# 3-D Modeling of an Outdoor Scene by Multi-baseline Stereo Using a Long Sequence of Images

Tomokazu Sato†, Masayuki Kanbara†, Naokazu Yokoya† and Haruo Takemura‡

†Graduate School of Information Science, Nara Institute of Science and Technology
8916-5 Takayama, Ikoma, Nara 630-0101, Japan
{tomoka-s, kanbara, yokoya}@is.aist-nara.ac.jp
‡Cybermedia Center, Osaka University

## Abstract

*Three-dimensional (3-D) models of outdoor scenes are widely used for object recognition, navigation, mixed reality, and so on. Because such models are often made manually with high costs, automatic 3-D modeling has been investigated. A 3-D model is usually generated by using a stereo method. However, such approaches cannot use several hundreds images together for dense depth estimation because it is difficult to accurately calibrate a large number of cameras. In this paper, we propose a 3-D modeling method that first estimates extrinsic camera parameters of a monocular image sequence captured by a moving video camera, and then reconstructs a 3-D model of a scene. We can acquire a 3-D model of an outdoor scene accurately by using several hundreds input images.*

## 1. Introduction

Three-dimensional (3-D) models of outdoor scenes are widely used for object recognition, navigation, mixed reality, and so on. Because such models are often made manually with high costs, automatic 3-D modeling is desired.

One of approaches to the problem is to use an image sequence that is called shape-from-motion [1, 4, 5, 7]. The method can automatically recover camera parameters and 3-D positions of feature points by tracking natural features in captured images. Factorization algorithm [7] is one of the well known shape from motion methods that can estimate a rough 3-D scene stably and efficiently by assuming an affine camera model. However, when the 3-D scene is not suitable for the affine camera model, estimated camera parameters are not reliable. Although there exist other reconstruction methods [1, 4, 5], most of the methods reconstruct only a limited scene from a small number of images and are not designed to obtain a dense model. A method [4] which recovers camera parameters and a dense scene can reconstruct only a simple outdoor scene without occlusion from a small number of images.

In order to reconstruct a complex outdoor scene densely and stably, we propose a new 3-D reconstruction method that first recovers extrinsic camera parameters of an input image sequence that consists of several hundreds images, and then generates a 3-D model of a scene by combining several hundreds depth maps together in a voxel space. In the first process, we use a camera parameter estimation method [6]. This method uses a small number of predefined markers of known 3-D positions and many natural features for stable and efficient estimation of extrinsic camera parameters. Dense depth maps are then computed by using an extended multi-baseline stereo method. The proposed method can reconstruct a complex outdoor scene densely and accurately by using several hundreds images of a long sequence.

## 2. Camera parameter estimation by tracking features

This section describes an extrinsic camera parameter estimation method which is based on tracking features (markers and natural features). Figure 1 illustrates the flow diagram of our algorithm. First, we must specify the positions of six or more markers in the first frame of input sequence, and extrinsic camera parameters in the first frame are estimated. Then extrinsic camera parameters in all the frames are sequentially determined by iterating the processes at each frame (A). Finally, extrinsic camera parameters are refined by minimizing the accumulation of estimation errors over the whole input (B). Using this approach, we can estimate extrinsic camera parameters efficiently and accurately regardless of the visibility of initial markers.

### 2.1 Initial camera parameter estimation

By iterating the following processes from the first frame to the last frame, initial extrinsic camera parameters and 3-D positions of natural features are determined.

**(a) Marker and natural feature tracking.** Markers are

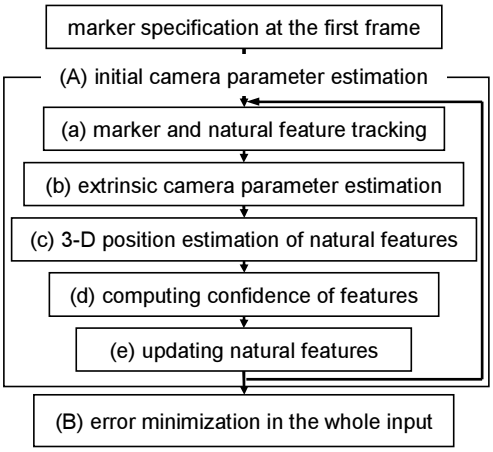



**Figure 1. Flow diagram of camera parameter estimation.**

tracked based on color and shape information. Natural features are tracked using a robust estimation approach by projecting the 3-D positions of natural features that are estimated until the previous frame [6]. Harris's interest operator is employed for robust tracking of natural features.

**(b) Extrinsic camera parameter estimation.** The re-projection error $R_{fp} = |\mathbf{x}_{fp} - \hat{\mathbf{x}}_{fp}|^2$ is used as a measure of estimation error, where $\mathbf{x}_{fp}$ is the tracked 2-D position of feature $p$ and $\hat{\mathbf{x}}_{fp}$ is the re-projected position of estimated 3-D position $\mathbf{S}_p$ of feature $p$ onto the image at the $f$-th frame by camera parameter $\mathbf{M}_f$. Then camera parameter $\mathbf{M}_f$ at the $f$-th frame is estimated by minimizing the estimation error $E_f = \sum_p W_{fp} R_{fp}$, where $W_{fp}$ is a weighting coefficient representing the confidence of the feature $p$ at the $f$-th frame.

**(c) 3-D Position estimation of natural features.** The position $\mathbf{S}_p$ of the natural feature $p$ in real world is estimated from multiple $\mathbf{x}_{fp}$ and $\mathbf{M}_f$ that have already been determined until the current frame. The position $\mathbf{S}_p$ is computed by minimizing the sum of squared distances between $\mathbf{S}_p$ and straight lines in 3-D that connect the centers of projection and the positions $\mathbf{x}_{fp}$ of feature $p$ in previous frames $f$.

**(d) Computing confidence of features.** We assume that the distribution of tracking errors can be approximated by a Gaussian probability density function. Then the confidence of feature, $W_{fp}$, is computed by the inverse of variance of re-projection error $R_{fp}$.

**(e) Updating natural features.** Feature candidates that satisfy all the following conditions are added to the set of natural features at every frame.

- The confidence $W_{fp}$ 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 estimated 3-D position of the feature candidate and centers of projection is more than a given threshold.

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

- The confidence $W_{fp}$ is under a given threshold.
- The matching error is more than a given threshold.

### 2.2 Error minimization in the whole input

In the final step, the accumulation of estimation errors defined as $E = \sum_f \sum_p W_p |\mathbf{x}_{fp} - \hat{\mathbf{x}}_{fp}|^2$ is minimized over the whole input with respect to the camera parameter $\mathbf{M}_f$ and natural feature positions $\mathbf{S}_p$. The camera parameter and feature positions that have already been estimated by earlier process for each frame are used as initial values. $W_p$ is a weighting coefficient for the feature $p$ in the final frame of the image sequence.

## 3. Dense 3-D reconstruction by hundreds-baseline stereo

In this section, we describe a dense 3-D reconstruction method using estimated camera parameters. First, a dense depth map for each image is computed by using a multi-baseline stereo method, then a 3-D model is reconstructed by combining obtained dense depth maps in a voxel space.

### 3.1. Dense depth estimation by multi-baseline stereo

A depth map is computed for each frame by using a multi-baseline stereo technique [3]. Depth value $z$ of pixel $(x, y)$ in the $f$-th frame is computed by using the $k$-th to the $l$-th frames ($k \leq f \leq l$) around the $f$-th frame. In the following expression, we assume the focal length as 1 for simplicity. Then, the 3-D position of the pixel $(x, y)$ can be expressed by $(xz, yz, z)$, and we can define the projected position $(\hat{x}_j, \hat{y}_j)$ of the 3-D position $(xz, yz, z)$ onto the $j$-th frame ($k \leq j \leq l$) as follows:

$$(a\hat{x}_j, a\hat{y}_j, a, 1)^T = \mathbf{M}_j \mathbf{M}_f^{-1} (xz, yz, z, 1)^T, \qquad (1)$$

where $a$ is a parameter. In the multi-baseline method, SSD (Sum of Squared Differences) is employed as an error function, that is computed as the sum of squared differences between the window $W$ in the $f$-th frame centered at $(x, y)$ and that in the $j$-th frame centered at $(\hat{x}_j, \hat{y}_j)$. We define the SSD function for the $j$-th frame in Eq. (2) using RGB components $(I_R, I_G, I_B)$.

$$\begin{aligned} SSD_{fj}(x, y, o_x, o_y) = & \\ \sum_{(u-o_x, v-o_y) \subseteq W} & \{(I_{Rf}(x+u, y+v) - I_{Rj}(\hat{x}_j + u, \hat{y}_j + v))^2 \\ & + (I_{Gf}(x+u, y+v) - I_{Gj}(\hat{x}_j + u, \hat{y}_j + v))^2 \\ & + (I_{Bf}(x+u, y+v) - I_{Bj}(\hat{x}_j + u, \hat{y}_j + v))^2\}, \end{aligned} \qquad (2)$$

where $o_x$ and $o_y$ are offsets of the window $W$ for $x$ and $y$ axes, respectively.

In the multi-baseline stereo method, the depth $z$ of $(x, y)$ is determined so as to minimize the SSSD (Sum of SSD) from the $k$-th frame to the $l$-th frame. We define a modified SSSD in Eq. (3) using the median of SSD because the template of window $W$ in the $f$-th frame may be occluded in other frames.

$$SSSD_f(x, y, o_x, o_y) = \sum_{j=k}^{l} \begin{cases} SSD_{fj}(x, y, o_x, o_y); \\ \quad SSD_{fj}(x, y, o_x, o_y) \leq T \text{ and } |j - f| > D, \\ 0; \quad \text{otherwise.} \end{cases} \quad (3)$$

where,

$$T = median(SSD_{fk}(x, y, o_x, o_y), \cdots, SSD_{f(f-D-1)}(x, y, o_x, o_y), SSD_{f(f+D+1)}(x, y, o_x, o_y), \cdots, SSD_{fl}(x, y, o_x, o_y)). \quad (4)$$

Note that images from the $(f - D)$-th frame to the $(f + D)$-th frame are not used for computing SSSD, because baselines in these frames are not long enough to estimate depth stably. Multiple centered window approach [2] is also used to reduce estimation errors around occlusion boundaries. SSSD is now extended to SSSDM as follows:

$$SSSDM_f(x, y) = \min_{(u,v) \subseteq W} (SSSD_f(x, y, u, v)). \quad (5)$$

We can estimate the depth value $z(x, y)$ correctly by minimizing SSSDM unless the pixel $(x, y)$ is occluded in more than $(l - k - 2D)/2$ frames. Additionally, we avoid a local minimum problem and achieve stable depth estimation using a multiscale approach [9].

### 3.2. 3-D model reconstruction in a voxel space

In this paper, a 3-D model is reconstructed in a voxel space by combining several hundreds dense depth maps. In the voxel space, each voxel has two values $A$ and $B$ which are voted by already estimated depth values and camera parameters. For each pixel $(x, y)$ in an image, both $A$ and $B$ are voted when the voxel is projected onto the pixel as follows. Value $A$ is voted if depth of the voxel in camera coordinate system is equal to $z$ of $(x, y)$. On the other hand, value $B$ is voted when depth of the voxel is equal to or less than $z$ of $(x, y)$. We use the ratio $A/B$ as a normalized voting value. A 3-D model is then reconstructed by selecting the voxel whose $A/B$ is more than a given threshold. Note that the color of the voxel is decided by computing a mean color of pixels that have been voted to the value $A$ of the voxel.

## 4. Experiments

We have conducted two experiments: One is a model reconstruction of a single building and the other is a reconstruction of a street scenery. Both scenes are complex and have many occlusions. In both experiments, we use a hand-held CCD camera (Sony VCL-HG0758) with a wide conversion lens (Sony VCL-HG0758). The intrinsic camera parameters are estimated by Tsai's method [8] in advance.

### 4.1. Reconstruction of building

In the first experiment, we captured a single building (Suzaku-mon) shown in Figure 2(a) by walking around the building viewing it at the center of image. This image sequence lasts 40 seconds and has 599 frames (720×480 pixels, progressive scan).

In this experiment, a dense depth map of the $f$-th frame is obtained by using every two frames from the $(f - 100)$-th to the $(f + 100)$-th frames excluding the $(f - 15)$-th to the $(f+15)$-th frames. Figure 2(b) shows computed dense depth maps in which depth values are coded in intensity. It is confirmed that correct depth values are obtained for most part of the images. Figure 3 shows a reconstructed 3-D model with textures obtained by combining 399 dense depth maps together in the way of voxel voting that is described in Section 3.2. In this experiment, the voxel space is constructed of 10cm cube voxels. It can be observed that a wall behind a column of the building is reconstructed even if the wall is occluded from time to time. We also observe that some positions are holed because they are not visible enough for sufficient precision in the image sequence.

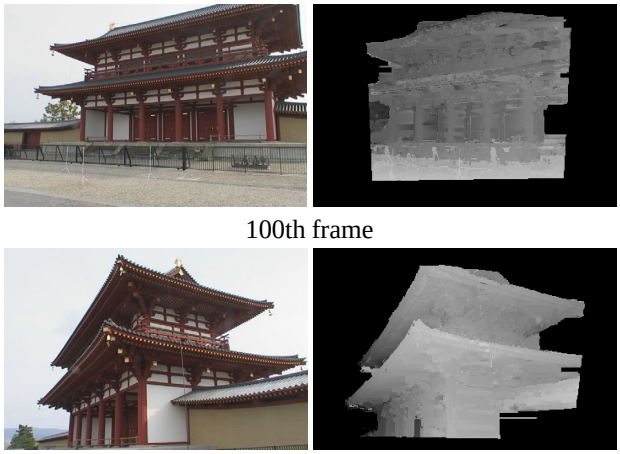
### 4.2. Reconstruction of street scenery

In the second experiment, a street is captured as shown in Figure 4(a) by the CCD camera put on a slowly moving car. This image sequence lasts 19 seconds and has 284 frames (720×480 pixels, progressive scan).

A dense depth map of the $f$-th frame is obtained by using 30 frames from the $(f + 6)$-th to the $(f + 35)$-th frames. As shown in Figure 4(b), it is confirmed that correct depth values are obtained for most part of the images even around the occlusion edges. However, there exist some incorrect depth values at the right of the trees because the areas are occluded by the trees during over 15 frames. Figure 5 shows a reconstructed 3-D model with textures obtained from 249 dense depth maps. In this experiment, the voxel space is constructed of 10cm cube voxels. Note that many parts of walls are holed around the windows of the buildings. We confirmed that it is difficult to reconstruct the reflective objects.

## 5. Conclusion

In this paper, a dense 3-D reconstruction method from a monocular image sequence captured by a hand-held video camera is proposed. In experiments, the dense 3-D scene reconstruction is successfully accomplished by using a long



100th frame

499th frame

(a) Input images (b) Dense depth maps

**Figure 2. Input images and estimated dense depth maps (Suzaku-mon).**

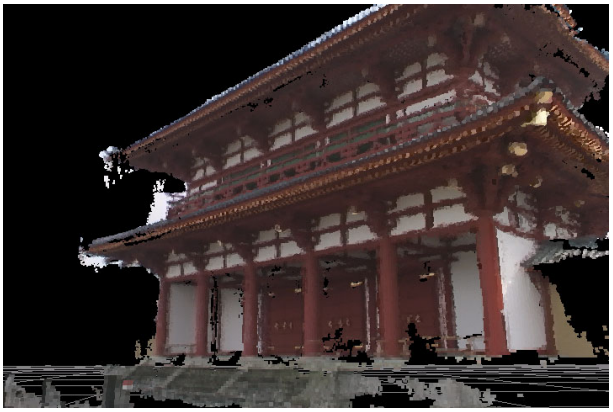
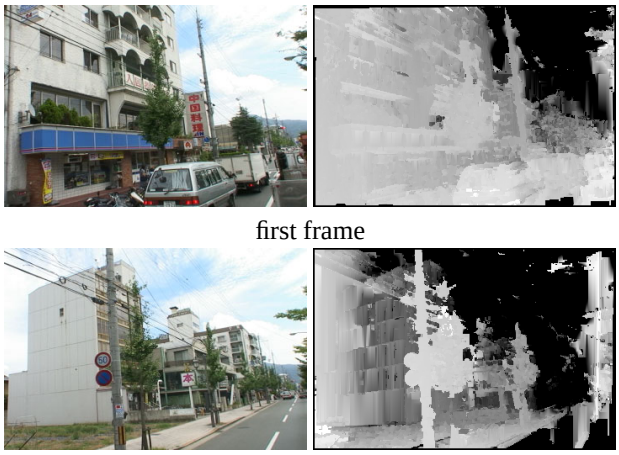


**Figure 3. Results of outdoor scene recovery (Suzaku-mon).**



first frame

249th frame

(a) Input images (b) Dense depth maps

**Figure 4. Input images and estimated dense depth maps (street scenery).**

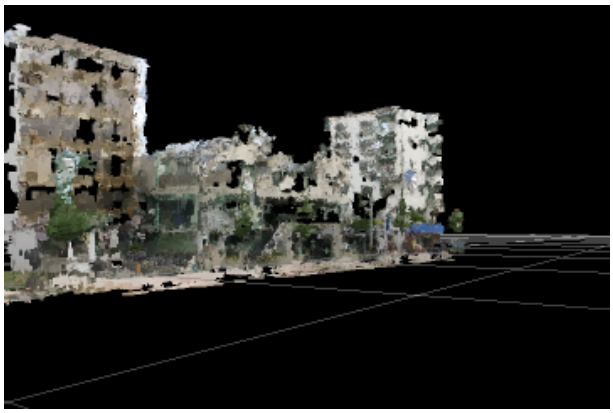


**Figure 5. Results of outdoor scene recovery (street scenery).**

sequence of images captured in complex outdoor environments. However, we observe that some parts of reconstructed models are holed. In future work, integration of 3-D models from multiple image sequences should be investigated for obtaining a complete surface model.

**Acknowledgments**

This work was supported in part by Internet Systems Research Laboratories, NEC Corporation.

## References

[1] P. Beardsley, A. Zisserman, and D. Murray. Sequential Updating of Projective and Affine Structure from Motion. In *Int. Jour. of Computer Vision*, Vol. 23, No. 3, pp. 235–259, 1997.

[2] R. Kumar, H. S. Sawhney, Y. Guo, S. Hsu, and S. Samarasekera. 3D Manipulation of Motion Imagery. In *Proc. Int. Conf. on Image Processing*, pp. 17–20, 2000.

[3] M. Okutomi and T. Kanade. A Multiple-baseline Stereo. In *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 15, No. 4, pp. 353–363, 1993.

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

[5] G. Roth and A. Whitehead. Using Projective Vision to Find Camera Positions in an Image Sequence. In *Proc. 13th Int. Conf. on Vision Interface*, pp. 87–94, 2000.

[6] T. Sato, M. Kanbara, H. Takemura, and N. Yokoya. 3-D Reconstruction from a Monocular Image Sequence by Tracking Markers and Natural Features. In *Proc. 14th Int. Conf. on Vision Interface*, pp. 157–164, 2001.

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

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

[9] N. Yokoya. Surface Reconstruction Directly from Binocular Stereo Images by Multiscale-multistage Regularization. In *Proc. 11th Int. Conf. on Pattern Recognition*, Vol. I, pp. 642–646, 1992.