3-D Modeling of an Outdoor Scene from Multiple Image Sequences by Estimating Camera Motion Parameters

Tomokazu Sato, Masayuki Kanbara, and Naokazu Yokoya

Graduate School of Information Science, Nara Institute of Science and Technology,
8916-5 Takayama, Ikoma, Nara 630-0192, Japan
{tomoka-s, kanbara, yokoya}@is.aist-nara.ac.jp
http://yokoya.aist-nara.ac.jp/~tomoka-s/research-e.html

Abstract. Three-dimensional (3-D) models of outdoor scenes can be widely used in a number of fields such as object recognition, navigation, scenic simulation, and mixed reality. Such models are often made manually with high costs, so that automatic 3-D reconstruction has been widely investigated. In related works, a dense 3-D model is generated by using a stereo method. However, such approaches cannot use several hundred images together for dense depth estimation of large constructs and urban environments because it is difficult to accurately calibrate a large number of cameras. This paper proposes a novel dense 3-D reconstruction method that uses multiple image sequences. First, our method estimates extrinsic camera parameters of each image sequence, and then reconstructs a dense 3-D model of a scene using an extended multi-baseline stereo and voxel voting techniques.

1 Introduction

Three-dimensional (3-D) models of outdoor scenes can be widely used in a number of fields such as object recognition, navigation, scenic simulation, and mixed reality. Because such models are often made manually with high costs, automatic and dense 3-D reconstruction is required. In the field of computer vision, there are many researches that reconstruct 3-D models from multiple images [1].

One of approaches to the problem is to use an image sequence that is called shape-from-motion [2–4]. The method can automatically recover camera parameters and 3-D positions of feature points by tracking natural features in captured images. However, most of the existing methods reconstruct only a limited scene from a small number of images and are not designed to obtain a dense model. Besides such problems, these approaches cannot unify the reconstructed scenes from multiple image sequences for the purpose of making a complete 3-D model or improving the precision of a reconstructed 3-D model, because they cannot reconstruct a scale factor and relationship among multiple image sequences.

In order to reconstruct a complex outdoor scene densely and stably, we propose a novel 3-D reconstruction method that uses multiple image sequences. First camera parameters of each input image sequence are estimated with information of a scale factor and relationship of sequences by using predefined markers of known 3-D position [5]. Then dense depth maps are computed from multiple image sequences by using a multi-baseline stereo technique. Finally, a 3-D model of a scene is reconstructed by unifying several hundred depth maps in a voxel space. The proposed method can reconstruct a complex outdoor scene densely and accurately by using several hundred images of multiple image sequences.

2 3-D Model Reconstruction from Multiple Image Sequences

In this section, we describe a 3-D model reconstruction method using multiple image sequences. First, camera parameters of each image sequence are estimated in a single coordinate system. Then, colors of multiple image sequences are adjusted, and a dense depth map for each image is computed by using an extended multi-baseline stereo method. Finally, a 3-D model is reconstructed by combining obtained dense depth maps in a voxel space.

2.1 Camera Parameter Estimation for Each Image Sequence

This section briefly describes a camera parameter estimation method [6] for each input image sequence which is based on tracking features (markers and natural features). First, we must specify the positions of six or more markers in the first frame of an 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 process at each frame. The predefined markers are not necessary to be visible throughout an input sequence because 3-D positions of natural features are detected in the process of estimation and are used instead of markers. Finally, extrinsic camera parameters are refined by minimizing the accumulation of estimation errors over the whole input. The last frame and some other frames of the input image sequence must contain markers for minimizing the accumulated estimation errors of camera parameters. It should be noted that intrinsic camera parameters (focal length, pixel size, center of image, radial distortion factor coefficient) must be estimated in advance.

Using this approach, we can estimate extrinsic camera parameters efficiently and accurately regardless of the visibility of initial markers. A scale factor and relationship among multiple image sequences are also obtained easily by unifying the 3-D coordinates of the predefined markers.

2.2 Color Adjustment among Multiple Image Sequences

It is necessary for accurate depth estimation to consider the difference in colors among image sequences caused by the change of lighting condition. We assume that all materials in a scene exhibit only diffuse reflection. From this assumption, the color change of these materials caused by motion of the sun and clouds is linearly approximated as follows:

$$(I'_R, I'_G, I'_B)^T = (s_r, s_g, s_b)(I_R, I_G, I_B)^T + (o_r, o_g, o_b), \tag{1}$$

where $(I'_R, I'_G, I'_B)^T$ and $(I_R, I_G, I_B)^T$ represent changed and original RGB color components of the same material, respectively. Parameters (s_r, s_g, s_b) and (o_r, o_g, o_b) represent the linear color transformation coefficients. These coefficients can be estimated by specifying least six points of the same material in multiple image sequences.

2.3 Dense Depth Estimation by Extended Multi-baseline Stereo

A depth map is computed for each frame of all image sequences by using a multi-baseline stereo technique [7]. In the multi-baseline method, the SSD (Sum of Squared Differences) is employed as an error function, and the depth z of each pixel (x, y) in the f-th frame is determined so as to minimize the SSSD (Sum of SSDs) for multiple images $\mathbf{S_f}$.

$$SSSD_f(x, y, z) = \sum_{i \subseteq \mathbf{S_f}} SSD_{fi}(x, y, z), \tag{2}$$

$$SSD_{fi}(x, y, z) = \sum_{(u,v) \subseteq W} \left\{ (I_f(x+u, y+v) - I_i(\hat{x_i}+u, \hat{y_i}+v))^2 \right\}. \tag{3}$$

where $(\hat{x_i}, \hat{y_i})$ denotes the projected position of 3-D position (x, y, z) to the i-th frame, and W represents a template window for calculating SSD.

In the following sections, some extensions of the original multi-baseline stereo are described for stable and accurate estimation of depth maps for complex outdoor scenes. A criterion of selecting the images $\mathbf{S_f}$ is also defined for applying the multi-baseline method to multiple image sequences.

(1) Detecting Depth Regions of Low Confidence

In traditional stereo methods, depth values are usually estimated only around edges and depth filling is applied for regions without edges because the confidence of estimated depth in the region without informative textures becomes very low. We also employ this approach in the multi-baseline stereo framework to detect depth regions of low confidence. These regions can be easily detected by using output values of an edge detector for the input image.

(2) Considering Occlusions and Scales

Modified SSSD is used for considering both occlusions and scales instead of the original SSSD. First, SSD is extended to SSD' for considering scales between frames by using RGB color components (I_R, I_G, I_B) as follows:

$$\begin{aligned} SSD'_{fi}(x,y,z) = \sum_{(u,v)\subseteq W} \big\{ & (I_{Rf}(x+u,y+v) - I_{Ri}(\hat{x}_i + s_{fi}u, \hat{y}_i + s_{fi}v))^2 \\ & + (I_{Gf}(x+u,y+v) - I_{Gj}(\hat{x_i} + s_{fi}u, \hat{y_i} + s_{fi}v))^2 \\ & \quad + (I_{Bf}(x+u,y+v) - I_{Bj}(\hat{x_i} + s_{fi}u, \hat{y_i} + s_{fi}v))^2 \big\}, \end{aligned} \quad (4)$$

where s_{fi} is the ratio of depths between the f-th and i-th frame centers of projection and 3-D position (x,y,z).

Modified SSSD is defined as the sum of selected SSD's that are smaller than the median of SSD's.

$$SSSD'_f(x,y,z) = \sum_{i \subseteq \mathbf{S_f}} \begin{cases} SSD'_{fi}(x,y,z); & SSD'_{fi}(x,y,z) \leq M \\ 0; & \text{otherwise} \end{cases}, \quad (5)$$

$$M = \underset{i \subseteq \mathbf{S_f}}{\text{median}} (SSD'_{fi}(x,y,z)). \quad (6)$$

Using this formulation, correct depth is estimated unless the pixel (x,y) in the f-th frame is occluded in more than half of images $\mathbf{S_f}$.

(3) Stable and Efficient Estimation of Depth Map

Detecting the depth value z that minimize the SSSD' function is a nonlinear problem. Therefore, there exist problems of computational cost and local minima. To avoid these problems, we employ a multi-scale approach [8] for stable and efficient estimation of depth maps.

In advance, multiple resolution images are generated for every image in $\mathbf{S_f}$. First, the lowest resolution image is used for depth estimation. The depth z is searched in a range decided by considering a scene depth from all cameras in advance. In the next resolution, depth values are searched in a limited range around the depth estimated in the previous resolution depth map. By iterating this estimation up to the highest resolution image, searching regions are limited gradually at each resolution and a total computational cost becomes much cheaper. Local minimum problems can also be avoided because the result of the lower resolution depth map eliminates uncertain searching regions.

(4) Image Selection from Multiple Image Sequences

For correct estimation of depth value z by minimizing the SSSD', a target object must be visible in more than half of images $\mathbf{S_f}$. In our method, the limited searching range described in the previous section is used in selection of images $\mathbf{S_f}$ for satisfying this condition.

As illustrated in Figure 1, the images in which the limited searching range is fully visible are selected for $\mathbf{S_f}$ and are used to estimate a depth of pixel (x,y). Note that the images $\mathbf{S_f}$ are automatically renewed in every resolution of multi-scale approach regardless of sequence.

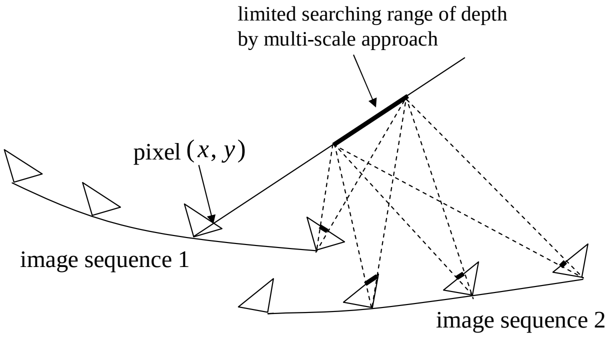


Fig. 1. Image selection from multiple image sequences in extended multi-baseline stereo

2.4 3-D Model Reconstruction in a Voxel Space

This section describes a method for reconstructing a 3-D model by integrating all the estimated depth maps in a single voxel space. The method is based on voxel voting from all the depth maps. Note that we have no need to consider each sequence separately because extrinsic camera parameters of each sequence are already unified by the method described in Section 2.1.

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. 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 textured 3-D model is then reconstructed by selecting the voxel whose A/B is more than a given threshold. 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.

3 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 CCD video camera (Sony VCL-HG0758) with a wide conversion lens (Sony VCL-HG0758). The intrinsic camera parameters are estimated by Tsai's method [9] in advance. The extrinsic camera parameters are estimated by the method described in Section 2.1 for each image sequence by using predefined markers of known 3-D position, that are measured by the total station (Leica TCR1105XR) in a single coordinate system.

(1) Reconstruction of Building

In the first experiment, a single building (Suzaku-mon) is captured as two image sequences shown in Figure 2 by walking around the building viewing it at the center of image. The image sequence of the front of the building has 747 frames, and the back has 982 frames (720×480 pixels, progressive scan).

First, camera parameters of every frame in the sequences are reconstructed. In this experiment, we specified the markers manually from the first frame to the 100th frame, the middle and the last frame for each sequence. Natural features are automatically detected and tracked throughout each sequence. The average number of tracked natural features is 190 points per frame. The squared average re-projection error of features in the image is 0.95 pixel. The lengths of reconstructed camera paths are 63m (front of building) and 76m (back of building).

Then dense depth maps of the entire input image are computed. Figure 3 shows computed dense depth maps in which depth values are coded in intensity. It is confirmed that correct depth values are obtained in most part of the images. However there exist some incorrect depth values around the roof of the building because there are no vertical edges and camera moves horizontally.

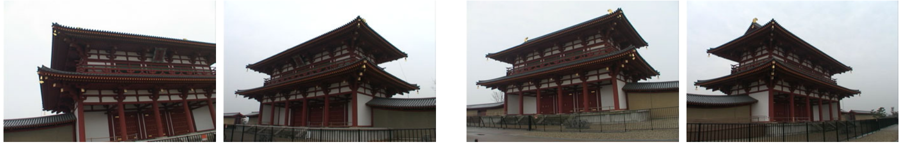
Figure 4 shows a reconstructed 3-D model with textures obtained by combining about 1700 dense depth maps together in the way of voxel voting described in Section 2.4. In this experiment, the voxel space is constructed of 10cm cube voxels (size: $450 \times 260 \times 240$). It can be observed that a wall behind columns of the building is successfully reconstructed even if the wall is occluded from time to time. We can also observe that some positions are holed because they are not visible enough for sufficient precision in the image sequence.

(2) Reconstruction of Street Scenery

In the second experiment, a part of street scenery is captured as two sequences by the CCD camera put on a slowly moving car. As shown in Figure 5, one sequence captures the forward view of moving car and the other captures the backward view. The forward and backward image sequences have 500 frames and 405 frames, respectively.

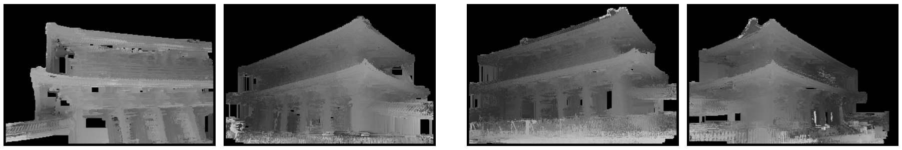
First, camera parameters of the sequences are recovered. The markers are specified manually from the first frame to the 50th frame, the middle and the last frame for each sequence. In this camera parameter recovery, the average number of tracked features is 110 points per frame, and the squared average re-projection error is 0.82 pixel. The lengths of reconstructed camera paths are 128m (forward) and 133m (backward).

Then colors of the sequences are adjusted by the method in Section 2.2 by specifying some points of the same materials in multiple image sequences by hand. As shown in Figure 6, it is confirmed that correct depth values are obtained in most part of the images even around occlusion edges. Figure 7 shows a reconstructed 3-D model with textures obtained from about 900 dense depth maps. In this experiment, the voxel space is constructed of 10cm cube voxels (size: $1000 \times 180 \times 110$). Note that many parts of walls are holed around the windows of the buildings mainly due to specular reflection. It is essentially difficult to reconstruct specular objects.



(a) front of building (b) back of building

Fig. 2. Input images (building)



(a) front of building (b) back of building

Fig. 3. Result of depth estimation (building)

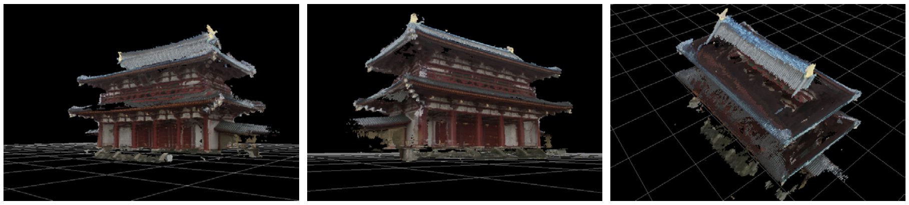
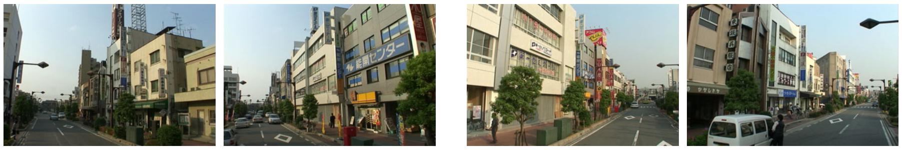
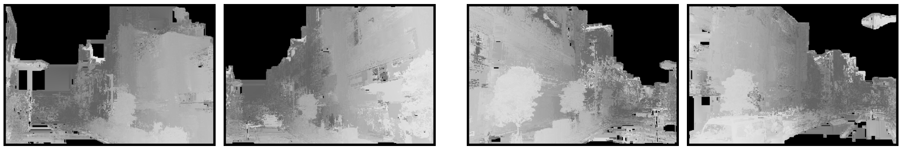


Fig. 4. Reconstructed 3-D model (building)



(a) forward view (b) backward view

Fig. 5. Input images (street scenery)



(a) forward view (b) backward view

Fig. 6. Result of depth estimation (street scenery)

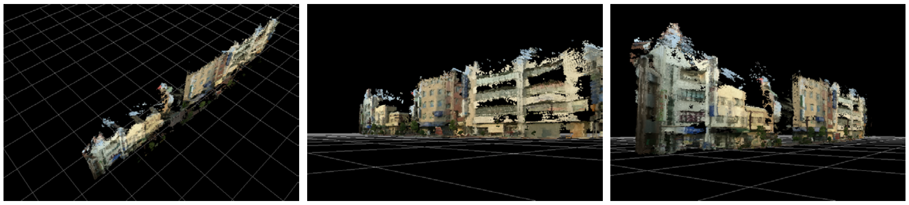


Fig. 7. Reconstructed 3-D model (street scenery)

4 Conclusion

In this paper, a dense 3-D reconstruction method from multiple image sequences captured by a video camera is proposed. In experiments, the dense 3-D scene reconstruction is successfully accomplished by using multiple long image sequences captured in complex outdoor environments. However, we observe that some parts of reconstructed models are holed due to specualr reflection or lack of visibility. In future work, omnidirectional sensors [10] will be used to acquire images for more accurate and stable estimation of complex outdoor scenes.

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. P. Beardsley, A. Zisserman and D. Murray: "Sequential Updating of Projective and Affine Structure from Motion," Int. Jour. of Computer Vision, Vol. 23, No. 3, pp. 235–259, 1997.
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. C. Tomasi and T. Kanade: "Shape and Motion from Image Streams under Orthography: A Factorization Method," Int. Jour. of Computer Vision, Vol. 9, No. 2, pp. 137–154, 1992.
5. T. Sato, M. Kanbara, N. Yokoya and H. Takemura: "Dense 3-D Reconstruction of an Outdoor Scene by Hundreds-baseline Stereo Using a Hand-held Video Camera," Int. Jour. of Computer Vision, Vol. 47, No. 1-3, pp. 119–129, 2002.
6. 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.
7. M. Okutomi and T. Kanade: "A Multiple-baseline Stereo," IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 15, No. 4, pp. 353–363, 1993.
8. 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.
9. 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.
10. S. Ikeda, T. Sato and N. Yokoya: "An Calibration Method for an Omnidirectional Multi-camera System," Proc. SPIE, Vol. 5006, 2003.