Dense 3-D Reconstruction of an Outdoor Scene by Hundreds-baseline Stereo Using a Hand-held Video Camera

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, masay-ka, 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 and dense 3-D reconstruction is widely investigated. In related work, a dense 3-D model is generated by using a stereo method. However, these methods 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 dense 3-D reconstruction method that first estimates extrinsic camera parameters of a hand-held video camera, and then reconstructs a dense 3-D model of a scene. We can acquire a model of the 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 and dense 3-D reconstruction is desired. In the field of computer vision, there are many researches that reconstruct 3-D models from multiple images [1].

One of the major approaches to 3-D reconstruction is to use static stereo. However, the methods cannot use a large number of images because it is difficult to calibrate a large number of cameras accurately. Therefore, these methods become sensitive to noise. Although many researchers often use a constraint of surface continuity to reduce noises, such an approach limits a target scene and may sometimes reduce accuracy of reconstruction.

One of other approaches is to use an image sequence. The methods that use an image sequence can automatically estimate camera parameters and 3-D positions of natural features by tracking features in captured images. Factorization algorithm [2] is one of the well known methods that can estimate a rough 3-D scene stably and efficiently from an image sequence by assuming an affine camera model. However, when the underlying scene is not suitable for the affine camera model, estimated camera parameters are not reliable. Therefore, this method is not suitable for reconstructing a dense 3-D model by stereo method. Although there exists another reconstruction method [3] that estimates camera parameters and a dense 3-D model from an image sequence, the method uses only a small number of images. Additionally, it seems to be difficult to reconstruct a complex outdoor scene because it uses the constraint of surface continuity in dense depth estimation.

In order to reconstruct an outdoor scene densely and stably, we propose a new 3-D reconstruction method that first estimates extrinsic camera parameters of an input image sequence, and then reconstructs a dense model of a scene. In the first process, we use a camera parameter estimation method [4] that is based on tracking both markers and natural features. Next, dense depth maps are computed by using a multi-baseline stereo method from hundreds images. Finally, depth maps are combined together in a voxel space. The proposed method can reconstruct a complex outdoor scene densely and accurately by combining several hundreds images of a long sequence without the constraint of surface continuity.

This paper is structured as follows. Section 2 describes a method of estimating camera parameters of a hand-held video camera by tracking markers and natural features. In Section 3, we describe a method of dense depth estimation and integration of these dense data in a voxel space. Then, we demonstrate experimental results of 3-D reconstruction from real image sequences to show the feasibility of the proposed method in Section 4. Finally, Section 5 describes conclusion and future work.

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 shows the flow diagram of our algorithm. First, we must specify the positions of six

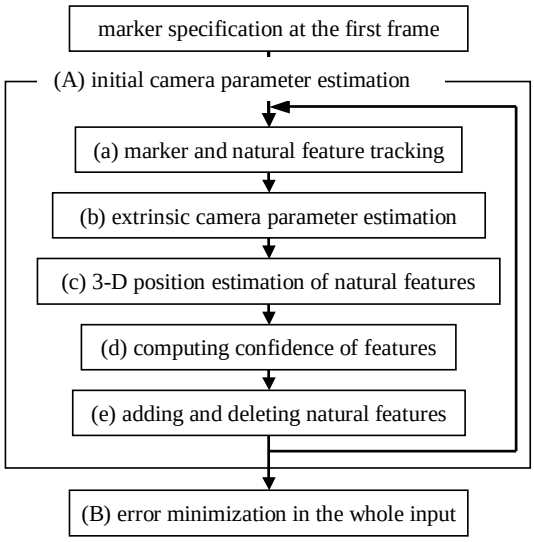


Figure 1: Flow diagram of camera parameter estimation.

or more markers in the first frame of input sequence to estimate extrinsic camera parameters in the first frame. Then initial extrinsic camera parameters in all the frames are determined by iterating the processes at each frame (A). However, the accumulation of estimation error exists. Therefore, 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 with automatically adding and deleting features regardless of the visibility of initial markers. It should be noted that intrinsic camera parameters must be estimated in advance.

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 tracked by using 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. Harris's interest operator is used for robust tracking of natural features.

(b) Extrinsic camera parameter estimation In this method, the re-projection error $R_{fp} = |\mathbf{x}_{fp} - \hat{\mathbf{x}}_{fp}|^2$ is used as a measure for 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 using 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 defined as follows:

$$E_f = \sum_p W_{fp} R_{fp}, \qquad (1)$$

where W_{fp} is a weighting coefficient for the feature p at the f-th frame and is computed by considering the confidence of the feature.

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

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

(d) Computing confidence of features We assume that the distribution of tracking error can be approximated by a Gaussian probability density function. Then the confidence of feature, W_{fp}, is defined by the inverse of variance of re-projection error R_{fp} in the frames where feature p has already been tracked.

(e) Adding and deleting 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 that 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.

2.2 Optimization in the whole input

In the final step, the accumulation of estimation error is minimized over the whole input. The accumulated estimation error is given by the sum of re-projection errors as in Eq.(2) and is minimized with respect to the camera parameter $\mathbf{M}_f$ and natural feature position $\mathbf{S}_p$.

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

The camera parameters and feature positions that have already been 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.

3. Dense 3-D reconstruction of scene

In this section, we describe a dense 3-D reconstruction method using camera parameters estimated by the method described in Section 2. First, a dense depth map for each image is computed by using a multi-baseline stereo method, then a dense 3-D model is reconstructed by combining obtained dense depth maps in a voxel space.

3.1. Depth mapping by multi-baseline stereo

A depth map is computed for each frame by using a multi-baseline stereo technique [5]. Depth value z of pixel (x, y) at the f-th frame is estimated by using the images from the $(f-k)$-th frame to the $(f+k)$-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 as follows:

$$\begin{pmatrix} a\hat{x_j} \\ a\hat{y_j} \\ a \\ 1 \end{pmatrix} = \mathbf{M}_j \mathbf{M}_f^{-1} \begin{pmatrix} xz \\ yz \\ z \\ 1 \end{pmatrix}, \quad (3)$$

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 in Eq.(4) using RGB components (I_{Rf}, I_{Gf}, I_{Bf}).

$$\begin{aligned} &SSD_{ij}(x, y, o_x, o_y) = \\ &\sum_{W \supseteq (u-o_X, v-o_Y)} \{(I_{Ri}(x+u, y+v) - I_{Rj}(\hat{x_j}+u, \hat{y_j}+v))^2 \\ &\quad + (I_{Gi}(x+u, y+v) - I_{Gj}(\hat{x_j}+u, \hat{y_j}+v))^2 \\ &\quad + (I_{Bi}(x+u, y+v) - I_{Bj}(\hat{x_j}+u, \hat{y_j}+v))^2\}, \end{aligned} \quad (4)$$

where o_x and o_y are offsets of the window W for x and y axes, respectively. We define a modified SSSD (Sum of SSD) in Eq.(5) using the median of SSD because the template of window W in the f-th frame may be occluded in other frames.

$$SSSD_i(x, y, o_x, o_y) = \sum_{i=f-k}^{f+k} \begin{cases} SSD_{fi}(x, y, o_x, o_y); \\ \qquad SSD_{fi}(x, y) \leq M \text{ and } |i-f| > C, \\ 0; \qquad \text{otherwise} \end{cases} \quad (5)$$

where,

$$\begin{aligned} M = median(&SSD_{f(f-k)}(x, y, o_x, o_y), \cdots, \\ &SSD_{f(f-C)}(x, y, o_x, o_y), SSD_{f(f+C)}(x, y, o_x, o_y), \\ &\cdots, SSD_{f(f+k)}(x, y, o_x, o_y)). \end{aligned} \quad (6)$$

Note that images from the $(f-C)$-th frame to the $(f+C)$-th frame are not used for computing SSSD, because baselines in these frames are not long enough to estimate depth stably. Multiple centered windows approach [6] is also used to reduce estimation error around occlusion boundaries. Then SSSD is extended to SSSDM as follows:

$$SSSDM_i(x, y) = \min_{W \supseteq (u,v)} (SSSD(x, y, u, v)). \quad (7)$$

We can estimate the depth value z correctly by minimizing SSSDM unless pixel (x, y) is occluded in more than k frames. Additionally, we avoid a local minimum problem and achieve stable depth estimation using a multiscale approach [7]. Note that we use the linear interpolation to compute z in the regions without informative textures because the confidence of estimated z is low in such regions.

3.2. 3-D reconstruction in 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 estimated depth data and camera parameters. Both A and B are voted when the voxel is projected onto a pixel (x, y) in a frame. Value A is voted if depth of the voxel in camera coordinate system equals z of (x, y). On the other hands, value B is voted when depth of the voxel is equal to or less than z of (x, y). The 3-D model is reconstructed by selecting the voxel whose A/B is more than a given threshold. Note that the color of the voxel is decided by mean color of pixels that has voted to the value A of the voxel.

4. Experiment

In experiment, we captured an outdoor scene as shown in Figure 2(a) by a hand-held CCD camera (Sony DSR-DP-150) with a wide conversion lens (Sony VCL-HG0758). This image sequence lasts 40 seconds and has 599 frames (720×480 pixels, progressive scan). The intrinsic camera parameters are estimated by using the Tsai's method [8] in advance, and the extrinsic camera parameters are estimated by the method described in Section 2. The curved lines in Figure 3 indicate the camera path and the quadrilateral pyramids indicate the camera postures drawn at every 50 frames.

Dense depth map of the f-th frame is estimated by using odd 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 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 as shown in this figure. However there exist some incorrect depth values between a column and a wall of the building because there are no textures around the wall of the building. Linear interpolation is used for determining depth values in these areas.

Figure 3 shows a 3-D model 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 consists of 10cm cube voxels. A wall behind a column of the building is reconstructed even if the wall is occluded from time to time. We also confirm that some positions are holed because these pixels are not visible enough for sufficient precision.

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 this method, first, extrinsic camera parameters are estimated by tracking markers and natural features. Then, at each frame, a dense depth map is computed by using already estimated camera parameters. Finally, a dense 3-D model is reconstructed by combining hundreds of dense depth maps in a voxel space.

In the experiments, the dense 3-D scene reconstruction is accomplished for a long image sequence captured in a complex outdoor scene successfully with stable dense depth estimation. However, some parts of reconstructed model have holes. In future work, more accurate model reconstruction will be explored using the confidence of depth value. Integration of dense 3-D models from multiple image sequences will also be investigated for obtaining a complete model.

References

[1] N. Yokoya, T. Shakunaga and M. Kanbara: "Passive Range Sensing Techniques: Depth from Images," IEICE Trans. Inf. and Syst., Vol. E82-D, No. 3, pp. 523–533, 1999.

[2] 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.

[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.

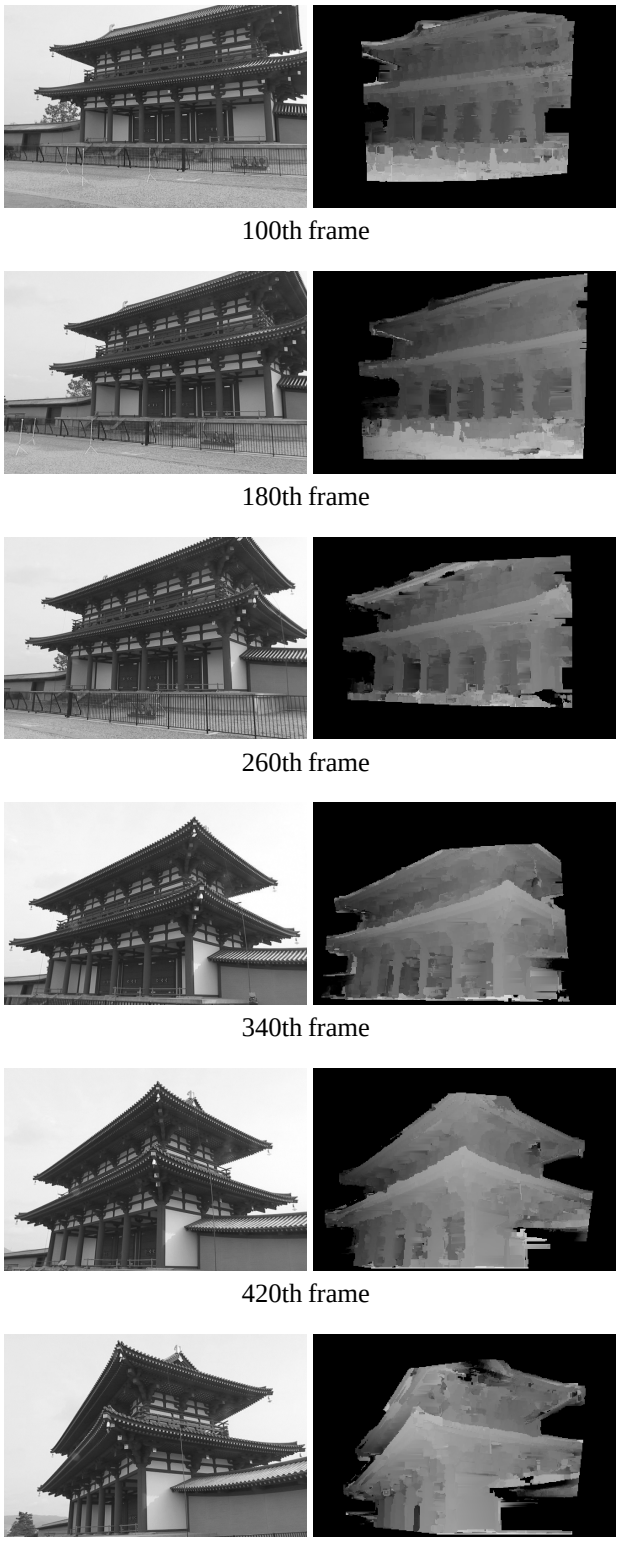
[4] T. Sato, M. Kanbara, H. Takemura and N. Yokoya: "3-D Reconstruction from a Monocular Image Sequence by Tracking Markers and Natural Features," Proc. 14th Int. Conf. on Vision Interface, pp. 157–164, 2001.

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

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

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

[8] 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.



(a) Input images (b) Dense depth map

Figure 2: Input images and estimated dense depth maps.

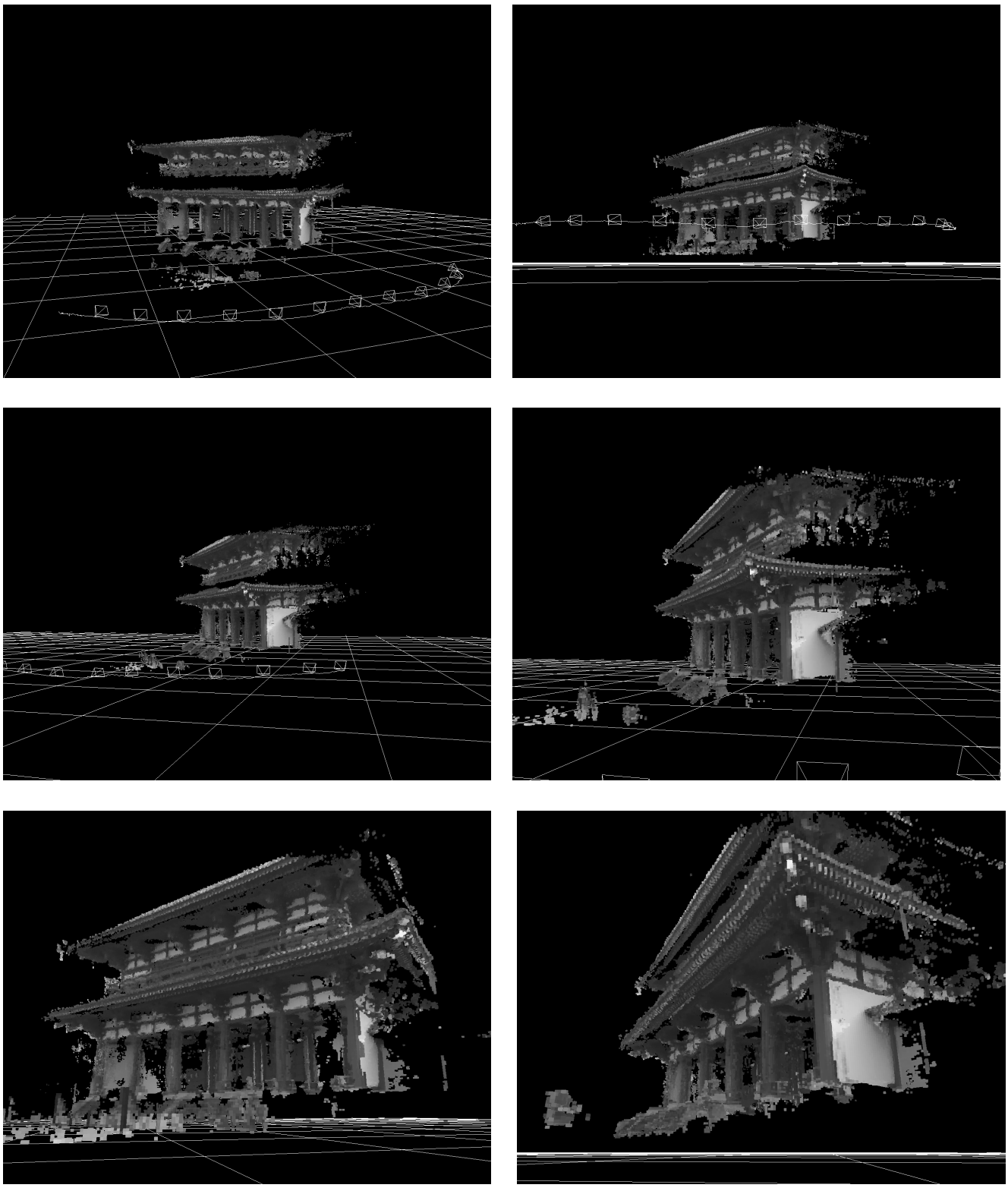


Figure 3: Results of outdoor scene recovery as well as estimated camera positions and postures.