Free-viewpoint video synthesis for soccer games

Davide Lusuardi
Department of information engineering and computer science
University of Trento
Trento, Italy
davide.lusuardi@studenti.unitn.it

***Abstract*—This paper presents some methods for free viewpoint synthesis that generate novel views of actions from any angle and allows viewers to virtually fly through real soccer scenes.**

I. Introduction

Nowadays, sport broadcasting plays a large role in current society. Therefore, it is important to provide a high quality and visually pleasing reporting of sports events. One common issue in sports is that incidents tend to be over very quickly. Slow-motion replays can be used to illustrate these incidents as clearly as possible to the viewer. Although time is stretched in these replays, there is no exploration of the spatial scene information, which is usually important for understanding the event. A system that allows a replay from any angle adds a lot of value to the viewer experience.

Free-viewpoint video (FVV) is one of the recent trends in the development of advanced visual media type that provides an immersive user experience and interactivity when viewing visual media. Compared with traditional fixed-viewpoint video, it allows users to select a viewpoint interactively and is capable of rendering a new view from a novel viewpoint [1]. Examples of physically impossible camera views which could be desirable are the goal keepers view, a player tracking camera or even a ball camera. FVV has been a research topic in the field of computer vision since a virtualized reality system was developed [2], ranging from static models for studio applications with a fixed capture volume, controlled illumination, and backgrounds [3] to dynamic object models for sports scenes. Live outdoor sports such as soccer involve several additional challenges for both acquisition and processing phases. The action take place over an entire pitch and video acquisition should be done at sufficient resolution in order to do analysis and production of desired virtual camera views.

In this paper we briefly present and compare the following methods to accomplish free-viewpoint video visualization for soccer scenes: the *iview* system [3], where automatic online camera calibration, segmentation and 3D reconstruction is performed; the work of Goorts et al. [1], where a plane-sweep approach is used; and the work of Ohta et al. [4], where a billboard-based representation is employed to simplify the 3D model.

II. Overview

In the field of computer vision, the techniques for synthesizing virtual view images from a number of real camera images have been studied since the 1990s. Free-viewpoint video in sports TV broadcast production is a challenging problem involving the conflicting requirements of broadcast picture quality with video-rate generation. FVV techniques for generating novel viewpoints using a multiview camera setup can be categorized into two classes [1]: 3D reconstruction and image-based rendering.

Using 3D reconstruction, it is possible to construct 3D models of objects that provides a geometric proxy which can be used to combine observations from multiple views in order to render images from an arbitrary viewpoint. Several approaches have been realized for reconstruction: visual-hull, photo-hull, stereo, and global shape optimisation [3]. The quality of the virtual view image generated by these methods depends on the accuracy of the 3D model. In order to produce an accurate model, a large number of video cameras surrounding the object should be used. Although 3D reconstruction is robust, artifacts such as ghosting objects can be introduced if a lot of objects are in the scene.

Image-based methods generate the image of the novel viewpoint directly without explicitly reconstructing the 3D scene structure. The quality of rendered views depends on the accuracy of alignment between multiple view observations [1], [3]. Usually, these methods are limited to rendering viewpoints between the camera views.

III. The iview system

In this section we present the DTI-funded collaborative project *iview* [5], a free-viewpoint video system which enables the production of novel desirable camera views. This system exploits the already placed live TV broadcast cameras as the primary source of multiple view video. Usually, soccer matches are covered by 12-20 high-definition cameras placed all over in the stadium providing wide-baseline views. Match cameras are manually controlled to follow the game play zooming in on events when occurs. However, only a fraction of these are focused on specific events of interest and can be used for production of free-viewpoint renders, the remaining cameras cover the pitch, crowd and coaches.

The *iview* system is composed of three main modules as shown in Figure 1: capture, processing and replay module.

Capture is performed using time synchronised acquisition from both auxiliary and match cameras. The minimal number of cameras is about four, but for good quality results a higher

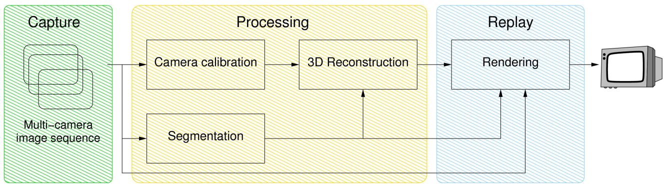

Figure 1. Overview of the *iview* free-viewpoint video system [3].

number is required. Camera synchronisation is achieved using standard genlock process.

The processing module computes a 3D model of the scene. This is done using segmentation of objects from the background and 3D reconstruction [6].

To allow the use of footage from match cameras and to avoid the need for prior calibration, automatic calibration of all cameras is performed using a line-based approach against the pitch lines of the captured footage, achieving a root-mean-square error of 1-2 pixels for moving cameras. The calibration is very fast and robust, capable of real-time operation for use during live match footage. Calibration estimates the extrinsic and intrinsic parameters of each camera including lens distortion.

The segmentation is needed to separate the foreground, i.e., the players from the background. Matting of players from the green pitch is performed using chroma-keying matting. The authors developed and tested a k-nearest neighbour approach for chroma-keying and evaluated two other known techniques, *Fast green subtraction* in RGB colour space and keying in HSV colour space. The k-nearest neighbour classifier is controlled by a GUI where the user has to click on position in the image that corresponds to the background. The process is repeated until the resulting segmentation is satisfying. A deeper explanation is present in the paper [6] where the authors present and evaluate also *Fast green subtraction* and keying in HSV.

The accumulation of errors from calibration and matting can cause large errors in the reconstruction of the scene, such as loss of limbs. Therefore, robust algorithms have been developed for scene reconstruction. One possible technique is called visual-hull (VH) and represents the maximum volume occupied by an object given a set of silhouettes from multiple views [7]. The visual-hull is a single global representation integrating silhouette information from all views. A polygonal mesh surface is typically extracted and texture mapped by resampling the captured multiple view video for rendering [8]. Due to accumulating errors in camera calibration and segmentation, visual-hull accuracy is reduced. A refinement of the view-dependent visual hull (VDVH) [9] using stereo correspondence to interpolate between captured views can be used to overcome these issues achieving the best alignment between adjacent views and hence improve visual quality. More information about *iview* 3D reconstruction can be found in [3], [6], [8].

Finally, the replay module renders the novel view of the scene using the computed 3D model together with the original camera images. Cameras closer to the virtual viewpoint are chosen to generate the novel viewpoint.

The method proposed achieves an image quality comparable to that of the input images and it is robust to the wide-baseline moving camera views at different resolutions. Calibration and segmentation are very important in order to obtain an overall good quality of the system. Degradation in image quality will also occur if there are insufficient views for reconstruction.

IV. Image-based rendering - extended plane sweeping

In this section we present the work of Goorts et al. [1], i.e., an image-based approach to generate virtual camera view interpolating real camera images. Instead of performing 3D reconstruction, image-based methods directly generate the image of the novel viewpoint. When multiple cameras are present, plane sweeping can be used [10] for both small and wide baseline setups. Plane sweeping has already been used for novel viewpoint in soccer scenes. Goorts et al. [11] present a method with two plane sweeps and a depth filtering step suitable for smaller baseline setups of about 1 meter. This method presents some problems like disappearing players when they overlap in the image.

The method presented here is fully automatic and employ GPU parallel processing to achieve fast processing speed. The system setup consists of 7 static cameras placed in a wide baseline setup, i.e., 10 meters between each camera. All cameras are synchronized on shutter level using a global clock. The generation of novel viewpoint consists of two steps as shown in Figure 2: a first off-line preprocessing phase and a real-time interpolation phase.

The preprocessing phase is responsible for camera calibration and background determination. Cameras are calibrated in order to acquire their position, orientation and extrinsic and intrinsic parameters. The calibration method of Svoboda et al. [12] has been employed for this purpose: SIFT features are extracted from a number of frames and the pairwise matching between them is calculated using the k-d tree algorithm; these matches are tracked across different image pairs obtaining point correspondences between multiple images [1]. The background of every image stream is determined using a per-pixel median approach applied to about 30 images per stream (2 seconds apart each other).

The real-time phase generates images for a chosen virtual camera position and a chosen time in the video sequence. First, camera images are debayered and segmented. These images are then used to process foreground and background independently. The foreground rendering uses a plane-sweep approach followed by depth filtering and a depth-selective plane sweep as explained in [1]. The foreground and background are then merged together according to the segmentation information.

Debayering consists of converting the raw images to its RGB representation. Segmentation is based on backgrounds obtained during off-line preprocessing and is performed on a per-pixel basis using the differences between the color values

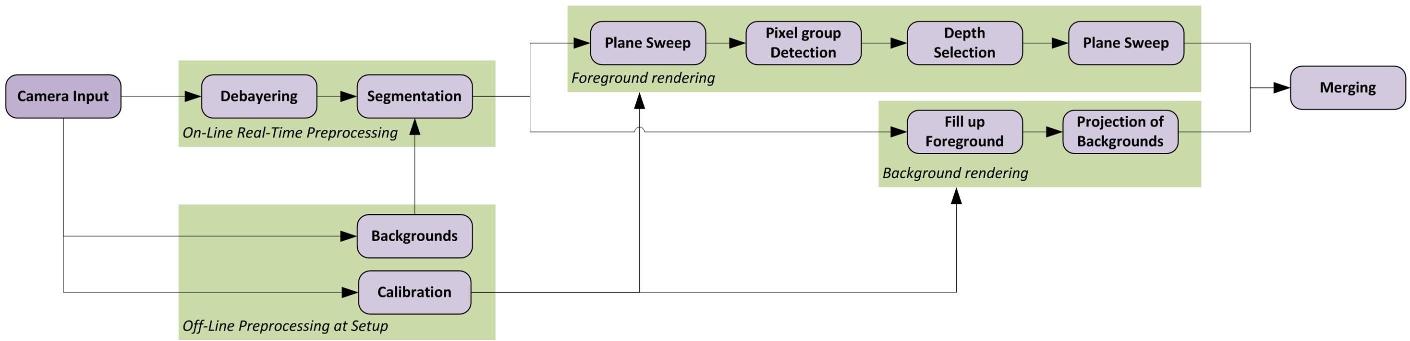

Figure 2. Overview of the extended plane sweeping method. Both the non real-time and real-time phase are shown [1].

and three thresholds. This allows fast segmentation in high quality.

In this way, the authors obtained high quality results using wide baseline setup and typical artifacts of normal plane sweeping, such as ghost players, are removed. Some other artifacts can still be present, such as ghost lines, caused by the simplified assumption of the geometry of the pitch.

V. Billboard-based visualization

In this section we present the work of Ohta et al. [4]. The authors use billboard representation to make a 3D model of each player. This method is simpler than full 3D reconstruction and require less computation. A player billboard is a small rectangle standing perpendicular to the ground and a 2D texture is shown on it. The difference between 3D reconstruction and billboard representation is shown in Figure 3: as we can see, the visual difference is clear at a close viewpoint but becomes very small at a distant one. For this reason, becomes particularly important to place player billboards at a right place and direction.

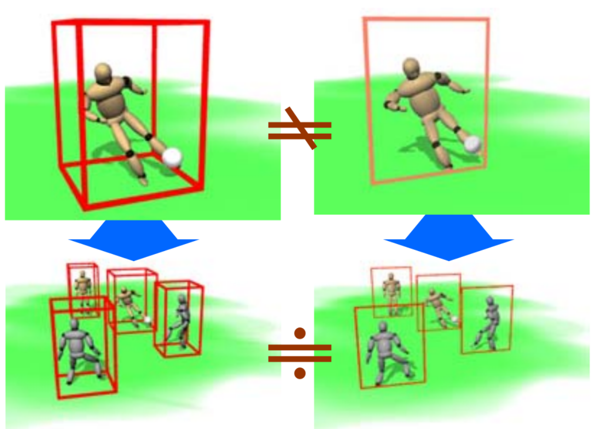

Figure 3. Appearance similarity between 3D reconstruction and billboard in close and distant view [4].

The system proceeds as follow: first extracts texture segments from camera videos, then selects appropriate textures according to the virtual viewpoint and finally layouts the player billboards in virtual space.

Texture extraction phase consists in obtaining the location of each player and extracting texture segments from every image video by projecting player location onto the image plane. Background region is removed in the texture by video capturing PC. Please note that camera calibration is done before this phase.

Texture selection phase selects a set of texture to be sent to each viewer based on his viewpoint. Given a viewpoint, the system finds the camera that minimizes the angle between the line from the viewpoint to the player location and the line from the camera to the player location. Then, a texture segment obtained by that camera is selected and placed so that the texture faces the viewpoint [4].

One possible problem happens when players are overlapped each other at a certain camera and billboard texture could include both players. To eliminate extra player region, authors used stereo based method [13] as explained in [4].

VI. Conclusion

We have briefly presented three free-viewpoint systems for soccer games: the *iview* system, an image-based method using plane sweeping, and a billboard-based method.

The *iview* system is a robust method using a multi-camera setup that performs 3D reconstruction exploiting a refinement of the view-dependent visual hull method.

The image-based method presented interpolates real camera images using a plane sweeping approach. High quality images using wide baseline setup are obtained and typical artifacts are removed.

The billboard-based method is simpler than full 3D reconstruction and require less computation. However, this method produces lower quality images, and some visual artifacts can be present.

References

[1] P. Goorts, S. Maesen, M. Dumont, S. Rogmans, and P. Bekaert, “Free viewpoint video for soccer using histogram-based validity maps in plane sweeping,” in *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, vol. 3, 2014, pp. 378–386.

[2] T. Kanade, P. Rander, and P. J. Narayanan, “Virtualized reality: constructing virtual worlds from real scenes,” *IEEE MultiMedia*, vol. 4, no. 1, pp. 34–47, 1997.

[3] A. Hilton, J.-Y. Guillemaut, J. Kilner, O. Grau, and G. Thomas, "Free-viewpoint video for tv sport production," in *Image and Geometry Processing for 3-D Cinematography*, 2010.

[4] Y. Kameda, T. Koyama, Y. Mukaigawa, F. Yoshikawa, and Y. Ohta, "Free viewpoint browsing of live soccer games," in *2004 IEEE International Conference on Multimedia and Expo (ICME) (IEEE Cat. No.04TH8763)*, 01 2004, pp. 747–750.

[5] "The iview project," http://www.bbc.co.uk/rd/projects/iview.

[6] O. Grau, G. Thomas, A. Hilton, J. Kilner, and J. Starck, "A robust free-viewpoint video system for sport scenes," in *2007 3DTV Conference*, 06 2007, pp. 1 – 4.

[7] Heung-Yeung Shum, Sing Bing Kang, and Shing-Chow Chan, "Survey of image-based representations and compression techniques," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 11, pp. 1020–1037, 2003.

[8] O. Grau, A. Hilton, J. Kilner, G. Miller, T. Sargeant, and J. Starck, "A free-viewpoint video system for visualisation of sport scenes," 2006.

[9] G. Miller and A. Hilton, "Exact view-dependent visual hulls," in *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 1, 2006, pp. 107–111.

[10] R. Yang, M. Pollefeys, H. Yang, and G. Welch, "A unified approach to real-time, multi-resolution, multi-baseline 2d view synthesis and 3d depth estimation using commodity graphics hardware." *Int. J. Image Graphics*, vol. 4, pp. 627–651, 10 2004.

[11] P. Goorts, C. Ancuti, M. Dumont, S. Rogmans, and P. Bekaert, "Real-time video-based view interpolation of soccer events using depth-selective plane sweeping," *VISAPP 2013 - Proceedings of the International Conference on Computer Vision Theory and Applications*, vol. 2, 04 2013.

[12] T. Svoboda, D. Martinec, and T. Pajdla, "A convenient multi-camera self-calibration for virtual environments," *Presence*, vol. 14, pp. 407–422, 08 2005.

[13] T. Koyama, I. Kitahara, and Y. Ohta, "Live mixed-reality 3d video in soccer stadium," in *The Second IEEE and ACM International Symposium on Mixed and Augmented Reality, 2003. Proceedings.*, 2003, pp. 178–186.