3DTV VIEW GENERATION USING UNCALIBRATED CAMERAS

Songkran Jarusirisawad and Hideo Saito

Department of Information and Computer Science, Keio University
3-14-1 Hiyoshi, Kohoku-ku, Yokohama, 223-8522, Japan
songkran,saito@ozawa.ics.keio.ac.jp

ABSTRACT

This paper proposes a method for synthesizing free viewpoint video which is captured by uncalibrated multiple cameras. Each cameras are allowed to be zoomed and rotated freely during capture. Neither intrinsic nor extrinsic parameters of our cameras are known. Projective Grid Space (PGS), which is the 3D space defined by the epipolar geometry of two basis cameras, is employed for calibrating dynamic multiple cameras, because geometrical relations among cameras in PGS are obtained from 2D-2D corresponding points between views. We utilize Keypoint Recognition for finding corresponding points in natural scene for registering cameras to PGS. Moving object is segmented via graph cut optimization. Finally, free viewpoint video is synthesized based on the reconstructed visual hull. In the experimental results, free viewpoint video which is captured by uncalibrated cameras is successfully synthesized using the proposed method.

Index Terms— Image synthesis, Calibration, Image registration

1. INTRODUCTION

In most of free viewpoint video creation from multiple cameras system, cameras are assumed to be fixed by mounting with the poles or tripods thoughtout the capturing. Calibration is only done before starting video acquisition. During video aquisition, cameras cannot be moved, zoomed or even changed view direction. Field of view of each camera in those systems must be wide enough to cover all the area in which the object moves. If the area is large, moving object's resolution in the captured video and also in the free viewpoint video will become very low.

Allowing cameras to be zoomed and changed view direction during capture is more flexible in terms of video acquisition. However, all cameras must be dynamically calibrated every frame. Doing strong calibration every frame with multiple cameras is possible by using some special markers. Marker's size should be large enough comparing to the scene to make calibration accurate. In case that capturing space is large, it's not suitable to use a huge artificial marker.

In this paper we propose method to synthesize free viewpoint video from uncalibrated cameras which allowing cameras to be zoomed and change view direction during capture. Our method does not require special markers or information about cameras parameters. For obtaining geometrical relation among the cameras, Projective Grid Space (PGS)[1] which is 3D space defined by epipolar geometry between two basis cameras is used. After initial frame, Fundamental matrices are reestimated automatically. Keypoint Recognition[2] is used for finding corresponding points between initial frame and the other frame for automatic homography estimation. We recover shape of objects by silhouette volume intersection[3] in PGS. The recovered shape in PGS provides dense correspondences among the multiple cameras, which are used for synthesizing free viewpoint images by view interpolation[4].

1.1. Related Works

Pioneering research in free viewpoint image synthesis of a dynamic scene is Virtualized Reality [5]. In that research, 51 cameras are placed around hemispherical dome called 3D Room to transcribe a scene. 3D structure of a moving human is then reconstruct for rendering from new view.

Many methods for improving quality of free viewpoint image have been proposed. Carranza et al. recover human motion by fitting a human shaped model to multiple view silhouette input images for accurate shape recovery of the human body [6]. Starck optimizes a surface mesh using stereo and silhouette data to generate high accuracy virtual view image [7].

Cameras in the mentioned systems are assumed to be calibrated and fixed thoughtout capturing. In this paper, we propose method to synthesize 3D video from uncalibrated cameras which can also be zoomed or rotated during capture. This situation can be apply to the case that cameras are place on tripod or the case that handy cameras are capture by man where the capturing position is not much change during capture as well.

978-1-4244-1755-1/08/$25.00 ©2008 IEEE

2. OVERVIEW

To reconstruct a 3D model without full camera calibration, we utilize Projective Grid Space (PGS)[1] which is weak calibration framework based on epipolar geometry. 3D space in PGS is defined by image coordinates of two basis cameras as Fig.1. All cameras can be weakly calibrated to PGS by fundamental matrices between basis cameras.

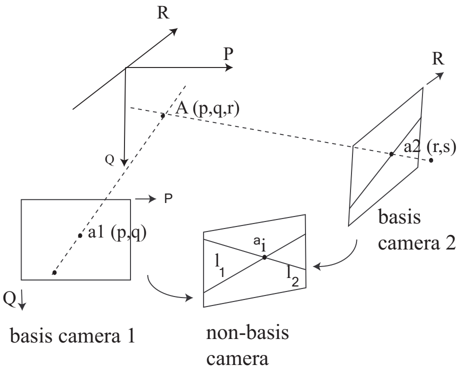


Fig. 1. Definition of Projective Grid Space.

Without any assumption about cameras movement, we must calibrate each frame to PGS by estimating fundamental matrices from 2D-2D correspondences between cameras. However, tracking or finding such corresponding points in 3D complex scene where viewpoint different is large is difficult to acheive robustly as shown in [8]. Two images from different views have very different appearance due to motion parallax.

To reduce this problem, we limit that capturing position of cameras is not much changed during capture but can zoom and rotate freely. At initial frame of each camera, we capture the whole background scene without moving object. We select two cameras for defining PGS and estimate initial fundamental matrices by assigning corresponding points manually. To recalibrate the current frames to PGS, fundamental matrices are needed to be reestimated. As we will show in section 3.2, we can reestimate fundamental matrices of each cameras from homography matrices between current frame and initial frame of the same camera. Because capturing position of initial frame and the current frames are the same, there is no motion parallax between these images. Two images are approximately 2D similarity. Accurate corresponding points can be found automatically as will be described in section 3.2 .

In our experiment, we use 4 hand-held cameras capturing from positions like Fig.2. All cameras are zoomed and rotated independently during capture. The overall process is illustrated in Fig.3.

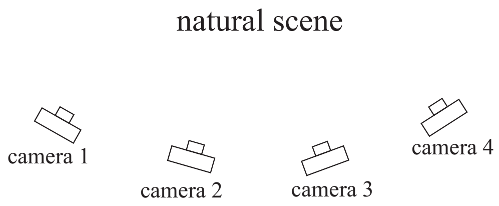


Fig. 2. Cameras Configuration.

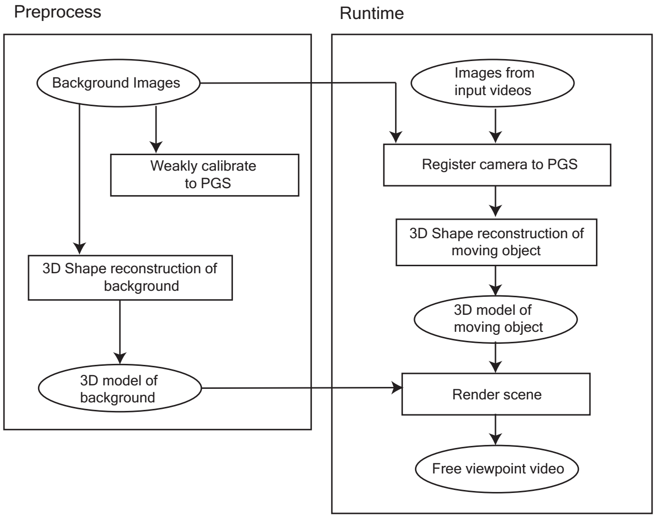


Fig. 3. Overall Process.

3. WEAK CALIBRATION

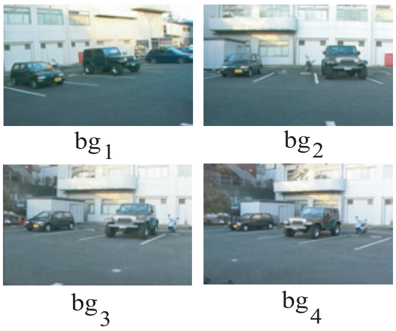
3.1. Preprocess

At initial frame, we zoom out all cameras to capture the whole area of a scene without object. We call this background image of camera i as bg$_i$. We select camera1 and camera4 as basis cameras defining PGS. 2D-2D Corresponding points for estimating fundamental matrices between basis cameras and other cameras are assigned manually on bg$_i$ image during preprocess. Once fundamental matrices are estimated, PGS is completely defined. These images will be used as reference image for register hand-held cameras to PGS as will be described in section3.2. Fig.4 shows background images of our experiment.

3.2. Runtime

During capture input video, object will move around a large space. Each camera is zoomed and rotated to capture moving object with high resolution in the image. View and focal length of each camera are changed from initial frame. Fundamental matrices are needed to be reestimated for redefining PGS.

Suppose that fundamental matrix of background images from camera i to camera j is F_{ij}. From the assumption that



bg_1 bg_2 bg_3 bg_4

Fig. 4. Background Images.

capturing position of each cameras is not much changed during capture, we can reestimate fundamental matrix $\mathbf{F}'_{ij}$ of the current frame between cameras i and j using equation 1

$$\mathbf{F}'_{ij} = \mathbf{H}_j^T \mathbf{F}_{ij} \mathbf{H}_i \tag{1}$$

where H_i is the homography matrix that transfer image coordinate of the other frame of camera i to bg_i image. H_j is also defined in the same way.

To estimate homography matrix, corresponding points between bg_i and the other frames are necessary. We employ Keypoint Recognition using Randomized Trees[2] for finding such corresponding points. In our previous work[9] SIFT (Scale Invariant Feature Transform)[10] was used. However Using Keypoint Recognition has more advantage in terms of computation time. Keypoint Recognition need long time in learning phase (several minutes on Pentium 4 cpu), but can finding corresponding points after training very fast compare to SIFT. In our case, template is fix (background image) so it's suitable to use Keypoint Recognition.

Example corresponding points that automatically found using Keypoint Recognition are shown in Fig.5. In Fig.5, the upper image is bg_i image and the bottom image is the other frame which will be registered to Projective Grid Space. The lines show corresponding points which will be used for estimating homography.

4. 3D RECONSTRUCTION AND RENDERING

To segment silhouette of moving object, virtual background image of the input frame are created by warping initial frame where there is no moving object using the same homography for registering cameras to PGS. Graph cut optimization is then used for silhouette segmentation.

3D shape of moving object is reconstructed by silhouette volume intersection[3] in PGS. The recovered shape in PGS provides dense correspondences among the cameras, which are used for synthesizing free viewpoint images by view interpolation[4].

In constrast to our previous work[9], they do not reestimate fundamental matrices but simply use the estimated homographies for warping projected points from PGS on background images to the current frames during voxel reconstruction. If we fix the voxels number in PGS, PGS defined on background images covers the whole scene, but PGS defined on current images covers only current interested area which gives more resolution in reconstructed model. Fig.6 show comparison between 3D model reconstructed from PGS defined by background images and defined by current images.

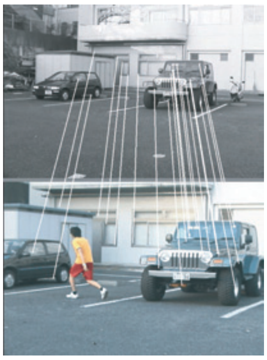
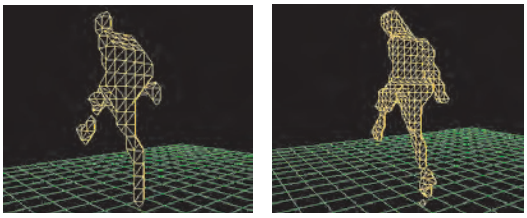


Fig. 5. Corresponding Points Found Using Keypoint Recognition for Estimating Homography.



(a) (b)

Fig. 6. Comparison of Reconstructed 3D Mesh Model of human. (a) PGS Defined by Background Images (b) PGS Defined by Current Images

5. EXPERIMENTAL RESULTS

In this section, we evaluate our proposed method by synthesizing free viewpoint images from the captured videos. The experimental environment is a large natural scene as Fig.4. We use 4 Sony-DV cameras with 720x480 resolutions in both experiments. All cameras are in front of the scene as in Fig.2.

During capture, each camera have been zoomed from 1X to 2X and changed view direction about -40 to +40 degree to capture moving object independently. There is no artificial marker placed in the scene.

Fig.7 shows one frame from input videos. The result free viewpoint images between camera2 and camera3 are shown in Fig.8. Ratio of virtual camera position between two views is written under each figure. We can see that the rendered background planes, static objects and moving object from both reference views are correctly aligned and merged in the free viewpoint images. Occlusion areas between two reference views, e.g. motorcycle, are also correctly rendered.

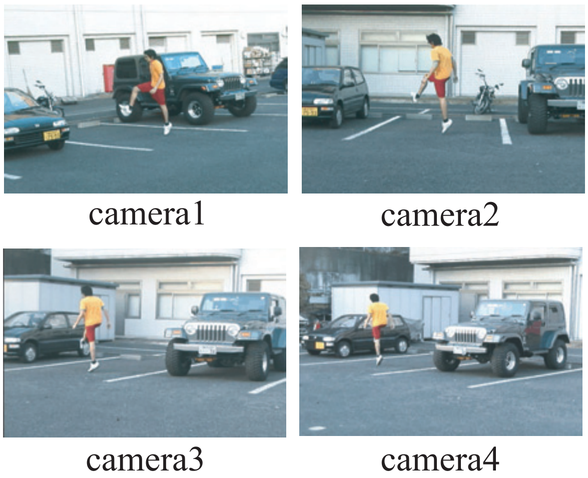


Fig. 7. One Frame from Four Input Videos.

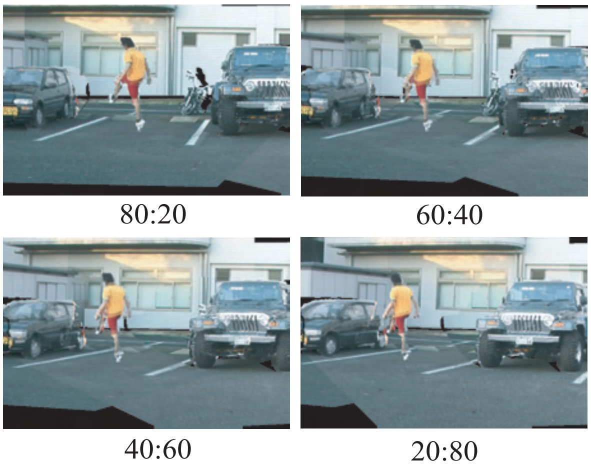


Fig. 8. Free Viewpoint Images Between Camera2 and Camera3.

6. CONCLUSIONS

We proposed a method for synthesizing free viewpoint video of a moving object in natural scene, which is captured by hand-held multiple cameras. Our method allows cameras to be zoomed and changed view direction during capture. Fundamental matrices are automatically estimated from 2D-2D corresponding points in the scene for calibrating multiple cameras to PGS [1]. Thus our calibration method is done without special markers and captured area can be more wider compare to the other methods which assume that cameras are fixed.

Acknowledgement

This work is supported in part by a Grant-in-Aid for the Global Center of Excellence for High-Level Global Cooperation for Leading-Edge Platform on Access Spaces from the Ministry of Education, Culture, Sport, Science, and Technology in Japan. The work presented in this paper is partially supported by CREST, JST (Research Area: Foundation of technology supporting the creation of digital media contents).

7. REFERENCES

[1] Hideo Saito and Takeo Kanade, "Shape reconstruction in projective grid space from large number of images," in *CVPR '99*, June 1999.

[2] Vincent Lepetit and Pascal Fua, "Keypoint recognition using randomized trees," *IEEE Trans. on Pattern Analysis and Machine Intelligence*.

[3] A. Laurentini, "The visual hull concept for silhouette based image understanding," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 16, no. 2, pp. 150–162.

[4] S. Chen and L. Williams, "View interpolation for image synthesis," in *SIGGRAPH'93*, pp. 279–288.

[5] T. Kanade, P. W. Rander, and P. J. Narayanan, "Virtualized reality: concepts and early results," in *IEEE Workshop on Representation of Visual Scenes*, 1995, pp. 69–76.

[6] J. Carranza, C. Theobalt, M. Magnor, and H.-P. Seidel, "Free-viewpoint video of human actors," in *SIGGRAPH'03*, pp. 569–577.

[7] J. Starck and A. Hilton, "Towards a 3D virtual studio forhuman appearance capture," *IMA International Conference on Vision, Video and Graphics (VVG)*, pp. 17–24, 2003.

[8] Pierre Moreels and Pietro Perona, "Evaluation of features detectors and descriptors based on 3d objects," *International Journal of Computer Vision*, vol. 73, no. 3, pp. 263–284, 2007.

[9] S. Jarusirisawad and H. Saito, "Free viewpoint video synthesis based on visual hull reconstruction from handheld multiple cameras," in *ACCV'07 Workshop on Multi-dimensional and Multi-view Image Processing*.

[10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, 2004.