Arbitrary viewpoint video synthesis from uncalibrated multiple cameras

Satoshi Yaguchi and Hideo Saito

Department of Information and Computer Science, Keio University
3-14-1 Hiyoshi Kouhoku-ku Yokohama 223-8522, Japan
Phone: +81-45-566-1753, Fax: +81-45-566-1747
{yagu, saito}@ozawa.ics.keio.ac.jp

ABSTRACT

In this paper, we propose a method for arbitrary view synthesis from uncalibrated multiple camera system, targeting large space such as soccer stadium. In Projective Grid Space (PGS), that is the three-dimensional space defined by epipolar geometry between the two basis cameras in the camera system, we reconstruct three dimensional shape models from the silhouette images. By using the three dimensional shape model reconstructed in PGS, we can obtain the dense point correspondence map between reference images. The obtained correspondence can synthesize the image of arbitrary view between the reference images. We also propose the method for merging the synthesized images with the virtual background scene in PGS. We apply the proposed method to image sequences taken by the multiple camera system, which is developed in a large space on a concert hall. The synthesized image sequences of virtual camera have enough quality to demonstrate effectiveness of the proposed method.

Keywords: Virtual view synthesis, Shape from multiple cameras, View interpolation, Projective geometry, Fundamental matrix, Projective grid space

1 INTRODUCTION

The new view synthesis from images is recently studied for enhancing the visual entertaining effect of the movie. One method for enhancing the visual effect is virtual movement of viewpoint so that the audience can virtually feel existence in the target scene. The recent applications of such effect are known as "Matrix" of Hollywood movie, EyeVision system in the SuperBowl 2001 broadcast by CBS. Virtualized Reality (TM) [4] is one of the frontier project that realizes such virtual viewpoint movement for dynamic scene by the use of computer vision technology. Although the "Matrix" and "EyeVision" are just a switching effect of multiple camera real images, the computer vision-based technology can synthesize arbitrary viewpoint images for the virtual viewpoint movement effect.

We aim to apply the virtualized reality technology to actual sports events, etc. The new view images are generated by

rendering pixel values of input images in accordance with the geometry of the new view and the 3D structure model of the scene, which is reconstructed from multiple-view images. The 3D shape reconstruction from multiple view requires camera calibration that is used for relating the camera geometry to the object space geometry. For calibrating cameras, 3D position in Euclidean space of several points and 2D position on each view images of those points must be measured precisely. For this reason, when there are many cameras, much effort is needed to calibrate every camera. Especially, in the case of large space, such as sports stadiums, it is difficult to set many calibrating points of which the position is precisely measured throughout the large area. For removing such effort to obtain calibration data, we have already proposed a new framework for shape reconstruction from uncalibrated multiple cameras in the "Projective Grid Space (PGS)" [6], in which the coordinate is defined by epipolar geometry between cameras.

In this paper, we present a method for generating arbitrary view movie from multiple uncalibreated camera image sequence. In this method, the shape from silhouette (SS) [2, 5] method is applied for reconstructing the shape model in the PGS. Then the dense corresponding relation between the images derived from the shape model is used for synthesizing intermediate appearance view image. We demonstrate the proposed framework by showing several virtual image sequences generated from multiple camera image sequences that are corrected in a large space of 110m x 50m x 25m.

2 PROJECTIVE GRID SPACE

Reconstructing 3D shape model from multiple view images requires relationship between the 3D coordinate of the object scene and the 2D coordinate of the camera image plane. Projection matrices that represent such relationship are estimated by the measurement of 3D-2D correspondences at some sample points. Since the 3D coordinate is defined independently from cameras, the 3D position of the sample points must be measured independently from the camera. This procedure is camera calibration [9]. Calibrating all the camera in the multiple camera system requires much labor [4, 10].

In our method, 3D point is related to 2D image point without estimating the projection matrices by "Projective Grid Space (PGS)[6]", which can be determined by only fundamental matrices[11] representing the epipoler geometry between two basis cameras. Because the 3D coordinate of PGS is dependently defined by the camera image coordinates, 3D position of any sample points does not have to be measured. Therefore, the PGS enables 3D reconstruction from multiple images without estimating the projection matrices of each camera.

Figure 1 illustrates a scheme of PGS. PGS is defined by camera coordinate of the two basis cameras. Each pixel point (p, q) in the first basis camera image defines one grid line in the space. On the grid line, grid node points are defined by horizontal position r in the second image. Thus, the coordinate P and Q of PGS is decided by the horizontal coordinate and the vertical coordinate of the first basis image, and the coordinate R of PGS is decided by the horizontal coordinate. Since fundamental matrix $\boldsymbol{F}_{21}$ limits the position in the second basis view on the epipolar line $\boldsymbol{l}$, r is sufficient for defining the grid point. In this way, the projective grid space can be defined by two basis view images, of which node points are represented by (p, q, r).

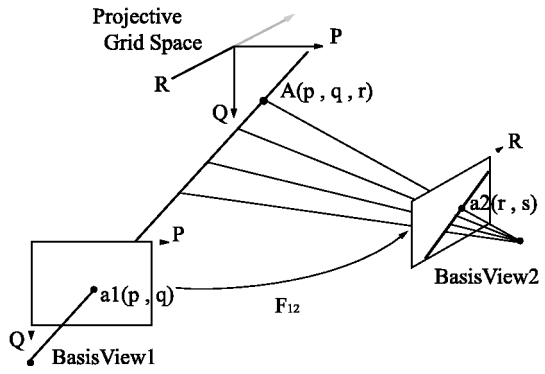

Figure 1: Definition of projective grid. The point $A(p, q, r)$ on the projective grid space is projected to $a_1(p, q)$ and $a_2(r, s)$ on the first and second basis image.

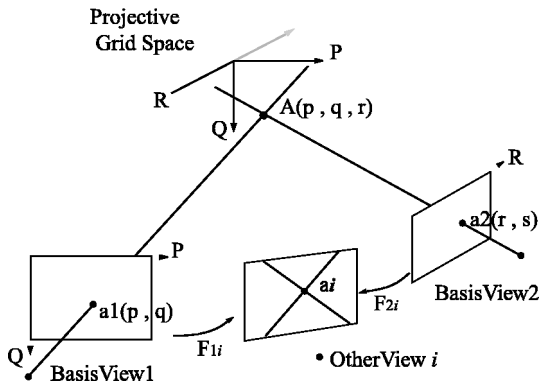

Figure 2: Projection of point in the space onto an image. The point $A(p, q, r)$ on the projective grid space is projected to the cross point of two epipolar lines in the image of view i.

3 MODEL RECONSTRUCTION

Under the framework of PGS, we reconstruct 3D shape model of the dynamic object by shape from silhouette (SS) method. We assume that the silhouette is previously extracted by background subtraction.

In the conventional SS method, each voxel in a certain Euclidean space is projected onto every silhouette image with projection matrices, which are calculated by strong calibration of every camera [2, 5], for checking if the voxel is included in the object region. For applying the SS method in the PGS, every point in PGS must be projected onto each silhouette image. As described in the previous section, the PGS is defined by two basis views, and the point in the PGS is represented as $A(p, q, r)$. The point $A(p, q, r)$ is projected onto $a_1(p, q)$ and $a_2(s, r)$ in the first basis image and the second basis image, respectively. The point a_1 is projected as the epipolar line l on the second basis view. The point a_2 on the projected line (figure 1), is expressed as

$$\boldsymbol{l} = \boldsymbol{F}_{21} \begin{bmatrix} p \\ q \\ 1 \end{bmatrix} \tag{1}$$

where $\boldsymbol{F}_{21}$ represents the fundamental matrix between the first and second basis images.

The projected point in ith arbitrary real image is determined two fundamental matrices, $\boldsymbol{F}_{i1}$, $\boldsymbol{F}_{i2}$ between two basis images and ith image. Since $A(p, q, r)$ is projected onto $a_1(p, q)$ in the first basis image, the projected point in the ith image must be on the epipolar line $\boldsymbol{l}_1$ of $a_1(p, q)$, which is derived by the $\boldsymbol{F}_{i1}$ as

$$\boldsymbol{l}_1 = \boldsymbol{F}_{i1} \begin{bmatrix} p \\ q \\ 1 \end{bmatrix} \tag{2}$$

In the same way, the projected point in the ith image must be on the epipolar line $\boldsymbol{l}_2$ of $a_2(r, s)$ in the basis image, which is derived by the $\boldsymbol{F}_{i2}$ as

$$\boldsymbol{l}_2 = \boldsymbol{F}_{i2} \begin{bmatrix} r \\ s \\ 1 \end{bmatrix} \tag{3}$$

The intersection point between the epipolar line $\boldsymbol{l}_1$ and $\boldsymbol{l}_2$ is the projected point $A(p, q, r)$ onto the ith image (figure2). In this way, every projective grid point is projected onto every image, where the relationship can be represented by only the fundamental matrices between the image and two basis images.

Outline of the process for reconstructing 3D shape model is as follows. First of all, two cameras are selected from as basis cameras, and then coordinate of PGS is determined. Every voxel in the certain region is projected onto each silhouette image with proposed scheme as shown in figure 1, 2. The voxel that is projected onto the object silhouette for all images is decided as existent voxel, while others are nonexistent. Thus the volume of the object can be determined in the voxel represented in PGS. In this process, the order of voxel existence checking is important for reducing the computation, because the cost for computing projection of voxel onto images is not the same for the different images in the proposed scheme. Since the vertical and horizontal coordinate of the first basis view image are equivalent to P and Q coordinate of the PGS, any calculation involving fundamental matrix is not required to project each voxel onto the first basis view image. In the second basis view image, the projected point is decided by calculating only one multiplication of fundamental matrix for determining epipolar line. This implies that projection calculation onto second basis view becomes half compared with projecting the other images. Therefore, the order of images on which each voxel is projected should be basis view 1, basis view 2, and the other view images.

After this existent voxel determination, implicit surface of the voxel representation of the object is extracted by "Marching Cubes". Finally, the object model is reconstructed as the surface representation in the PGS.

4 VIRTUAL VIEW SYNTHESIS

There are two ways to generate the arbitrary view image from 3D shape model, that is texture mapping on the 3D Shape model [4, 10], and the morphing from the point correspondence of some reference images that is calculated by the model [1, 3, 8, 7]. In the former procedure, the texture of the images are projected onto the 3D shape model, then re-projected onto the image again. In this procedure, however, the generated images are likely to be suffered from rendering artifact caused by the inaccuracy of 3D shape. Therefore we apply the latter procedure to generate arbitrary view images.

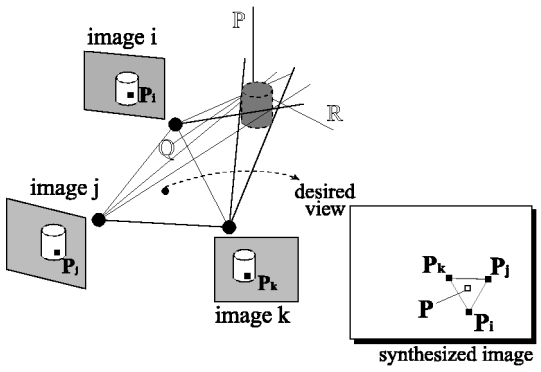

Figure 3: Synthesis of desired view from neighboring three view images.

4.1 Arbitrary View Synthesis

Arbitrary view image is synthesized as intermediate images of selected neighboring two or three reference real images. If two reference images are selected, virtual viewpoint can be taken on the line between two reference real viewpoints. In case of three reference images, virtual viewpoint can be taken inside of the triangle that is formed by the three real viewpoints. Therefore, if a number of cam-

eras are mounted on half spherical surface around the target space and each cameras formed triangle effectively, virtual viewpoint can be moved freely all around the half sphere.

For synthesizing arbitrary view image, intermediate image are synthesized by interpolating two or three reference images. The interpolation is based on the related concepts of view interpolation [3]. At reference views, depth in PGS are rendered at the reference views with z-buffer algorithm. To apply the z-buffer algorithm to render the 3D model at the reference views, the surface of 3D model that reflect on the input views is decided. In the z-buffer of each pixel, the distance between the viewpoint and the closest surface point is stored, and the stored distances. Then these rendered depth images generate the correspondence maps between the reference images. For the correspondence points between the reference images, the following equations are applied to the interpolation:

$$\boldsymbol{P} = w_1\boldsymbol{P}_i + w_2\boldsymbol{P}_j + w_3\boldsymbol{P}_k \qquad (4)$$

$$I(\boldsymbol{P}) = w_1 I_i(\boldsymbol{P}_i) + w_2 I_j(\boldsymbol{P}_j) + w_3 I_k(\boldsymbol{P}_k) \qquad (5)$$

$\boldsymbol{P_i}$, $\boldsymbol{P_j}$ and $\boldsymbol{P_k}$ are the position of the corresponding points in the three reference images (Figure3), $I_i(\boldsymbol{P_i})$, $I_j(\boldsymbol{P_j})$ and $I_k(\boldsymbol{P_k})$ are the colors of the corresponding points, and $\boldsymbol{P}$ and $I(\boldsymbol{P})$ are the interpolated position and color. The interpolation weighting factors are represented by w_1 , w_2 and w_3 ($w_1 + w_2 + w_3 = 1$, $0 \leq w_1, w_2, w_3 \leq 1$). Changing the weighting ratio, virtual viewpoint can move inside of the triangle.

However, there is the case that some points in one image cannot be seen in another image. In this case, the position on the interpolating image of such point is calculated by equation 4, and the color is decided on the value of visible point. Therefore, even if certain region is occluded on the one image, the region can be synthesized on the novel image, as long as the region is not occluded on the other images.

5 EXPERIMENTAL RESULTS

We constructed multi-camera movie capturing system in large concert hall, "B-con Plaza hall", located in Beppu City, Oita, Japan. The concert hall is 110 meters (L) $\times$ 50 meters (W) $\times$ 25 meters (H). We mounted 16 cameras on the wall and 2 cameras on the ceiling, capturing PC were connected every neighboring 2 cameras, and all of the PC could be controlled by system control PC. All of those cameras are fully synchronized with a common signal, all the video signals could be digitized and captured as color image (640 $\times$ 480 BMP format) sequence at full video rate (30 frame per second).

Figure 4: The B-con Plaza hall

The inside of the concert hall is shown figure 4, and the outline of the each camera position of the camera system is shown figure 5.

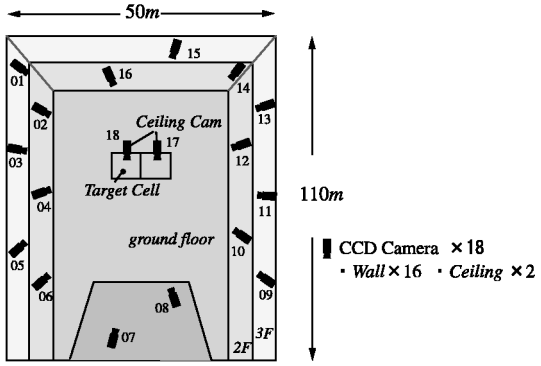

Figure 5: Outline of camera placement of our system.

We applied our method to image sequence taken by the cameras system. 3D shape models were reconstructed each frame, and arbitrary view images were synthesized as interpolated image between any three or two images.

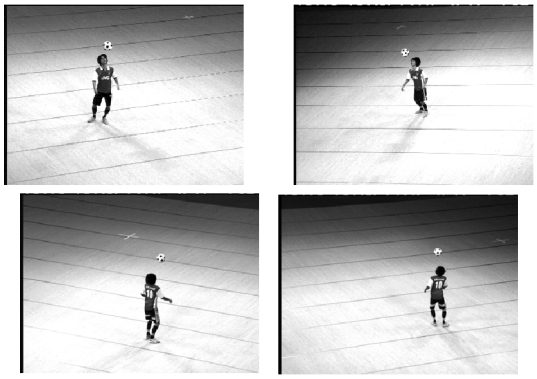

Figure 6: The example of input real images.

The example of original real images are shown figure 6. The silhouette images were generated background subtraction, and 3D shape model was reconstructed in the PGS by the proposal method. The 3D shape model in the representation of orthographic grid space, which is seen from the arbitrary view, are shown figure 7.

Figure 8 shows synthesized image of a scene to which two men are doing lifting. Three images of the vertex of the triangle are the selected reference images of the

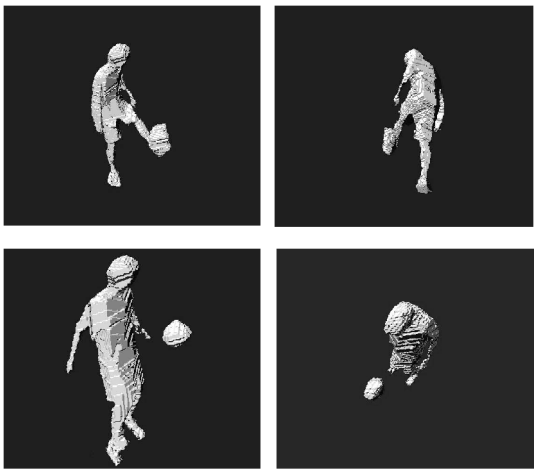

Figure 7: Reconstructed projective shape in the representation of orthographic grid space.

same frame. Arbitrary view images are synthesized from those images by changing the weighting ratio. The images on the line are interpolated from two reference images of both ends, and center of the image is interpolated from three reference images.

Figure 9 shows the sequence of synthesized virtual moving camera. This is the example to apply our method to image sequence synthesized by interpolating 4 reference views.

6 CONCLUSION

We proposed a method for reconstructing the 3D shape model in the projective grid space and synthesizing the arbitrary view image from the multiple image sequences taken with uncalibrated cameras. The projective grid space can be defined with two basis views, whose relationship is represented by a fundamental matrix. The grid points in the space are related to an arbitrary image by fundamental matrices between the image and the two basis views. In the projective grid space, the shape from silhouette

(SS) method is applied for reconstructing the shape model, which provides dense corresponding points between the images for synthesizing intermediate appearance view image. We demonstrated the proposed framework by showing several virtual image sequences generated from multiple camera image sequences that are corrected in a large space of 110m x 50m x 25m.

ACKNOWLEDGEMENT

The authors thank the members of the consortium for virtualized reality experimental project in Oita, Japan, including Prof. T.Kanade(CMU) and Prof. Y.Ohta (University of Tsukuba), for their co-operative efforts to capturing the image sequences.

REFERENCES

[1] Beier,T. and Neely,S.: Feature-Based Image Metamorphosis, *Proc. of SIGGRAPH '92*, pp. 35–42, 1992.

[2] Chein,C.H. and Aggarawal,J.K.: Identification of 3D Objects from Multiple Silhouettes using Quadtrees/Octrees, *Conputer Vision, Graphics and Image Processing*, vol. 36, pp. 100–113, 1986.

[3] Chen,S. and Williams,L.: View Interpolation for Image Synthesis, *Proc. of SIGGRAPH '93*, pp. 279–288, 1993.

[4] Kanade,T., Rander,P.W., Vedula,S. and Saito,H.: Virtulized Reality : Digitizing a 3D Time-Valying Event As Is and in Real Time, *International Symposium on Mixed Reality(ISMR99)*, pp. 41–57, 1999.

[5] Potmesil,M.: Synthesizing Octree Models of 3D Objects from Their Silhouettes in a Sequense of Images. *Computer Vision, Graphics and Image Processing*, Vol. 40 pp. 277–283, 1987.

[6] Saito,H. and Kanade,T.: Shape Reconstruction in Projective Grid Space from Large Number of Images, *IEEE Proc. Computer Vision and Pattern Recognition*, Vol. 2, pp. 49–54, 1999.

[7] Saito,H., Baba,S., Kimura,M., Vedula,S. and Kanade,T.: Apperance-Baced Virtual View Generation of Temporally-Varying Events from Multi-Camera Images in 3D Room, *Second International Conference on 3-D Digital Imaging and Modeling (3DIM99)*, pp. 516–525, 1999.

[8] Seitz,S.M. and Dyer,C.R.: View Morphing,*proc. of SIGGRAPH '96*, pp. 21–30, 1996.

[9] Tsai,R.: A Versatile Camera Caribration Technique for High- Accuracy 3D Machine Vision Metrology Using Off-the-Shelf Tv Cameras and Lenses, *IEEE Journal of Robotics and Auto mation* RA-3,4, pp. 323–344, 1987.

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

[11] Zhang,Z.: Determining the Epipolar Geometry and its Uncertainly: A Review. *INRIA reserch report*, Vol. 2927, 1996.

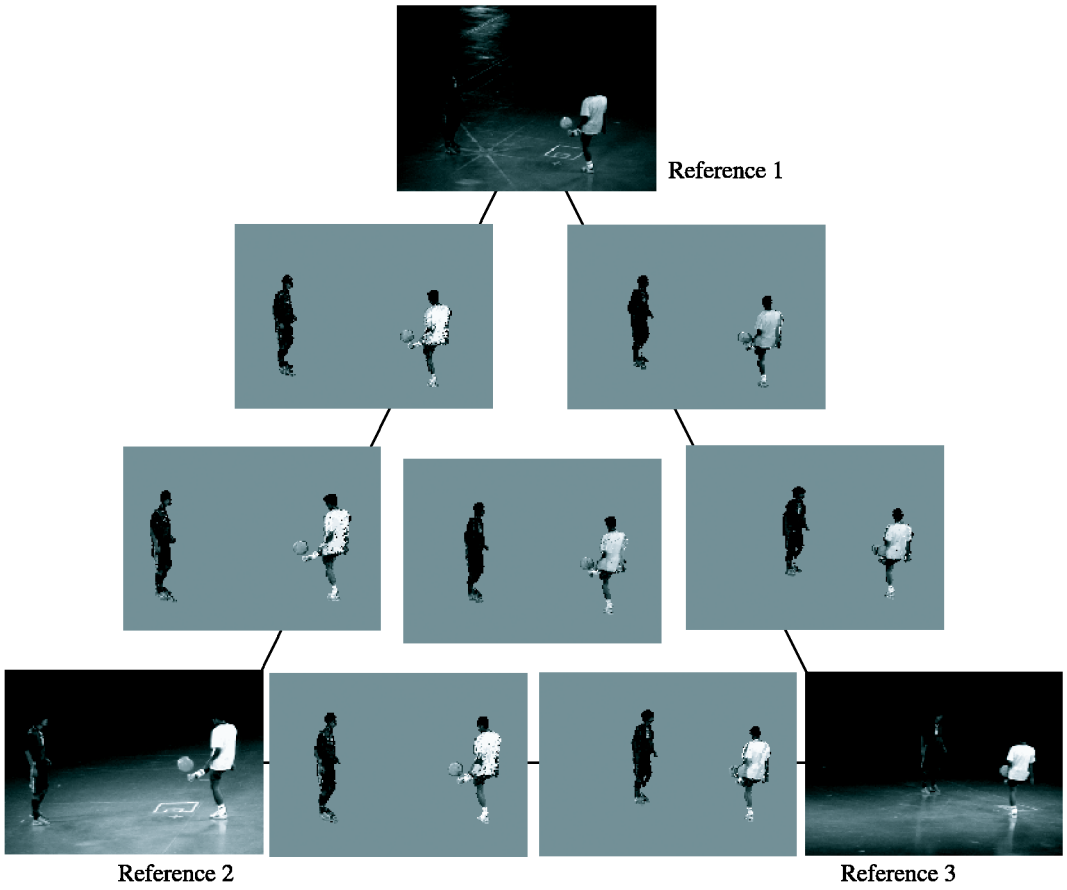

Figure 8: Synthesized intermediate view images among three images.

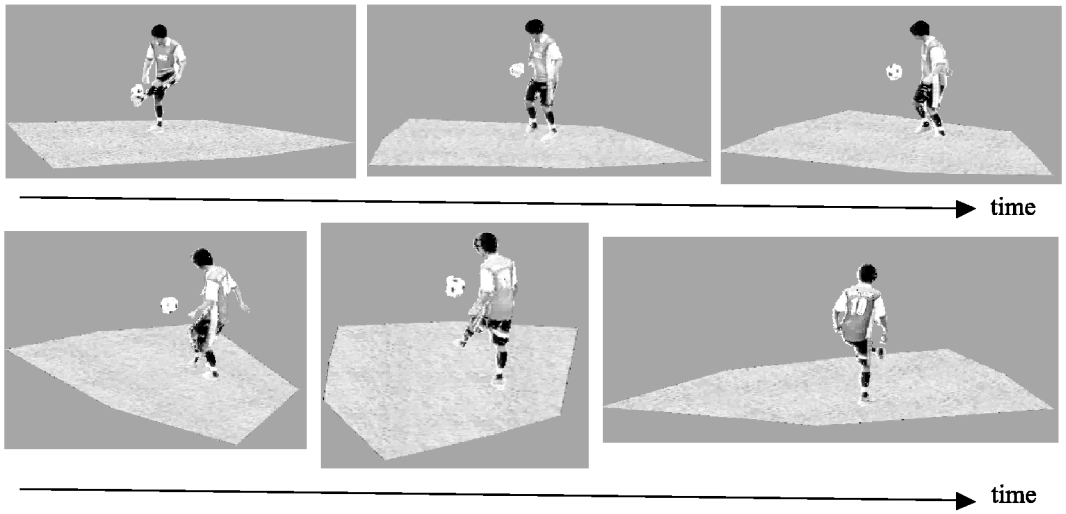

Figure 9: Image sequence at virtually moving view points for the object with synthesized lawn floor. Since the shape model is reconstructed in the PGS, the relationship of the occlusion is correctly detected.