Player Viewpoint Video Synthesis Using Multiple Cameras

Kenji Kimura*, Hideo Saito†

Department of Information and Computer Science Keio University, Yokohama, Japan
*k-kimura@ozawa.ics.keio.ac.jp, †saito@ozawa.ics.keio.ac.jp

Keywords: player viewpoint, virtual view, epipolar geometry, Image-Based Rendering, multiple view

Abstract

In this paper, we propose a new method for synthesizing player viewpoint images from multiple view videos in tennis. Our method is divided into two key techniques: virtual-view synthesis and player's viewpoint estimation. In the former, we divide the object tennis scene into sub regions, which are background, tennis court ground, players, and ball. For each sub region, a virtual viewpoint image is synthesized by considering the geometrical condition of each region. Then the virtual viewpoint images of the sub regions are merged into a virtual viewpoint image of whole scene. In virtual viewpoint image generating, "view interpolation", which is a technique of synthesizing images in an intermediate viewpoint from the real images in two viewpoints, restricts a viewpoint position between two viewpoints. To avoid this restriction, we propose a new method of the ability to set up a viewpoint position freely. In the latter,viewpoint position is computed from the center of gravity of a player region in captured videos using epipolar geometry. By applying the computed player's viewpoint to the former method, we can synthesize player viewpoint images. Experimental results demonstrate that the proposed method can successfully provide tennis player's viewpoint video from multiple view video images.

1 Introduction

New visual effects have recently been given to videos in broadcasts such as a sport. For example, there is a system[5] which makes it possible to watch a tennis match with a free viewpoint by showing CG images of the locus of the position of the tracked tennis ball. "Eye Vision" system is also an example of the sysytem that gives the special image effect by changing continuously 30 or more cameras which capture the same scene in the Super Bowl broadcasting of American football. On the other hand, in the field of Computer Vision, there are researches[4] which aim at generating a virtual viewpoint image from the multiple view videos in a sport.

The technique of generating a virtual viewpoint image from two or more camera images is called Image Based Rendering (IBR). IBR can be categorized into "Model Based Approach" and "Transfer Based Approach". In the former[4, 3], a 3D shape model is reconstructed and a virtual viewpoint image is generated by texture mapping to its surface.

In the latter, a virtual viewpoint image is synthesized by transferring of corresponding between real view images. View Interpolation[1], View Morphing[9], etc. are the examples of this approach. The pixel-wise correspondence between real view images is required for this approach. In Transfer Based Approach, there is a technique[8] which reconstructs a 3D model of an object in order to estimate these correspond automatically from multiple view. Zitnick[11] also has succeeded in generation of a high-quality virtual view of dynamic scenes by using segmentation-based stereo approach. This method segments the image into small regions to the stereo computation. These methods generate the high-quality virtual view by assuming dence camera arrangement indoors. The drawbacks of such research is that it is necessary to perform strong calibration, which estimates the projective relation between world coordinate system and camera coordinate system. Therefore, it becomes more difficult as the target space becomes large. In order to reduce such difficulty in strong calibration of cameras, the technique[10] of generating a virtual view is also proposed using the 3D model estimated from the weak calibration, which estimates only the epipolar geometry between multiple cameras.

R.J.Radke[7] proposed the method of estimating correspondence points to a large-scale space in applying the Transfe Based Approach. In this method, which is applicable to the rotating source cameras, the correspondence points are estimated by exploiting temporal continuity and suitably constraining the correspondences. On the other hand, Inamoto[2] also has succeeded in generating a virtual viewpoint image, by applying IBR based on this Transfer Based Approach to the whole soccer scene. In its method, after classifying a soccer scene into dynamic regions and static regions, the appropriate projective transformation is applied respectively. This is easily applicable also to the

sport scene in large-scale space like soccer, since it is only required to perform weak calibration, which estimates a projective geometry between multiple views. This method is based on dividing a scene according to the geometric character, generating a virtual view image for every domain, and merging them. By taking such scene dividing strategy, it has succeeded in generating a virtual view image of the whole soccer scene.Inamoto's method[2], which assumes the fixed cameras, has succeeded in generating a more high-quality virtual viewpoint image than Radke's method[7].

In this paper, we extend the Inamoto's method[2] for tennis, and propose the technique for generating a virtual view image at tennis player viewpoint. In our method, the position of tennis player is estimated with multiple cameras, and a virtual view image in this player position is generated. Finally, by applying our method to the multiple views in tennis, we show the availability of this.

In Inamoto's method[2], a virtual view image is generated based on View Interpolation[1] that is one of the techniques of the Transfer Based Approach. This technique has the drawback that viewpoint position is restricted between two viewpoints in order to set up it by the relative weights to them. This implies that a virtual view at an estimated player's position cannot be generated only by applying this to tennis as it is. For eliminating such restriction, we proposed the new technique, which is able to give a viewpoint position freely and set up it directly without doing by relative weights. In this method, a virtual view image is synthesized by using the F-Matrix between the reference viewpoint and a virtual viewpoint given by the user. Based on the way for specifying the virtual viewpoint, the virtual view image can be generated at the position of the tennis player. In addition, in our method, the standard point on tennis court is used in order to estimate F-Matrix between reference viewpoint and virtual viewpoint.

Player-viewpoint image under a game is one of visual effects, by virtual view generating system from multiple views in sport scene. However, there are no existing methods, which enable generation of player-viewpoint image actually. Generating player-viewpoint image from the real image is one of the features in our research.

2 Geometrical Transforms

2.1 Projection Matrix

In camera model as shown in Fig.1, the relation between $\mathbf{X}$ which is the coordinate of a point in world coordinate system and $\mathbf{m}$ which is homogenous coordinate of it on the image is defined by projection matrix $\mathbf{P}$.

$$\tilde{\mathbf{m}} = \mathbf{P}\tilde{\mathbf{X}_\mathbf{w}} \tag{1}$$

$$\mathbf{P} = \begin{pmatrix} f & 0 & u_0 \\ 0 & f & v_0 \\ 0 & 0 & 1 \end{pmatrix} [\mathbf{R}|\mathbf{T}] \tag{2}$$

where, f is focal length, (u_0, v_0) are the coordinates of image center, $\mathbf{R}$ is rotation matrix, and $\mathbf{T}$ is translation matrix.

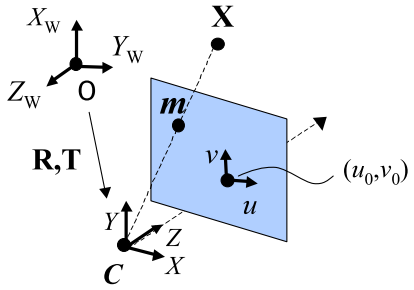


Figure 1: Camera model

2.2 F-Matrix

Epipolar geometry can define the relative position and posture between two cameras. As shown in Fig.2, a 2D line drawn by projecting a 3D line including 3D point $\mathbf{X}$ and $\mathbf{m}$ on View 2 includes the matching point on View 2 for $\mathbf{m}$, which is a point on View 1. This 2D line is called epipolar line, which restricts search area for matching of point. An Epipolar line $\mathbf{l}'$ is estimated as below.

$$\mathbf{l}' = \mathbf{F}\tilde{\mathbf{m}} \tag{3}$$

where, F is called Fundamental Matrix, which is computed by several known matching points.

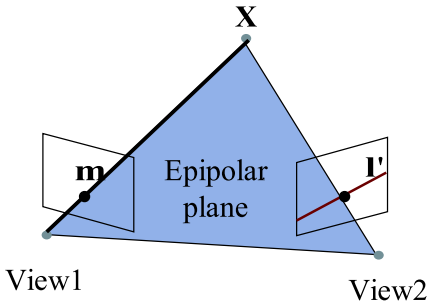


Figure 2: Epipolar geometry

2.3 homography

For a point on a plane in 3D space, a unique matrix $\mathbf{H}$ to the plane gives one to one matching relation between two views. (Fig.3).

$$\tilde{\mathbf{m}}' = \mathbf{H}\tilde{\mathbf{m}} \tag{4}$$

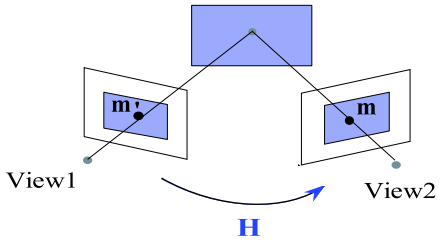


Figure 3: homography

where, each of $\mathbf{m}$ and $\mathbf{m}'$ is one of a pair of matching point between two views. This matrix $\mathbf{H}$ is called homography, which is computed from known point correspondences on the plane between two views.

3 Proposed Method

The flowchart of the proposed method is shown in Fig.4. In our method, after dividing a tennis scene into dynamic regions (player and ball) and static regions (court and background), we synthesize a virtual view for every domain. By merging them, we can generate a virtual view image of the whole tennis scene. In our synthesis process, two reference views near a virtual viewpoint position are chosen from three input views. In addition, we generate a player viewpoint image, by given a player viewpoint position which is separately estimated from player-tracking cameras, as input of our virtual view method.

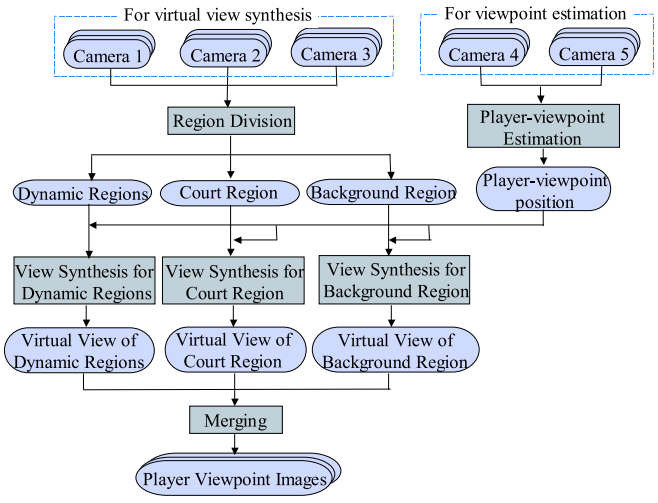


Figure 4: Flowchart of proposed method

3.1 View Synthesis for Dynamic Region

In dynamic region, we generate a virtual view by the synthesis method of Transfer Based Approach. View Interpolation[1], which is a technique of synthesizing images in an intermediate viewpoint from the real images in two viewpoints, is one of Transfer-Based methods. In View Interpolation, virtual view position is given by the relative weights to two viewpoints, and the correspondence points between reference images are interpolated by this weighting factor, so the positions of corresponding points in the virtual view can be estimated. These positions are estimated to all over the correspondence pixels between the reference images, and then virtual view is generated by given a color of the reference images to this positions.

As shown in Fig.5, View Interpolation has the drawback that the position of viewpoint is restricted between two viewpoints in order to set up the virtual viewpoint by the relative weights to them. On the other hand, in our method, the position and posture of a virtual view should be able to be set up according to the estimated player position, rather than the relative weights to the reference images. Therefore we come up with a method for synthesizing a virtual viewpoint image by F-Matrix between virtual view and reference view, which is estimated to a set up virtual view position.

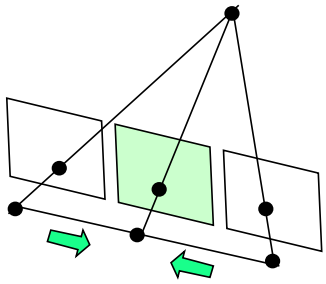


Figure 5: Transfer of corresponding points in View Interpolation

As shown in Fig.6, giving two F-Matrices between the virtual view and two reference views, two epipolar lines are drawn on the virtual view from a pair of correspondence points on the reference views. Then the two epipolar lines have an intersection point, which is also corresponding to the point pair on the reference views. In this way, the two F-Matrices give mapping of corresponding points between two reference views onto the virtual view by giving the color of all correspondence points on the reference views to these intersecting pixels on the virtual view.

As above, F-Matrices between virtual view and reference view and correspondence maps between the reference images are needed, in order to generate a virtual view for dynamic region. The F-Matrices are estimated using the standard parts of tennis court. The correspondence maps are estimated each domain (player and ball). Details of the estimation methods of these are given as follows.

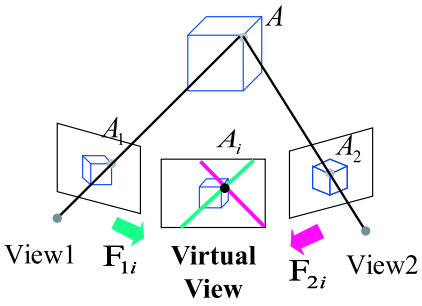


Figure 6: Transfer of corresponding points in the proposed method

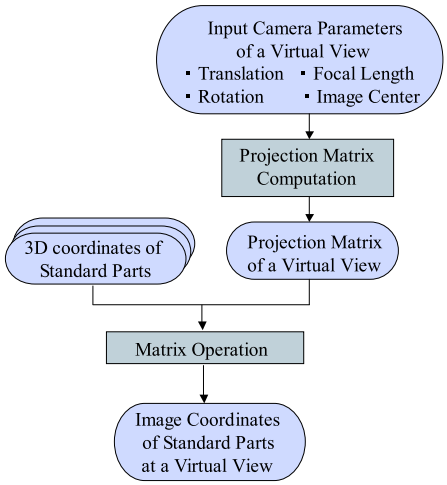


Figure 7: Estimation of image coordinates of standard parts at a virtual view

3.1.1 Estimation of F-Matrix between Reference View and Virtual View

Giving the position and posture of a virtual view, $\mathbf{R}$ and $\mathbf{T}$ of the equation (2) are determined. A focal length f of a virtual view is given arbitrarily. By this process, a projection matrix of a virtual view is computed. On the other hand, we assume that 3D positions of some points on the standard parts of tennis court (Fig.8) are known. Then, the coordinates on a virtual view of these points are estimated using the projection matrix of a virtual view. Meanwhile, we estimate the correspondences of those points on the standard parts between virtual view and reference view, by giving the positions of them on the reference views manually. By these correspondences, the F-Matrix between a virtual view and the reference view is estimated. Our method enables the viewer to control back/forth, left/right, up/down and pan/tilt/roll of a virtual camera, by giving the position and posture of the virtual view directly. The flow of view synthesis for dynamic domain is summarized in Fig.9.

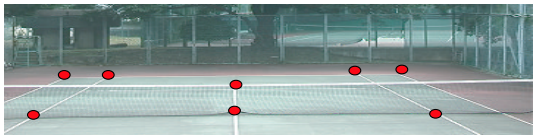


Figure 8: Standard parts on tennis court

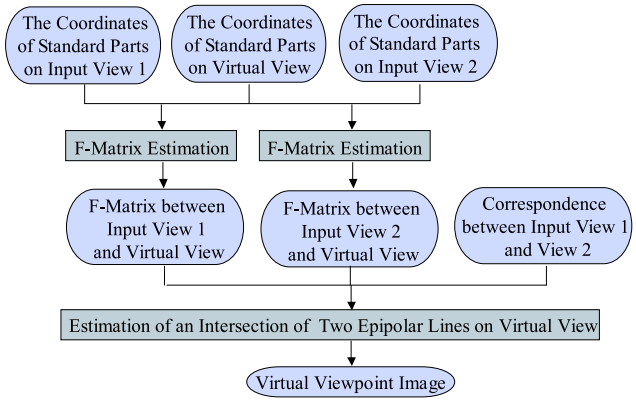


Figure 9: Flow of processing for dynamic domain

3.1.2 Estimation of Pixel Correspondence for Dynamic Region

The dynamic regions of tennis scene include player region and ball region. In these regions, correspondence map pixel by pixel every frames, and virtual view is also synthesized every frame.

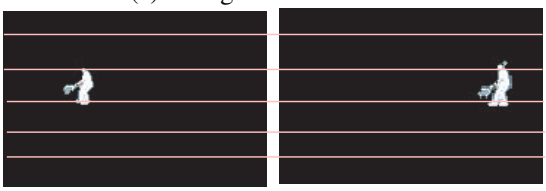
[Correspondence for Player Region]

In order to extract player region, we obtain a player's silhouettes by background subtraction. Next, a pair of the silhouette images is rectified so that epipolar lines are horizontal scan lines. As shown in Fig.10, after matching a person edges on the scan lines [6], correspondences inside of the silhouettes is estimated pixel by pixel by performing linear interpolation to edge correspondences.

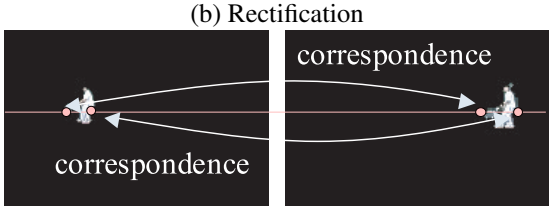
[Correspondence for Ball Region]

In order to extract ball region, we perform frame subtractions between previous-current frames and between current-next frames respectively. We can obtain the silhouette image, which doesn't include static regions, by performing AND operator to two subtraction images obtained by this process. However, regions other than a ball region still remain in this image. Therefore, in reference to color pixel value of the silhouette in reference images, the rigion with the closest color to the ball (yellow) is extracted as a ball region. Finally, the center of gravity of a ball domain is computed, and they are correspondence points in a ball region (Fig.11).

(a) Background subtraction

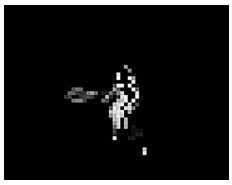


(b) Rectification

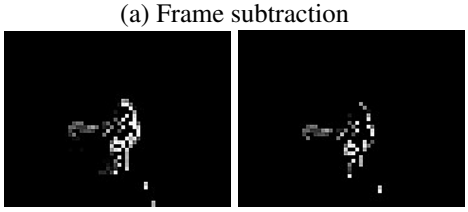


(c) Matching using person edges

Figure 10: Matching player domain



(a) Frame subtraction



(b) AND operator

(c) Extraction

Figure 11: Extraction of ball region

3.2 View Synthesis for Static Region

In static region (Fig.12), a virtual view is synthesized using the standard parts of tennis court.

First, in court and net region, the positions of these vertices on a virtual view are estimated by the same method as **3.1, 3.1.1**, because the 3D positions of their standard parts of tennis court are known. The positions of these points on the reference view are extracted manually. Because court and net regions are planar, we can synthesize virtual views of them by transferring correspondences to virtual view using homography, which is estimated from the correspondences between virtual view and reference view Fig.13). In the net region, the homography is applied to only white belt parts except the reticulation part for synthesizing virtual views. In addition, we processed the reticulation part by the method of lowering the brightness of the back pixels seen through the reticulation.

On the other hand, a virtual view of background region can be synthesized using an estimated homography in the same way, by assuming that background exists far away so that it can consider that a background region is also a plane.

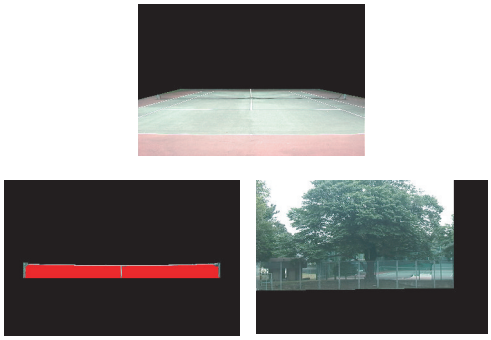


Figure 12: Divided static regions

3.3 Estimation of Player Viewpoint Position

A player viewpoint position is estimated from the images captured with two player-tracking cameras. In our research, all camera are not strongly calibrated. This is because it is very troublesome and difficult to perform strong calibration of cameras in large-scale space like a tennis court. In our method, player viewpoint positions are estimated using only projective relationship between captured images. Meanwhile, it is difficult to track the correct direction of the player's gaze, and the 3D position of player's eyes with player-tracking cameras. Consequently, under the assumption that a player's eyes are in fixed height to a tennis court surface and always turn player's gaze in the direction of an opponent, the player viewpoint position can be estimated from images captured with player-tracking cameras.

Based on such discussion, we estimate a player viewpoint position as follows. First, the player positions on the ground plane are estimated after extracting the player region by background subtraction. They are estimated using homography between the images captured with the player-tracking camera and the ground plane image, which is computed from correspondences in the standard parts on tennis court between the former and the latter. Finally, the player viewpoint position is determined as the 3D position of

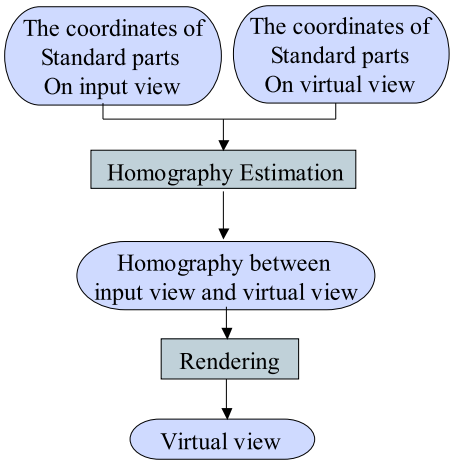


Figure 13: Flow of processing for static domain

the estimated player's position on the ground plane with the pre-determined constant value of player's height.

Next, we discuss how to estimate the player positions on the ground plane in detail. Assuming simply that these input image are the bottom point of the domain extracted by background subtraction, the following problems occur. (1) The error generated when a part of leg is not able to be extracted in background subtraction. (2) The error generated when the player has jumped. In addition, since the leg sides detected with each frame may differ respectively, it may be tracked as a direction contrary to the actual direction where the player is moving between frames. On the other hand, even if we take the method of detecting both legs and setting the average of two bottom points of them to these in input image, a racket may be incorrectly recognized to be a leg. Thus, in our method, the player positions on the ground plane are estimated from the center of gravity of a player with two player-tracking cameras, which is considered to be the element estimated most stably in a player domain.

As shown in Fig.14, we assume that player-tracking camera capture a player so that the axis of the player is almost parallel to the vertical axis of the image plane. This assumption is equivalent to player's standing vertically to the ground plane and player-tracking camera's being parallel to the ground plane, so it is realistic enough in actual captured images. By this assumption, we can consider that player positions on the ground plane are on the vertical line from the center of gravity of a player, and these positions in two player-tracking cameras are correspondence points between these cameras. Therefore, as shown in Fig.15, a vertical line is taken down from the center of gravity on the image, and this line is projected on a ground plane image using homography between a player-tracking camera and the ground plane image (Fig.16). The intersection of these line estimated in two player-tracking cameras gives a player position on the ground plane.

In this technique, the influence to the above-mentioned case of (1) is small. This is because the position of the center of gravity hardly changes though the shape of the extracted silhouette changes somewhat. Moreover, it is not influenced to the case of (2), because it can estimate the position which is in contact with the ground from the center of gravity directly.

As mentioned above, by this technique, an actual position of the player is not estimated since it assumes that a player's eyes are in fixed height to a tennis court surface. However, our method realizes the much smoother motion in player viewpoint video than the method which estimates a player viewpoint including height element directly.

Figure 14: An input image that fills assumption

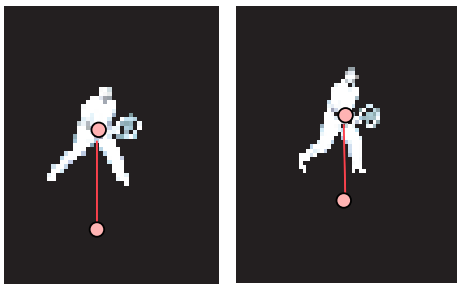


Figure 15: Vertical line from the center of gravity

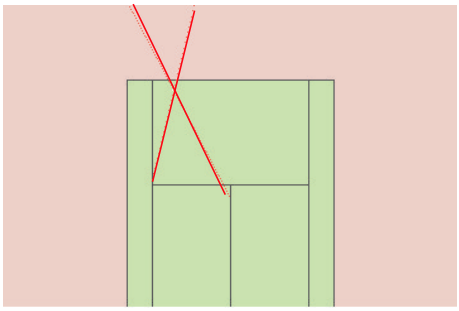


Figure 16: Intersection of the line projected on ground plane

4 Experimental Results

We performed experiments to generate a player viewpoint image from multiple cameras, which capture tennis scene. As shown in Fig.17, three cameras for virtual view synthesis and

two cameras for player viewpoint estimation were set up, and a static image which captures the court/background was also used. The former cameras were set up to capture a opponent and the background/court on the opponent side. On the other hand, the latter cameras were set up to capture a player who gives virtual viewpoints and court on his side. (It is not necessary to capture the background.)

The captured videos were converted to BMP format image sequences, which were composed of 720/480 pixels, 24-bit-RGB color images. The example input image are shown in Fig.18.

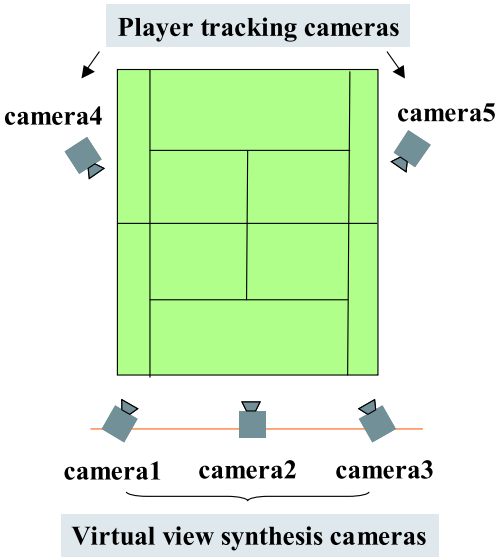
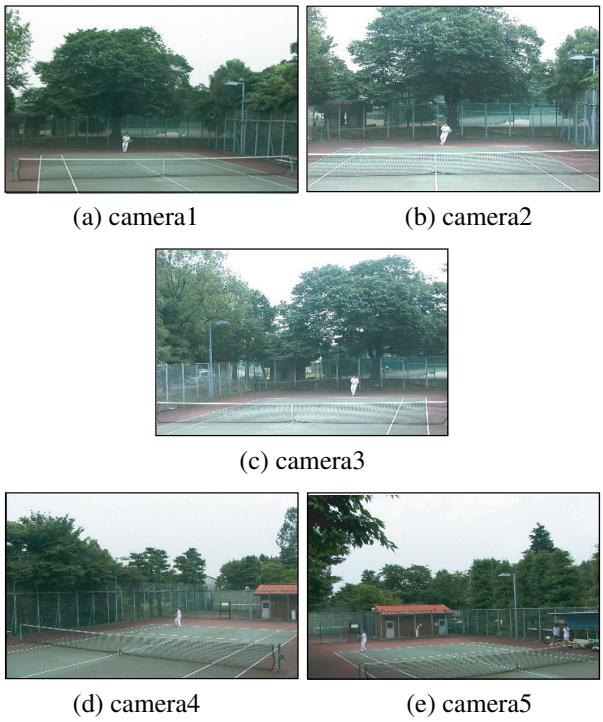


Figure 17: Locations of cameras

First, we verify the result of the player viewpoint estimation (Fig.19). As shown in Fig.19(b), detected motion of the player is different from actual motion that can be observed in the video sequence. In this scene, the player is moving to the diagonal right from the player. In addition, estimated player's trajectory seems ramdom. This is because that detecteted side of the leg is randomly switched in frames. On the other hand, in our method, the trajectory according to the player's motion can be obtained by estimating a player position using the center of gravity of a player domain.

Next, we verify the result of the virtual viewpoint image generated by our method. For comparison, the virtual viewpoint images generated using View Morphing[2], which is the technique of extending View Interpolation are shown in Fig.20. View Morphing is the technique of synthesizing images after rectifying the reference images as a preprocessing, in order to avoid the defect of generating of a geometric distortion in View Interpolation. However, View Morphing is essentially equivalent to View Interpolation, and a virtual viewpoint positions in this method are specified by the relative weight between reference views. As shown in Fig.20, you could see that View Morphing makes it difficult to generate the player viewpoint image, since a virtual viewpoint position can be



(a) camera1 (b) camera2

(c) camera3

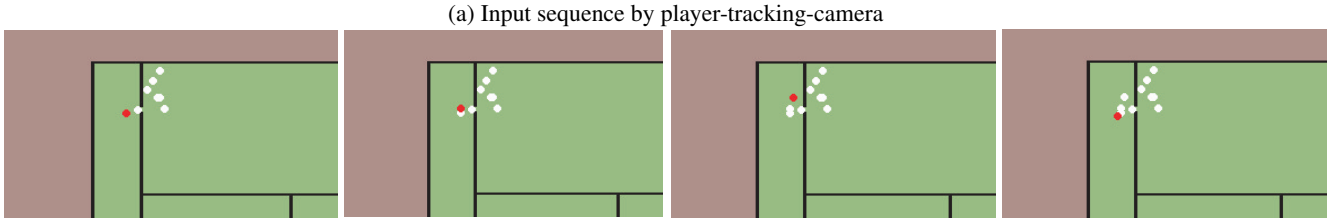
(d) camera4 (e) camera5

Figure 18: Input images

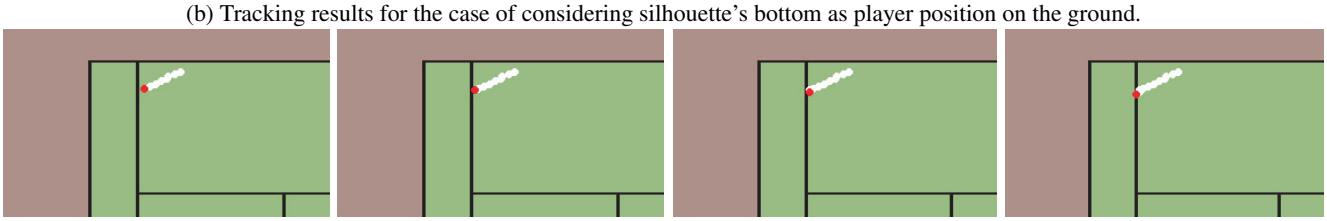
specified only on the straight line which connected the input view position. On the other hand, as shown in Fig.21, which are results generated the player viewpoint images using our method, we can see the virtual view in the viewpoint of the player on the court. Furthermore, the virtual viewpoint images in the various viewpoints on a court, which are generated by our method, is shown in Fig.22. Thus, we can confirm that the virtual viewpoint images in the free position on the court are generable by our technique. Finally, the virtual views generated by view morphing and the proposed method are arranged to Fig.23 to compare the qualities of these views. The proposed method is not inferior to view morphing in image quality, and has generated the virtual view with high quality.

Meanwhile, our method is implemented as a system that processes the recorded video images. This method requires the processing time 3 seconds per frame. However, in off-line, it is needed only to give the correspondence points between input images. So we think that there is a possibility that this technique becomes a real-time system enough.

(a) Input sequence by player-tracking-camera



(b) Tracking results for the case of considering silhouette's bottom as player position on the ground.

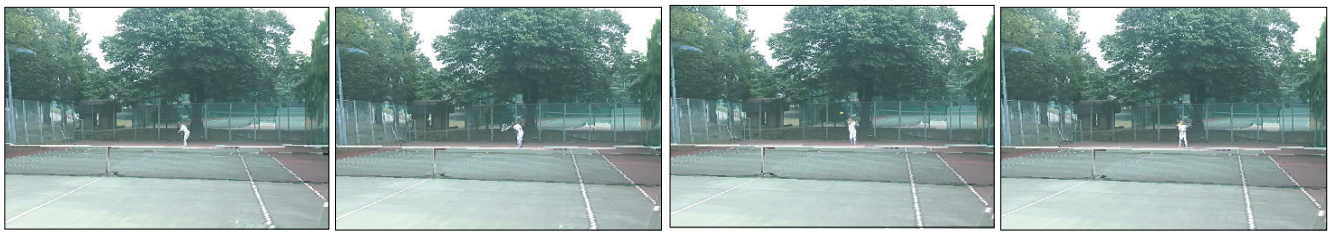


(c) Tracking results for the case of estimating player position on the ground using the center of gravity.

Figure 19: Results of player-viewpoint estimation (white dot: past trajectory, black dot: current position)

(a) 1:9 (b) 4:6 (c) 6:4 (d) 9:1

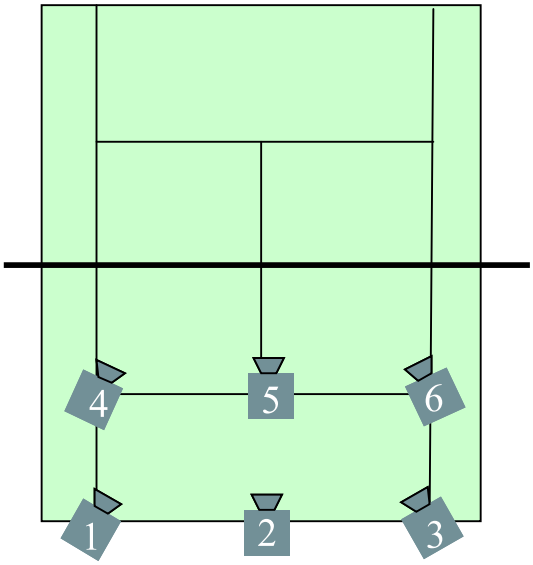
Figure 20: Virtual views using View Morphing (The above-mentioned ratios show the relative positions of the virtual viewpoint to input images 1 and 3)



(a) Player viewpoint images generating by our method

(b) The image which expanded near the player

Figure 21: Player-viewpoint images using our technique



(a) virtual view positions set up

virtual view 4

virtual view 5

virtual view 6

virtual view 1

virtual view 2

virtual view 3

(b) virtual views in these positions

Figure 22: Virtual viewpoint images using our method

(a) view morphing

(b) proposed method

Figure 23: Comparison of image qualities of view morphing and proposed method

5 Conclusion

We generate a player viewpoint image from multiple cameras, which capture tennis scene. In this paper, we propose a novel synthesis method, which can set up virtual viewpoint position freely, to avoid restriction of it in View Interpolation. In addition, we also proposed an efficient and robust method of estimating player viewpoint position using projective geometry. It made it possible to obtain the stable results by estimating from the center of gravity, which is the stable element in a player domain.

This experiment is conducted in the minimum number of cameras required for our method to apply. Therefore, it is a future subject to examine the possibility of improvement in the quality of the image by the addition of the number of camera.

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] S.E. Chen and L. Williams. View interpolation for image synthesis. In *Proceedings of SIGGRAPH '93*, pages 279–288, 1993.

[2] N. Inamoto and H. Saito. Intermediate view generation of soccer scene from multiple videos. In *Proceedings of ICPR '02*, volume 2, pages 713–716, 2002.

[3] T. Kanade, P.J. Narayanan, and P.W. Rander. Virtualised reality: Concepts and early results. In *IEEE Workshop on Representation od Visual Scenes*, pages 69–76, 1995.

[4] I. Kitahara and Y. Ohta. Scalable 3d representation for 3d video display in a large-scale space. In *IEEE Virtual Reality Conference 2003*, pages 45–52, 2003.

[5] Gopal Pingali, Agata Opalach, and Yves Jean. Ball tracking and virtual replays for innovative tennis broadcasts. In *ICPR'00*, volume 4, page 4152, 2000.

[6] S. Pollard, M. Pilu, S. Hayes, and A. Lorusso. View synthesis by trinocular edge matching and transfer. In *Image and Vision Computing 18*, pages 749–757, 2000.

[7] R.J. Radke, P.J. Ramadge, S.R. Kulkarni, and T. Echigo. Efficiently synthesizing virtual video. In *IEEE Transactions on Circuits and Systems for Video Technology*, pages 325–337, February 2003.

[8] H. Saito, S. Baba, and T. Kanade. Appearance-based virtual view generation from multicamera videos captured in the 3-d room. In *IEEE Trans. on Multimedia*, volume 5, pages 303–316, 2003.

[9] S.M. Seitz and C.R. Dyer. View morphing. In *Proceedings of SIGGRAPH '96*, pages 21–30, 1996.

[10] S. Yaguchi and H. Saito. Arbitrary viewpoint video synthesis from multiple uncalibrated cameras. In *IEEE Trans. on Systems, Man and Cybernetics, PartB*, volume 34, pages 430–439, February 2004.

[11] C.L. Zitnick, S.B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. High-quality video view interpolation using a layered representation. In *SIGGRAPH '04*, pages 600–608, 2004.