SYNTHESIZING FREE-VIEWPOINT IMAGES FROM MULTIPLE VIEW VIDEOS IN SOCCER STADIUM

Kunihiko Hayashi, Hideo Saito
Department of Information and Computer Science, Keio University
{hayashi,saito}@ozawa.ics.keio.ac.jp

Abstract

We propose a new method for synthesizing free-viewpoint images from multiple view videos in soccer stadium. The previous method[1] can synthesize virtual viewpoint images at only intermediate viewpoint between real cameras. The previous method estimates only pixel-wise correspondences between input images for synthesizing virtual viewpoint images without a 3D reconstruction. In this paper, we propose a new method that can synthesize virtual viewpoint images at free-viewpoints which are not limited at intermediate viewpoint by reconstructing simple 3D models. To reconstruct the object soccer scene, we perform camera calibration without marker objects because we can save labor to put such objects on the soccer ground. For reconstructing of dynamic regions, we employ a billboard representation of dynamic regions. The position of the billboard is determined by tracking the players. The billboard representation can realize a simple 3D reconstruction and transform textures into the appropriate appearance of the aspect of dynamic regions at virtual viewpoint. We also introduce the choice of optimal cameras for rendering on the billboard. We choose the optimal camera that has the most similar texture at virtual viewpoint. By applying our method to a soccer scene, we can synthesize free-viewpoint images by employing only pixel values of real images.

Keywords— **image processing, free-viewpoint, multiple view videos, homography, projective matrix, calibration**

1 Introduction

The various kinds of visualization are given to broadcastings such as sports. For example, a virtual offside line is inserted to soccer live videos[2]. An approximate distance from a set-point to the goal in a free-kick is inserted to soccer live videos too. Moreover, a record line is inserted to live videos while a swimming race is played.

In a similar manner, there are much requirements that a user can observe sports games at free-viewpoint in broadcastings. Then, many researchers are dealing with view synthesis of dynamic scene [3, 4, 5, 6, 7, 8].

Virtualized reality[9] reconstructs a full-3D model of a scene to synthesize free-viewpoint images. However, because a large number of cameras are necessary to recover a precise 3D model, this method requires a great number of computation and targets basically local spaces.

On the other hand, there are synthesizing methods without recovering a explicit 3D model but transfering correspondences or recovering a simple 3D model in a large scale space to save the amount of the calculation[1, 10, 11].

Inamoto at al. proposed a method for synthesizing intermediate view images between two cameras in a soccer scene [1]. In their method, after classifying a soccer scene into dynamic regions and static regions, an appropriate projective transformation is applied to each region for synthesizing intermediate images. However, this method can synthesize only intermediate viewpoint images between two cameras. There are two reasons why this method can synthesize only intermediate images. One is that this method uses a weak calibration. The other is that this method does not recover 3D shape information in soccer stadium.

Kameda at al. solve this restriction and propose a method that a user can observe from his/her favorite free-viewpoint in a soccer scene[10]. Furthermore, this method enables real-time transmission and display. However, in this method, static regions except dynamic regions are rendered by using only synthesized 3D model, which causes lacks of reality.

Kimura at al. proposed a method for synthesizing player-viewpoint in a single tennis match scene by employing the only pixel values of real images[11]. However, a single tennis match scene only includes dynamic regions of two players and a ball.

In this paper, we propose a new method to synthesize free-viewpoint images from multiple view videos in soccer stadium. Our method is based on the previous method[1], but can synthesize not only intermediate images but also extra-intermediate images which are at user's favorite free-viewpoint by caribrating cameras to reconstruct simple 3D models and reconstructing simple 3D models. We divide a soccer scene into dynamic regions and static regions for reconstructing simple 3D models in each region. For virtual

Proceedings of the International Conference on Computer Graphics, Imaging and Visualisation (CGIV'06)
0-7695-2606-3/06 $20.00

viewpoint synthesis of dynamic regions, which are a ball and players, we employ two important ideas which are the billboard representation and the choice of optimal cameras for rendering the billboard. The choice of optimal cameras can choose the optimal camera has the most similar texture at virtual viewpoint. The billboard representation can realize a simple 3D reconstruction and transform textures of the optimal cameras into the appropriate appearance of the aspect of dynamic regions at virtual viewpoint. For virtual viewpoint synthesis of static regions, which are a ground region and a background region, we apply a homography in each region. By applying our method to a soccer scene, we can synthesize free-viewpoint images by employing only pixel values of real images.

2 The proposed method

2.1 Video capturing environment

The four cameras are located in an half of the soccer field like Figure 1. Under such a video capturing environment, we record a soccer match and use this multiple videos as input images. All input images are 720x480 pixels, 24-bit-RGB color images. The examples are listed in Figure 2.

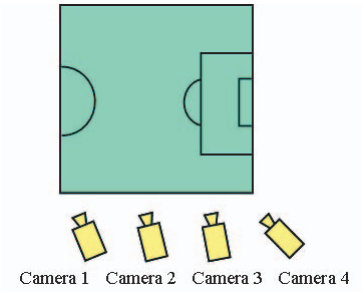


Figure 1: Video capturing environment

2.2 Calibration

For the improvement of the generally method[1], we need to get camera parameters. In previous calibration method, we need to put marker objects with known 3D shape information. However, it takes a lot of trouble with putting such marker objects. So we perform camera calibration without such marker objects.

First, we can get a focal length which is a component of intrinsic parameters without such objects but using vanishing points made from intersections of the white markers like Figure 4 [12].

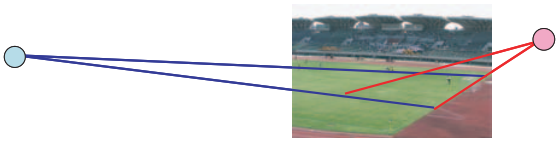
Next, we compute extrinsic parameters by makeing use of two important processes. One is getting a homography between input images and a virtual top-view image like Figure 7(a). The other is getting rotation matrices of input cameras.

A homograpy $\boldsymbol{H}$ is a planer transformation. $\boldsymbol{H}$ can be computed by at least four correspondences between image

(a) camera1 (b) camera2

(c) camera3 (d) camera4

Figure 2: input images



, : Vanishing points

Figure 4: The used calibration method

I_1 and image I_2 like the equation (1), where x_1, x_2 are homogenous coordinates on the image I_1, I_2

$$\boldsymbol{x_2} \simeq \boldsymbol{H}\boldsymbol{x_1} \tag{1}$$

Because the width of the soccer field is previously determined, we can get coordinates of the feature points which are intersections of the white markers on the soccer field. On the othre hand, we can manually obtain coodinates of the feature points in input images. Therefore, we can obtain a homography between an input image and a virtual top-view image which has the same coordinate system as the soccer field, by a number of corresponding which are the feature points.

There is a nature of a rotation matrix $\boldsymbol{R}$ of a projective matrix like the equation (2),(3).

$$\boldsymbol{R} = \left[\begin{array}{ccc} \boldsymbol{r_1} & \boldsymbol{r_2} & \boldsymbol{r_3} \end{array} \right] \tag{2}$$

$$\boldsymbol{r_1} \times \boldsymbol{r_2} = \boldsymbol{r_3} \tag{3}$$

We get a rotation matrix by using this nature and this homography between an input image and a virtual top-view image. The homography has information about a translation matrix. So we can get extrinsic parameters by a matrix theory.

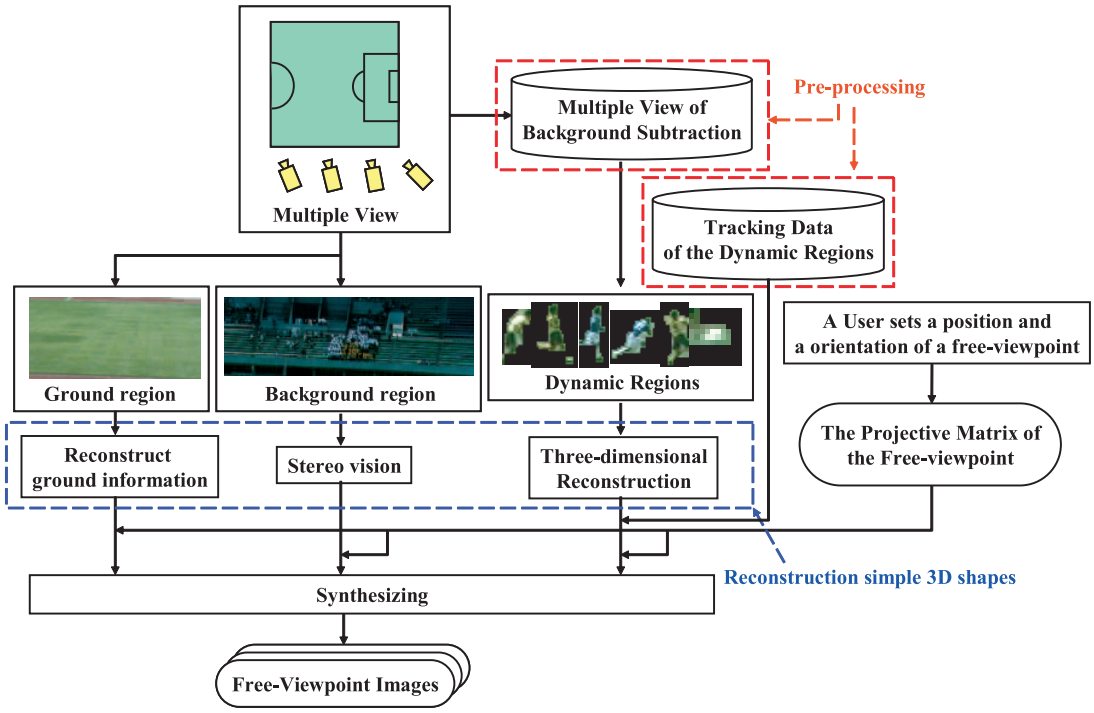


Figure 3: The flowchart of the proposed method

Finally, we can get camera parameters because we get intrinsic parameters and extrinsic parameters.

2.3 The overview of the proposed method

We describe a overview in our method. The flowchart is shown in Figure 3. First, we compute a projective matrix at the free-viewpoint by determining a position and an orientation of the free-viewpoint, and fixing the focal lengths. Second, we manually divide input images into dynamic regions and static regions. The dynamic regions include player's textures and a ball texture. The static regions include the ground region and the background region. Next, after recovering a simple 3D shape information in each region process, we synthesize each region of the free-viewpoint image with making use of the projective matrix of the free-viewpoint. For virtual viewpoint synthesis of dynamic regions, we employ two important ideas which are the billboard representation of the dynamic regions and the choice of optimal cameras for rendering the billboard. The choice of optimal cameras can choose the optimal camera has the most similar texture at virtual viewpoint. The billboard representation can realize a simple 3D reconstruction and transform textures into the appropriate appearance of the aspect of dynamic regions at virtual viewpoint. Finally, three rendered regions are merged and the free-viewpoint images are synthesized.

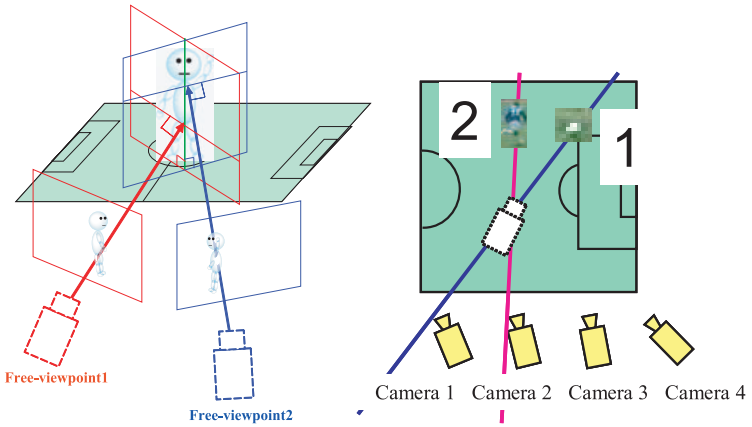
2.4 View synthesis for dynamic regions

In this process, we need coordinates of the dynamic regions which are players and a ball in the input images. We construct a database that includes such the coordinates. Since the primary purpose is not player tracking but new viewpoint synthesis, we track the dynamic region almost manually, but automatic tracking methods such as [13] also can be applied.

Moreover we perform background subtraction beforehand so as to get only the textures of dynamic regions. We construct a database that includes such the background subtraction images.

For rendering the dynamic regions, we need to consider two important issues: how to transform textures of the dynamic regions from the real camera to the virtual camera by using the billboard representation, and how to select an optimal camera for rendering the billboard.

First, we describe the method for the billboard representation that appropriately transform the textures of dynamic regions. After obtaining a player's texture by background subtraction, we transform the textures of dynamic regions by assuming the dynamic region can be represented as a billboard, which is a plane vertical to the ground with the shape of the silhouette of the dynamic region like Figure 5(a). Then, the billboard is also vertical to the orienta-



(a) The billboard representation (b) The choice of the optimal camera

Figure 5: Synthesizeing the dynamic region

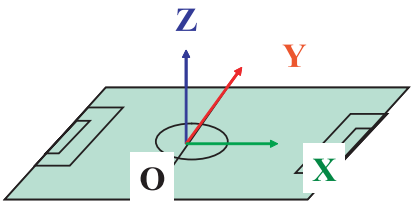
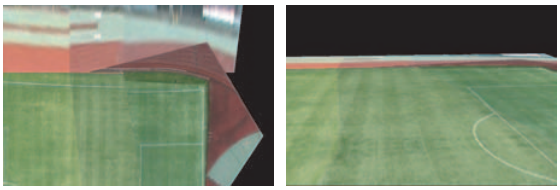
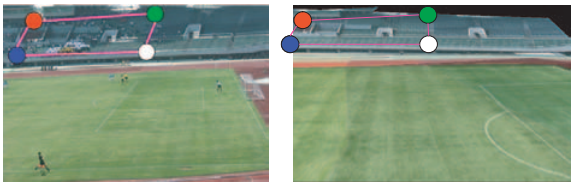


Figure 6: The world coordinate system



(a) a virtual top-view image (b) The ground region at a free-viewpoint

Figure 7: Synthesizeing the ground region



(a) camera2 (b) The background region at a free-viewpoint

Figure 8: Synthesizeing the background region

tion vector from the free-viewpoint to the dynamic region like Figure 5(a). This billboard representation can realize a simple 3D reconstruction and transform textures into the appropriate appearance of the aspect of dynamic regions at virtual viewpoint. We can synthesize the dynamic regions by using the projective matrix and the simple 3D models.

Second, we describe the method of selecting the optimal camera for rendering the billboard. We consider lines from the free-viewpoint location to the dynamic regions as shown in Figure 5(b) by using a database that includes coordinates of the dynamic regions. For rendering each dynamic region, we select the closest camera to the line connected the dynamic region with the free-viewpoint. When the dynamic region is not captured in the selected camera, we select the second nearest camera.

2.5 View synthesis for ground region

We can get the homographies between input images and a virtual top-view image in Section 2.2. Then, we apply the homographies to the ground region of the input images to get the ground image which is the virtual top-view image like Figure 7(a) because the ground region can be considered as a single plane in a soccer scene. As a result, we obtain the ground image like Figure 7(a). Finally, we can acquire the ground region of the free-viewpoint like Figure 7(b) by using the projective matrix of the free-viewpoint.

2.6 View synthesis for background region

First, we manually give four correspondences like Figure 8(a) between two input images. Second, we compute the 3D positions of the four correspondences with the triangulation method. Third, we project the four correspondences to the free-viewpoint as shown in Figure 8(b) with the projective matrix of the free-viewpoint. Because the four correspondences between an input image and a free-viewpoint image are computed, a homography can be computed. Finally, the pixel values of the background region can be inserted by this homography like Figure 8(b).

3 Experimental results

We set the world coordinate system like Figure 6. We regard the original point of the world coordinate system as the center of the soccer field. We synthesize free-viewpoint images at a lot of free-viewpoints for demonstrating the effectiveness of our method in the experiment.

3.1 Synthesizing free-viewpoint images

Based on the proposed method, we can also synthesize virtual viewpoint images at user's favorite free-viewpoints by giving the positions and the orientations of the free-viewpoints like Figure 9. The synthesized images are shown in Figure 11. The captions of Figure 11 indicate the positions of the free-viewpoints in the world coordinate system. All the virtual cameras are looking at the same point (X=20m, Y=36.9m, Z=0m) in Figure 11. We synthesize images are similar to intermediate images like Figures 11(a),11(b),11(c),11(d). We successfully synthesize at the inside of the soccer field like Figures 11(e),11(f),11(g),11(h). We also synthesize at higher positions than input images like Figures 11(i),11(j),11(k),11(l). These images can not be synthesized by the previous methods[1].

3.2 Synthesizing player-viewpoint images

If we assume that the player sees the ball, we can get a projective matrix at player-viewpoint. So we can also synthesize player-viewpoint images from multiple cameras in soccer scene. Figures 10(b), 10(c), 10(d) represent the virtual viewpoint images at the view point of the player shown in Figure 10(a). The caption of the result images indicate a focal length which is a component of the intrinsic parameters at the player-viewpoint. We can successfully synthesize player-viewpoint images at lower positions than input images. The previous method can not synthesize such player-viewpoint images, but our method can.

4 Conclusion

We propose the new method to synthesize free-viewpoint images from multiple view videos in soccer stadium. In the experimental results, we realize to synthesize free-viewpoint images of not only intermediate images but also extra-intermediate images from multiple view videos which previous methods couldn't realize in soccer stadium. In addition, we realize to freely set free-viewpoint. The textures of dynamic regions are appropriately selected from the optimal camera because we make use of our original method. The automatic tracking of the dynamic regions, making higher precision images and resolution images, and the real time rendering are the problem to resolve in future.

Acknowledgement

This work has been supported in part by a Grant in Aid for the 21st century COE for Optical and Electronic Device Technology for Access Network from the MEXT in Japan.

References

[1] N.Inamoto, H.Saito, *"Intermediate View Generation of Soccer Scene from Multiple Views,"* ICPR2002, Vol.2, pp.713-716, 2002.

[2] CyberPlay, URL: http://www.orad.co.il.

[3] I.Kitahara, Y.Ohta, H.Saito, S.Akimichi, T.Ono, T.Kanade, *"Recording Multiple Videos in a Large-scale Space for Largescale Virtualized Reality,"* Proc. Of International Display Workshops (AD/IDW'01), pp.1377-1380, 2001.

[4] D.Snow, O.Ozier, P.A.Viola, W.E.L.Grimson, *"Variable Viewpoint Reality,"* NTT R&D, Vol.49, pp.383-388, 2000.

[5] 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.

[6] S. Moezzi, L.C.Tai, P.Gerard, *"virtual View Generation for 3D Digital Video,"* IEEE Multimedia, Vol 4, pp.18-26, 1997.

[7] 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, Vol 14, pp.370-380, 2004.

[8] J. Starck, A. Hilton, *"Model-based multiple view reconstruction of people,"* IEEE International Conference on Computer Vision, pp.915-922, 2003.

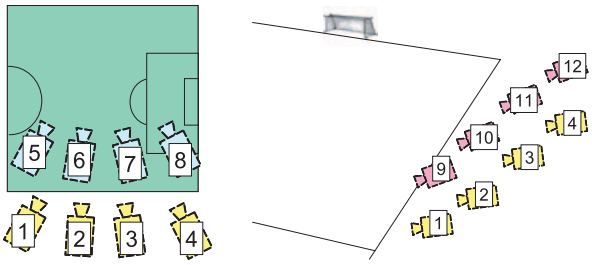
[9] T.Kanade, P.J.Narayanan, P.W.Rander, *"Virtualized reality:concepts and early results,"* Proc. of IEEE Workshop on Representation of Visual Scenes, pp.69-76, 1995.

[10] Y.Kameda , T.Koyama , Y.Mukaigawa , F.Yoshikawa, Y.Ohta, *"Free Viewpoint Browsing of Live Soccer Games,"* ICME2004, Vol.1, pp.747-750, 2004.

[11] K.Kimura, H.Saito, *"Video synthesis at tennis player-viewpoint from multiple view videos,"* IEEE VR2005, pp.281-282, 2005.

[12] G.Simon, A.W.Fitzgibbob, A.Zisserman, *"Markerless tracking using planar structures in the scene,"* Proc. of the International Symposium on Augmented Reality, pp.120-128, 2000.

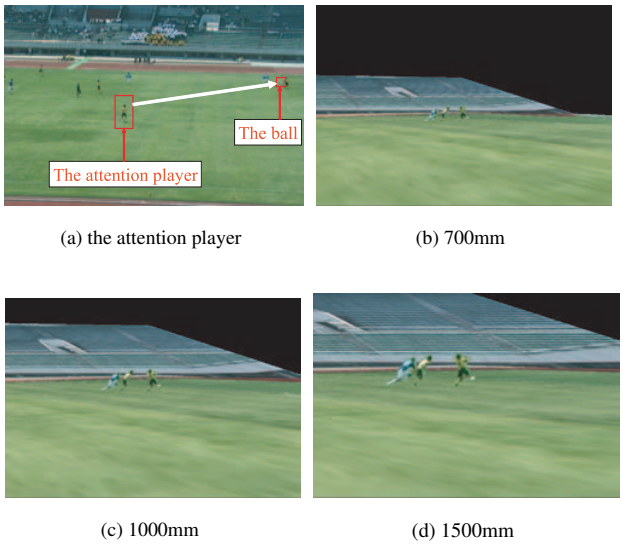
[13] T.Misu, M.Naemura, W.Zheng, Y.Izumi, K.Fukui, *"Robust tracking of soccer players based on data fusion,"* ICPR2002, Vol.1, pp.556-561, 2002.



(a) The camera locations of the free-viewpoints No.1

(b) The camera locations of the free-viewpoints No.2

Figure 9: The camera locations of the free-viewpoints



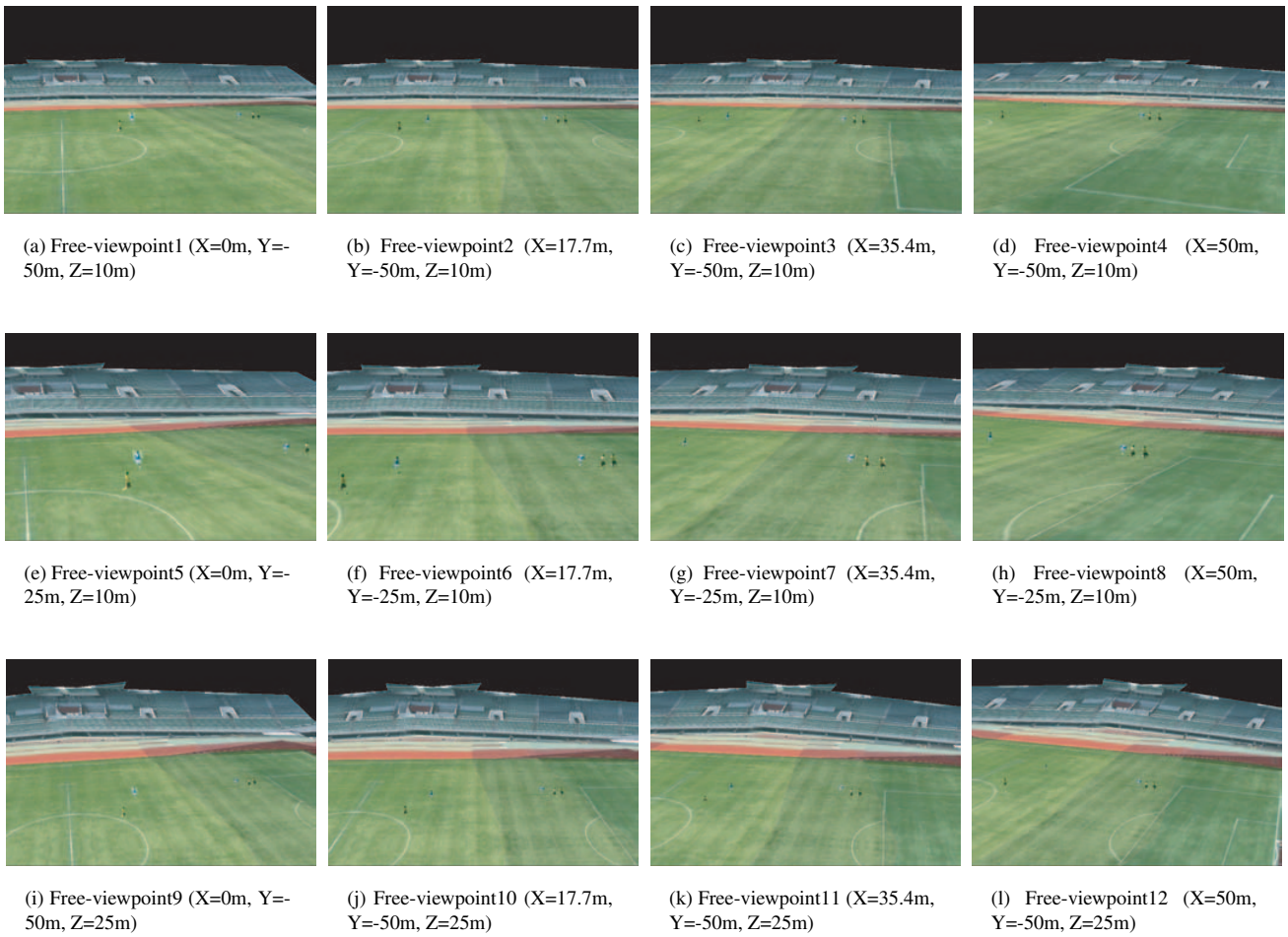
(a) the attention player

(b) 700mm

(c) 1000mm

(d) 1500mm

Figure 10: The results of the player-viewpoint images



(a) Free-viewpoint1 (X=0m, Y=-50m, Z=10m)

(b) Free-viewpoint2 (X=17.7m, Y=-50m, Z=10m)

(c) Free-viewpoint3 (X=35.4m, Y=-50m, Z=10m)

(d) Free-viewpoint4 (X=50m, Y=-50m, Z=10m)

(e) Free-viewpoint5 (X=0m, Y=-25m, Z=10m)

(f) Free-viewpoint6 (X=17.7m, Y=-25m, Z=10m)

(g) Free-viewpoint7 (X=35.4m, Y=-25m, Z=10m)

(h) Free-viewpoint8 (X=50m, Y=-25m, Z=10m)

(i) Free-viewpoint9 (X=0m, Y=-50m, Z=25m)

(j) Free-viewpoint10 (X=17.7m, Y=-50m, Z=25m)

(k) Free-viewpoint11 (X=35.4m, Y=-50m, Z=25m)

(l) Free-viewpoint12 (X=50m, Y=-50m, Z=25m)

Figure 11: The results of the free-viewpoint images