DIMINISHED REALITY VIA MULTIPLE HAND-HELD CAMERAS

Songkran Jarusirisawad, 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 novel method for calibrating multiple hand-held cameras target for Diminished Reality application. Our method does not require any special markers or information about camera parameters. Projective Grid Space (PGS) which is 3D space defined by epipolar geometry of two basis cameras is used for dynamic cameras calibration. Geometrical relations among cameras in PGS are obtained from 2D-2D corresponding points between views. We utilize Scale Invariant Feature Transform (SIFT) for finding corresponding points in natural scene for registering cameras to PGS. Moving object is segmented via graph cut optimization. Finally, the reconstructed visual hull is used to synthesize free viewpoint video in which unwanted or occluding object is deliberately removed. In the experimental results, free viewpoint video without unwanted object which is captured by hand-held cameras is successfully synthesized using the proposed method.

Index Terms— Diminished Reality, Projective Grid Space, Cameras Calibration, 3D Reconstruction

1. INTRODUCTION

The objective of Diminished Reality is to deliberately remove some unwanted object from the image sequence or video. This kind of research has become popular topic recently. In this paper we introduce a combination between free viewpoint video synthesis and diminished reality. Our goal is to synthesize free viewpoint video in which unwanted object is also removed. Our chalenge is multiple hand-held cameras, which are need to be dynamically calibrated, are used to capture the input video.

In this paper we present a novel method for calibrating multiple hand-held cameras for diminished reality application. Our method of cameras calibration 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. All other cameras can be related to the PGS by fundamental matrices. Fundamental matrices for relating every cameras are estimated once at initial setting. After that, all cameras are dynamically registered to PGS via homography matrices. SIFT[2] is used for finding corresponding points between initial frame and the other frame for automatic homography estimation. We recover shape of the moving objects by silhouette volume intersection[3] in PGS. The recovered shape in PGS provides dense correspondences among the multiple cameras for synthesizing free viewpoint video in which unwanted object is also removed.

1.1. Related Works

One of the earliest researches for free viewpoint image synthesis of a dynamic scene is Virtualized Reality [4]. In that research, 51 cameras are placed around hemispherical dome called 3D Room to transcribe a scene. 3D structure of a moving human is extracted using multi-baseline stereo (MBS) [5]. Then free viewpoint video is synthesized from the recovered 3D model.

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]. Saito et al. propose appearance-based method [8], which combines advantage from Image Based Rendering and Model Based Rendering.

In the mentioned systems and most of the previous researches on free viewpoint image synthesis, they propose the systems that use calibrated fix cameras. Cameras in those systems are arranged to the specified positions around a scene and calibrated before capturing. During video acquisition, camera parameters must be the same, so cameras can not be moved or zoomed. Field of view (FOV) of all cameras must be wide enough to cover the whole area in which object moves. If the object moves around a large area, the moving object's resolutions in the captured video will not enough to synthesize a good quality free viewpoint image. One possible way to get enough resolution of input video is using hand-held cameras which allow to be zoomed or changed view direction during capture. However all cameras must be dynamically calibrated.

In terms of Diminished Reality, there are several researches proposed methods for removing unwanted object from input video. Mann et al. [9] proposed a method for removing the a planar object from input video and replacing with another texture. Wang et al. [10] proposed a method for segmenting a sequence of video images into multiple layers and rendering the same video but one layer is removed. Lepetit et al. [11] proposed a semi-automatic method for tracking a boundary of moving images and detecting the occlusion to remove the object from the scene. Most of research on Diminished Reality are dealing with one camera input. Using multiple cameras give us more information about the scene.

In this paper, we introduce a combination system between free viewpoint video synthesize and diminished reality from multiple hand-held cameras. Our goal is to synthesize free viewpoint video in which unwanted object is also removed. Main contribution of this paper is the method for calibrating hand-held cameras in natural scene without special markers.

2. OVERVIEW

There are two main difficulties of using hand-held cameras to capture the scene. The first one is cameras must be dynamically calibrated where there is no any special marker. Weak calibration technique like Projective Grid Space (PGS) [1], require only 2D-2D correspondences between cameras. Using PGS, there is no need of special markers. However, tracking or finding such corresponding points in 3D complex scene is difficult to acheive robustly as shown in [12]. Two images from different views have very different appearance due to motion parallax. To make the system practical for synthesizing a long video sequence, these calibration task must be done automatically.

The second difficulty is silhouette segmentation of moving object for 3D shape reconstruction. If cameras are static, background scene can be captured beforehand, so it is trivial to get silhouette image using simple background subtraction. In the case that hand-held cameras are used, background image of these cameras cannot be captured before hand because it is impossible to recapture the scene with the same trajectory and zoom.

To reduce these problems, we assume that capturing position of hand-held cameras is not much changed during capture. Even such assumption, we can still give a flexibility of allowing hand-held cameras to be zoomed and/or changed viewpoint to capture moving object, e.g. camera is hold by man and rotated/zoomed freely. By using this assumption we can resolve two stated problems as following.

At initial frame of each input video, we capture the whole background scene without moving object. We select two cameras for defining PGS and weakly calibrate initial frames to PGS by assigning corresponding points manually. To register the other frames to PGS, homographies which relate those frames to initial frames are estimated automatically. Because capturing position of initial frame and the other frames are almost the same, there is no motion parallax between these images. Two images are approximately 2D similarity. Accurate corresponding points can be found automatically using SIFT as will be described in section 4.2 .

For silhouette segmentation of moving object, background image of moving cameras are created by warping initial frame where there is no moving object using the same homography for registering moving cameras to PGS.

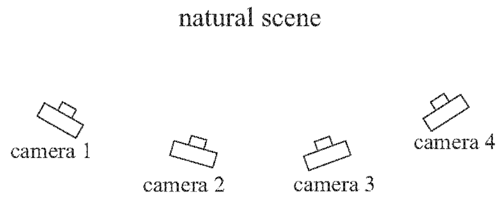


Fig. 1. Cameras Configuration.

In our experiment, we use four hand-held cameras capturing from positions like Fig.1. All cameras are zoomed and rotated independently during capture. The overall process is illustrated in Fig.2.

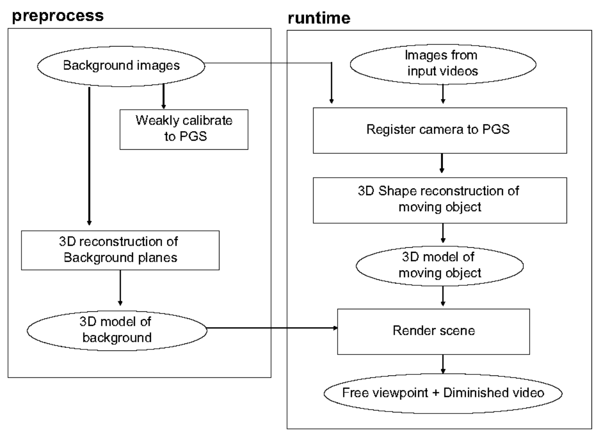


Fig. 2. Overall Process.

3. PROJECTIVE GRID SPACE

Reconstructing 3D model for diminished reality requires a relation between 3D world coordinate and 2D image coordinate. Projection matrix that represents this relation can be estimated by strong camera calibration which requires 3D-2D correspondences. Measuring 3D-2D corresponding points requires a lot of work. Moreover, in case of a large natural scene, it is difficult to precisely measure calibrating points throughout all the area.

To remove effort of obtaining strong calibration data, we use a weak calibration framework, called Projective Grid Space

(PGS) [1], for shape reconstruction. 3D coordinate in PGS and 2D image coordinate is related by epipolar geometry using fundamental matrices. To estimate fundamental matrices between views, only 2D-2D correspondences which can be directly measured from input videos are required.

3D space in PGS is defined by image coordinates of two arbitrarily cameras. These two cameras are called the basis camera1 and the basis camera2. The nonorthogonal coordinate system P-Q-R is used in PGS. The image coordinates x and y of basis camera1 corresponds to the P and Q axis in PGS. Image coordinate x of the basis camera2 corresponds to the R axis.

Fig.3 illustrates how PGS is defined. 3D coordinate A (p,q,r) in PGS is projected on image coordinate a_1 (p,q) of the basis camera1 and on image coordinate a_2 (r,s) of the basis camera2. a_2 is the point on epipolar line of point a_1 where image coordinate x equals to r.

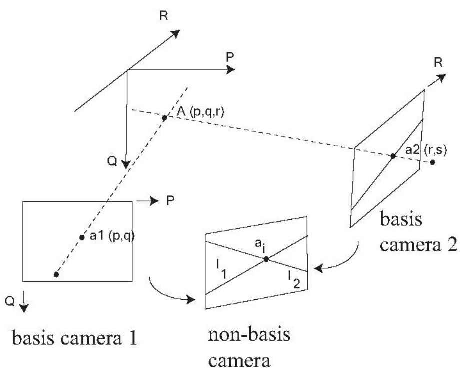


Fig. 3. Definition of Projective Grid Space.

Other cameras can be related to PGS by fundamental matrices between 2 basis cameras. Finding such fundamental matrices required only 2D-2D correspondences. So, it is relatively easy comparing to full calibration which required 3D-2D correspondences. 3D coordinate A (p,q,r) in PGS is projected onto non-basis camera at point a_i which is the intersection between epipolar line l_1 and l_2 as shown in Fig.3

4. WEAK CALIBRATION

4.1. 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, 3D coordinate in PGS can be project to all bg_i images. These images will be used for generate virtual background for background subtraction and also used as reference image for register moving cameras to PGS as will be described in section4.2. Fig.4 shows background images of our experiment.

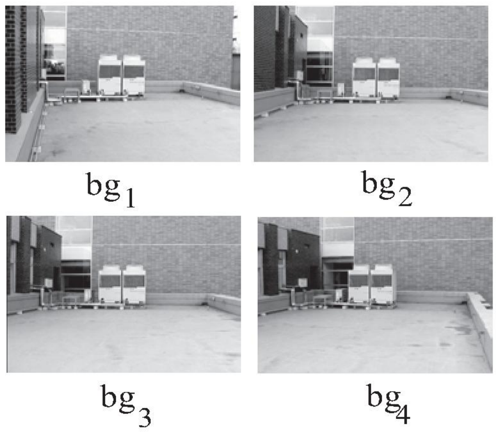


Fig. 4. Background Images.

4.2. Runtime

During capture input video, moving objects 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 estimated during preprocess can not be used to project 3D coordinate in PGS to 2D coordinate of the other frames. In this section we will explain method for calibrating other frames to PGS. We denote the frame that is going to be calibrated to PGS of camera i as img_i.

From the assumption that capturing position of each cameras is not much changed during capture, 2D coordinate of bg_i can be transformed to 2D coordinate of img_i using homography matrix. To estimate homography matrix, corresponding points between bg_i and the img_i are necessary.

We employ SIFT (Scale Invariant Feature Transform)[2], which is the method for extracting features from images that can be used to perform reliable matching, for finding such corresponding points. SIFT is robust for finding corresponding points between images which there are different in scale and 2D rotation. However, reliable of matching will decrease if there are much different in view appearance between two images [12]. In our case, two images are captured from approximately the same position. There is no motion parallax between these images. Two images are approximately 2D similarity regardless of complexity of a scene. Therefore, SIFT is robust for using in our system.

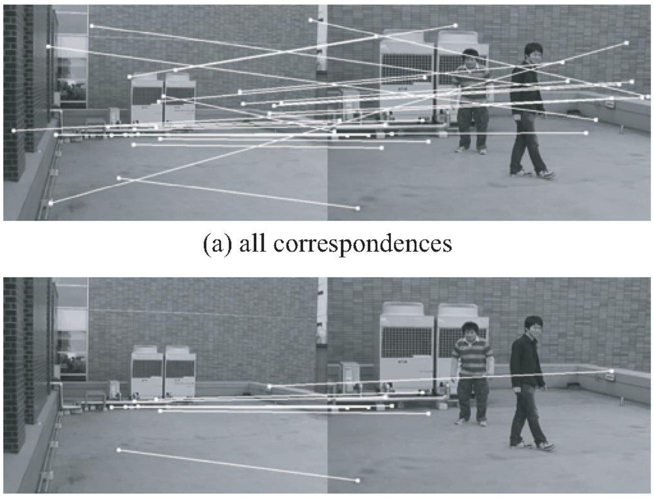
Coorresponding point initially found by SIFT include some outlier. We employ RANSAC (RAndom Sample Consensus)[13] to remove those outliers. Only inliers are used for estimating homography.

3D coordinate $A(p, q, r)$ in PGS which is projected on (x_{bg}, y_{bg}) of bg_i image is projected to img_i image at (x_{img}, y_{img}) by equation 1.

$$s \begin{pmatrix} x_{img} \\ y_{img} \\ 1 \end{pmatrix} = H_i \begin{pmatrix} x_{bg} \\ y_{bg} \\ 1 \end{pmatrix} \tag{1}$$

where H_i is homography matrix between bg_i image and img_i

Example corresponding points that automatically found using SIFT are shown in Fig.5. In Fig.5, the left image is bg_i image and the right image is img_i image which will be registered to Projective Grid Space. Fig.5(a) shows all corresponding points found by SIFT while Fig.5(b) shows only inliers after outliers are removed by RANSAC.



(a) all correspondences

(b) inliers

Fig. 5. Corresponding Points Found Using SIFT for Estimating Homography.

4.3. Homography Refinement

In this section we describe method for improving accuracy of homography H_i which is initially estimated in section4.2. First, we warp bg_i image using homography H_i to create new image called $\mathrm{warpedbg}_i$. Optical flow between $\mathrm{warpedbg}_i$ and img_i is then computed. If homography H_i is accurate, magnitude of optical flows between $\mathrm{warpedbg}_i$ and img_i where there is no foreground object should be around zero which means that $\mathrm{warpedbg}_i$ is perfectly align with img_i.

Accuracy of homography H_i between bg_i and img_i depends on number of correspondences found by SIFT and also location of corresponding points in images as well (correspondences should distribute all over image area to give accurate homography).

Fig.6 shows optical flows between $\mathrm{warpedbg}_i$ and img_i (the same frame as Fig.5). Comparing with Fig.5(b), magnitude of optical flows are almost zero where there are corresponding points for estimating initial homography H_i and become larger as it is far away.

$\mathrm{warpedbg}_\mathrm{i}$ img_i

Fig. 6. Optical flows between warped background and current frame. (Image is cropped to show detail)

To improve accuracy of H_i, homography $H_{refine\ i}$ which transforms coordinate from $\mathrm{warpedbg}_i$ to img_i is estimated by fitting correspondences from optical flows using RANSAC. More accurate homography $H_{accurate\ i}$ for registering hand-held camera to PGS is then computed from equation 2. Finally, $H_{accurate\ i}$ is then used instead of H_i by replacing H_i in equation 1 with $H_{accurate\ i}$.

$$H_{accurate\ i} = H_{refine\ i} H_i \tag{2}$$

5. SEGMENTATION

To reconstruct 3D model of moving objects, each moving object's silhouette is needed to be segmented labeling. Background image of camera i during runtime is generated by warping bg_i image using homography which is estimated automatically in section4.2. Example generated background images for moving camera are shown in Fig.7. After generating virtual background for hand-held cemera, silhouettes are segmented for reconstructing visual hull.

Using background subtraction by simply set thresholding, post processing like morphological operation are necessary. Such morhological operation can remove hole in silhouette but tends to enlarge silhouette to be bigger than the real shape. Accuracy of silhouette segmentation is a crucial step for reconstruct visual hull.

To get accurate visual hull, we use graph cut optimization for silhouette segmentation. We use energy function proposed in Background Cut [14]. Fig.7 shows segmentation result. Silhouette segmentation using graph cut give more accurate result compared to simple thresholding with postprocessing (morphological operation).

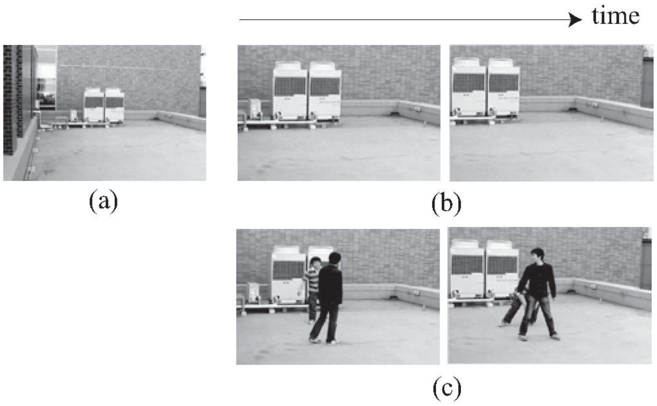


Fig. 7. (a) Background Image. (b)Automatically Generated Background Images from (a). (c) Some Frames from Input Video of the Same Camera as (a).

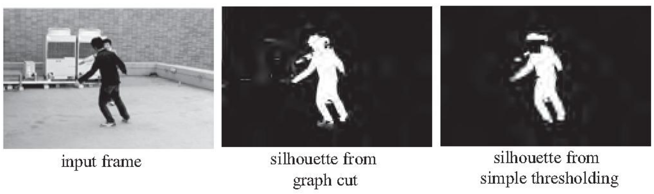


Fig. 8. Silhoutte Segmentation.

After silhouette segmentation of all moving objects, each object in blob must be labeled so that we can know which 3D model should be removed in the final diminished image. Several object tracking algorithms under occlusion in video sequence can be use for automatic tracking [15]. In this works we emphasize on our novel the method for calibrating multiple hand-held cameras, so we do silhouette labeling by manually.

6. 3D RECONSTRUCTION

We consider that objects in input video consist of

- Background plane
- Moving object (Human)

which will have the different way to recover the 3D information for rendering free viewpoint image. Background plane is the real plane like a floor or the scene which is far away so that can be approximated as planar scene. Fig.9 shows how background scene is catagorized. Background scene is not changed during video capture, so 3D information of these are estimated only once during preprocess while 3D shape of moving object is reconstructed automatically every frame.

Fig. 9. Background Scene Consist of Planes and Static Objects.

6.1. Preprocess

During preprocess we reconstruct 3D position of several points which lie on a plane by assigning corresponding points between background image of two basis cameras defining PGS manually. Let $a_1(p,q)$ be 2D coordinate of point in basis camera1 and $a_2(q,r)$ is 2D coordinate of point in basis camera 2, 3D position of this point is (p,q,r) from definition of PGS in section3. These 3D position in PGS will be used for rendering planes in free viewpoint video as will be explained in section7.1.

6.2. Runtime

Visual hull of moving object is reconstructed using silhouette volume intersection method[3]. Voxels in PGS are projected onto bg_i image first, then the projected 2D coordinate is transfered to current frame using equation 1. Voxel is considered to be in a 3D model volume if projected points of all cameras are in silhouette. Surfaces of 3D voxel model are extracted to 3D triangular mesh model using Marching Cube algorithm[16]. Fig.10 shows 3D model of moving object reconstructed in PGS.

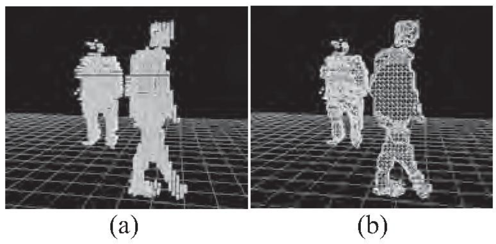


Fig. 10. 3D Model of Human in PGS. (a) Volumetric Representation (b) Triangular Mesh Representation.

7. RENDERING

To render free viewpoint image without unwanted moving object, output image is rendered in two steps. First we synthesize background scene without any object. Only interested moving object is then rendered overlay to the rendered background. The following subsections explain the detail of two rendering phase.

7.1. Background Scene Rendering

During preprocess, 3D position of points which lie on planes are already reconstructed. These 3D position in PGS are projected onto both reference views. 2D positions of these points on free viewpoint image are determined using linear interpolation as equation 3

$$\begin{pmatrix} x \\ y \end{pmatrix} = w \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + (1-w) \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} \tag{3}$$

where w is a weight, ranging from 0 to 1, defining the distance from virtual view to second reference view. $(x_1, y_1)^T$ and $(x_2, y_2)^T$ are corresponding points on the first reference view and the second reference view respectively.

Corresponding points between background image of reference view and virtual view are used for estimating homography. Plane in background image which is segmented manually during preprocess is warped and blended to virtual view. Fig.11 illustrates how the plane is rendered in free viewpoint image.

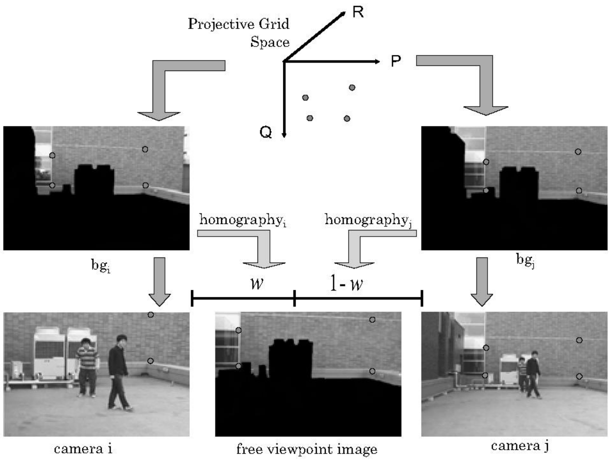


Fig. 11. Rendering Plane on Free Viewpoint Image.

7.2. Objects Rendering

Once background planes are rendered, all moving objects are diminished from output video. Only interested moving object will be renderd overlay to the rendered background. 3D triangular mesh model of each moving object is combined together. The combined 3D model is used for making depth map to test occlusion between reference images. After making depth map, only 3D model of the interested moving object is used for making dense corresponding points for view interpolation [17].

To test occlusion triangular patches, Z-Buffer of each camera is generated. All triangle patches of a combined 3D model are projected onto Z-Buffer of each camera. Pixel value of Z-Buffer is store the 3D distance from camera's optical center to the projected triangle patch. If some pixels are projected by more than one patch, the shortest distance is stored.

To synthesize free viewpoint image, each triangle mesh of the interested moving object is projected onto two reference images. Z-Buffer is used to test occlusion. The patches which are occluded in both two input views will not be interpolated in a free viewpoint image. Position of a warped pixel in free viewpoint image is determined by equation 3.

To merge two warped triangular patch, RGB colors of the pixel are computed by the weighted sum of the colors from both warped patch. If a patch is seen from both input view, weight that use for interpolating RGB color is the same for determining position of a patch. In case that patch is occluded in one view, weight of occluded view is set to 0 while the other view is set to 1. Fig.12 shows example of free viewpoint image of interested moving object.

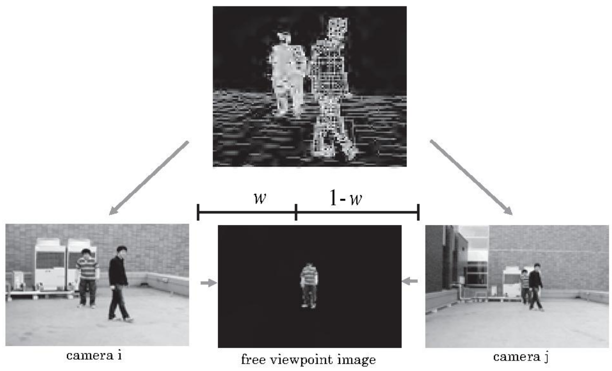


Fig. 12. Rendering Moving Object on Free Viewpoint Image.

8. EXPERIMENTAL RESULTS

In this section, we show and evaluate the result from our proposed method. The experimental environment is a large natural scene as Fig.4. We use four Sony-DV cameras with 720x480 resolutions in both experiments. All cameras are in front of the scene as in as in Fig.1.

During video capture, hand-held cameras have been zoomed and changed view direction to capture high resolution moving object independently. There is no artificial marker placed in the scene. Only natural feature are used for cameras calibration. We select several frames to calibrate multiple hand-held cameras by our proposed method. Our method can correctly register all frames to PGS for doing 3D reconstruction with out manual operation.

Silhouette segmentation by using Graph cut optimization is better than using simple thresholding but still has some artifact if color between foreground and background are very similar. For our input video, background and foreground color are almost the same in some views. The result shown in this section was calibrated automatically but segmented and labeled manually.

Fig.13 shows sampled input frames for synthesizing free viewpoint video where unwanted object is removed.

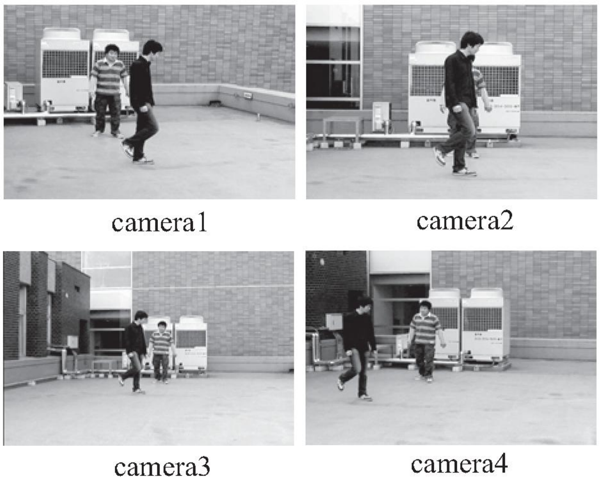


Fig. 13. One Frame from Four Input Videos.

The result free viewpoint images in which unwanted object is removed between camera1 and camera2 are shown in Fig.14 and Fig.15. Ratio between two views is written under each figure. We can see that the rendered background planes and moving object from both reference views are correctly aligned and merged in the free viewpoint images while unwanted object is also completely removed.

Computation time of four input videos on 2.0 GHz CPU, from registering camera to PGS until 3D model reconstruction, takes about 30 seconds per frame. To render new viewpoint video, our system takes about 2 seconds per viewpoint.

There are some blured texture of static objects and moving object in synthesized image due to inaccuracy of 3D model. This accuracy can be improved easily by increasing number of cameras in the system. In terms of reconstruction algorithm, other technique can be used together with shape from silhouette to improve quality of reconstructed 3D model as well [18].

9. CONCLUSIONS

In this paper we introduce a combination between free viewpoint video and diminished reality. Possible applications is for surveillance system in which one can monitoring the scene without occluding object. Multiple hand-held cameras in our system are automatically calibrated. Projective Grid Space (PGS) [1] which is 3D space defined by epipolar geometry of two basis cameras is used for visual hull reconstruction. By using PGS, geometrical relationship among cameras can be obtained from 2D-2D corresponding points between views. We use SIFT [2] to find corresponding points in natural scene for dynamically registering cameras to PGS. Graph cut optimization is used for silhouette segmentation.[14] In the experiment, multiple hand-held cameras are correctly calibrated and diminished video from virtual viewpoint are successfully synthesized by our proposed method.

10. REFERENCES

[1] Hideo Saito and Takeo Kanade, "Shape reconstruction in projective grid space from large number of images," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '99)*, June 1999.

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

[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] 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.

[5] M. Okutomi and T. Kanade, "A multiple-baseline stereo," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 4, pp. 353–363.

[6] J. Carranza, C. Theobalt, M. Magnor, and H.-P. Seidel, "Free-viewpoint video of human actors," in *Proceedings of 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] H. Saito, S. Baba, and T. Kanade, "Appearance-based virtual view generation from multicamera videos captured in the 3-d room," *IEEE Transactions on Multimedia*, vol. 5, no. 3, pp. 303–316.

[9] James Fung Steve Mann, "Video orbits on eye tap devices for deliberately diminished reality or altering the visual perception of rigid planar patches of a real world scene," in *International Symposium on Mixed Reality (ISMR2001)*, March 2001.

[10] E. Adelson J. Wang, "Representing moving images with layers," *IEEE Trans. on Image Processing*, vol. 3, pp. 625–638, September 1994.

Fig. 14. Diminished Occluded Object Between Camera1 and Camera2.

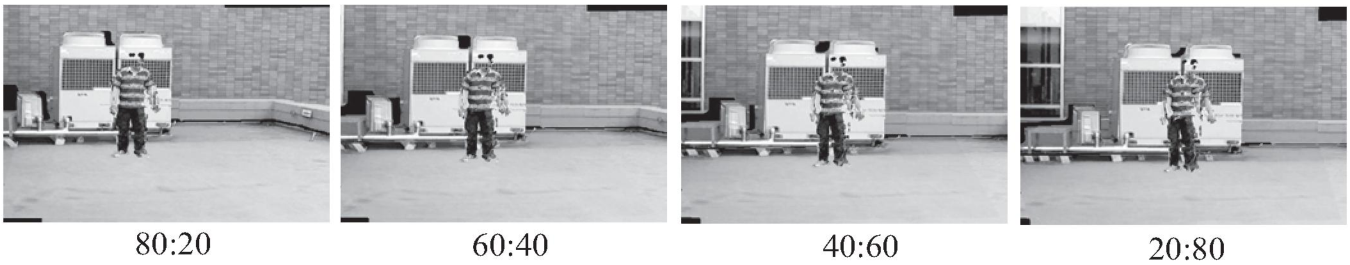


Fig. 15. Diminished Occluding Object Between Camera1 and Camera2.

[11] Vincent Lepetit and Marie-Odile Berger, "A semi automatic method for resolving occlusions in augmented reality," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, June 2000.

[12] 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.

[13] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with application to image analysis and automated cartography," *Communication Association and Computing Machine*, vol. 24, no. 6, pp. 381–395, 1981.

[14] Jian Sun, Weiwei Zhang, Xiaoou Tang, and Heung-Yeung Shum, "Background cut," in *ECCV 2006*, 2006, vol. 2, pp. 628–641.

[15] Justus Piater Andr Genon Pierre Gabriel, Jacques Verly, "The state of the art in multiple object tracking under occlusion in video sequences," in *Proceedings of the Advanced Concepts for Intelligent Vision Systems*, 2003, pp. 166–173.

[16] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3d surface construction algorithm," in *SIGGRAPH '87: Proceedings of the 14th annual conference on Computer graphics and interactive techniques*, 1987, pp. 163–169.

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

[18] S. Yaguchi and H. Saito, "Improving quality of free-viewpoint image by mesh based 3d shape deformation," in *Proceedings of The 14th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision(WSCG2006)*, February 2006.