Multiple Video Camera Calibration using EPI for City Modeling

Hiroshi KAWASAKI Atsunori MIYAMOTO Yutaka OHSAWA
Saitama University, 255 Shimo-Okubo, Sakura, Saitama 338-8570, JAPAN
Shintaro ONO Katsushi IKEUCHI
University of Tokyo, 6–4–1 Komaba, Meguro-ku, Tokyo, 153-8505, Japan

Abstract

In this paper, we present an Epipolar Plane Image (EPI)-based multiple video camera calibration method which enables simultaneous estimation of the multiple video cameras' parameters and the synchronization of video data. Using our proposed method, in order to capture a large scale scene's texture image, a user is only required to install multiple video cameras on top of his/her car without setting exact configurations, and can take a video without using a special external synchronization device. Since the optical centers of cameras are usually inside the cameras, and bringing them together is impossible, distortions inevitably occur in fused images. However, with our proposed method, the optical centers of the multiple cameras are made to coincide at one point in spatio-temporal space and thus we can fuse these images into a single one with no distortions. Another strength of our method is that we do not need any overlapping areas between images for calibration, therefore large scale scene can be efficiently captured.

1 Introduction

The image synthesis of a large-scale environment, such as an entire town or city, and the efficient acquisition of such large-scale data are considered to be important research areas in computer vision (CV) and computer graphics (CG).

So far, in terms of synthesis, model-based rendering (MBR) is usually adopted for this purpose. A MBR uses the geometry and surface attributes of objects to construct images from a given viewpoint. However, in practice, it is still difficult to efficiently acquire the precise geometric model and the surface reflectance model for large scale scenes.

On the other hand, image-based rendering (IBR) technique has recently become a major research topic in CV and CG. Since much research has been done on IBR techniques, few actual applications have been developed. One significantly important reason for this is because it is difficult to acquire such huge amounts of data efficiently.

Based on these facts, efficient data acquisition is crucial for both major synthesis methods. Using an omni-directional camera for large coverage of space and a video recorder for dense sampling, can be considered to be a realistic and straightforward solution. In this paper, we propose a new omni-directional video camera which consists of multiple cameras and can capture high-resolution images without distortions.

Since our target is large-scale environments, such as urban cities, we mount the omni-directional video camera on top of the car and capture the cities by running along the street; it is usually hard to analyze such huge data efficiently. Therefore, to make analysis robust and simple, we apply an EPI analysis to our method. With our proposed EPI-based multiple camera calibration method, simultaneous estimation of the parameters of multiple video cameras and the automatic synchronization of video data can be robustly achieved.

This paper is organized as follows. In Section 2, we describe related studies, and in Section 3, we explain the algorithm used in our method. Section 4 presents our analysis of error in fused image and Section 5 contains the results of experiments. Section 6 concludes our method.

2 Related Works

2.1 Large Scale Scene Modeling

There have been many research projects for modeling large scale scenes such as those found in urban cities. So far, the model-based method is mainly applied for city modeling purposes. Fruh and Zakhor[3] and Teller et al.[1] have proposed a unique technique to capture and reconstruct urban areas. However, the techniques used in these studies still have difficulties in capturing fine texture images for large scale scenes.

On the other hand, generating a 3-D virtual world directly from real scene images, referred to IBR, is considered a promising technique. "Aspen Movie Map"[4] was the pioneering work of this IBR technology. Takahashi et al.[7] worked on rendering large-scale scenes using the IBR technique. However, the previous IBR research mainly concentrated on data representation, not on techniques for capturing data.

2.2 Omni-directional Camera

A simple and easy omni-directional camera system to use is a single camera with rotational mirrors[6][8][9]. However,

these cameras usually capture whole panoramic scenes in just one image and therefore, resolution of the final image is inevitably low.

Another method for capturing omni-directional images with high resolution is to arrange multiple cameras cylindrically and stitch them together[7]. Although this type of omni-directional camera can solve the resolution problem, the optical centers of every camera cannot be brought together at one point, and stitched images usually contain distortions. Furthermore, the synchronization of multiple cameras is also difficult if there is no external synchronization device.

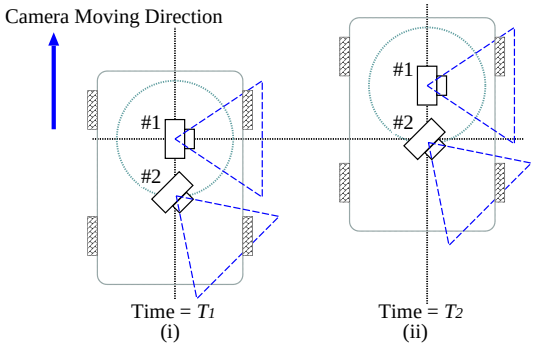


Figure 1: Optical synchronization

3 Multiple Camera Calibration in Spatio-Temporal Space

In this section, we explain how we fuse multiple camera images without distortions. we first explain how we employ the camera calibration in spatio-temporal space and the reason why our proposed technique can efficiently remove the distortions. Then we propose an EPI-based calibration method which enables efficient calibration in spatio-temporal space and eliminates distortions in fused image.

3.1 Spatio-Temporal Synchronization

Since we cannot make the optical centers of multiple cameras coincide at one point, inevitable distortion occurs when fusing multiple camera images. Certainly, if we know the shape of the object and the camera parameters, we can remove the distortions. However, acquiring such data for large scale scenes is usually difficult. In this paper, to remove the distortions which derived from inconsistency of optical center, we propose a new calibration technique which considers the spatio-temporal space.

Synchronization in spatio-temporal space can be explained as follows. As shown in Fig.1-(i), camera # 1 captures an image at a 3D point (x_0, y_0, z_0) at time T_1. Then, at time T_2, camera # 2 captures an image at the same 3D point (x_0, y_0, z_0) (Fig.1-(ii)). Therefore, if we know the time T_1 and T_2, we can cause the cameras' optical centers to coincide at one point. In the following section, we explain how we estimate these T_1 and T_2 values automatically in detail.

Distortion Free Camera Configuration

Obviously every camera should path through the same 3D point for our synchronization technique; therefore, the ideal configuration of multiple cameras is as shown in Fig.2-(a). However, the pixel difference derived from the camera position's difference perpendicular to the camera's moving direction is smaller than that of the parallel direction (details in sec 4). Therefore, the cameras does not need to be strictly aligned along a straight line and our actual implementations are Fig.2-(b) and (c).

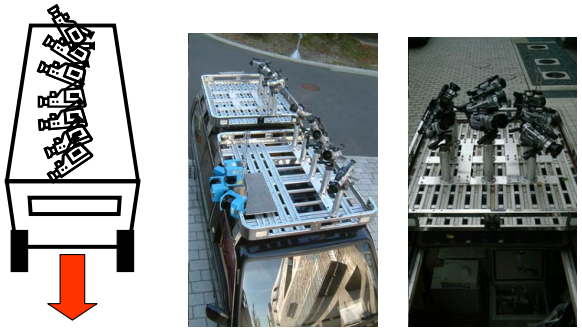

(a) (b) (c)

Figure 2: (a)Distortion free camera configuration. (b)(c)Our actual implementation of multiple cameras

In addition, since our method does not need large overlapping for camera calibration, the whole hemisphere can be efficiently covered by a small number of cameras; therefore, configuration of multiple cameras is usually not difficult in practice for both Fig.2-(b) and (c).

3.2 Multiple Camera Calibraiotn using EPI

To achieve the spatio-temporal synchronization efficiently, we propose an EPI-EPI matching method. Epipolar plane image (EPI) was first proposed by [2]; EPI can be produced by accumulating epipolar line in each frame of video data along the time axis (Fig.3). In our camera configuration, cameras are set to arbitrary directions, therefore image rectification is needed before accumulation to make EPI.

3.2.1 Image Rectification

Usually, buildings and other objects in cities are composed of two kinds of lines, vertical and horizontal lines in relationship to the ground. Therefore, there are two Vanishing

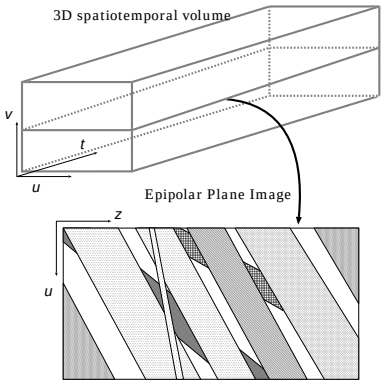


Figure 3: Definition of EPI

Point(VP)s in an image: one for the vertical direction and the other for the horizontal direction. Theoretically, all lines related to a VP cross at just one point. In actuality, they do not because of noise and pixel resolution. We use the average values of these cross points instead.

Using these two VPs, a rotation matrix and focal length can be calculated as follows. First, we define r_1 and r_2 as:

$$r_i = \left(u_i, v_i, f \right)^t (i = 1, 2) \tag{1}$$

where the VP on image coordinate is (u_i, v_i), and f denotes the focal length of each camera (in pixel). These vectors are parallel to the vertical and horizontal directions of the real-world [5], thus r_1 and r_2 are vertically crossed and their inner products should be 0. Therefore, f can be calculated as:

$$f = \sqrt{-(u_1, v_1) \cdot (u_2, v_2)} \tag{2}$$

After normalizing r_1 and r_2 to $\hat{r}_1$ and $\hat{r}_2$, we can calculate $\hat{r}_3$ as:

$$\left[\hat{r}_3 = \hat{r}_1 \times \hat{r}_2 \right] \tag{3}$$

Therefore, the rotation matrix R can be described as:

$$R = \left(\hat{r}_1 \quad \hat{r}_2 \quad \hat{r}_3 \right) \tag{4}$$

By using this matrix R and the focal length, rectification can be accomplished by projective transformation. After rectification of every captured image, we can make the EPI by accumulating the epipolar line, which describes the horizontal line in a rectified image, to the time direction.

3.2.2 EPI-EPI Matching

To achieve the efficient synchronization, we propose an EPI-EPI matching method. We first make EPIs for both two cameras at **a** and **b** as shown in Fig.4. Then we put these two EPIs on the same image plane and consequently we have Fig.4(i).

Fig.4(ii) shows the 1 dimensional EPI-EPI matching result. This is equivalent to images mosaicing using homographic transformation. Most of lines on EPIs are smoothly connected; however, we cannot connect all lines smoothly with this 1 D matching because depth of the all objects is not the same and optical centers of cameras do not meet at one point.

Fig.4(iii) shows a 2 dimensional EPI-EPI matching result. We can see that all lines on EPIs are smoothly connected. This is because that matching result z represents the synchronization parameter (T_1 and T_2 in Sec. 3.1) for temporal direction and, by using z, we can make all video cameras' optical centers coincide at one point. Therefore, we can employ image synthesis without any distortions by directly using the 2D EPI matching result.

Fig.5 show the actual result of 2D EPI matching. We can see that all edges on multiple EPIs are smoothly connected in spite of the small overlapping areas; this feature is quite important to reduce number of cameras to cover the whole hemisphere.

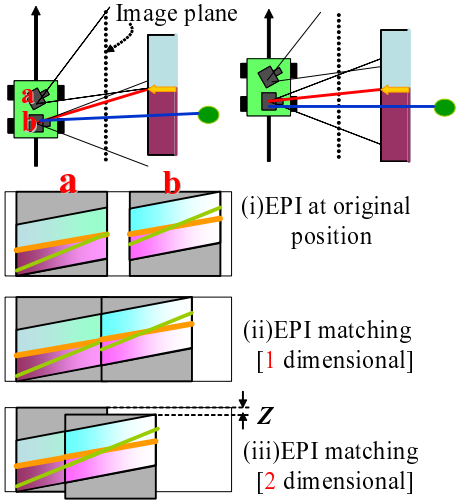


Figure 4: EPI matching

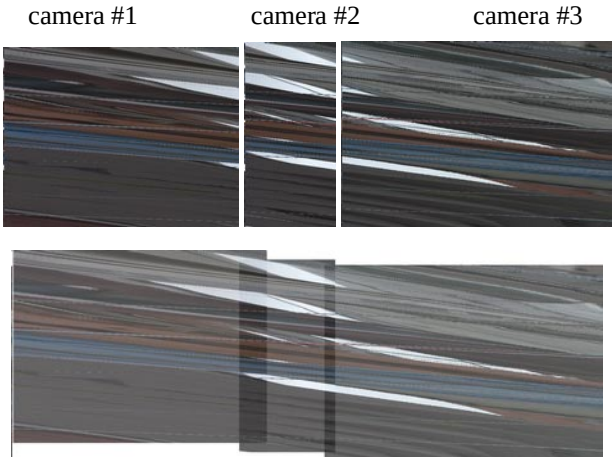


Figure 5: Result of EPI-EPI matching

4 Error Analysis of Image Fusion

To fuse multiple images into a single one by using estimated camera calibration parameters, we apply homographic transformation to each image and warp them onto the single depth plane. As stated in Sec.3.1, accurate alignment of cameras is practically difficult, therefore, inevitable distortion occurs when fusing multiple camera images. In this section, we evaluate actual pixel error values derived from such misalignment of cameras.

The pixel difference in the synthesized image produced by depth difference between actual depth d and assumed depth D can be calculated as:

$$diff \propto \frac{\Delta y - \Delta x \tan\theta}{D} \frac{d - D}{d - \Delta x} \tag{5}$$

where Δx and Δy represent the differnce between cameras. To make analysis simple, we define 2 camera positions as shown in Fig.6 (A) and (B). Calculated pixel differences for each camera position are shown in Fig.7 (A) and (B) respectively.

By examining these figures, we can see that pixel difference (error value) is drastically changed dependent on d and D and the error value at position (B) tends to be larger than the value at position (A). Fig.7 also tells that, if we define the plane depth close to actual depth, pixel difference becomes small. Therefore, if there are no objects along the street, we can fuse multiple images with small distortions by simply defining the depth of the plane as that of front face of buildings.

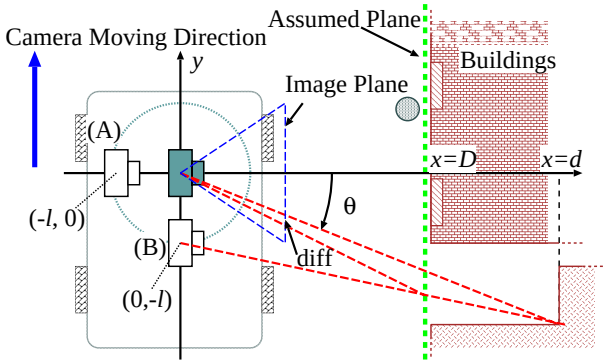
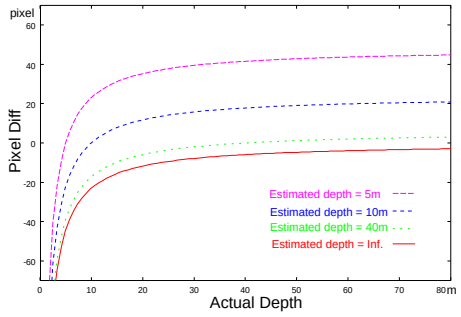
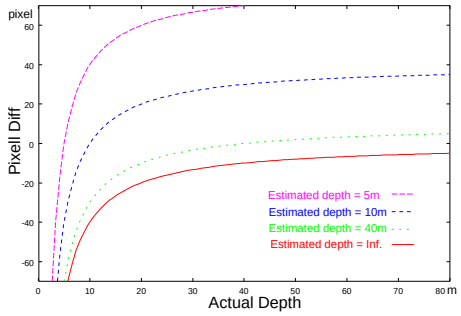


Figure 6: Configuration of Multiple Cameras



(A)



(B)

Figure 7: Calculation of pixel differences

5 Experiments

We performed several experiments to test the effectiveness of our method. In the following two experiments, we used an outdoor scene, located in the landscape of a city, with images captured by multiply-configured cameras on the car.

5.1 Capturing path

EPI analysis requires a straight path and a constant velocity; however, streets do not lie exactly on a straight line, and the car cannot move at a fixed speed. Therefore, in our experiment, first we used GPS, and a gyro-scope sensor to detect the straight path intervals and divide the video sequence into segments. Then we adjusted each segmented video data as if it had been taken with constant speed by using actual car speed; car speed can be estimated by either a speed sensor. Then we can directly apply our method to the data.

5.2 Mosaic multiple cameras images

Fig.8 and Fig.9 are the results of fusing 6 images taken by 6 video cameras as shown in Fig.2(c) .

Fig.8(a - f) shows the original images, and Fig.8(g) the resulting image. On this synthesized image, we can see only small distortion. On the other hand, in Fig.9 , some distortion exists, unlike in Fig. 8. The objects which were located close to the camera, such as the utility pole, were distorted in the image. This is because all cameras do not lie on a straight line, therefore, we can not eliminate the distortion derived from a large depth difference between object and assumed depth.

Fig.10 shows the result of fusing 9 images captured by 9 video cameras as shown in Fig.2(b). With this camera configuration, we can make optical centers coincide at one point; therefore, there is no distortion on either the electric pole or the buildings.

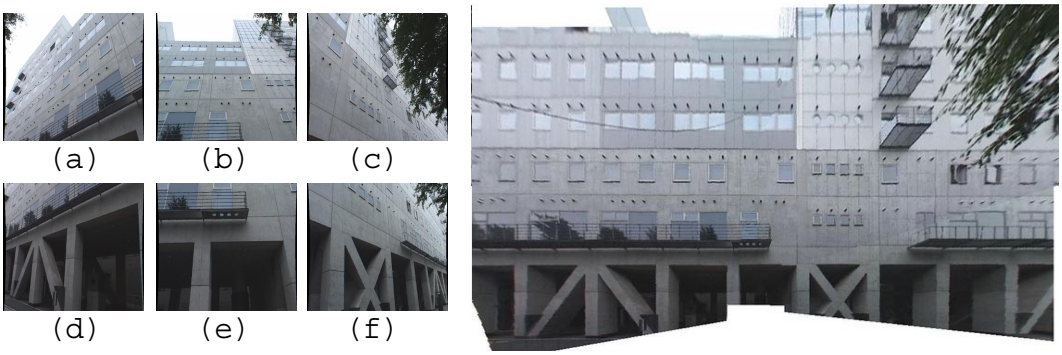


Figure 8: Left (a – f): Captured image of each cameras configured as Fig.2-(c), Right: synthesized image by stitching 6 images. Note that depth variation of the building is quite small and we can not see any distortions in fused image.

Figure 9: Another result of stitching 6 images taken by omni-directional camera configured as Fig.2-(c). Note that the depth of the buildings and the electric pole is largely different, therefore, distortion occurs in fused image.

Figure 10: Result of stitching 9 images taken by omni-directional camera configured as Fig.2-(b). Because of the spatio-temporal synchronization technique, there are no distortions on either the electric pole and the buildings.

5.3 View Dependent Rendering

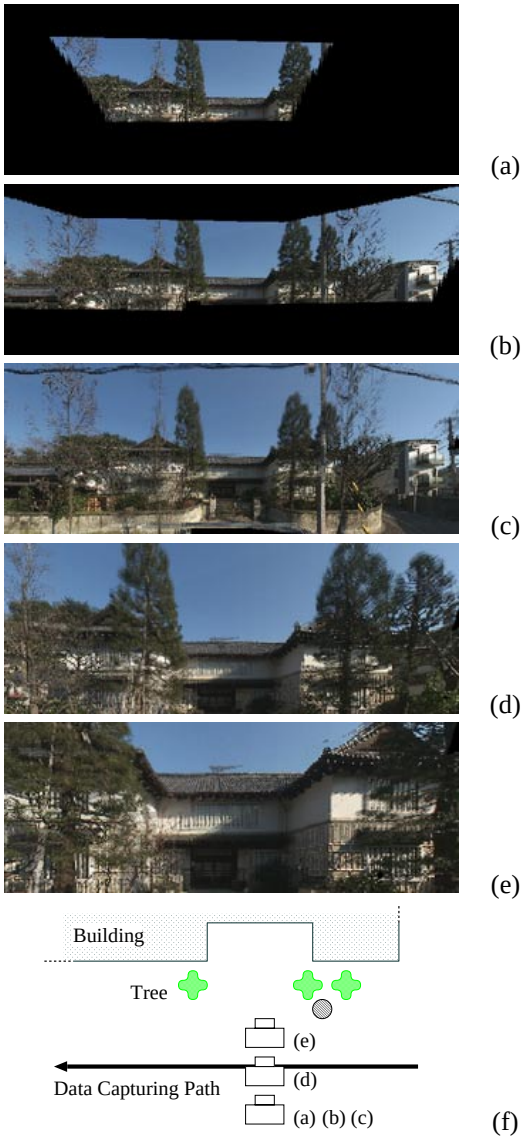


Figure 11: (a)only 1 camera used. (b) 3 cameras used. (c) 9 cameras used. (d)and(e) view dependent images rendered from novel position. (f) camera and scene configuration.

In this experiment, we synthesized images from a novel viewing position using the IBR method proposed by Takahashi et.al.[7]. The new viewing positions and other object locations are shown in Fig.11(f). Fig.11(a),(b) and (c) show a rendered image using a single, 3 and 9 cameras at the same position respectively. Figs.11(d) and (e) shows a series of rendered images viewed from the virtual camera positions and directions. When looking at the images(c)(d) and (e), we cannot find any noticeable distortion in the resulting image. Also, we can see that the trees are rendered at the correct position and occlude the building correctly dependent on the camera position.

6 Conclusions

In this paper, we have proposed an EPI-based multiple camera calibration method for the purpose of city modeling. Using our proposed method in order to capture a large scale scene, users are only required to install multiple video cameras on top of a car, with no exacting configuration, and can then freely scan along a street.

The proposed method can make the optical centers of multiple cameras coincide at one point in spatio-temporal space, thereby enabling distortion-free image fusion. Further, our techniques were efficiently achieved by EPI-EPI matching method, thus, we did not need any external synchronization device and large overlapping areas between images to estimate precise parameters for image fusion; this was the great advantage of our method.

To test the effectiveness of our proposed method, we conducted several experiments using real world sequences. The results of the experiments show the effectiveness of our proposed method to synthesize panorama images and novel view images without distortions.

References

[1] M. Antone and S. Teller. Scalable, absolute position recovery for omni-directional image networks. In *Computer Vision and Pattern Recognition*, 2001.

[2] R. Bolles, H. Baker, and D. Marimont. Epipolar plane image analysis:an approach to determining structure from motion. *Int.J.of Computer Vision*, 1:7–55, 1987.

[3] C. Fruh and A. Zakhor. 3d model generation for cities using aerial photographs and ground level laser scans. In *Computer Vision and Pattern Recognition*, volume 2, pages 31—38, 2001.

[4] A. Lippman. Movie-maps. an application of the optical videodisc to computer graphics. In *Proceedings of ACM SIGGRAPH '80*, pages 32–43, 1990.

[5] S. D. Ma. A self-calibration technique for active vision systems. *IEEE Trans. RA*, 12:114–120, 1996.

[6] Y. Onoue, K. Yamasawa, H. Takemura, and N. Yokoya. Telepresence by real-time view-dependent image generation from omnidirectional video streams. *Computer Vision and Image Understanding*, 71(2):154–165, Aug. 1998.

[7] T. Takahashi, H. Kawasaki, K. Ikeuchi, and M. Sakauchi. Arbitrary view position and direction rendering for large-scale scenes. In *Computer Vision and Pattern Recognition*, volume 2, pages 296–303, June 2000.

[8] K. Yamasawa, Y. Yagi, and M. Yachida. New real-time omnidirectional image sensor with hyperboloidal mirror. In *Proc. 8th Scandinavian Conf. on Image Analysis*, pages 1381–1387, May 1993.

[9] J. Y. Zheng and S. Tsuji. Panoramic representation of scenes for route understanding. In *International Conference on Pattern Recognition*, pages 161–167, June 1990.