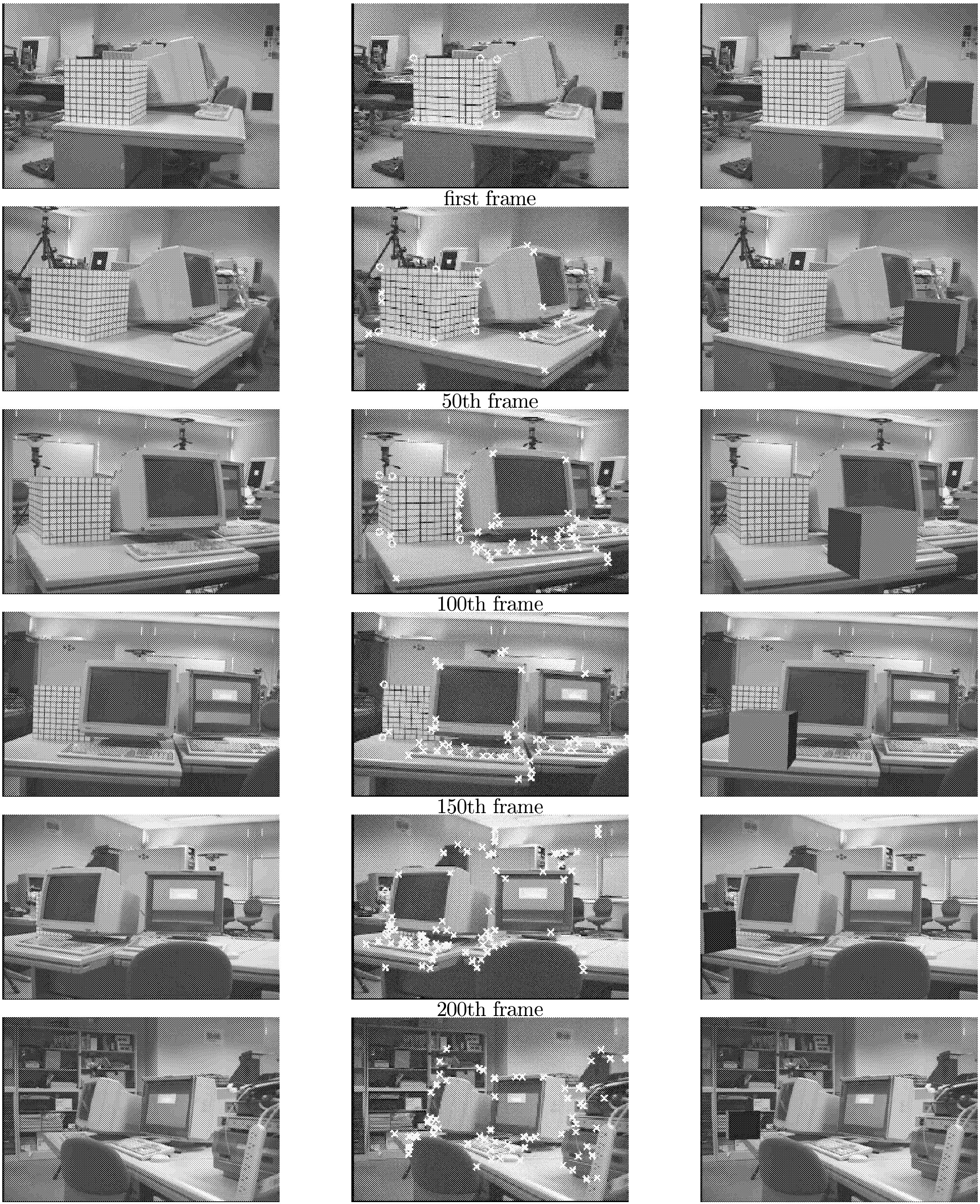
- The confidence is under a given threshold.
- The matching error is more than a given threshold.

2.5 Error minimization in the whole input

By using the method described above, the camera parameters and the natural feature positions in real world can be estimated over the whole frames. However, the accumulation of estimation error occurs. Therefore, in the final step, the accumulation of estimation error is minimized by optimization over the whole input. The accumulated estimation error is given by the sum of re-projection errors as follows and is minimized by optimizing the camera parameter $\mathbf{M}_f$ and natural features position $\mathbf{S}_p$ over the whole input.

$$E = \sum_f \sum_p W_p |\mathbf{x}_{fp} - \hat{\mathbf{x}}_{fp}|^2. \quad (9)$$

The camera parameters and feature positions that are already estimated by earlier process for each frame are used for initial values in the minimization. W_p is a weighting coefficient for the feature p in the final frame of the image sequence. Note that, when the feature p is deleted in the f-th frame, $W_{(f-C)p}$ is used as W_p, and the positions of feature p from the $(f-C)$-th frame to f-th frame are not used for this optimization, where C is a constant, since the features during the period are considered to be unreliable. Because the initial values of parameters are considered to be close to the true values, the error E is expected to be globally minimized efficiently.



first frame

50th frame

100th frame

150th frame

200th frame

266th frame

(a) Input images

(b) Results of feature tracking

(c) Match Move

Figure 3. Experimental results with an indoor scene.

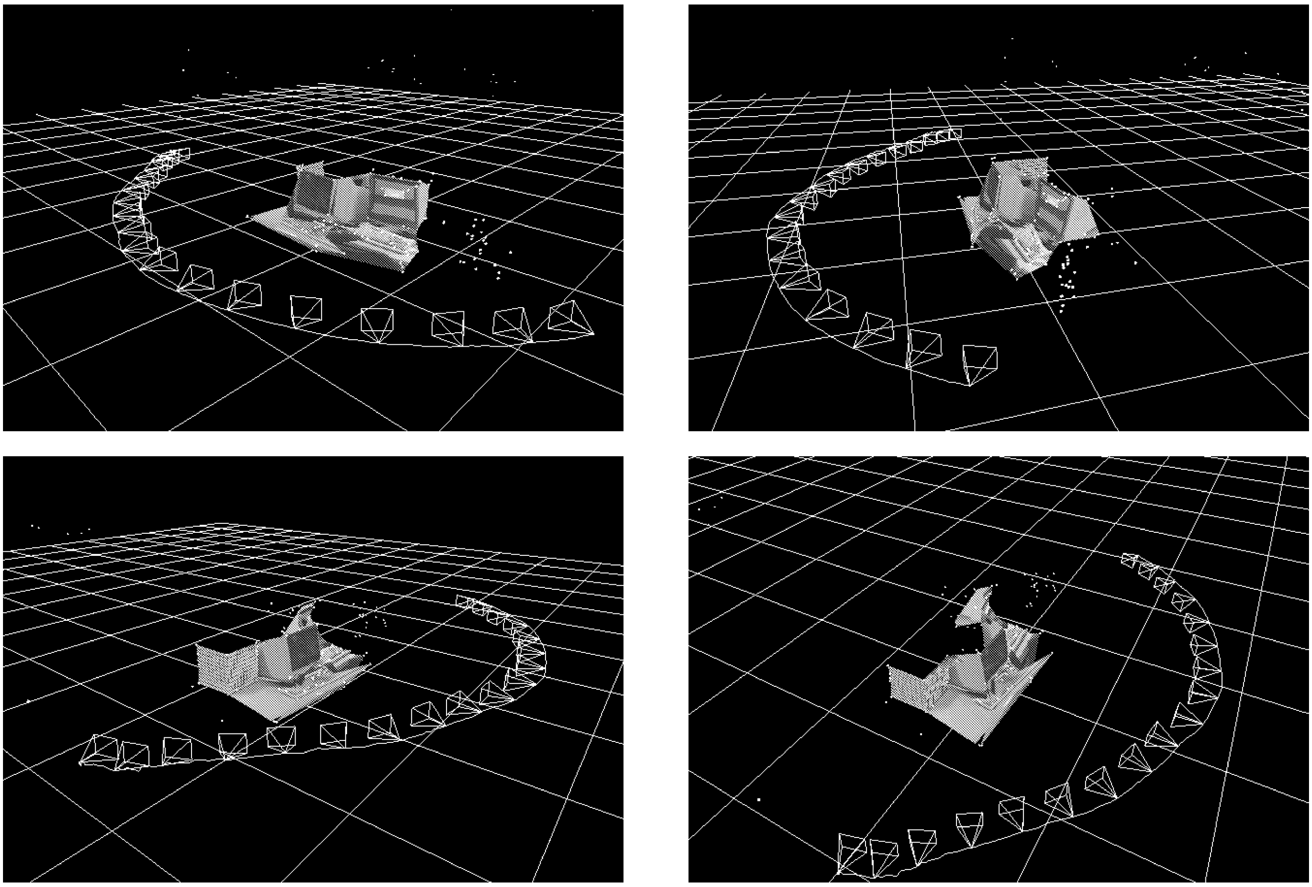


Figure 4. Results of camera parameter estimation, feature position estimation and indoor scene recovery.

3 Experiments

We have conducted two experiments: One is 3-D reconstruction from a short image sequence of an indoor scene and the other is from a long image sequence of an outdoor environment. The intrinsic camera parameters are estimated by Tsai's method [7] in advance, and the extrinsic camera parameters and the positions of natural features are estimated by the proposed method.

3.1 Experiment with indoor scene

In this experiment, we captured an indoor scene as shown in Figure 3(a) with a hand-held CCD camera (Sony DCR-VX 1000). This image sequence lasts 8.9 seconds and contains 267 frames (720×240 pixels). Although markers can be tracked by using color and shape information in our method, in this experiment markers are tracked manually.

Figure 3(b) shows the results of natural feature tracking. The markers and natural features are represented by white circles and white crosses, respectively. As shown in Figure 3(b), natural features are detected at corners and cross-points of edges. Figure 3(c) shows the results of Match Move. The Match Move is a demonstration to show the validity of estimated camera parameters by drawing a computer graphics (CG) object in the input sequence. Since the CG object stays still at the fixed position, we can conclude that camera parameters are correctly estimated.

Estimated camera parameters, feature positions and a reconstructed indoor scene shape are shown in Figure 4. Note that in Figure 4 the estimated camera path, posture and scene shape are illustrated assuming four different view points. The curved lines in this figure indicate the camera path and the quadrilateral pyramids indicate the camera postures drawn at every 15 frames. The reconstructed indoor scene is rendered by combining the polygons that are made from some camera positions and features in the image by using the Delaunay's triangulation method [8].

As shown in Figure 4, the estimated camera path is smooth. It should be noted that the camera path and posture are estimated even when the markers cannot be tracked. However, some errors are found in the estimated camera positions in early frames because there are few natural features in these frames. The shape of target object looks natural, even when it is viewed from the position that is far from original camera positions.

This experiment is carried out on a PC (CPU: Pentium III 1GHz, Memory: 512MB). The sum of calculation time for processes at each frame is 92 seconds for 8.9 seconds of the input sequence in Figure 3(a). The calculation time for global optimization is 130 seconds. The error E expressed in Equation (9) becomes 61.2% of its initial value with 500 iterations. When 1200 seconds are spent E becomes 58.9% with 5000 iterations. This shows that the almost optimal parameters are obtained with 500 iterations.

3.2 Experiment with outdoor scene

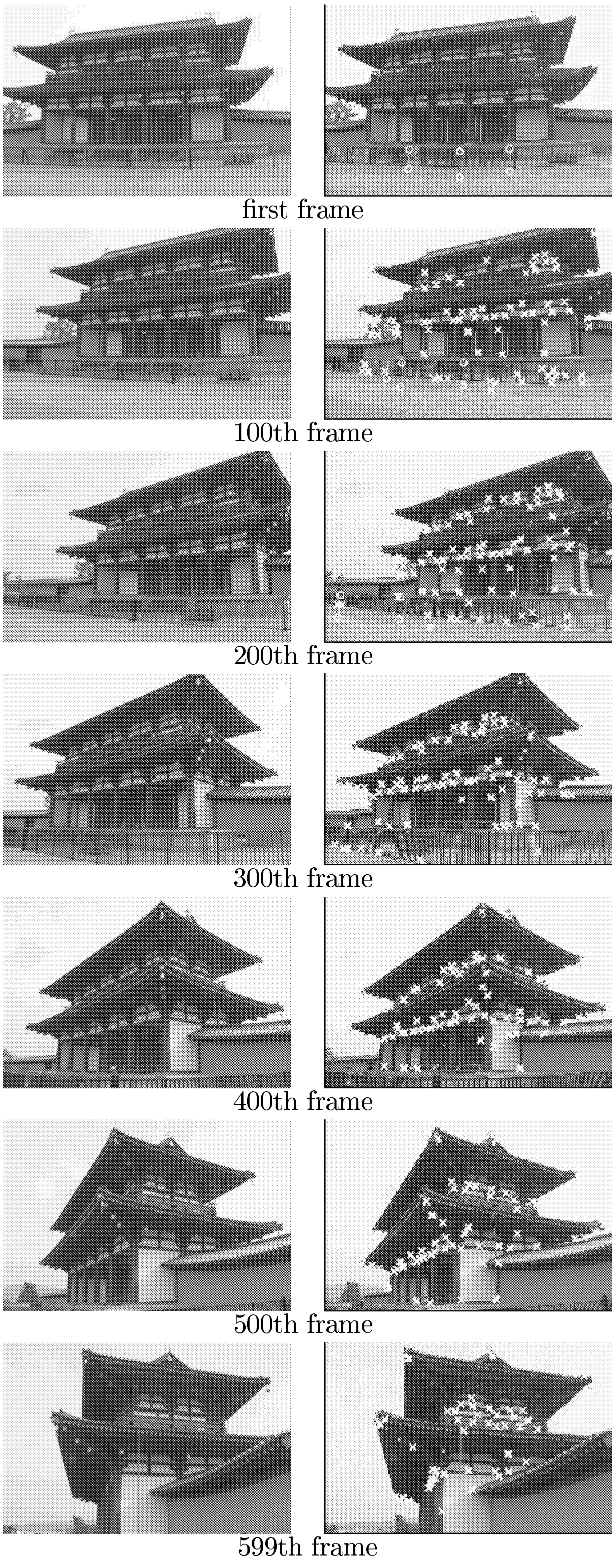
In this experiment, we captured an outdoor scene as shown in Figure 5(a) with a hand-held CCD camera (Sony DSR-DP-150) with a wide coversion lens (Sony VCL-HG0758). This image sequence lasts 40 seconds and has 599 frames (720×480 pixels, progressive scan).

Figure 5(b) shows the results of feature tracking. The markers and natural features are represented by white circles and white crosses, respectively. Note that six markers represented by circles are defined in the first frame as shown in Figure 5. Corners and cross-points of edges are automatically chosen as natural features with stable tracking in subsequent frames.

The estimated camera parameters and recovered scene shape are illustrated in Figure 6. The curved lines in this figure indicate the camera path and the quadrilateral pyramids indicate the camera postures drawn at every 50 frames. Note that the reconstructed shape is distorted around the roof of the building because there are no detected natural features on the roof. A simple triangulation using the Delauney's method is not suitable for a complex scene like this.

This experiment is carried out on the same PC as in Section 3.1. The sum of calculation time at each frame is 514 seconds for 40 seconds of the input sequence in Figure 5(a). The calculation time for global optimization is 580 seconds. The error E becomes 81.3% of its initial value with 500 iterations. It becomes 80.0% by 5000 iterations.

Figure 7 illustrates average reprojection errors of



first frame

100th frame

200th frame

300th frame

400th frame

500th frame

599th frame

(a) Input images

(b) Result of tracking features

Figure 5. Experimental results with an outdoor scene.

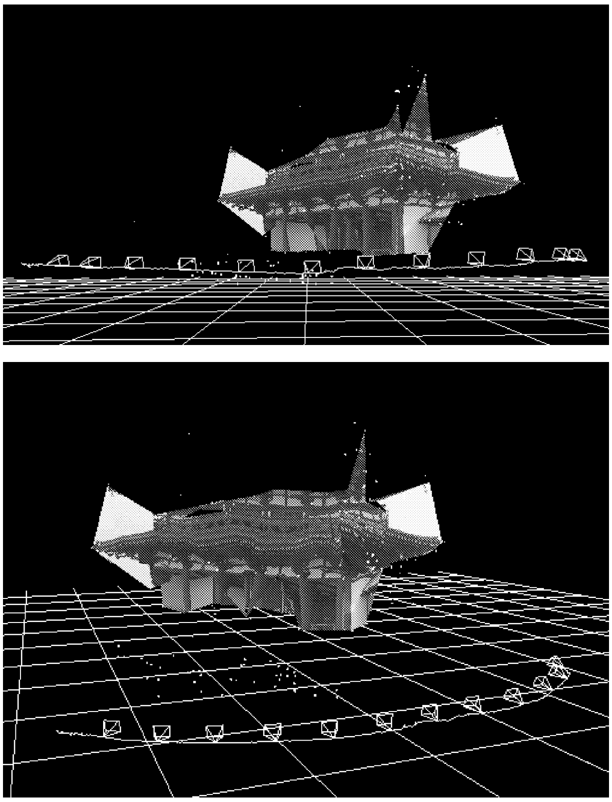


Figure 6. Results of camera parameter estimation, features position estimation and outdoor scene recovery.

the features after 500 iterations. The horizontal and vertical axes represent the frame number and average reprojection error, respectively. It is confirmed that the average re-projection errors are not accumulated so much; that is, the errors are less than 2.4 pixels throughout the sequence.

4 Conclusion

In this paper, a 3-D reconstruction method from a monocular image sequence is proposed. In this method, first, given markers in the first image frame are used so that initial camera parameters can be estimated. Then, at each frame, 3-D positions of natural features and camera parameters are estimated efficiently by tracking both markers and natural features automatically. Finally, the accumulation of estimation errors is minimized over the image sequence.

In the experiments, the 3-D scene reconstruction is accomplished for the image sequences captured in both indoor and outdoor environments successfully with automatically adding and deleting features regardless of the visibility of initial markers. However, when the target scene is complex, it is found that the reconstruction of shape by Delauney's method is not suitable. In future work, a dense depth mapping by multi-ocular stereo matching will be explored in order to reconstruct a 3-D shape. The objective and quantitative evaluation of the method is also planned.

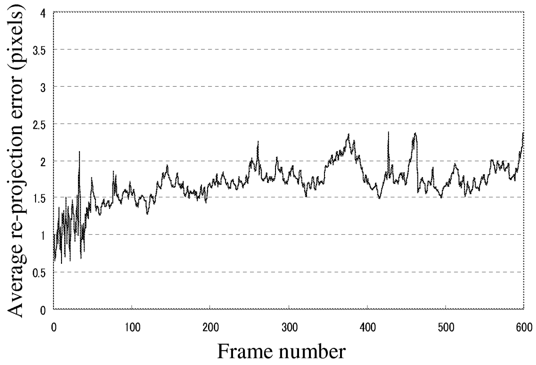


Figure 7. Average re-projection error of features after global optimization with an outdoor scene.

References

[1] C. Tomasi and T. Kanade: "Shape and Motion from Image Streams under Orthography: A Factorization Method," Int. Journal of Computer Vision, Vol. 9, No. 2, pp. 137–154, 1992.

[2] H. S. Sawhney, Y. Guo, J. Asmuth and R. Kumar: "Multi-view 3D Estimation and Application to Match Move," Proc. IEEE Workshop on Multi-view Modeling and Analysis of Visual Scenes, pp. 21–28, 1999.

[3] M. Pollefeys, R. Koch, M. Vergauwen, A. A. Deknuydt and L. J. V. Gool: "Three-dimentional Scene Reconstruction from Images," Proc. SPIE, Vol. 3958, pp. 215–226, 2000.

[4] T.Mukai and N.Ohnishi: "The Recovery of Object Shape and Camera Motion Using a Sensing System with a Video Camera and Gyro Sensor," Proc. 7th Int. Conf. on Computer Vision, pp. 411–417, 1999.

[5] W.Niem and J.Wingbermühle: "Automatic Reconstruction of 3D Objects Using a Mobile Monoscopic Camera," Proc. Int. Conf. on Recent Advances in 3D Imaging and Modeling, pp. 173–180, 1997.

[6] C. Harris and M. Stephens: "A Combined Corner and Edge Detector," Proc. Alvey Vision Conf., pp. 147–151, 1988.

[7] R. Y. Tsai: "An Efficient and Accurate Camera Calibration Technique for 3D Machine Vision," Proc. IEEE Conf. on Computer Vision and Pattern Recognition, pp. 364–374, 1986.

[8] P. Heckbert Ed.: Graphics Gems IV, pp. 47–59, Academic Press, 1994.