Virtual Omnidirectional Video Synthesis with Multiple Cameras for Sports Training

Mariko Isogawa, Dan Mikami, Kosuke Takahashi and Akira Kojima
NTT Media Intelligence Laboratories, 1-1 Hikarino-oka, Yokosuka, Kanagawa, Japan

Keywords: Inpainting, Virtual Omnidirectional Video.

Abstract: This paper proposes a new method to synthesize an omnidirectional video at a viewpoint inside a sports ground, with the goal of sports training. If athletes could virtually experience real games from a player's viewpoint, they might possibly be able to exhibit higher performance in an actual game. A head mounted display, which makes it possible to watch intuitive and interactive omnidirectional video from a 360-degree player's view, together with head direction tracking, leads to further enhanced effectiveness of training. However, it is difficult to put an omnidirectional camera on the field during a real game. Therefore, techniques for synthesizing an omnidirectional video at a player's viewpoint (virtual viewpoint) with the cameras outside the field are required. With this aim in mind, we propose a fast and stable omnidirectional video synthesis technique with image inpainting, which removes unwanted occluders between the virtual viewpoint and the cameras.

1 INTRODUCTION

Many athletes adopt the approach of watching videos as a type of scouting method for sports training. A particularly effective approach is to watch videos of an opponent one has never faced or one who could be considered a "difficult" opponent. The effectiveness could be even further increased by using immersive videos that would show instances in a game from a player's viewpoint, as if one were actually playing.

Watching omnidirectional video with a head mounted display (HMD) is one of the easiest ways to experience such video-based scouting. Displays of this type give users a full 360-degree view based on their head position. The higher immersion HMDs provide enhances the effectiveness of training.

To obtain omnidirectional video from a player's viewpoint, a camera inside a field is needed. However, it is difficult to keep cameras in certain positions inside a stadium during an actual game. Therefore, techniques are required for synthesizing an omnidirectional video from a player's viewpoint (virtual viewpoint) with the cameras outside the field. In this paper, we refer to an omnidirectional video from a virtual viewpoint as a "virtual omnidirectional video". Because this technique is used for scouting as an approach to sports strategy, it must provide fast and robust synthesis that does not rely on captured scenes or the positions and behavior of moving players.

Many studies have addressed these technical requirements to synthesize virtual omnidirectional video. However, with existing methods the synthesis fails when the players are overlapped(Guillemaut and Hilton, 2011), or else heavy calculation cost is incurred(Inamoto and Saito, 2007), or else many cameras are needed to obtain each and every light ray(Levoy and Hanrahan, 1996).

One possible answer to these problems is an approach based on the work done by Levoy and Hanrahan, which uses not only light rays passing through a virtual viewpoint but neighboring light rays to reduce the amount of cameras. This is a fast and stable method that does not depend on the scene and positions of players. However, the field side's appearance from the virtual viewpoint may be shielded if players are located between the virtual viewpoint and the outside cameras. Therefore it is impossible to correctly synthesize the appearance at the virtual viewpoint.

The work described in this paper solves that problem by introducing the technique of image inpainting. This technique removes and synthesizes unwanted occluders from images/videos. With this technique, even if the field side's appearance from a virtual viewpoint is occluded as a consequence of the player's position or other factors, it becomes possible to synthesize the desired appearance by removing the affected area.

The remainder of this paper is structured as fol-

Isogawa, M., Mikami, D., Takahashi, K. and Kojima, A..
Virtual Omnidirectional Video Synthesis with Multiple Cameras for Sports Training.
In *Proceedings of the 3rd International Congress on Sport Sciences Research and Technology Support (icSPORTS 2015)*, pages 271-275
ISBN: 978-989-758-159-5
Copyright © 2015 by SCITEPRESS – Science and Technology Publications, Lda. All rights reserved

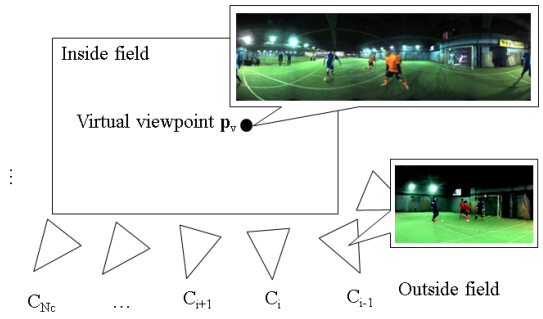

Figure 1: System configuration.

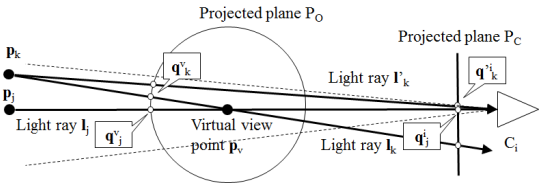

Figure 2: Optical configuration.

lows. Section 2 introduces existing methods to generate virtual omnidirectional video and completion methods. Section 3 describes our proposed method, a method to generate virtual omnidirectional video that removes unwanted occluders by image inpainting. Section 4 shows synthesized results obtained with the proposed method. Section 5 concludes the paper with a summary and a mention of future work.

2 RELATED WORK

One of the common approaches to synthesizing a virtual omnidirectional video is clipping and connecting videos captured by multiple cameras whose appearance is similar to that obtained from the virtual viewpoint. This approach has significant advantages: fast rendering is achieved because of low calculation cost, and robust synthesis is achieved that does not depend on scenes or player's positions. Thus, this technique is quite suitable as a means to our goal, i.e., scouting based on sports contents with omnidirectional video. This section first explains the pipeline of this common technique in 2.1. However, it is known that this technique involves a problem, which is described in 2.2.

2.1 Common Approach to Synthesizing a Virtual Omnidirectional Video

The configuration of the system was use is shown in Fig.1. Here, N_c is the number of cameras placed outside the field to synthesize virtual omnidirectional video at virtual viewpoint p_v. The cameras C_i are synchronized and the videos they capture include p_v.

The optical configuration is shown in Fig.2. The omnidirectional 3D plane at the center of virtual viewpoint p_v is denoted as projected plane P_O. The 2D plane the camera captures is denoted as P_C. To generate virtual omnidirectional images, an appearance obtained with l_j, a light ray that passes through p_v should be projected onto q^v_j which is the crosspoint between P_O and p_v. If light ray l_j passes through the optical center of C_i, the appearance obtained with l_j is projected at q^i_j which is at P_C and captured by C_i. The appearances of q^i_j are the same as those of q^v_j. Thus, if every light rays passes through the p_v captured by multiple cameras outside a field, virtual omnidirectional video from a virtual viewpoint p_v can be synthesized. However, in order to obtain an ideal omnidirectional image, the technique requires a large amount of cameras, i.e., an amount equal to the number of pixels in the omnidirectional image. This is not feasible from the standpoint of practicality.

On the other hand, there is a more practical technique that uses a light ray passing through the vicinity of the virtual viewpoint. In the same way as for the technique just described, light ray l_j passing through virtual viewpoint p_v is captured with camera C_i. In addition, a neighboring light ray l_j is denoted as l_k, which also passes through virtual viewpoint l_j (see Fig.2). Note that camera C_i does not capture l_k passing through p_v, but captures l'_k passing through the neighboring point of p_v and projects it at the optical center of camera C_i. We approximate light ray l_k as l'_k. That is, we use the luminance value at q'^i_k as q^v_k. This approximation enables us to reduce the amount of cameras by the number of approximated l_k.

Fig.3 shows a pipeline of the technique. First, partial region S_i, which captures the information of light ray l_j and the neighboring light ray, is clipped. Then, affine transform with affine matrix A_i is performed to S_i to generate S'_i. Here, A_i is an affine matrix from the C_i coordinate to the omnidirectional image coordinate. After that, S'_i is rendered to omnidirectional image B. Finally, blending is performed to obscure the boundaries of the pasted areas.

2.2 Technical Problem with Common Approach to Synthesizing Virtual Omnidirectional Video

This section describes a technical problem encountered with the approach described in 2.1. The technique's main advantage is that fast and robust rendering is achieved regardless of the captured scene and

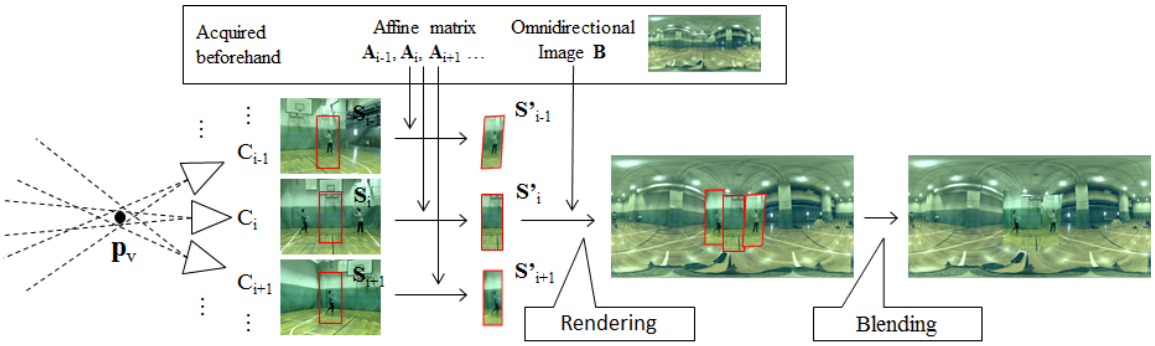

Figure 3: Pipeline with a common approach to synthesizing virtual omnidirectional video.

whether or not the players are moving. However, this advantage is not realized if the field side's appearance from the virtual viewpoint is occluded, e.g., by players located between the viewpoint and outside the cameras.

Fig.4(a) shows an image generated in the case where there are unwanted objects, such as players crossing between the virtual viewpoint and the camera, an example of which is shown in Fig.4(b). In this case, the field side's appearance from the virtual viewpoint was not captured by the cameras and an unwanted occluder that should not be visible from the virtual viewpoint was captured instead. As a result, it became impossible to synthesize the correct visual image of the virtual omnidirectional videos. This is highly disadvantageous and thus there is a strong need for a method that can solve this problem.

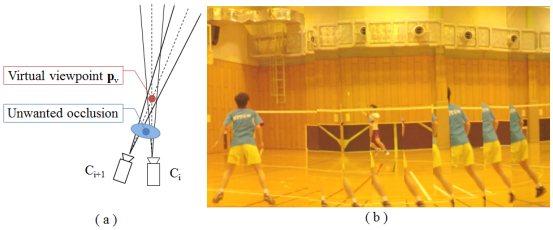

Figure 4: Problem with method given in 2.1.

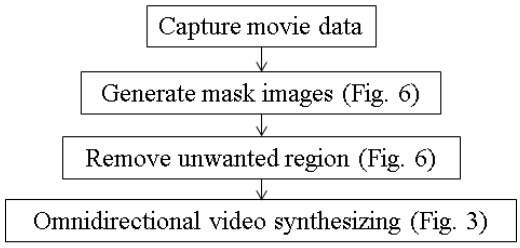

Figure 5: The pipeline of the proposed method.

3 PROPOSED METHOD

We propose a new method to generate virtual omnidirectional videos with correct visuals even if there are any occluders. To achieve this, we introduce image inpainting, which removes unwanted occluders between the virtual viewpoint and the cameras. In 3.1 we overview the method and in 3.2 we introduce a way to synthesize omnidirectional video while removing unwanted regions, which is the main contribution of this paper.

3.1 Method Overview

The pipeline of the proposed method is shown in Fig.5. Here, N_c is the number of cameras C_i set outside the field, all of them synchronized. Mask regions M_i are manually generated to indicate unwanted regions. Then, masked regions are removed by image inpainting, We explain this in detail in 3.2.

As a synthesizing technique, we use the method described in 2.1. First, partial region S_i which captures the information of light ray l_j and neighboring light rays, is clipped from the inpainted images. Then, affine transform with affine matrix A_i is performed to S_i to generate S'_i. Here, A_i is an affine matrix from the C_i coordinate to the omnidirectional image coordinate. After that, S'_i is rendered to omnidirectional image B. Finally, blending is performed to obscure the boundaries of the pasted areas.

3.2 Removing Unwanted Regions

Many image inpainting methods have been proposed that remove unwanted regions(He and Sun, 2014; Huang et al., 2014; Criminisi et al., 2004). With our proposed method, however, image inpainting has a unique configuration problem, not the same as that occurring with general inpainting methods. We will therefore describe the inpainting procedure after

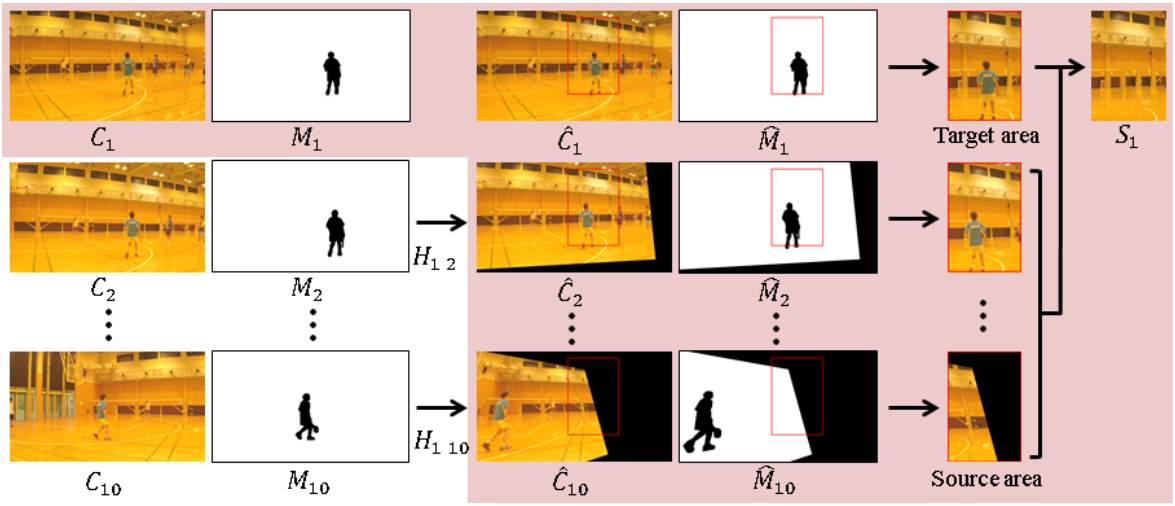

Figure 6: Inpainting procedure.

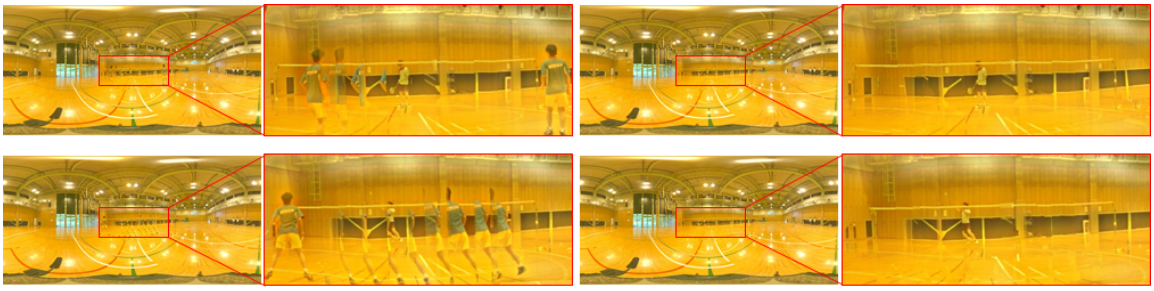

Figure 7: Synthesized virtual omnidirectional images generated by the method described in 2.1 (a) and the proposed method (b).

first describing the characteristics of the configuration problem.

With our configuration, multiple cameras are located side-by-side and synchronized with each other. This means that even if the field side's appearance from the virtual viewpoint is temporarily occluded in one of the cameras, the other cameras might capture the appearance. In that case, the occluded appearance can be synthesized by using the information obtained from the other cameras. However, the appearance will change depending on the cameras' positions, and thus camera calibration is needed. Therefore, we first generate the same viewpoint images by using homography transformation based on the relative position of each camera.

The inpainting procedure is shown in Fig.6. First, masked images M_i is manually annotated to indicate unwanted regions. To generate inpainted partial region S_i, homography transformation is performed for each camera's captured image C_j and masked image M_j to generate $\hat{C}_j$ and $\hat{M}_j$, with homography matrix H_{ij}. Here, H_{ij} is a homography matrix from the C_i coordinate to the C_j coordinate, with the latter acquired beforehand. Then, $\hat{S}_j$ was cropped from each camera $\hat{C}_j$ excepting C_i. These cropped regions were set as the source area and S_i was inpainted with the source region.

4 EXPERIMENT

This section shows synthesized results we obtained in an experiment with our proposed method. In the experiment, we set 10 GoPro cameras (1920×1080 [pixels]) outside a badminton court to synthesize an area with a 180° horizontal angle. We used a Ricoh Theta camera (1920×960 [pixels]) to obtain the background texture before the actual game. To perform the experiments we used a desktop PC of Intel Core i7 3.40GHz CPU, 32GB memory.

The inpainting method, we used was that proposed by He et al.'s, which is known to be fast and effective. Masked images were manually generated. Each homography matrix to generate $\hat{C}_j$ and $\hat{M}_j$ was calculated from 30 corresponding points.

The resulting images generated by the method described in 2.1 and the proposed method are respectively shown in Fig.7(a) and (b). In (a) a badminton player wearing a blue uniform is located between the virtual viewpoint and the cameras, which as a result produced an incorrect rendering. In contrast, in (b) the player is removed from the scene and as a result the proposed method synthesized visually correct omnidirectional images from the virtual viewpoint.

5 CONCLUSION

This paper presented a new method we propose to synthesize virtual omnidirectional video even if the field side's appearance from a virtual viewpoint is occluded as a consequence of the player's position or other factors. To remove unwanted occluders, we used image inpainting with source regions from other cameras located side-by-side. We confirmed that the proposed method works well with the contents captured in an actual sports scene.

However, it is known that there is a problem with the proposed method. In cases where the positions of captured subjects (such as players) are closer to or farther from the camera than the calibrated position, for which an affine matrix was calculated, the subjects may be rendered multiple times or simply disappear. In future work, we will attempt to devise a synthetic technique to address the problem. We also plan to investigate how effective the virtual omnidirectional videos the proposed technique produces are when applied to sports training.

REFERENCES

Criminisi, A., Perez, P., and Toyama, K. (2004). Region filling and object removal by exemplar-based inpainting. *IEEE Transactions on Image Processing*, 13(9):1200–1212.

Guillemaut, J.-Y. and Hilton, A. (2011). Joint multi-layer segmentation and reconstruction for free-viewpoint video applications. *International Journal of Computer Vision (IJCV)*, 93(1):73–100.

He, K. and Sun, J. (2014). Image completion approaches using the statistics of similar patches. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 36(12):2423–2435.

Huang, J.-B., Kang, S. B., Ahuja, N., and Kopf, J. (2014). Image completion using planar structure guidance. *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2014)*, 33(4):129:1–129:10.

Inamoto, N. and Saito, H. (2007). Virtual viewpoint replay for a soccer match by view interpolation from multiple cameras. *IEEE Transactions on Multimedia*, 9(6):1155–1166.

Levoy, M. and Hanrahan, P. (1996). Light field rendering. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 31–42. ACM.