Intermediate View Generation of Soccer Scene from Multiple Videos

Naho INAMOTO Hideo SAITO
Department of Information and Computer Science
Keio University, Yokohama, Japan
{nahotty, saito}@ozawa.ics.keio.ac.jp

Abstract

This paper introduces a novel method for generating an intermediate view of soccer scene taken by multiple video cameras. In the proposed method, soccer scene is classified into dynamic regions, a field region, and a background region. Using epipolar geometry in the first region and homography in the second, dense correspondence is obtained to interpolate views. For the third region, partial area images are extracted from the panoramic image compounded from the background of multiple views. Finally synthesizing them completes intermediate view images of the whole object. Applying this method to actual scenes of a soccer match captured at the stadium, we succeeded in generating natural intermediate view videos.

1. Introduction

Recently there has been great deal of interest in making system that enables an audience to view sports event from any arbitrary viewpoint. The rapid development of networks and computers will soon make it possible to provide a form of real-time video that allows the viewer to select any desired view of the action. Video processing technology is not enough, however, so many researchers are dealing with the problem of view synthesis of dynamic scene [4, 7, 8].

The several approaches to the view synthesis problem that have been proposed may be categorized into two groups. In the first group are those in which a full-3D model of an object is constructed to generate the desired view [3, 8]. The quality of the virtual view image then 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, this approach requires large amounts of calculation and information. Once the 3D model has been constructed, however, we are able to flexibly change the viewpoint to present another view. In the second group, the arbitrary view image is synthesized without an explicit 3D model; instead, some form of image warping, such as transfer of correspondences [1, 2, 5, 6] is used. While this approach doesn't require too much calculation or information, it is only possible to move the viewpoint within a limited area.

Our purpose is to generate arbitrary 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, the camera calibration that is necessary to compute 3D positions is very difficult in a real stadium. Therefore, it's not practical to apply a method that requires 3D models to the generation of views of soccer scenes. Instead of using 3D models, we classify the scene into several regions. An image of the required arbitrary view is generated for the respective regions and superimposition then completes the desired view of the whole object. In this paper, we describe how to generate intermediate view images/videos of a soccer scene, and in addition, by applying this method to actual scenes of a soccer match captured at the stadium, we ensure the utility.

2. Generating intermediate view images

Figure 1 gives an overview of the approach. First of all, the fundamental matrices between the viewpoints of the cameras and the homographic matrices between the planes in different views are estimated from multiple view images. Next, the scene is classified into dynamic regions, in which the shape or position changes over time, and a static region. In a soccer scene, the former corresponds to players and the ball and the latter to the ground, goal, and background. The static region is further classified into two regions. One is a field region, which we can approximate as sets of planes, and the other is a background region, which we can approximate as an infinitely distant plane. The object scene has now been classified into the dynamic regions, field region, and background region. Intermediate view images of the respective regions are then generated. Dense correspondence is obtained to interpolate views by using fundamental matrix (F-Matrix) in the dynamic regions and homography in the field region. For the background region, partial area im-

1051-4651/02 $17.00 © 2002 IEEE

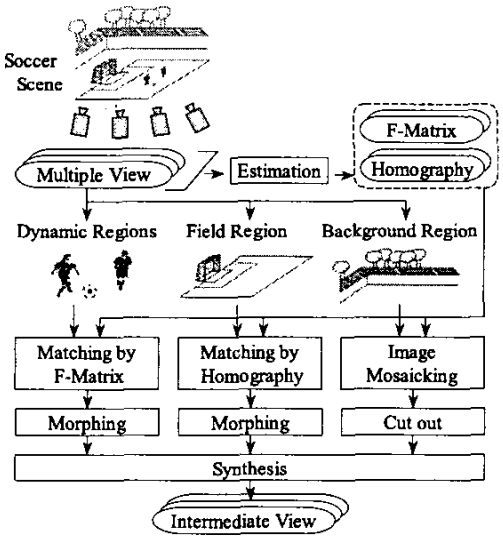

Figure 1. Overview of the approach.

ages are extracted from the panoramic image compounded from the background of multiple views. Finally the whole image is synthesized by their superimposition. The method for each region is described in this section.

2.1. Players and ball

A single scene usually contains several players and a ball, so we deal with these objects separately. Firstly, all dynamic regions are extracted by subtracting the background from the image. Considering color (RGB) vectors as well as intensity leads to more accurate extraction of these regions. After the silhouettes have been generated by binarization, labeling process segments each player and the ball. The correspondence between the segmented silhouettes of the players and those in images from other viewpoints is obtained by applying homography to the ground as shown in figure 2. 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.

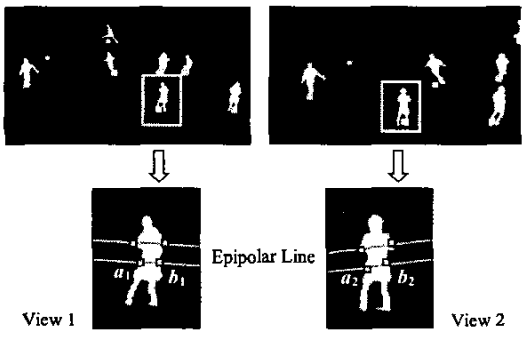

Figure 2. Dense correspondence.

Next, each pair of silhouettes is extracted to obtain the pixel-wise correspondence within the silhouette. Epipolar lines are drawn between images in the different views, view 1 and view 2, by using a fundamental matrix. On each epipolar line, the correspondences of intersections with boundaries in the silhouettes, such as a_1 and a_2, b_1 and b_2 of figure 2, 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 values are transferred from the source images of view 1 and view 2 to the destination image by image morphing, i.e., by linear interpolation according to the displacement of the pixel positions. The location in the synthesized view is given by

$$\dot{p} = (1 - \alpha)p_1 + \alpha p_2 \qquad (1)$$

where p_1, p_2 are the coordinates of matching points in images I_1, I_2, and α defines the relative weights given to the respective actual viewpoints. All correspondences are used in the transfer to generate a warped image. Here two transfers are required, one from view 1 and the other from view 2. 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.

For all of the extracted players and the ball, a pixel-wise correspondence as described above is established and additionally these pixel values are transferred for the rendering of intermediate views. Finally, synthesizing them in or-

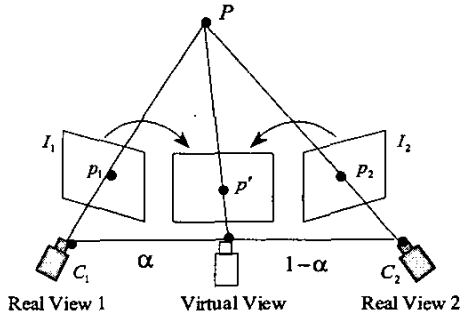

Figure 3. Transfer of correspondence.

der of distance from the viewpoint completes an intermediate view image of the dynamic regions. However, the segmentation doesn't work well when the players occlude each other. So we are working on the improvement on the method for the segmentation.

2.2. Ground and goal

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.

$$\boldsymbol{p}_2 \cong \boldsymbol{H}\boldsymbol{p}_1 \tag{2}$$

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. 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 pixel values are then transferred to corresponding points by image morphing to complete the destination images in the same way as for the players and ball. Figure 4 presents examples of generated images of the ground and goal regions, where the virtual viewpoint has been placed at the center of the pair of actual viewpoints.

2.3. Background

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

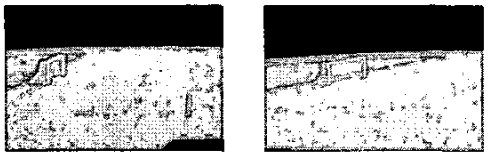

Figure 4. Intermediate view images of the ground and goal.

Figure 5. Synthesized mosaic image.

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.

$$\acute{\boldsymbol{H}}_b = (1-\alpha)\boldsymbol{E} + \alpha \boldsymbol{H}_b^{-1} \tag{3}$$

where α is the weight and $\boldsymbol{E}$ is the unit matrix. Figure 5 presents an example of mosaic images thus synthesized.

3. Generating intermediate view videos

When the methods described above are applied to the synthesis of an image sequence, the ground, goal, and background may be regarded as stable, so that frame-by-frame generation of these elements is inefficient. Instead, the stable regions are generated in advance for all possible virtual viewpoints; the dynamic regions are then synthesized for each frame.

First of all, the two cameras nearest the virtual viewpoint given by the user are selected and the required pairs of images are obtained. Next, the dynamic regions are extracted from each of the images, and the intermediate views of these regions are generated according to the method introduced in the previous section. These images are then composed with the stable regions from the same virtual viewpoint. This completes the image of the whole scene from the desired viewpoint.

For example, we can produce a video that gives the viewer the feeling of flying over the field with a viewpoint that moves with the ball. Another possibility is a video that create a three-dimensional effect of walking around the action like the Hollywood movie "The Matrix".

4. Experimental results

We have applied this method to scenes of an actual soccer match that were taken by multiple video cameras at the stadium. A set of four cameras is placed to one side of the field to capture the penalty area mainly. All input images are 720×480 pixels, 24-bit-RGB color images.

In such a real stadium, 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 fundamental matrices between the viewpoints of the cameras. We are easily able to obtain these through correspondences between several feature points in the images. In this

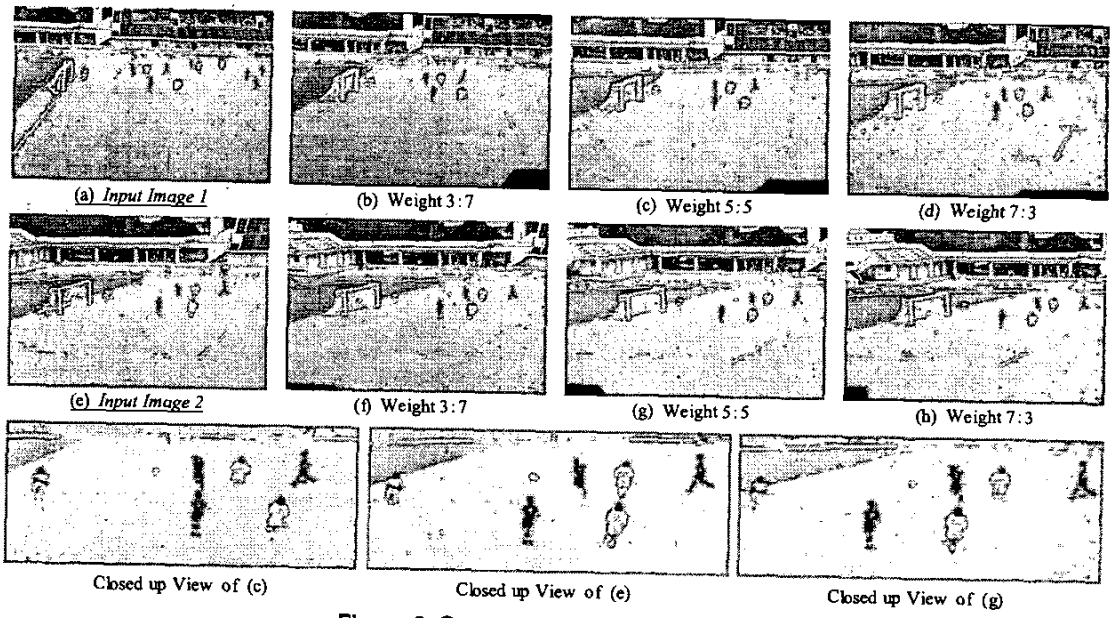

Figure 6. Generated intermediate views.

experiment, we manually selected about 20 corresponding feature points in the input images.

Figure 6 presents some results of generated intermediate view images. From (a) to (h), position of players and location of the background gradually change with the angle of the virtual view. Three images on the bottom especially present closed up views of dynamic regions of (c), (e), (g). 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. Full-color versions of these images and generated intermediate view videos are available at

http://www.ozawa.ics.keio.ac.jp/~nahotty/research.html

5. Conclusions

This paper has presented a novel method for the generation of intermediate view videos of a soccer game. The key idea behind the proposed method is to start by segmenting the image into object regions according to the properties of a soccer scene and to then apply the appropriate projective transformations and generate the desired view. This enables us to successfully generate arbitrary view videos from actual images of soccer matches that were captured at a stadium. The method will lead to the creation of completely new and enjoyable ways to present and view soccer games.

We are currently investigating the reduction of the number of manual operations, such as when the correspondences between player regions are obtained in complex scenes where the silhouettes cross each other.

References

[1] S.Avidan, 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] S.E.Chen, L.Williams,"View Interpolation for Image Synthesis," Proc. of SIGGRAPH '93, pp.279-288, 1993.

[3] T.Kanade, P.J.Narayanan, P.W.Rander, "Virtualised reality:concepts and early results," Proc. of IEEE Workshop on Representation of Visual Scenes, pp.69-76, 1995.

[4] I.Kitahara, Y.Ohta, H.Saito, S.Akimichi, T.Ono, 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.

[5] S.Pollard, M.Pilu, S.Hayes, A.Lorusso, "View synthesis by trinocular edge matching and transfer," Image and Vision Computing, Vol.18, pp.749-757, 2000.

[6] S.M.Seitz, C.R.Dyer,"View Morphing," Proc. of SIGGRAPH '96, pp.21-30, 1996.

[7] D.Snow, O.Ozier, P.A.Viola, W.E.L.Grimson, "Variable Viewpoint Reality," NTT R&D, Vol.49, No.7, pp.383-388, 2000.

[8] S.Yaguchi, 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.