NEW VIEWPOINT VIDEO SYNTHESIS IN NATURAL SCENE USING UNCALIBRATED MULTIPLE MOVING CAMERAS

Songkran Jarusirisawad Hideo Saito

Department of Information and Computer Science, Keio University,
3-4-1 Hiyoshi, Kohoku-ku, Yokohama, 223-8522 JAPAN
E-mail: {songkran,saito}@ozawa.ics.keio.ac.jp

ABSTRACT

In most conventional systems for new viewpoint video synthesis of a moving object, calibrated multiple fixed cameras are used. Field of view of every camera must cover all the area in which the object moves. If the area is large, object size in the captured images and also in the new viewpoint video will become very small. In this paper, we propose a novel method to synthesize new viewpoint video of a moving object in natural scene, which is captured by uncalibrated moving cameras. During capture, cameras can be moved and zoomed to capture only part of a scene where there is a moving object. Projective Grid Space (PGS) which is 3D space defined by two cameras is used for object's shape reconstruction. We use SIFT for finding corresponding points in natural scene for registering moving cameras to PGS. In the experiment which is performed for demonstrating the efficacy of the proposed method, new viewpoint video is successfully synthesized from multiple moving cameras without manual operation.

1. INTRODUCTION

New viewpoint image synthesis is the problem of creating a novel viewpoint image of a scene, from many input images, as it would be seen from the novel viewpoint. By using this technique in movie or broadcasting, interesting visual effect can be created. This kind of research is one of popular topics in computer vision.

In this paper we propose a new method for synthesizing new viewpoint video in natural scene using uncalibrated multiple moving cameras. There is no other research that has been done under this constraint before. The difficulty of using moving camera in natural scene is dynamically calibrating each camera where there is no any special marker. In our method, Projective Grid Space (PGS)[1] which is 3D space defined by 2 basis cameras is used for object's shape reconstruction. During preprocess, one frame from each camera are weakly calibrated to PGS. All other frames are automatically registered to PGS by homography. SIFT[2] is used for finding corresponding points in natural scene to estimate such homography. 3D model of object is then reconstructed in PGS using silhouette volume intersection. New viewpoint video is synthesized based on view interpolation[3] using dense correspondences from 3D model in PGS. Our method is robust enough, so that new viewpoint video can be created without manual operation after preprocessing.

1.1 Related Works

One of the earliest researches for new view synthesis is Virtualized Reality[4]. They apply multibaseline stereo for recovering accurate 3D shape from 50 cameras[5]. Moezzi et al. also synthesize free viewpoint video by recovering visual hull of objects from silhouette images using 17 cameras[6]. Many methods for improve quality of new 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[7]. Starck optimizes a surface mesh using stereo and silhouette data to and generate high accuracy virtual view image[8]. Saito et al. propose appearance-based method[9], which combines advantage from Image Based Rendering and Model Based Rendering. Real-time systems for synthesize free viewpoint video also have recently been developed[10][11].

Most of previous researches on new view synthesis propose a system using calibrated fixed cameras. During capture, cameras can not be moved or zoomed. If the area in which the object moves is large, object size in the captured images will become small. Resolution and quality of synthesized new viewpoint image depends on resolution of the object in the input images. Therefore, resolution of the object in new viewpoint image is usually insufficient and unsatisfied.

Ito et al. try to overcome this problem by proposing a system that can synthesize new viewpoint image using moving cameras[12][13]. Using their method high resolution new viewpoint image can be obtained. However, they demonstrate only in the case that a clear homogeneous color background with some artificial markers placed around the scene because of the difficulty of the feature point tracking for the moving camera. In some situations, high resolution new viewpoint images are desired in a natural scene without any marker, for example for sports or outdoor events. This paper purpose a novel method for synthesis new viewpoint video in such environment.

2. PROJECTIVE GRID SPACE

Reconstructing 3D model for synthesizing new viewpoint image requires a relation between 3D coordinate of the

scene and 2D coordinate in image frame. Projection matrix that represents this relation can be estimated by full calibration which requires 3D-2D correspondences. Finding 3D-2D corresponding points requires a lot of work and also not suitable in a natural scene.

Projective Grid Space (PGS)[1] is a 3D space 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.1 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.

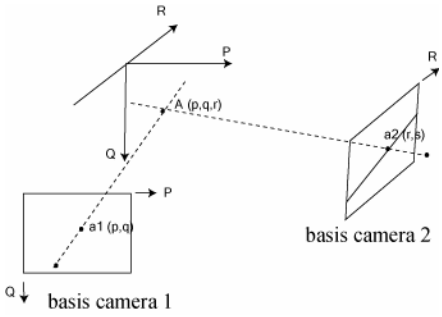

Figure 1. *Projective Grid Space is defined by 2 basis cameras. Point A (p,q,r) is projected to a1 (p,q) and a2 (q,r) on the first and second basis image, respectively.*

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.2.

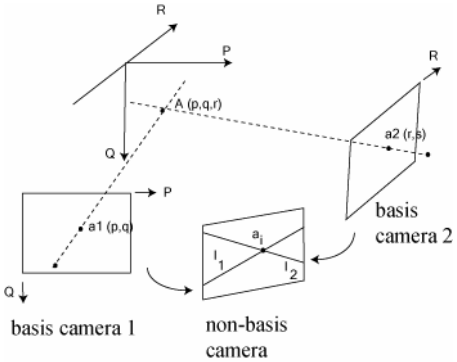

Figure 2. *Point A (p,q,r) in PGS is projected to a_i where epipolar line of a1 intersect epipolar line of a2*

Epipolar line l_1 and l_2 are calculated from the following equations

$$l_1 = F_{1i}\begin{bmatrix} p \\ q \\ 1 \end{bmatrix} \quad \textbf{(1)}$$

$$l_2 = F_{2i}\begin{bmatrix} r \\ s \\ 1 \end{bmatrix} \quad \textbf{(2)}$$

where F_{1i} and F_{2i} are fundamental matrix from basis camera1 and from basis camera2 to non-basic camera respectively.

3. PREPROCESS

Our system environment consists of five cameras in large natural scene as shown in Fig.3. During preprocess we zoom out all cameras to capture the whole area of a scene without object. We call this background image for each camera i as bg_i. We select camera1 and camera5 as basis cameras defining PGS. 2D-2D Corresponding points for estimate 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 for relate moving cameras to PGS as will be described in the next section. Fig.4 shows background images of our experiment.

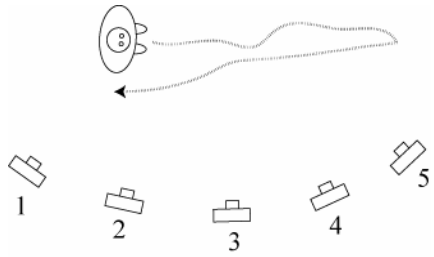

Figure 3. *Top view of system configuration. System consist of 5 cameras taking video from different views. Camera1 and camera5 are two basis cameras defining PGS.*

Figure 4. *Images of a whole area without object captured by all cameras.*

4. REGISTER MOVING CAMERAS TO PGS

During capture, object will move around a large space. Each camera is zoomed and rotated to capture the object with high resolution in the image. Because view and focal length of camera is changed, fundamental matrices estimated during preprocess can not be used to relate 3D position in PGS to 2D position in captured image anymore. We assume that all moving cameras are zoomed and rotated freely but don't change much position during operation. Therefore, 2D coordinate in bg_i can be transformed to 2D coordinate of capturing camera i using homography matrix.

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 candidate corresponding points between bg_i and captured object image. We also employ RANSAC (RAndom Sample Consensus)[14] using homography constraint to remove miscorresponding points. Only inliers are used for finding accurate homography. Example corresponding points that automatically found using SIFT are shown in Fig.5

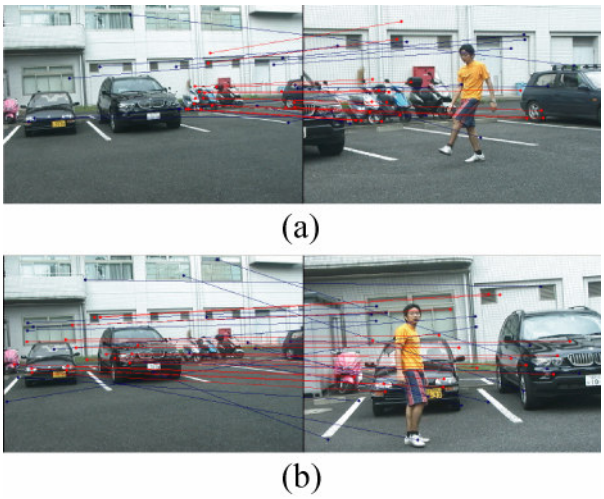

Figure 5. *Example corresponding points between bg_i image and current image of camera i. The red line show correct corresponding points which will be used for estimating homography. The blue line show ourliers which are removed by RANSAC.*

SIFT is robust for matching between images of the same object with different scale and 2D rotation. However, reliable of matching will decrease if there are perspective distortion between matched images. In our case, two images are captured from approximately same position but zoom and change view direction. There is no parallax between these images. Two images are approximately 2D similarity regardless of complexity of a scene. Therefore, SIFT is very robust for using in our system.

3D coordinate A (p,q,r) in PGS which is projected on (x_{bg},y_{bg}) of bg_i image is projected on (x_{cap},y_{cap}) of the same camera at object capturing time by the following equation

$$s\begin{bmatrix} x_{cap} \\ y_{cap} \\ 1 \end{bmatrix} = H_i \begin{bmatrix} x_{bg} \\ y_{bg} \\ 1 \end{bmatrix} \quad \textbf{(3)}$$

where H_i is homography matrix for camera i which is estimated automatically every frame.

5. 3D MODEL RECONSTRUCTION

Every frame, 3D model of moving object is reconstructed using silhouette volume intersection method[15]. If cameras are static, background scene can be captured beforehand, so it is trivial to get silhouette image using simple background subtraction. In our case that moving cameras are used, background image of moving camera cannot be captured before or after capturing object because it is impossible to recapture the scene with the same trajectory and zoom.

In our method, background image of moving cameras in any natural scene are generated automatically. Every frame, background image of camera i is generated by warping bg_i image using homography obtained in section 4. Example generated background images for moving camera are shown in Fig.6. After generating background, silhouette image can be created by background subtraction as in Fig.7.

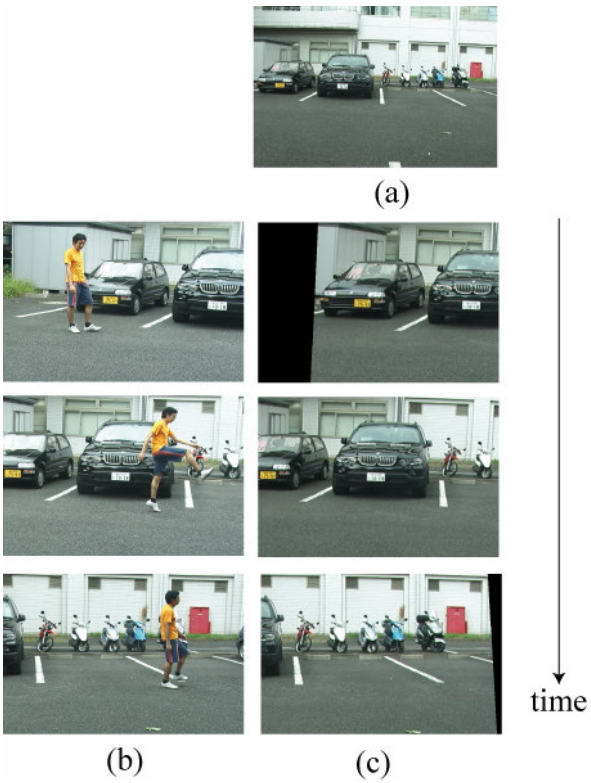

Figure 6. (a) *Background image captured before capturing object.* (b) Object images captured from the same camera as (a) but rotated and zoomed. (c) Automatically generated background images from (a), for background subtraction with (b).

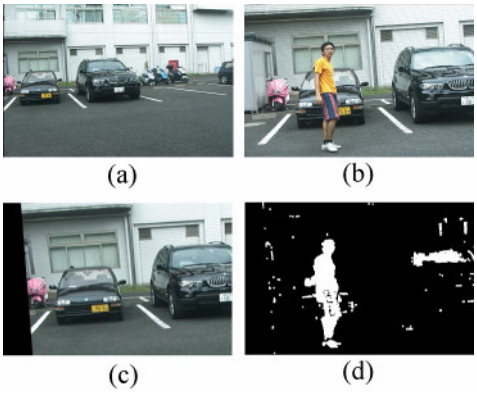

Figure 7. *(a) background image of large area scene. (b) object image capture from the same camera but change view direction and zoom parameter. (c) warped background image from (a). (d) silhouette image for (b)*

Voxels in PGS are projected onto each silhouette image using method described in section 2 and equation (3). Voxel is considered to be in a 3D model volume if projected points of all cameras are in silhouette. Surface of 3D voxels model are extracted to 3D triangle mesh model using Marching Cube algorithm[16]. Fig.8 shows 3D model reconstructed in PGS.

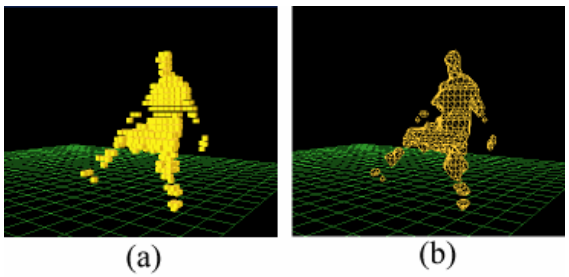

Figure 8. *3D model of human in PGS. (a) volumetric representation (b) triangle mesh representation*

6. NEW VIEWPOINT SYNTHESIS

New viewpoint images are synthesized by an image-based rendering method. 3D triangle mesh model is used for making a dense correspondence and also for testing occlusion between reference images. Each corresponding triangle mesh will be warped to new view point image, based on view interpolation method[3][12].

6.1 Z-Buffer Generation

To test occlusion between views, Z-Buffer of each camera is generated from 3D triangle mesh model. All triangle patches in PGS are projected onto each Z-Buffer in the similar manner in section 2. Each pixel of Z-Buffer is stored the 3D distance from camera's optical center to projected triangle patch. If some pixels are projected by more than one patch, the shortest distance is stored. In Fig.9, 3D camera position of the basis camera1 in PGS is ($C1_x$, $C1_y$, $e12_x$), where ($C1_x$, $C1_y$) is camera center in the basis camera1, and ($e12_x$, $e12_y$) is epipole of basis camera1 in basis camera2. In the same way, camera position of the basis camera2 is ($e21_x$, $e21_y$, $C2_x$), where ($e21_x$, $e21_y$) is the epipole of the basis camera2 in the basis camera1, and ($C2_x$, $C2_y$) is camera center in basis camera2. For non-basis camera, 3D camera position in the PGS is ($e1_x$, $e1_y$, $e2_x$) where ($e1_x$, $e1_y$) and ($e2_x$, $e2_y$) are epipoles on basis camera1 and basis camera2, respectively.

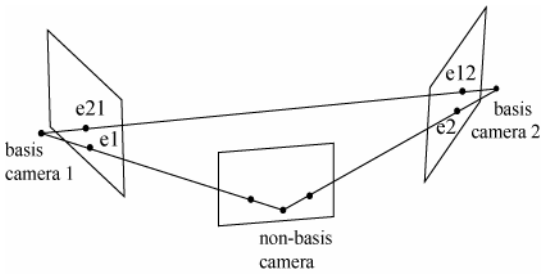

Figure 9. *Camera position in Projective Grid Space*

The distance of point ***a*** (p_1,q_1,r_1) and ***b*** (p_2,q_2,r_2) in PGS is defined as the following.

$$D = \sqrt{(p_1 - p_2)^2 + (q_1 - q_2)^2 + (r_1 - r_2)^2} \qquad \textbf{(4)}$$

6.2 Rendering

Each triangle mesh of 3D model in PGS is projected onto 2 neighboring images. Z-Buffer is used to test occlusion. Patch whose distance from input camera focal point is different from the value stored in the Z-Buffer is decided to be occluded. In the case that a patch are occluded in both 2 input view, this patch will not be interpolated in a new viewpoint image. If a patch is seen from either or both input views, this patch will be warped and merged into a new viewpoint image. Position of a warped pixel in new viewpoint image is determined by

$$\begin{bmatrix} x \\ y \end{bmatrix} = w \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} + (1-w) \begin{bmatrix} x_2 \\ y_2 \end{bmatrix} \qquad \textbf{(5)}$$

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

To merge two warped images, RGB colors of the pixel are computed by the weighted sum of the colors from both warped images. 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 and other view are set to 0 and 1 respectively.

7. EXPERIMENTAL RESULTS

In this section we will show our result of the proposed method. In our experiment, we use 5 Sony-DV cameras with 720x480 resolutions. The camera setting is as Fig. 3.

All cameras are placed on tripods. In this experiment we synthesize new viewpoint video of human motion. The experiment environment is like Fig. 4. During human move around the scene, we rotate and zoom 3 non-basis cameras to capture high resolution human images while 2 basis cameras are not changed view direction and zoom from preprocess. In our proposed method, all cameras include basis cameras can rotate and zoom freely. However, in experiment there are only 3 people capturing. Fig.10. is an example frame of input video from all 5 cameras.

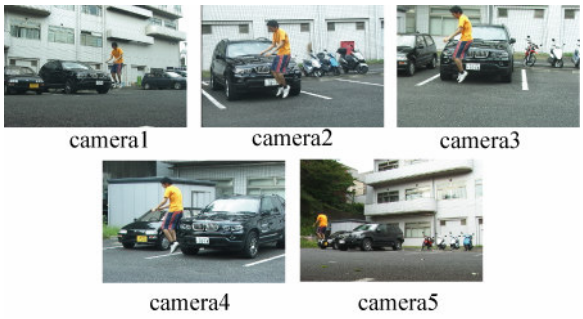

Figure 10. *Example frame of input videos. Camera1 and Camera5 are used as basis cameras defining PGS*

7.1 New viewpoint image of one input frame

We synthesize new viewpoint images using one frame of input videos as Fig.10. Free viewpoint images can be successfully synthesized as in Fig.11. Some miscorresponding texture of warped image can be seen due to inaccuracy of 3D model. This can not be avoided by using only shape from silhouette. In future work some other technique can be used together with shape from silhouette to improve quality of 3D model. [17]

7.2 New viewpoint video from several input frames

To test robustness of our proposed method for registering moving camera to PGS, we synthesize new viewpoint video from consecutive 118 frames from 5 video inputs. During 118 frames, cameras have been rotated approximately 40 degrees. Our method can register all images to PGS correctly without manual operation. In Fig.12, show some frames from the result new viewpoint video. Processing time for 5 cameras per one frame on CPU 1.60 GHz, start from register camera to PGS until generating depth buffer, takes about 30 seconds. For render new viewpoint video, our system takes about 2 second per frame.

In Fig.12, we can see some part of leg is missing. This happens because of incomplete silhouette image from some view. Because now we work in natural scene, a completely clear silhouette from background subtraction is hard to achieve.

7.3 Comparison with conventional system

To compare our proposed method to conventional system, we synthesize new viewpoint video in the same scene using fixed cameras like conventional method. All fixed cameras must be zoomed out so that whole area of the scene that the object will move like Fig.4 can be seen. Fig.13 show result of new viewpoint image from conventional system and from our method. Object size in free viewpoint image obtained from conventional system is small compare to our method. This different can be more visible if the relative size of the whole scene and object become larger.

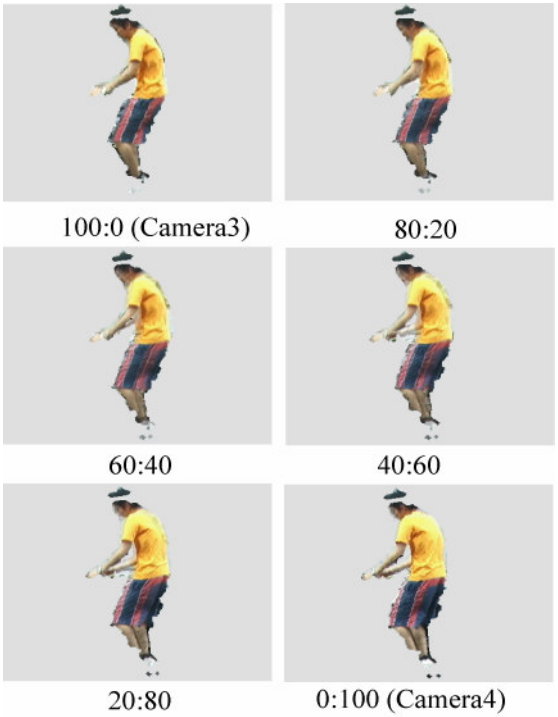

Figure 11. *New viewpoint images between camera3 and camera4. Ratio between two view is written under each figure. These images are cropped from real-size images to show more detail.*

Figure 12. *Sample frame from new viewpoint video. These images are cropped from real-size images to show more detail.*

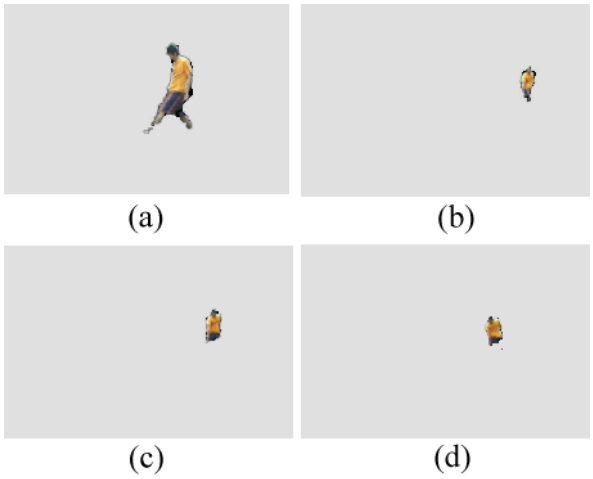

Figure 13. *(a) new view point image from our proposed method. (b)-(d) New viewpoint images in the same scene obtained from conventional system.*

8. CONCLUSION

From experiment results, cameras which are zoomed and rotated in natural scene without any marker can be automatically registered to PGS. Background image of each moving camera can be automatically generated from preprocess background image. Missing part of object in new viewpoint video is from inaccurate silhouette image. In future work we will improve quality of new viewpoint video and synthesize background scene together with moving object.

We propose a novel method to synthesize new viewpoint video of a moving object in natural scene, which is captured by uncalibrated multiple moving cameras. There is no other research that has been done under this constraint before. Projective Grid Space[1] which is 3D space defined by two cameras is used for object reconstruction. Every frame, homography is used to register moving cameras to PGS. We employ SIFT[2] for finding corresponding point to estimate homography automatically. In our experiment that is performed in natural scene without any special marker for demonstrating the efficacy of the proposed method, new viewpoint video can be successfully synthesized without manual operation.

9. REFERENCES

[1] Hideo Saito, Takeo Kanade, “Shape Reconstruction in Projective Grid Space from Large Number of Images”, IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'99), Fort Collins, CO, June 1999.

[2] David G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, 60, 2 (2004), pp. 91-110.

[3] S. Chen and L. Williams, “View interpolation for image synthesis”, in Proc. of SIGGRAPH ’93, pp. 279–288,1993.

[4] T. Kanade, P. W. Rander, and P. J. Narayanan, “Virtualized reality: concepts and early results,” IEEE Workshop on Representation of Visual Scenes, pp.69-76,1995.

[5] S. Vedula, P. W. Rander, H. Saito, and T. Kanade “Modeling, Combining, and Rendering Dynamic Real-World Events From Image Sequences,” Proc. 4th Conf. Virtual Systems and Multimedia, Vol. 1, pp. 326-322, 1998.

[6] S.Moezzi, L.C.Tai, P.Gerard, “Virtual View Generation for 3D Digital Video,” IEEE Multimedia, 4, 1, pp.18–26, 1997.

[7] J.Carranza, C.Theobalt, M.Magnor, H.-P. Seidel, ”Free-Viewpoint Video of Human Actors,” ACM Trans. on Computer Graphics, vol. 22, no. 3, pp. 569-577, July 2003.

[8] J. Starck and A. Hilton. Towards a 3D virtual studio for human appearance capture. IMA International Conference on Vision, Video and Graphics, Bath, 2003.

[9] 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, September, 2003, pp. 303 - 316.

[10] O. Grau, T. Pullen, G.A. Thomas, “A Combined Studio Production System for 3D Capturing of Live Action and Immersive Actor Feedback,” IEEE Trans. Circuits and Systems for Video Technology, March 2004.

[11] B.Goldlucke and M.Magnor, “Real-time microfacet billboarding for free-viewpoint video rendering,” in Proc. IEEE International Conference on Image Processing (ICIP’03), Barcelona, Sept. 2003, pp. 713–716.

[12] Yosuke Ito and Hideo Saito, “Free viewpoint image synthesis using uncalibrated multiple moving cameras”, Computer Vision / Computer Graphics Collaboration Techniques and Applications (MIRAGE2005), pp.173-180, March, 2005.

[13] Yosuke Ito, Hideo Saito, Free-viewpoint image synthesis from multiple-viewimages taken with uncalibrated moving cameras, The IEEE International Conference on Image Processing (ICIP05), Italy, Sep, 2005

[14] 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, 24(6), pp.381-395, 1981.

[15] 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, 1994.

[16] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. In M.C. Stone, editor, Proceedings of the SIGGRAPH, pages 163-- 169, Anaheim, CA, July 1987. in Computer Graphics, Volume 21, Number 4.

[17] S.Yaguchi , H.Saito, "Improving Quality of Free-Viewpoint Image by Mesh Based 3D Shape Deformation," The 14th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision(WSCG2006),February, 2006