A COMPARISON BETWEEN TWO 3D FREE-VIEWPOINT GENERATION METHODS - PLAYER-BILLBOARD AND 3D RECONSTRUCTION -

Tetsuya Shin, Nozomu Kasuya, Itaru Kitahara, Yoshinari Kameda, and Yuichi Ohta

Graduate School of Systems and Information Engineering, University of Tsukuba

ABSTRACT

This paper investigates the optimal 3D modeling solution for making free-viewpoint video in a soccer stadium. We compare a player-billboard method and a 3D reconstructing method that exploits a shape-from-silhouette approach. To examine the influence of noise and the number of cameras used to make the free-viewpoint video, we produce a CG simulation of a soccer player in action and conduct subject-based evaluation in a pair-wise comparison of the videos made by these two methods. The evaluation is made under the method of limits. From the results of the evaluation, the player-billboard method is more robust to noise than the shape-from-silhouette method, and it also better represents smooth motion parallax as the number of capturing cameras is increased, although representing parallax has traditionally been a weak point of the player-billboard method. Our results show that the player-billboard method is superior to the shape-from-silhouette method for generating 3D free-viewpoint video.

Index Terms **—** Computer vision, Computer graphics, 3D free-viewpoint video, Player-billboard, Shape-from-silhouette

1. INTRODUCTION

With the increased performance of computers and the technical development of video media, free-viewpoint video, specifically the 3D free-viewpoint video (3DFV) technique [1]-[5], is one of the most actively discussed topics in computer vision and graphics. It captures a real object with a set of video cameras and virtually represents the object's appearance from arbitrary viewpoints in a 3D world. In the related literature, the methodology of representing an object's appearance is always a major topic. We chose a soccer game for our application, so the objects to be captured were soccer players, who often run and move fast. Therefore, we needed to carefully consider the viewer's impression of player motions in selecting the best solution for the 3DFV-generation method.

The methods of representing an object can be classified into two groups: a player-billboard representation and a full-3D representation. In general, the player-billboard method is recognized for its ability to run at high-speed and its robustness against noise, but it suffers from low quality. Consequently, it has been used in the presence of large noise and when people watch the players at a distance. On the other hand, the 3D reconstructing method is known to have a large processing cost but also the ability to achieve high quality. Accordingly, it has been used when there is small noise and people watch the players at a close distance. However, these methods have not been compared quantitatively in the same environment, especially for noise and the number of cameras used, factors that actually influence the generated video quality.

In this paper, we present a user study comparing two videos, one from the player-billboard method and the other from a shape-from-silhouette method, which is one of the major 3D reconstruction methods. The comparison focuses on changes in the amount of noise and the number of cameras used for video capturing. From the experimental results, the player-billboard method achieved better scores than the shape-from-silhouette method, especially in various practical situations.

2. 3D MODELING METHODS IN FREE-VIEWPOINT VIDEO

2.1 Player-billboard method

A player billboard represents each player with a single rectangle called a "billboard," which displays a texture selected from a set of image segments taken by multiple cameras [5].

In this method, we simply map the image segment taken from a camera onto the billboard without any processing, except for warping, and thus the generated video reproduces a natural appearance of player actions. When a viewpoint is moved drastically, the image segment is changed to one taken from another camera, whose direction is now closer to that of the new viewpoint on the player. In other words, the player-billboard method sometimes has to flip through textures during large movements of the viewpoint.

One of the difficulties of this method is to accurately represent the motion parallax of the player's appearance because the shape of the player is actually represented in two dimensions. Furthermore, people may experience a strange feeling when the mapping texture is changed, thus causing a sudden change in appearance.

On the other hand, it has many good points for actual usability. Only a small amount of data is needed to represent the objects because the required data is just the player's 2D textures and positions. Furthermore, it has a lower processing cost because it does not cross reference video images and the rendering step is rather simple.

2.2 3D reconstruction method

The 3D reconstruction method represents players by using reconstructed shapes in three dimensions. Techniques of 3D reconstruction include shape-from-silhouette [6]-[7], shape-from-motion [8], and shape-from-shading [9].

In this paper, we take up soccer as our application, and thus our target scene is an outdoor event in a large-scale space. We choose the shape-from-silhouette method, which is considered robust against changes in illumination due to variations in sunlight, because it can exploit a state-of-the-art foreground detec-

978-1-4244-6378-7/10/$26.00 ©2010 IEEE

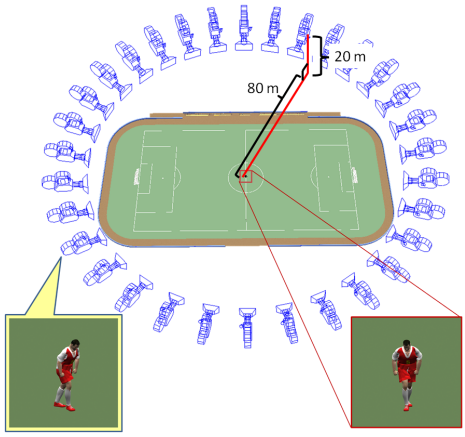

Figure 1. Placement of cameras in virtual environment

tion method to segment the foreground regions and it never needs to use intensity/color information later.

One advantage of this method is its ability to show the motion parallax of the player's appearance, since the object's shape is reconstructed in 3D. However, since it cannot reconstruct a concave shape, it is usually difficult to visualize the player precisely if there is an insufficient number of cameras. The accuracy and quality of the reconstructed 3D shape largely depend on the precision of camera calibration, which cannot be easily achieved for outdoor scenes in large-scale spaces. Furthermore, the necessary data size and processing cost with the shape-from-silhouette approach are larger than those of the player-billboard method.

3. CONFIGURATION FOR COMPARISON

To estimate the influence of noise precisely, we captured image data in a virtual environment. We controlled the amount of noise and changed the number of the cameras used up to a large number. We used 3ds Max 2010 of Autodesk to set up the virtual environment.

3.1 Virtual environment

Under the assumption that cameras are set up in the stands in a soccer stadium, we placed the cameras in the virtual environment as shown in Figure 1. The height of the camera position is 20 m and the distance from target object is 80 m. The maximum number of cameras is 32, and the size of the image regions in which the target object is captured is about 500 x 500 pixels, which in practice is nearly the finest possible resolution for taking video of players in a stadium. These conditions were determined based on our rich experience in recording soccer games in real scenes, which we have done more than 50 times.

The generic light condition in the virtual environment was set as a weak environment with non-directional light that does not cast a shadow. This simulates a cloudy day when the foreground region can be extracted precisely.

3.2 Target object and its motion

In this paper, our target object is a soccer player. Therefore, we prepared a 3D human model wearing a uniform. To be able to evaluate the quality of textures, the front side of the uniform is red, the back side is white, and both sides have yellow numbers as shown in Figure 1.

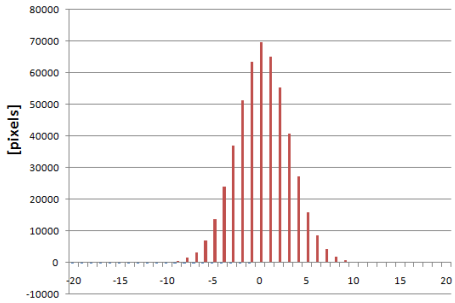

Figure 2. The histogram of white noise in actual environment

We prepared two sets of motion for the player model: active motion in a ball-handling scene and minimal motion in a dead-ball situation.

3.3 Video sequence

In order to make the comparison as fair as possible, we set the video sequence to follow a specific camera's motion pattern, although our system can actually generate images from any viewpoint. The camera motion pattern is set under the assumption that a user prefers to watch the player only, and it smoothly changes the direction to the player in order to see the action from different angles.

4. NOISE MODEL

Major factors that affect the quality of 3DFV in an actual environment are "white noise," "motion blur," "synchronization," and "calibration error." By applying these noises to the system while capturing images, we evaluated which technique was more practical.

4.1 White noise

White noise inevitably occurs when we capture images in an actual environment. To generate white noise by artificial means, we estimated the distribution of white noise in the images captured in the actual soccer environment. Figure 2 shows a histogram of the variations in RGB values of a pixel in the background regions at the size of 700 pixels by 700 pixels, from which the player and player-related areas are excluded. Since we took the images within a very short time period, the cause of all variations should be white noise. This histogram can be approximated by a Gaussian distribution, so we prepared the images for capture while including the white noise of a Gaussian distribution.

4.2 Motion blur

Motion blur occurs due to movement of objects within a camera's exposure time. In the virtual environment, we simulated motion blur by overlying many non-blurred images that were imaged virtually at a very short time interval within the exposure time, since 3ds max cannot produce ideal motion blur. We assumed that a normal camera operates at 30 fps, and here we first virtually captured images at 900 fps, 30 times faster. Then, For example, if the shutter speed is set to 1/60 seconds, a motion blur image is generated by overlying 15 images around the target frame, since 1/60 seconds can include 15 images at 900 fps.

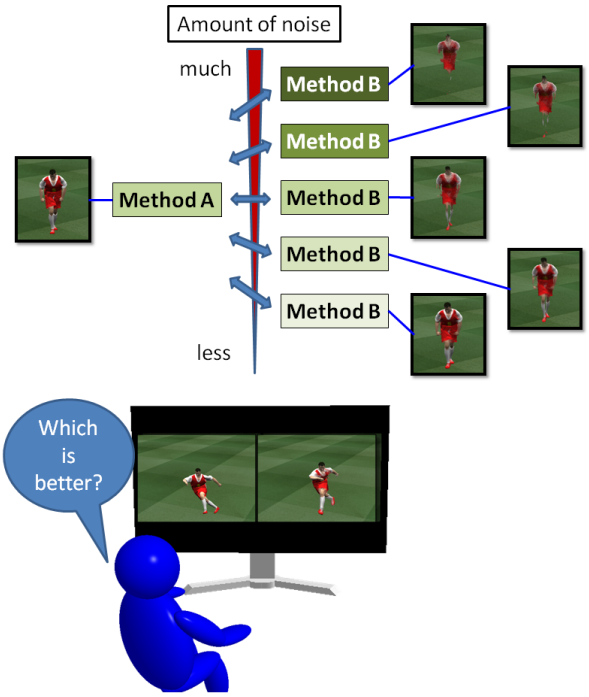

Figure 3. Experiment procedure

4.3 Synchronization

It is generally difficult to install a synchronization mechanism for cameras in a large-scale space. Therefore, the gaps in the shutter timing between the cameras are determined by the power-on time of the cameras. The power-on time of the cameras occurs randomly when the system is actually set up, so we assume that this has a uniform distribution.

Consequently, we simulate the synchronization problem by choosing the starting frame randomly in 900-fps images. The chosen frames are set to be the same for generating both of the videos for comparison.

4.4 Calibration error

Camera calibration is necessary to generate 3DFV. To calibrate cameras, the points whose 3D coordinate positions are known must be placed widely throughout the captured images. Thus, we typically put the calibration markers on the field, but unfortunately they need to be removed before a game starts. This means that we could not get calibration images before every capture.

Accordingly, there is always the chance that the calibration result becomes inaccurate for capturing. We call this "calibration error." When we captured half of a soccer field with the cameras at a resolution of 640 by 480 pixels in an actual soccer stadium, the cameras jolted out of alignment by about a pixel on average throughout the experiment. This implies the existence of a gap corresponding to approximately 18 pixels in the virtual environment as the cameras zoom up on the player. Here, gaps are simulated by translating the captured image, since the distances between the cameras and the captured target are far in relation to the player's size.

We assume that the calibration error can be approximated by a Gaussian distribution, and thus we produce an image with calibration errors by translating the image using Gaussian random numbers.

5. EVALUATION

5.1 Comparison method

We evaluated the quality of video produced by different 3DFV methods in the experiment. The comparison was made with a pair-wise video showing, and we adopted the method of limits. Figure 3 shows the experiment's procedure. To compare two videos produced by method-A and method-B, we fixed the amount of noise in method-A and changed incrementally the amount of noise in method-B. Method-A and Method-B are either player-billboard method or shape-from-silhouette method. The subjects watched the videos generated by method-A and method-B simultaneously and selected a response from "left is better," "there is almost no difference," and "right is better" based on their own impressions. They did not know which video was generated by which method. For method-A, the given noise amount is set to half of the noise range given to method-B, and we call this noise level standard stimulation. When increasing the amount of noise, the value at which the evaluation changes from difference to no difference is called the bottom-limit threshold, the value at which the evaluation changes from no difference to difference is called the upper-limit threshold, and the average value of the bottom- and upper-limit thresholds is called the equivalence value. Consequently, when the standard stimulation was between the upper- and bottom-limit thresholds, it showed no difference between the two methods.

We asked the subjects two questions. The first was "Which is superior in image quality for the player region by comparing the two videos?" This was aimed at evaluating the image quality of the generated 3DFV. The other question was "Which has a smoother motion of viewpoint by comparing the two videos?" This was aimed at evaluating the appearance of motion parallax.

5.2 Results

Either the player-billboard method or the shape-from-silhouette method could be assigned as method-A, and, conversely, the same holds for method-B. In fact, after the experiment, we found that both show the same tendencies. Thus, we show only the results when the player-billboard method is taken as method-A.

Figure 4 and Figure 5 illustrate the results. The heavy line in the middle of each graph shows the value of standard stimulation. The values on the vertical axis show the varying noise amounts given to method-B. If the equivalent value is under the line, it indicates that player-billboard is better than shape-from-silhouette. On the other hand, if the equivalent value goes over the line, it means that shape-from-silhouette is better than player-billboard.

5.2.1 Image quality

Figure 4 shows the results for image quality. When there were eight capturing cameras, the obtained results showed that the player-billboard method could generate better image quality than the shape-from-silhouette method, even if no or very small noise were given to shape-from-silhouette. By increasing the number of capturing camera, the equivalent amount of noises for shape-from-silhouette became larger for white noise. This indicates that the advantage of player-billboard becomes smaller when more cameras can be used, but player-billboard is still better if we consider the effect of white noise.

On the other hand, for the other three noises (i.e., motion blur, synchronization, and calibration error), there is little improvement in image quality even as more cameras become avail-

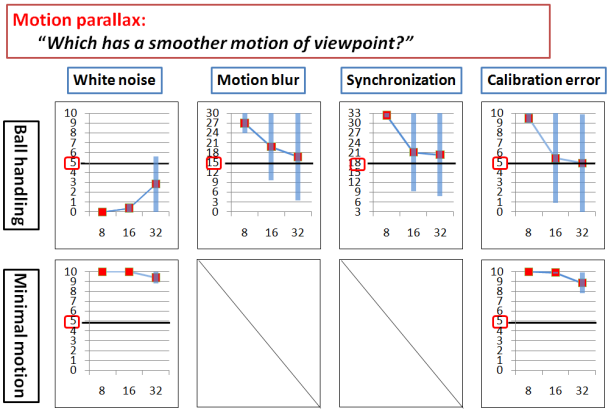

Figure 4. Results of the question about image quality

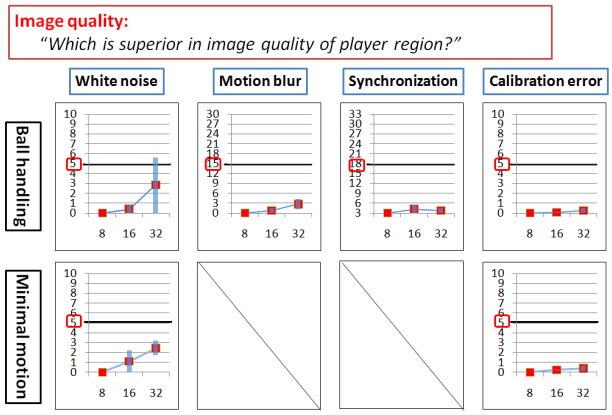

Figure 5. Results of the question about motion parallax

able. This is probably because the resulting 3D shape by shape-from-silhouette could be extensively deformed even with a small amount of these noises or gaps.

5.2.2 Motion parallax

Figure 5 shows the results for the motion parallax evaluation. When there were eight capturing cameras for ball-handling action, the shape-from-silhouette method could generate smoother appearance changes than could the player-billboard method because the latter has to switch textures suddenly when the viewpoint is rotated around the player. However, as more cameras become available, the difference disappears because the gap between the upper-limit threshold and the bottom-limit threshold becomes wider.

On the other hand, for videos with little action, there is obvious advantage to the shape-from-silhouette method with any set-up. Once again, we think it is due to the texture switching, and this could be easier to see when the target object is not in motion.

5.3 Discussion

When players are moving, the player-billboard method is robust to noise and it can produce better-quality videos. In addition, if a large number of cameras are available, the motion parallax will not greatly matter. On the other hand, when players are not moving so much, the player-billboard method cannot surpass the shape-from-silhouette method based on the evaluation of motion parallax. However, we do not think this is a serious problem for the player-billboard method because one of the typical situations that users see a player in minimal motion is to watch a free kick scene where they prefer to see the player from a specific direction.

Shape-from-silhouette can represent motion parallax regardless of the number of cameras and improves the quality of generated 3DFV by increasing the number of cameras used. However, it is sensitive to the various kinds of noises that are inevitable in capturing images.

Therefore, from a practical viewpoint, the player-billboard method performs better than the shape-from-silhouette method in generating 3DFV in a large-scale space.

6. CONCLUSION

This paper compared videos made by two major 3DFV methods, the player-billboard method and the shape-from-silhouette method. They were evaluated under fairly equal conditions so that we could clearly identify which method is better in the presence of noise and gaps, which are inevitable in practical video capturing. To control the amount of noise and the number of cameras, we prepared a virtual environment, to which noises were added based on noise models having a range typically found in actual video capturing. We conducted the evaluation using the method of limits. From the comparison results, the player-billboard method, due to its robustness against noise, achieved a higher quality of generated videos than did the shape-from-silhouette method.

In future work, we are going to compare the methods using scenes featuring more than one player as well as scenes in actual soccer games.

7. REFERENCES

[1] T. Kanade, P. Rander, and P. J. Narayanan, "Virtualized Reality: Constructing Virtual Worlds from Real Scenes," IEEE Multimedia 1997, Vol. 4, No. 1, pp. 34-47.

[2] W. Matusik, C. Buehler, R. Rasker, S. J. Gortler, and L. McMillan, "Image-Based Visual Hulls," ACM SIGGRAPH 2000, pp. 369-374.

[3] M. Waschbüsch, S. Würmlin, and M. Gross, "3D Video Billboard Clouds," Proceedings of Eurographics, Computer Graphics Forum, vol. 26, no. 3, pp. 561-569, 2007.

[4] H. Kim, D. Min, S. Choi and K. Sohn, "A 3D Modeling and Arbitrary view generation System Using Environmental Stereo Cameras," International Journal of Imaging Systems and Technology (SCIE), vol. 17, no. 6, pp. 367-378, 2008.

[5] T. Koyama, I. Kitahara, and Y. Ohta, "Live Mixed-Reality 3D Video in Soccer Stadium," ISMAR 2003, pp. 178-187.

[6] A. Laurentini, "The Visual Hull Concept for Silhouette-Based Image Understanding," IEEE Trans. Pattern Anal. Mach. Intell. 16(2): 150-162, 1994.

[7] German K. M. Cheung, S. Baker, and T. Kanade, "Shape-from-silhouette of articulated object and its use for human body kinematics estimation and motion capture," In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, pp. 77-84, 2003.

[8] D. Strelow, J. Mishler, S. Singh, and H. Herman. "Extending shape-from-motion to noncentral omnidirectional cameras," IEEE/RSJ International Conference on Intelligent Robots and Systems, 2001.

[9] R. Zhang, P. S. Tsai, J. Cryer, and M. Shah, "Shape from Shading: A Survey," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 21, no. 8, pp. 690-706, 1999.