Fly-through viewpoint video system for multi-view soccer movie using viewpoint interpolation

Naho Inamoto and Hideo Saito

Department of Information and Computer Science

Keio University, Yokohama, Japan

ABSTRACT

This paper presents a novel method for virtual view generation that allows viewers to fly through in a real soccer scene. A soccer match is captured by multiple cameras at a stadium and images of arbitrary viewpoints are synthesized by view-interpolation of two real camera images near the given viewpoint. In the proposed method, cameras do not need to be strongly calibrated, but epipolar geometry between the cameras is sufficient for the view-interpolation. Therefore, it can easily be applied to a dynamic event even in a large space, because the efforts for camera calibration can be reduced. A soccer scene is classified into several regions and virtual view images are generated based on the epipolar geometry in each region. Superimposition of the images completes virtual views for the whole soccer scene. An application for fly-through observation of a soccer match is introduced as well as the algorithm of the view-synthesis and experimental results.

Keywords: view interpolation, soccer match, dynamic event, large space, epipolar geometry

1. INTRODUCTION

Many kinds of visual effects can be seen in movies or TV broadcasts. One way of enhancing such effects is virtual movement of viewpoint. Recent applications of these techniques are found in the movie "The Matrix", and "Eye Vision" system used at Super Bowl XXXV broadcast by CBS. In the Eye Vision system, the multiple video streams are captured with more than 30 cameras. The sequences of video images from different angles are then used to create a three-dimensional visual effect such that the viewpoint turns around the object event at a temporally freezed moment. Whereas "The Matrix" and "Eye Vision" employ the switching effect of real camera images, computer vision based technology can synthesize intermediate view images between neighboring real cameras for the virtual viewpoint movement.

We have been studying the method for synthesis of arbitrary views from multiple video images targeting dynamic events at a large space.[5,6] Although several approaches to the view synthesis problem have been proposed,[8,13,15] the applicable method for the whole scene of large-scale real events does not exist.

In generally, view synthesis techniques can be categorized into two groups. In the first group, a 3D shape model of an object is reconstructed to generate the desired view.[7,10,12,15] The quality of the virtual view image depends on accuracy of the 3D model. As a large number of video cameras or range scanners are typically used to construct an accurate model. In the second group, arbitrary view image is synthesized without an explicit 3D model; instead, some form of image warping, such as transfer of correspondences is used.[1,3,9,11]

In this paper, we propose a fly-through viewpoint video system for soccer match playbacks. View interpolation reconstructs the whole soccer scene from any intermediate viewpoints between real cameras. Using the system, an user can freely select the preferred viewpoint. He/she may focus on a specific player in closed-up view or may track the ball movement through zoom-out virtual cameras.

Our purpose is to generate virtual views of scenes in a soccer match. Since the movements of each player are complex, it's almost impossible to reconstruct an accurate 3D model of the scene. Furthermore, strong camera calibration[14] that is necessary to compute 3D positions is very difficult in a real stadium. Therefore, it's not

E-mail: {nahotty, saito}@ozawa.ics.keio.ac.jp

Telephone: +81–45–563–1141

Address: 3–14–1 Hiyoshi, Kouhoku-ku, Yokohama 223-8522 Japan

Visual Communications and Image Processing 2003, Touradj Ebrahimi, Thomas Sikora, Editors, Proceedings of SPIE Vol. 5150 (2003) © 2003 SPIE · 0277-786X/03/$15.00

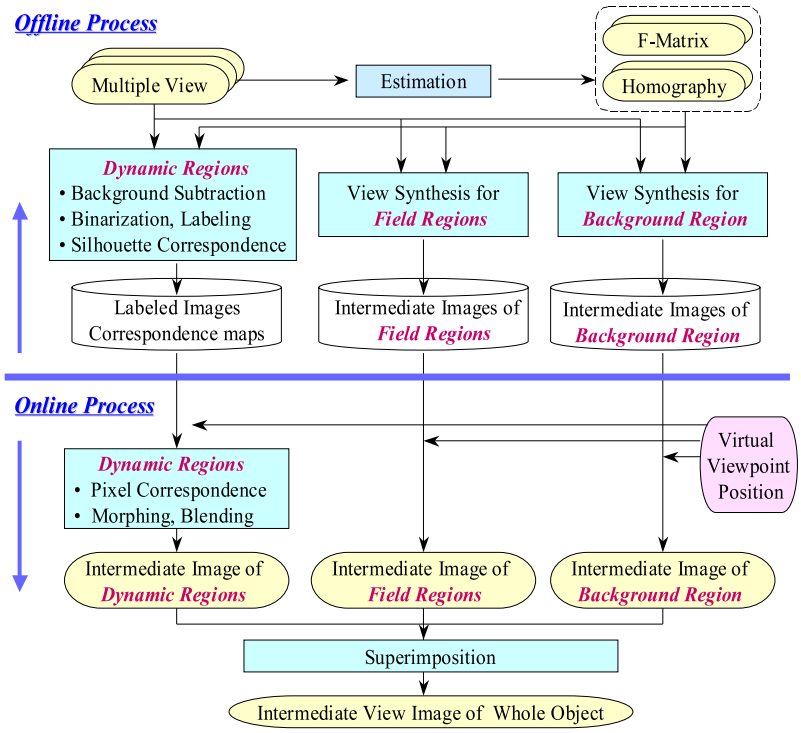

Figure 1. Overview of the proposed method

practical to apply a method that requires 3D models to the generation of virtual views of soccer scenes. Instead of using 3D models, we use projective geometry between multiple cameras, which can easily be obtained by images themselves. After classifying a soccer scene into dynamic regions and static regions, we apply the appropriate projective transformation respectively. Separating offline analysis of static scene parts from online processing of dynamic objects, our proposed system enables viewers to watch soccer scenes as a form of nearly to real-time.

2. OVERVIEW

Figure 1 describes an overview of the proposed method. Our approach obeys as follows. First of all, the fundamental matrices[4] between the viewpoints of the cameras and the homographic matrices[4] between the planes in different views are estimated from multiple view images. Then, soccer scene is divided into dynamic regions, field regions, and a background region for view interpolation. As offline process, virtual view images of static regions, such as field regions and a background region, are synthesized for all intermediate viewpoints. For dynamic regions, every player region is segmented and labeled. The labeled regions of the same player in the neighboring view images are corresponded by using homography of the ground plane between the views. Although it's necessary to generate the virtual views of the dynamic regions at each frame, the above process is done offline to render the scene efficiently. As online process, intermediate view images for the dynamic regions are synthesized based on a position of the virtual viewpoint given by the user. Finally superimposition completes virtual view images for the whole scene.

3. VIEW INTERPOLATION

3.1. Static regions

The method for view interpolation in each region is described below. Firstly as for the static regions, they are classified into two regions. One is field regions, which we can approximate as sets of planes, and the other is a background region, which we can approximate as an infinitely distant plane. Field regions correspond to the

ground and the soccer goal, where dense correspondence is obtained by applying homography to each plane. Their positions are transferred by means of image morphing.[2] For the background region including stands or stadium, partial area images are extracted from the panoramic image compounded from the background of multiple views. As intermediate viewpoint position is defined by the relative weights to two real camera viewpoints, changing the weights gradually synthesize virtual view images of static regions for all possible viewpoints.

3.1.1. Field regions

In a soccer scene, the ground and soccer goal can be considered as a single plane and a set of planes, respectively. We then apply homography to the planes to obtain the correspondences required for the generation of intermediate view images. The following equation gives the pixel-wise correspondence for two views of a plane.

$$\boldsymbol{p}_2 \cong \boldsymbol{H}\boldsymbol{p}_1 \tag{1}$$

where $\boldsymbol{H}$ is the homographic matrix that represents the transformation between the planes, and p_1, p_2 are homogenous coordinates on the images I_1, I_2 of different views (in figure 2). The homographic matrices of the plane that represents the ground and the planes of the soccer goal provide the dense correspondence within these regions. The position and the value of the pixels are then transferred by image morphing to complete the destination images as described by the following equations.

$$\acute{\boldsymbol{p}} = (1-\alpha)\boldsymbol{p}_1 + \alpha\boldsymbol{p}_2 \tag{2}$$

$$I(\acute{\boldsymbol{p}}) = (1-\alpha)I(\boldsymbol{p}_1) + \alpha I(\boldsymbol{p}_2) \tag{3}$$

where $\boldsymbol{p}_1$, $\boldsymbol{p}_2$ are the coordinates of the matching points in images I_1, I_2, and $I(\boldsymbol{p}_1)$, $I(\boldsymbol{p}_2)$ are the value of the matching points in images I_1, I_2 as well. $\acute{\boldsymbol{p}}$ is the interpolated coordinates and $I(\acute{\boldsymbol{p}})$ is the interpolated value. α defines the relative weights given to the respective actual viewpoints. All correspondences are used in the

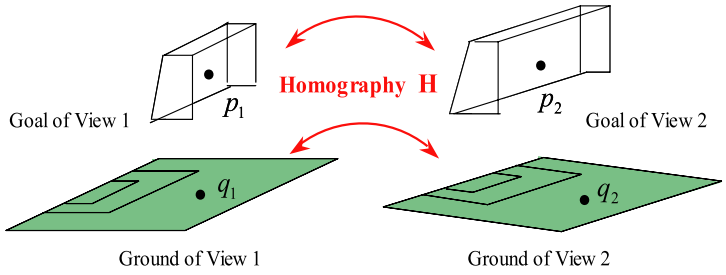

Figure 2. Correnspondence for the field regions.

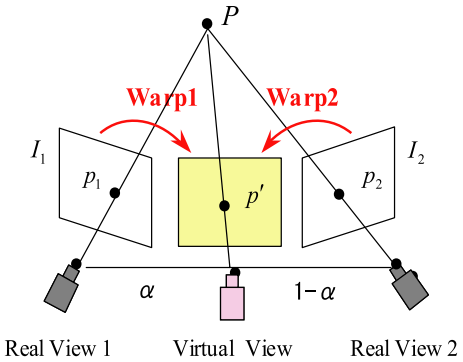

Figure 3. Image morphing for transfer of the correspondence.

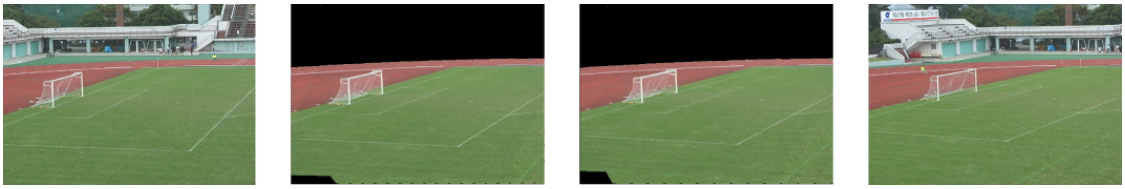

Figure 4. Examples of the intermediate images for the field regions

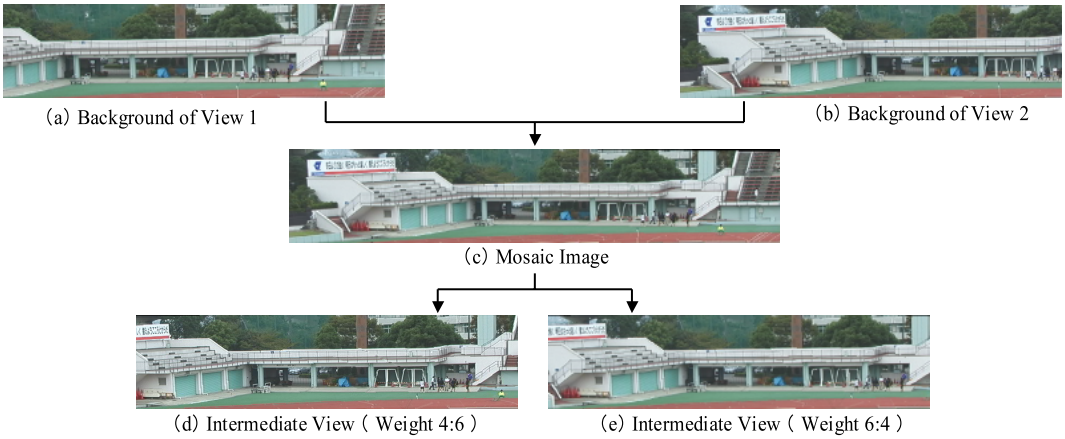

Figure 5. Examples of the intermediate images for the background.

transfer to generate a warped image. Here two transfers are required, one from view 1 and the other from view 2 as shown in figure 3. Two warped images are thus generated; they are then blended to complete the image of the virtual view. If the color of a pixel is different in the two images, the corresponding pixel in the virtual view is rendered with the average of the colors; otherwise the rendered color is from either actual image.

Figure 4 presents examples of generated intermediate images for the field regions. (a) and (d) are real camera images, and (b) and (c) are interpolated images from (a) and (d). The relative weight of the virtual view to the real views is 4 to 6 in (b) and 6 to 4 in (c).

3.1.2. Background region

The background may be considered as a single infinitely distant plane, so we are able to compose images from each of the two input viewpoints to make mosaics that are the respective panoramic images of the background. Intermediate views of this region are extracted from these panoramic images.

In composition, we start by integrating the coordinate systems of the two views in the homographic matrix $\boldsymbol{H}_b$ for the background. Next, blending the pixel values of the overlap area so that pixel colors at junction areas can be smoothed connects the two backgrounds. Partial area that is necessary for each virtual view image is cut out from the mosaic image thus synthesized. The following homographic matrix, $\acute{\boldsymbol{H}}_b$, is then used in the transformation of coordinates to complete the intermediate view of the background region.

$$\acute{\boldsymbol{H}}_b = (1-\alpha)\boldsymbol{E} + \alpha \boldsymbol{H}_b^{-1} \tag{4}$$

where α is the weight and $\boldsymbol{E}$ is the unit matrix.

Figure 5 presents examples of generated intermediate images for the background. (a) and (b) are background regions in real camera images, and (c) is a mosaic image composed of (a) and (b). (d) and (e) are interpolated background regions, whose relative weight is 4 to 6 in (d) and 6 to 4 in (e).

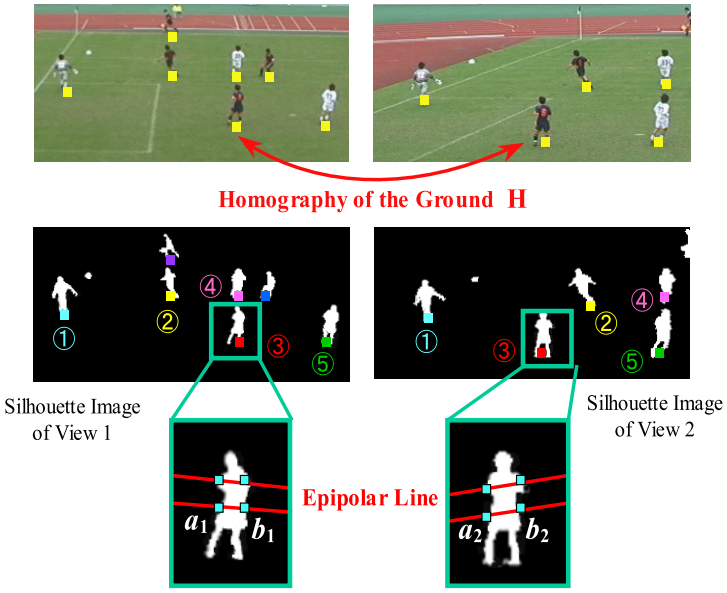

Figure 6. Correspondence for the dynamic regions.

3.2. Dynamic regions

Secondly, the method of view interpolation for the dynamic regions is introduced. The process is categorized into offline and online parts. At the start of offline process, subtracting the background from two original view images extracts all dynamic regions. A single scene usually contains several players and a ball, so we deal with these objects separately. After the silhouette images are generated by binarization, labeling process segments each player and the ball. If occlusion is detected, features of the previous frame are used for estimation of features of the current frame. The correspondence between silhouettes of the players is obtained by using the homography of the ground plane as shown in figure 6. This is based on the fact that all feet of the players are attached on the ground. Even when a player is jumping off the ground, the error caused by the jumping is sufficiently small, so the homographic matrix of the plane that represents the ground can still find corresponding silhouettes. Both labeled images and the correspondence maps are stored every two neighboring viewpoints for online process.

Then in online part, process goes based on the information about two reference viewpoints near the given virtual viewpoint. Drawing epipolar lines in the different views, view 1 and view 2, by using a fundamental matrix obtains dense correspondence inside of the silhouettes. On each epipolar line, the correspondences of intersections with boundaries, such as a_1 and a_2, b_1 and b_2 of figure 6, are made first. The correspondences between the pixels inside the silhouette are obtained by linear interpolation of the points of intersection. After a dense correspondence for the whole silhouette is obtained, the pixel positions and values are transferred from the source images of view 1 and view 2 to the destination image by image morphing in the same way as the field regions. Now we can use the following equation instead of equation (2).

$$\acute{\boldsymbol{p}} = (1-\alpha)\{(\boldsymbol{p}_1 - \boldsymbol{c}_1)\frac{\acute{f}}{f_1} + \boldsymbol{c}_1\} + \alpha\{(\boldsymbol{p}_2 - \boldsymbol{c}_2)\frac{\acute{f}}{f_2} + \boldsymbol{c}_2\} \tag{5}$$

where $\boldsymbol{c}_1$, $\boldsymbol{c}_2$ are the coordinates of the principal points in images I_1, I_2 and also f_1, f_2 are the focal lengths of camera 1, 2. $\acute{f}$ is the focal length of the virtual camera. This equation makes it possible to zoom in or out by changing the ratio of the focal length of the real camera to the focal length of the virtual camera. The pixels value is transferred by the same equation (3). Blending two warped images generate virtual views. The above algorithm is applied to every pair of silhouettes and synthesized in order of distance from the viewpoint completes view interpolation for dynamic regions. Finally superimposition of the images generated in static and dynamic regions completes the virtual view image of the whole scene for the given viewpoint.

4. EXPRIMENTAL RESULTS

In order to make sure the utility, we have applied this method to scenes of an actual soccer match that were taken by multiple video cameras at a soccer stadium, Edogawa athletics stadium in Tokyo, Japan. As figure 7 shows, a set of 4 fixed cameras was placed to one side of the soccer field to capture the penalty area mainly. The captured videos were converted to BMP format image sequences, which were composed of 720×480 pixels,

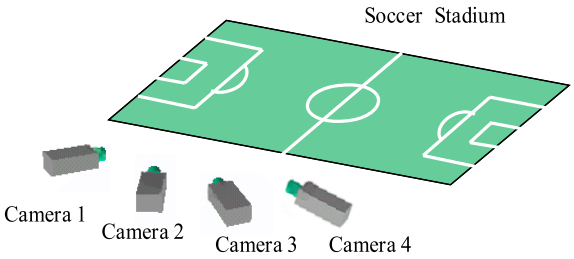

Figure 7. Configulation of the cameras at the soccer stadium.

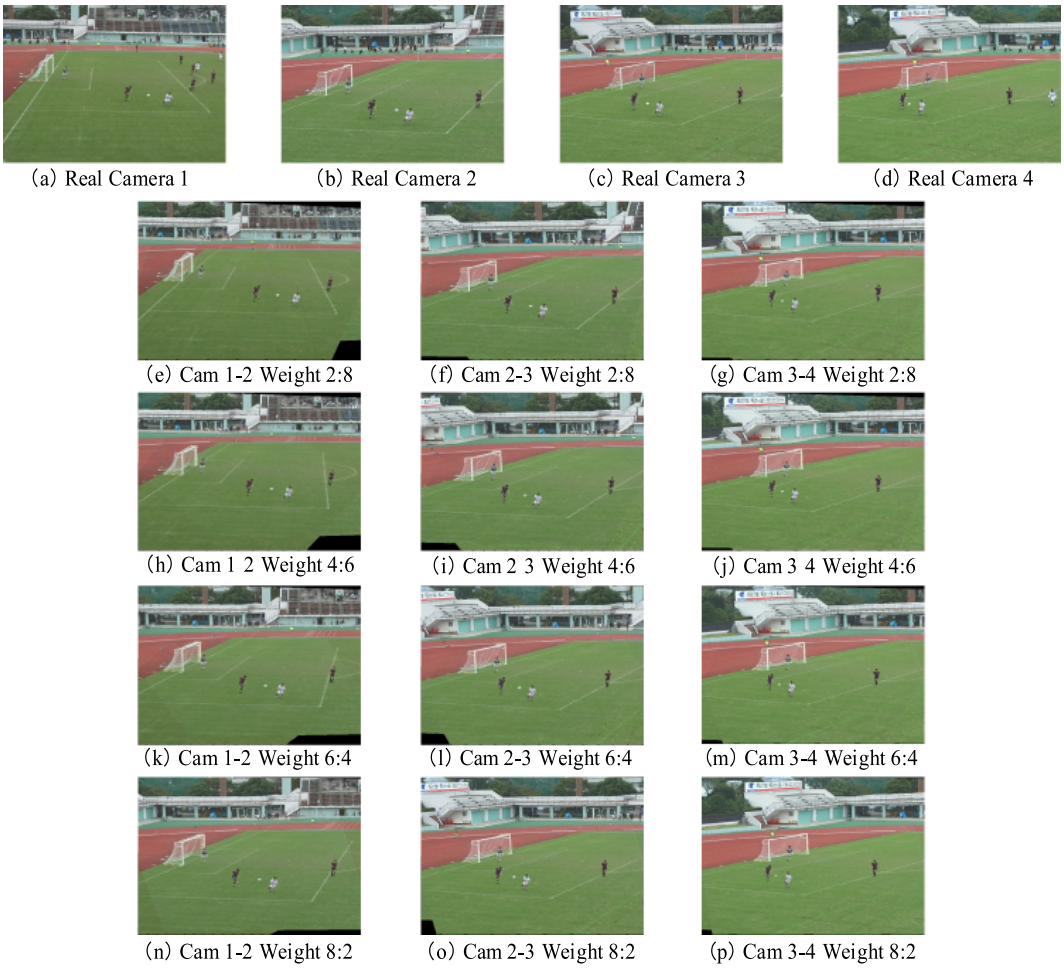

Figure 8. Synthesized virtual view images for the whole soccer scene about the same frame. (a),(b),(c),(d) are real camera images. (e),(h),(k),(n) are interpolated view images between camera 1 and 2. (f),(i),(l),(o) are between camera 2 and 3, and (g),(j),(m),(p) are between camera 3 and 4 as well.

24-bit-RGB color images, and then used for virtual view synthesis.

In such a real stadium, strong camera calibration that is sufficiently accurate for the estimation of camera rotation and position is almost impossible. The proposed method does not require accurate calibration; instead, it takes advantage of epipolar geometry. Fundamental matrices between the viewpoints of the cameras and homographic matrices between the planes in neighboring views can be computed easily through correspondences between several feature points in the images. In this experiment, we manually selected about 20 corresponding feature points in the input images.

Figure 8 presents some results of generated intermediate view images. (a),(b),(c),(d) are captured images by real cameras and the others are generated virtual view images.

Position of players and location of the background gradually change depending on the angle of the virtual viewpoint, which is determined by the relative weights to two real camera viewpoints. For example, the virtual viewpoint of (e) is placed at the position whose relative weight is 2 to 8 between camera 1 and camera 2. Comparing the results with the input images, we see that we were able to successfully obtain realistic images without distortion. Although the method involves the rendering of separate regions, the synthesized images look so natural that the boundaries between the regions are not visible.

In addition, we have produced a video* that gives viewers the impression of fly-through over the soccer field or playing together in the soccer match by changing positions of the viewpoint with the ball movement. Another example is a video that creates a three-dimensional effect of walking around the action like the movie "The Matrix". Rotating the virtual camera around one player focused on realize such an effect. Not only for entertainment but also for training soccer players, a kind of videos analyzing specific players can be given through the proposed method.

5. VIEWPOINT ON DEMAND SYSTEM

As application of the proposed method, we have constructed the system called "Viewpoint on Demand System", which allows viewers to watch soccer match from a favorite viewpoint. Figure 9 shows the interface of the system. At the center of the window, generated virtual view images are drawn in accordance with the position and the

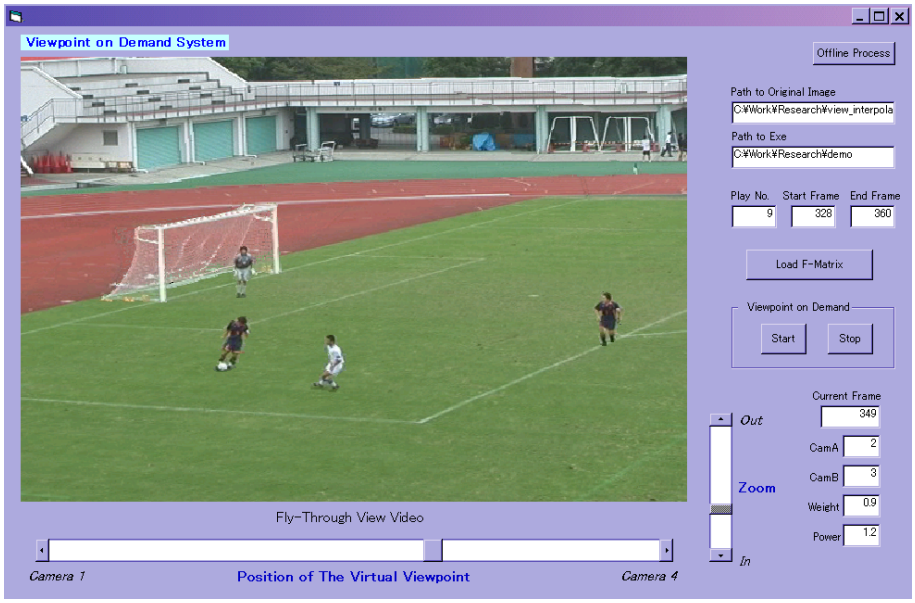

Figure 9. The interface of the Viewpoint on Demand System.

*The fly-through view videos are available at the following web site.
http://www.ozawa.ics.keio.ac.jp/~nahotty/research.html

(a) Frame 322 Cam 3-4 Weight 4:6 Zoom 0.8

(b) Frame 329 Cam 2-3 Weight 8:2 Zoom 1.2

(c) Frame 335 Cam 2-3 Weight 1:9 Zoom 1.5

Figure 10. Examples of the image window rendered soccer scene from various angles and focuses.

zoom ratio of the virtual camera. The position of the virtual camera, which means the relative weight α in equation (5), is decided by the horizontal slide bar at the bottom of the window. The zoom ratio, which means $\frac{f}{f_1}$ and $\frac{f}{f_2}$ in equation (5), is determined by the vertical slide bar on the right of the window. When users select favorite scenes, rendering of the virtual view, whose position and the zoom ratio is defined initially, starts. While watching the video, they can change the viewpoint anytime using two slide bars. Figure 10 presents examples of the images shown on the window of the system. (a), (b), and (c) are different scenes reconstructed from different angles with different focuses. (a) is the scene of frame number 322 where the virtual viewpoint is placed at the relative weight 4 to 6 between camera 3 and 4, and the zoom ratio of the virtual camera to the real camera is 0.8. In the same way, (b) is the scene of frame number 329 where the weight is 8 to 2 between camera 2 and 3, and zoom ratio is 1.2. (c) is the scene of the frame number 335 where the weight is 1 to 9 between camera 2 and 3, and zoom ratio is 1.5. This application offers a new framework of presenting a soccer match on demand.

6. CONCLUSION

This paper has presented a novel method for virtual view generation for fly-through viewpoint observation of a soccer match. A soccer scene is classified into three regions and epipolar geometry is employed for view interpolation in each region. Dividing offline process and online process realizes render a whole soccer scene effectively. Without reconstructing 3D models, taking advantage of projective geometry between cameras accomplish view synthesis targeting a dynamic event in a large space like a soccer stadium.

As well as the techniques of view synthesis, an application of the proposed method, "Viewpoint on Demand System" is introduced. Through the system, an audience can watch a soccer match from the favorite angle with the preferred zoom ratio and may change them anytime while watching the match. This framework will lead to the creation of completely new and enjoyable ways to present and view entertainments or sporting events as well as soccer games.

REFERENCES

1. S. Avidan and A. Shashua, *Novel View Synthesis by Cascading Trilinear Tensors*, IEEE Trans. on Visualization and Computer Graphics, Vol.4, No.4, pp.293-306, 1998.
2. T. Beier and S. Neely, *Feature-Based Image Metamorphosis*, Proc. of SIGGRAPH '92, pp.35-42, 1992.
3. S. E. Chen and L. Williams, *View Interpolation for Image Synthesis*, Proc. of SIGGRAPH '93, pp.279-288, 1993.
4. R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2000.
5. N. Inamoto and H. Saito, *Fly Through View Video Generation of Soccer Scene*, International Workshop on Entertainment Computing (IWEC2002), Workshop Note, pp.94-101, May 2002.
6. N. Inamoto and H. Saito, *Intermediate View Generation of Soccer Scene from Multiple Videos*, Proc. of International Conference on Pattern Recognition (ICPR2002), August 2002.

7. T. Kanade, P. J. Narayanan, and P. W. Rander, *Virtualised reality:concepts and early results*, Proc. of IEEE Workshop on Representation of Visual Scenes, pp.69-76, 1995.
8. I. Kitahara, Y. Ohta, H. Saito, S. Akimichi, T. Ono, and T. Kanade, *Recording Multiple Videos in a Large-scale Space for Large-scale Virtualized Reality*, Proc. of International Display Workshops (AD/IDW'01), pp.1377-1380, 2001.
9. S. Pollard, M. Pilu, S. Hayes, and A. Lorusso, *View synthesis by trinocular edge matching and transfer*, Image and Vision Computing, Vol.18, pp.749-757, 2000.
10. H. Saito, S. Baba, M. Kimura, S. Vedula, and T. Kanade, *Appearance-based Virtual View Generation of Temporally-Varying Events from Multi-Camera Images in 3D Room*, Proc. of the Second International Conference on 3-D Imaging and Modeling (3DIM99), pp.516-525, 1999.
11. S. M. Seitz, and C. R. Dyer, *View Morphing*, Proc. of SIGGRAPH '96, pp.21-30, 1996.
12. S. M. Seitz, and C. R. Dyer, *Photorealistic Scene Reconstruction by Voxel Coloring*, Proc. Computer Vision and Pattern Recognition (CVPR1997), pp.1067-1073, 1997.
13. D. Snow, O. Ozier, P. A. Viola, and W. E. L. Grimson, *Variable Viewpoint Reality*, NTT R&D, Vol.49, No.7, pp.383-388, 2000.
14. R. Y. Tsai, *A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology Using Off-the-Shelf TV Cameras and Lenses*, IEEE Journal of Robotics and Automation, vol.RA-3, no.4, pp.323-344, August 1987.
15. S. Yaguchi, and H. Saito, *Arbitrary View Image Generation from Multiple Silhouette Images in Projective Grid Space*, Proc. of SPIE Vol.4309 (Videometrics and Optical Methods for 3D Shape Measurement), pp.294-304, 2001.