Free Viewpoint Video Synthesis based on Visual Hull Reconstruction from Hand-Held Multiple 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 novel method for synthesizing free viewpoint video which is captured by hand-held multiple cameras. Cameras in our system are all uncalibrated. Neither intrinsic nor extrinsic parameters are known. 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, free viewpoint video is synthesized based on the reconstructed visual hull. In the experimental results, free viewpoint video which is captured by hand-held cameras is successfully synthesized using the proposed method.**

I. INTRODUCTION

In most of free viewpoint video creation from multiple cameras system, fixed cameras are usually used to capture input videos. Cameras in those systems are mounted with the poles or tripods. Calibration is necessary to be 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.

Using multiple hand-held cameras is more flexible in terms of video aquisition and cameras setting. Hand-held cameras need no stable place to be mounted and can change viewpoint/zoom during capture. However, hand-held cameras must be dynamically calibrated every frame. Doing strong calibration with multiple hand-held cameras is possible by using some special markers[8]. 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 a novel method for synthesizing free viewpoint video in natural scene from hand-held multiple cameras. Our method does not require special markers or information about cameras parameters. For obtaining geometrical relation among the cameras, Projective Grid Space (PGS)[17] 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[11] 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[9] 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[2].

A. Related Works

One of the earliest researches for free viewpoint image synthesis of a dynamic scene is Virtualized Reality [7]. 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) [15]. Then free viewpoint video is synthesized from the recovered 3D model. Moezzi et al. synthesize free viewpoint video by recovering visual hull of the objects from silhouette images using 17 cameras [12]. Their approach creates true 3D models with fine polygons. Each polygon is separately colored thus requiring no texture-rendering support. Their 3D model can use standard 3D model format such as VRML (Virtual Reality Modeling Language) delivered though the internet and viewed with VRML browsers.

Many methods for improve 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 [1]. Starck optimizes a surface mesh using stereo and silhouette data to generate high accuracy virtual view image [18]. Saito et al. propose appearance-based method [16], which combines advantage from Image Based Rendering and Model Based Rendering.

In terms of computation time, real-time systems for synthesizing free viewpoint video have also been developed recently [5], [4], [14]. 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 acqui-

sition, 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.

Ito et al. overcome this problem by proposing a system for synthesizing free viewpoint image using moving cameras [6]. They show a possibility of synthesizing high resolution free viewpoint image by capturing a target with sufficient resolution. However, they demonstrate only in the case that a clear homogeneous color background with some artificial markers are placed around the scene because of the difficulty of feature point tracking for the moving camera. In this paper, we proposed a method for synthesizing free viewpoint video of a human captured by hand-held cameras in natural scene without any special markers.

II. OVERVIEW

There are two main difficulties of using hand-held cameras in free viewpoint video system. The first one is cameras must be dynamically calibrated where there is no any special marker. Weak calibration technique like Projective Grid Space (PGS) [17], 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 [13]. 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 IV-B .

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.

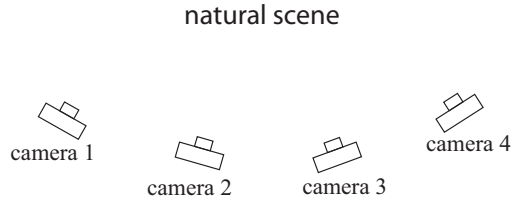

Fig. 1. Cameras Configuration.

In our experiment, we use 4 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.

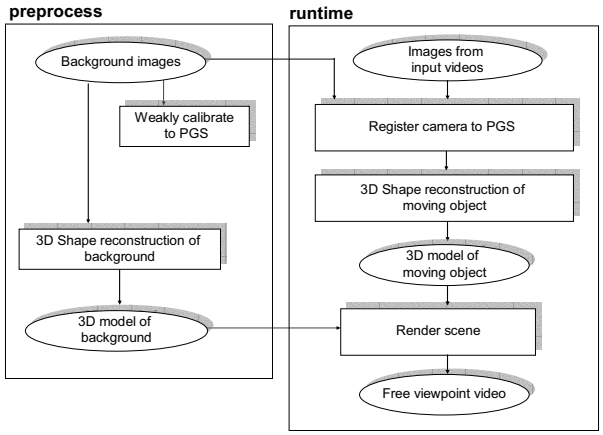

Fig. 2. Overall Process.

III. PROJECTIVE GRID SPACE

Reconstructing 3D model for synthesizing free viewpoint image 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) [17], 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.

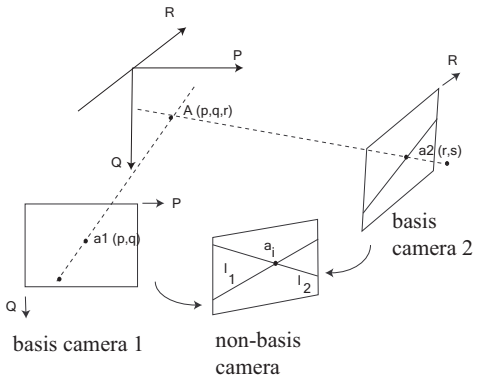

Fig. 3. Definition of Projective Grid Space.

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

IV. WEAK CALIBRATION

A. 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 sectionIV-B. Fig.4 shows background images of our experiment.

B. 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 estimated during preprocess can not be used to project 3D coordinate in PGS to 2D coordinate of the other frames.

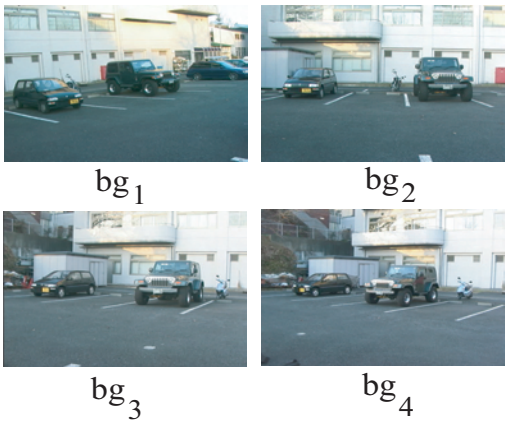

Fig. 4. Background Images.

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 the other frames of the same camera using homography matrix.

To estimate homography matrix, corresponding points between bg_i and the other frames are necessary. We employ SIFT (Scale Invariant Feature Transform)[11], 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 [13]. 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)[3] using homography constraint to remove those outliers. Only inliers are used for finding accurate homography.

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

$$s \begin{pmatrix} x_{other} \\ y_{other} \\ 1 \end{pmatrix} = H_i \begin{pmatrix} x_{bg} \\ y_{bg} \\ 1 \end{pmatrix} \quad (1)$$

where H_i is homography matrix between bg_i image and the other frame.

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 the other frame which will be registered to Projective Grid Space. The green lines show inliers which will be used for estimating homography. The red line show ourliers which are removed by RANSAC.

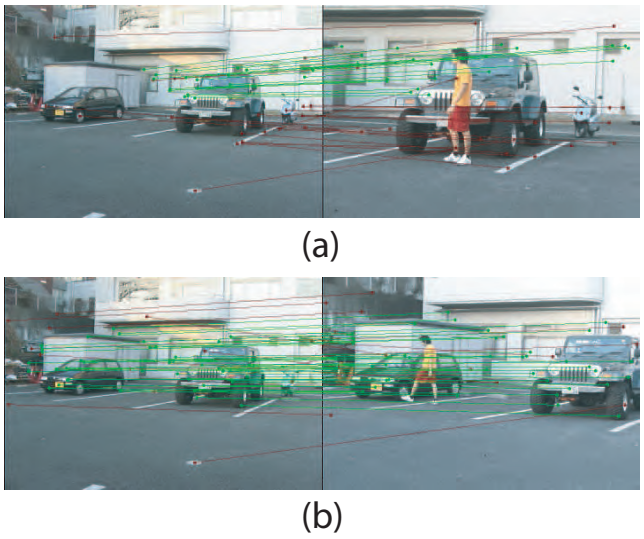

Fig. 5. Corresponding Points Found Using SIFT for Estimating Homography.

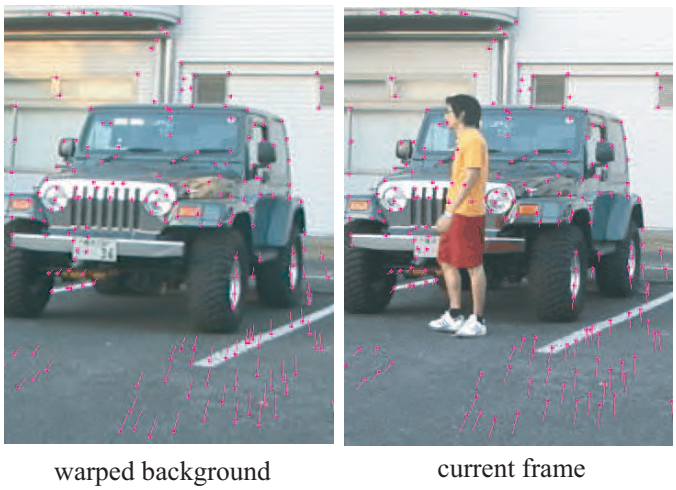

Fig. 6. Optical flows between warped background and current frame. (Image is cropped to show detail)

C. Homography Refinement

Corresponding points for estimating homography found using SIFT as described in sectionIV-B are reliable but usually not distributed all over the image area. Accuracy of homography is decreased as it is far from the corresponding points.

To improve accuracy of homography, we warp background image using H_i in equation 1 then compute optical flow between current frame. Initial optical flows include several outliers due to the present of foreground moving object. RANSAC is employed using homography constraint to remove such outliers.

Fig.6 shows optical flow between warped background and current frame (the same frame as Fig.5(a)). Comparing with Fig.5(a), 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. We estimate homography $H_{optical\ i}$ which transforms coordintate from warped background to current frame. More accurate homography $H_{refined\ i}$ is computed from equation 2. Finally, H_i in equation 1 is then replaced with $H_{refined\ i}$.

$$H_{refined\ i} = H_{optical\ i} H_i \tag{2}$$

V. 3D RECONSTRUCTION

We consider that objects in input video consist of

- Background planes
- Background static objects
- 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.7 shows how background scene is catagorized. Static objects and background planes are 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. 7. Background Scene Consist of Planes and Static Objects.

A. Preprocess

From bg_i images of all cameras, visual hull of static objects are reconstructed using silhouette volume intersection method[9]. Silhouette images of each static object are segmented manually. Surfaces of 3D voxel model are extracted to 3D triangular mesh model using Marching Cube algorithm [10]. This 3D triangular mesh model will be used for making dense correspondences for view interpolation.

During preprocess we reconstruct 3D position of several points which lie on a plane by assigning corresponding points between two basis cameras defining PGS. 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 sectionIII. These 3D position in PGS will be used for render planes in free viewpoint video as will be explained in sectionVI-A.

B. Runtime

Visual hull of moving object is reconstructed using silhouette volume intersection method in the same way as static object. The difference is visual hull of moving object is reconstructed every frame automatically. Silhouette of moving object needs to be segmented from input frames. Background image of camera i during runtime is generated by warping bg_i image using homography estimated automatically

in sectionIV-B . Example generated background images for moving camera are shown in Fig.8. After generating virtual background for hand-held cemera, silhouette image is segmented for reconstruct visual hull for rendering free viewpoint image.

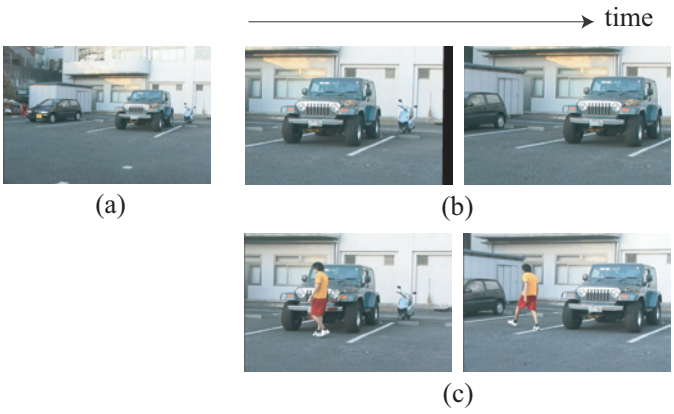

Fig. 8. (a) Background Image. (b)Automatically Generated Background Images from (a). (c) Some Frames from Input Video of the Same Camera as (a).

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. Our energy function is modified from equation proposed in *Background Cut* [19]. In that paper, a per-pixel propability function that a pixel r which has color $\boldsymbol{I}_r$ will be background, is defined by single isotopic Gaussian distribution $p_B(\boldsymbol{I}_r)$ in RGB space. $p_B(\boldsymbol{I}_r)$ is learned from several background images. Using this probability function is sensitive to illumination change. Our background images (warped background images) have different illumination from input image due to the sun light change, most of background regions will be classified as foreground region.

We change this probability function by first computing color and intensity disimilarity of every pixel between input image and warped background using equation 3 and 4 respectively

$$\theta = \arccos\left(\frac{\boldsymbol{I}_{input} \cdot \boldsymbol{I}_{bg}}{|\boldsymbol{I}_{input}||\boldsymbol{I}_{bg}|}\right) \tag{3}$$

$$d = |\boldsymbol{I}_{input} - \boldsymbol{I}_{bg}| \tag{4}$$

where $\boldsymbol{I}_{input}$ and $\boldsymbol{I}_{bg}$ is RGB color of a pixel in input image and warped background respectively.

Then we get distribution of 2-dimensional disimilarity vectors $\boldsymbol{D} = (\theta, d)^T$ as Fig.9. We represent this probability distribution as Gaussian Mixture Models:

$$p(\boldsymbol{D}) = \sum_{1 \le i \le N} w_i N(\boldsymbol{D}|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{5}$$

where $N(\cdot)$ is a Gaussian distribution and $(w_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ is the weight, the mean and the covariance matrix of the ith component of the GMMs. Typical number of components is 2-3.

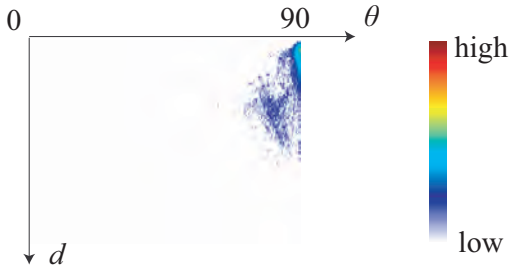

Fig. 9. Distribution of 2D disimilarity vectors.

Pixels which belong to background area will have similar θ and d. Because foreground area is small comparing to background, component which has maximum weight well represent background pixels. We replace per-pixel probability $p_B(\boldsymbol{I}_r)$ with a new function $w_k N(\boldsymbol{D}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ where maximum weight in GMMs is w_k.

Fig.10 shows segmentation result. Silhouette segmentation using graph cut give more accurate result compared to simple thresholding with postprocessing (morphological operation).

Fig. 10. Silhoutte Segmentation.

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[10]. Fig.11 shows 3D model of moving object reconstructed in PGS.

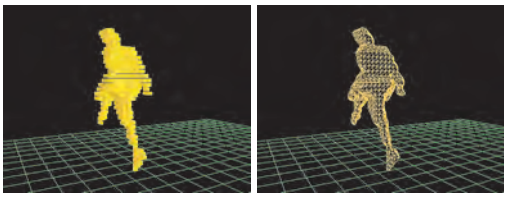

Fig. 11. 3D Model of Human in PGS. (a) Volumetric Representation (b) Triangular Mesh Representation.

VI. FREE VIEWPOINT RENDERING

Our method can synthesize free viewpoint between any two reference views. Free viewpoint image are rendered in two

steps. Background planes in scene are rendered first. Moving and static objects are then rendered overlay to synthesized planes. The following subsections explain the detail of two rendering phase.

A. Planes 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 6

$$\begin{pmatrix} x \\ y \end{pmatrix} = w \begin{pmatrix} x_1 \\ y_1 \end{pmatrix} + (1-w) \begin{pmatrix} x_2 \\ y_2 \end{pmatrix} \quad (6)$$

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 during preprocess is warped and blended to virtual view. Fig.12 illustrates how the plane is rendered in free viewpoint image.

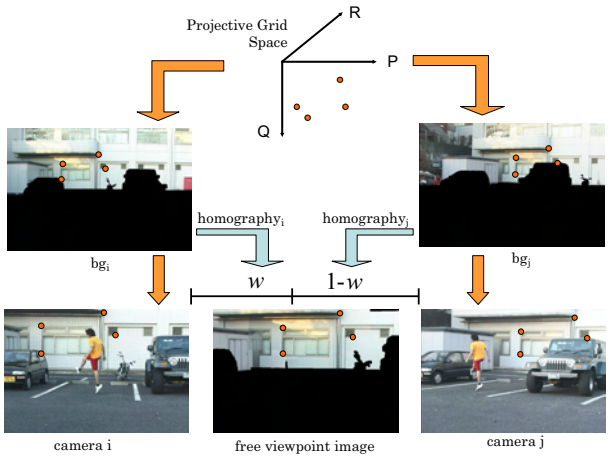

Fig. 12. Rendering Plane on Free Viewpoint Image.

B. Objects Rendering

Free viewpoint images of static and moving objects are synthesized by an image-based rendering method. 3D triangular mesh model of static objects and moving object are combined together. Combined 3D model is used for making a dense correspondence and also for testing occlusion between reference images. Each corresponding triangular mesh is warped to virtual viewpoint image based on view interpolation method[2].

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 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 6.

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.13 shows example of free viewpoint image of static and moving objects.

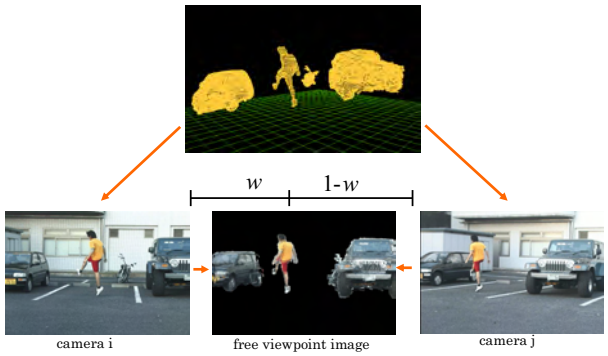

Fig. 13. Rendering Static and Moving Objects on Free Viewpoint Image.

VII. EXPERIMENTAL RESULTS

In this section we synthesize free viewpoint video from two input videos sets. The experimental environment of both input sets is a large natural scene as Fig.4. We use 4 Sony-DV cameras with 720x480 resolutions. All cameras are in front of the scene as in Fig.1. The first set of input videos are captured using hand-held cameras while the second one are captured using fixed cameras.

In both inputs, human moves over the same area (from the left most car to the right most car). Fig14 shows example frames of one camera from both input videos sets. We can see that hand-held camera has advantage of allow zooming to capture only some part of a scene where there is interested object (human).

In sectionVII-A, we use our proposed method for calibrating hand-held cameras in the first input set then synthesize free viewpoint video.

In sectionVII-B, we synthesize free viewpoint video using the second input set where fixed cameras are used. We use the same 3D reconstruction and rendering algorithm but do not use our method for calibrating hand-held cameras. The result shows that human's resolution from fixed cameras system is lower than using hand-held cameras which allow zooming and changing view direction.

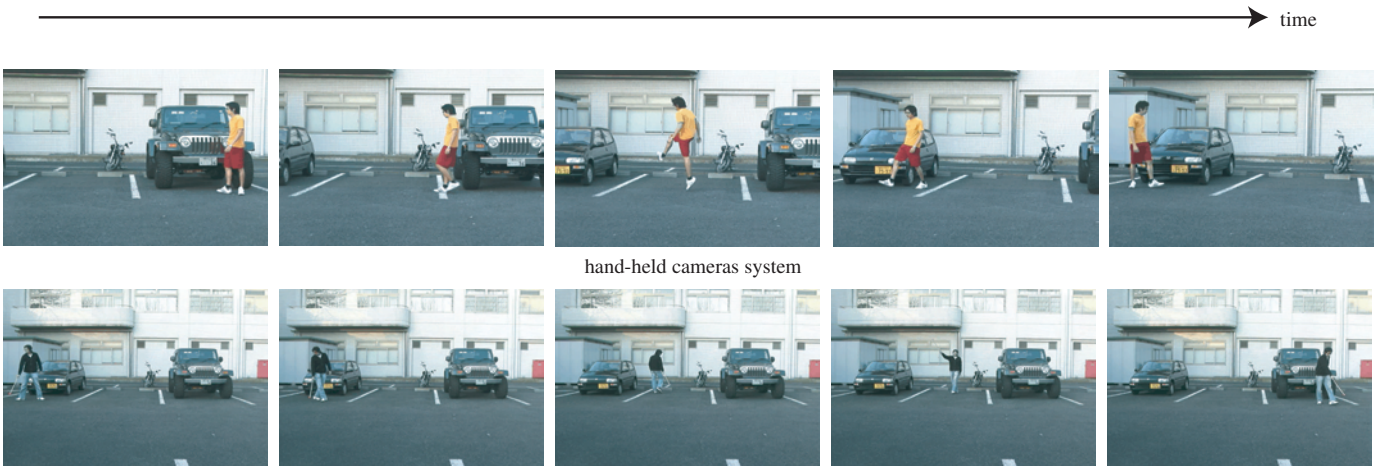

Fig. 14. Example Input Frames of One Camera from Both Input Sets.

A. Free Viewpoint Video using Hand-Held Cameras

We synthesize free viewpoint video from consecutive 200 frames by our proposed method. During 200 frames, hand-held cameras have been zoomed and changed view direction to capture high resolution human independently. There is no artificial marker placed in the scene. Only natural feature are used for finding corresponding points. Our method can correctly register all frames to PGS and synthesize free viewpoint video without manual operation. Fig.15 shows some example frames from the result free viewpoint video.

Fig. 15. Example Free Viewpoint Video from Consecutive 200 Frames.

To compare quality of interpolated images with the original images, we select one frame from the input video as Fig.16 to synthesize free viewpoint images at several weight ratios. The result free viewpoint images between camera2 and camera3 are shown in Fig.17. Ratio 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.

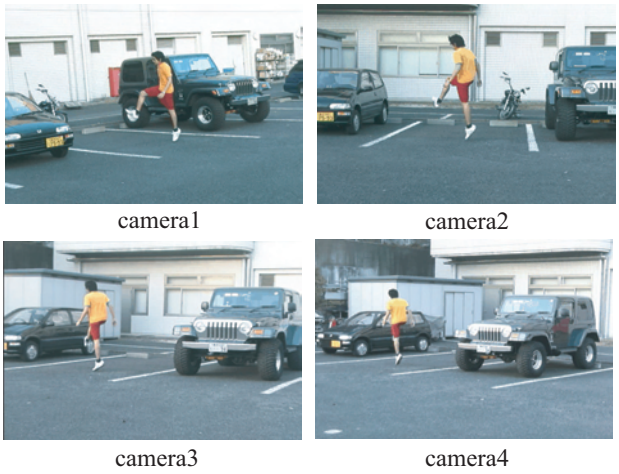

Fig. 16. One Frame from Four Input Videos.

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 [20].

B. Free Viewpoint Video using Fixed Cameras

To compare our proposed method that can synthesize free viewpoint video from hand-held cameras input with fixed cameras system, we synthesize free viewpoint video from the second input videos set. All fixed cameras must be zoomed out so that field of views are wide enough to cover the area in which the human moves.

Fig.18 shows the result of free viewpoint images from fixed cameras system while result from our moving cameras system are already shown in section VII-A. Moving object's size in free viewpoint image obtained from fixed cameras system is

Fig. 17. Free Viewpoint Images Between Camera2 and Camera3.

small (due to low resolution in input videos) and have less texture detail compared to hand-held cameras system. This difference can be more visible if the relative size between the scene and moving object become larger.

Fig. 18. Free Viewpoint Images from Fixed Cameras System.

VIII. CONCLUSIONS

We propose a novel method to synthesize free viewpoint video of a moving object in natural scene, which is captured by hand-held multiple cameras. Projective Grid Space (PGS) [17] 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 [11] to find corresponding points in natural scene for dynamically registering cameras to PGS. Graph cut optimization is used for silhouette segmentation.[19] In the experiment, free viewpoint video is successfully synthesized from hand-held multiple cameras without manual operation.

REFERENCES

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

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

[3] 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.

[4] B. Goldluecke and M.Magnor, "Real-time microfacet billboarding for free-viewpoint video rendering," in *Proceedings of the IEEE International Conference on Image Processing*, 2003, pp. 713–716.

[5] O. Grau, T. Pullen, and G. Thomas, "A combined studio production system for 3d capturing of live action and immersive actor feedback," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 3, pp. 370–380, March 2004.

[6] Y. Ito and H. Saito, "Free-viewpoint image synthesis from multiple-viewimages taken with uncalibrated moving cameras," in *The IEEE International Conference on Image Processing (ICIP05)*, September 2005.

[7] 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.

[8] H. Kato and M. Billinghurst, "Marker tracking and hmd calibration for a video-based augmented reality conferencing system," in *Proceedings of the 2nd IEEE and ACM International Workshop on Augmented Reality*, 1999, pp. 85–94.

[9] 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.

[10] 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.

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

[12] S. Moezzi, L.C.Tai, and P.Gerard, "Virtual view generation for 3d digital video," *IEEE MultiMedia*, vol. 4, no. 1, pp. 18–26, 1997.

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

[14] V. Nozick, S. Michelin, and D. Arques, "Real-time plane-sweep with local strategy," *Journal of WSCG*, vol. 14, no. 1-3, pp. 121–128, 2006.

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

[16] 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.

[17] H. Saito and T. 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.

[18] 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.

[19] J. Sun, W. Zhang, X. Tang, and H.-Y. Shum, "Background cut," in *ECCV 2006*, vol. 2, 2006, pp. 628–641.

[20] 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.