A FREE-VIEWPOINT VIDEO SYSTEM FOR VISUALISATION OF SPORT SCENES

Oliver Grau[1], Adrian Hilton[2], Joe Kilner[2], Gregor Miller[2]
Tim Sargeant[1], Jonathan Starck[2]

[1]BBC Research, UK
Oliver.Grau | Tim.Sargeant @ rd.bbc.co.uk
[2]Centre for Vision, Speech and Signal Processing, University of Surrey, UK
Adrian.Hilton | JStarck | Joe.Kilner | Gregor.Miller @ surrey.ac.uk

ABSTRACT

This contribution introduces a new approach of using multi-camera images of an event to provide a free-viewpoint video visualisation of sport scenes. This allows a virtual camera to be moved freely around and to view the action from any angle. The system is based on 3D reconstruction techniques previously developed for studio use. A number of different 3D representations, including billboards, visual hulls and view-dependent geometry are evaluated for the purpose.

First results show the potential of the new approach for action replay and strategy analysis of sport scenes. Current limitations of the 3D representations are discussed in the context of a practical use of the system.

INTRODUCTION

In television sports coverage, most interesting incidents tend to be over very quickly. Sports producers may use techniques such as slow-motion replays to illustrate these incidents as clearly as possible for the viewer. Although time is stretched in these replays, there is no enhancement of the spatial scene information, which could be important for understanding the event.

The work presented in this paper is part of the DTI-funded collaborative project ‘iview’ (1) with the goal to develop a system that allows the capture and interactive free-viewpoint replay of live events. Free-viewpoint video uses the input from a set of cameras to simulate novel, virtual camera viewpoints for visualisation. A method often used is to freeze time and then move the virtual camera in space. These effects were used in films like "The Matrix" but required intensive manual post-production work and the camera positions of the replay were fixed and covered only a small area. The “Eye Vision” system, developed for sports broadcast applications, uses cameras mounted on robotic heads that are controlled by an operator and follow the action. Because of the fixed camera positions, the system adds an interesting visual effect but cannot be used to visualize the scene from any angle.

A system that allows limited, but free camera movement is the Piero system (13), which is commercially available. The key to this is that Piero creates a 3D representation of the scene (12). This is done from just one camera image and a flat polygon per player is produced. However, the system is restricted in that that the virtual camera can be moved only very little from the original view-point without degradation of the image quality.

The technology being developed in iview is based on a flexible multi-camera capture system. It provides fully automatic algorithms for 3D reconstruction and texture mapping. The image-based 3D view-synthesis is at interactive refresh rates. Moreover the viewpoint of the camera can be chosen freely. First results of the project demonstrate its use in

football scenarios for production of TV sport coverage. A long-term goal of the project is to demonstrate how the techniques can be used on an interactive consumer platform like a games console.

The rest of this paper is organized such that the next section gives an overview of how the system would be used in a broadcast environment including camera set-up and calibration. The following section discusses different 3D representations. Finally, some results are presented with conclusions.

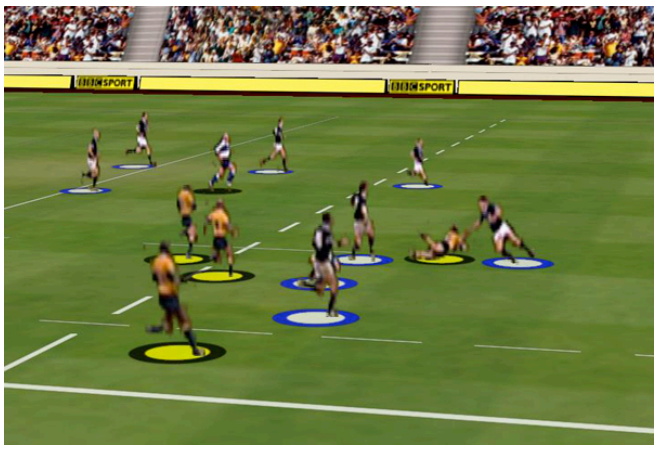


Figure 1 – Piero creates a flat 3D representation of an action

SYSTEM SET-UP AND CALIBRATION

This section discuses the importance of the camera set-up and how the system is integrated into the broadcast coverage system. The camera calibration is also briefly explained.

Camera configuration

The 3D reconstruction techniques discussed in subsequent sections need several images taken from different angles as input. We have identified two basic options for the camera set-up: First to use only the broadcast coverage cameras and second use statically mounted, unmanned cameras that cover the area of interest, either as the only image source, or in conjunction with images from the broadcast coverage cameras. Both options have implications on the 3D reconstruction techniques used and the practicality of the system.

Currently the question of the 'optimal' camera configuration is being studied. The first project results presented in this paper are based on two separate data captures - standard broadcast material captured from a 14-camera broadcast, and a test shoot with 15 cameras specifically positioned in a quarter circle of the football pitch. All cameras were operated in standard definition.

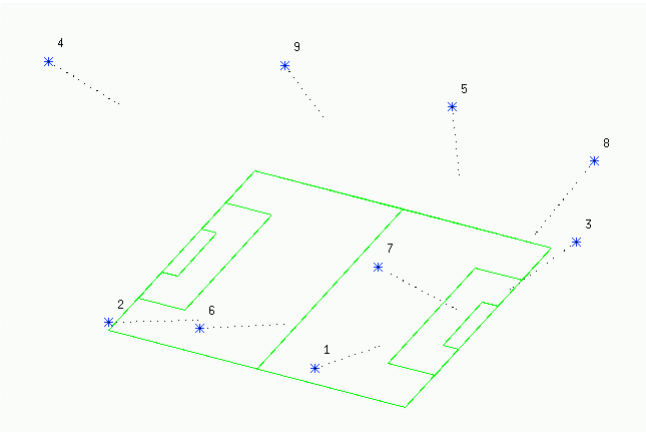


Figure 2 - Match camera configuration

Broadcast camera configuration

Typically, a football outside broadcast will consist of 8 -16 cameras, depending on the importance of the match. For our particular test sequence – that of a penalty shot – images from nine of the cameras are useful, as illustrated in Figure 2. The other cameras are either framed too tightly to be suitable for our processing, or are pointing away from the action (e.g. to record the reaction of a team coach or the crowd). For flowing sequences of play (e.g. a shot on goal, as opposed to a penalty, where play is stopped and camera operators have time to frame a shot), even fewer cameras generally offer acceptable images.

First observations from this test include:

- For typical sequences, only a few cameras from a multi-camera shoot can be expected to provide useful images for modelling.
- The pre-cabled camera positions around football grounds offer a good range of viewpoints (both high & low and from a variety of directions).
- The fact that cameras are operated usually gives a good coverage of all areas of the pitch.

Using only footage from broadcast cameras means that there are no additional rigging costs and the methods can be even used on archived material. On the other hand the number of cameras that contribute to the 3D reconstruction might be too low to create high quality visualisation from a desired viewpoint. It may be possible to improve the 3D reconstruction by supplementing the broadcast cameras with additional cameras in specific locations.

Narrow baseline configuration

We carried out a second test shoot, capturing a sequence using a narrow-baseline set-up of 15 cameras in an approximately 120° arc around the goal area. The camera positions and viewpoints with respect to the pitch are shown in Figure 3.

This test shoot provided useful results, with the following observations:

- Rigging so many cameras in non-standard places was high-cost and time-consuming.
- Because the cameras were fixed, 3D reconstruction was restricted to events that happened within a small area of the pitch, at one end of the stadium.
- The high density of cameras allows for good quality virtual camera moves along the arc.

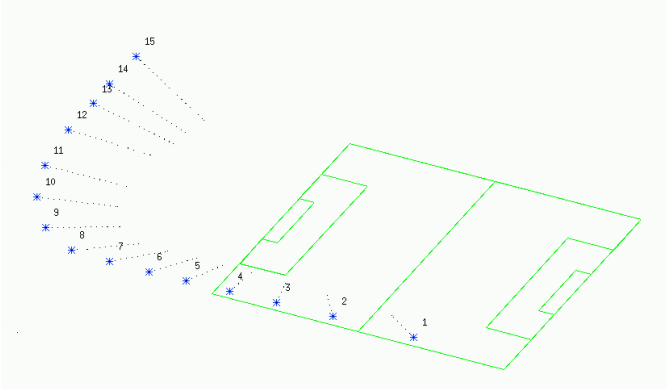


Figure 3 - Narrow baseline camera configuration

In all cases, it is important that cameras are positioned such that players can be easily segmented from the pitch (background clutter such as advertising hoardings and crowds should be avoided where possible).

Calibration

Before using images in the modelling process, it is necessary to calibrate each camera used to determine the relationship between objects appearing in the image plane and their location in the 3D world. In the case where static cameras are used, this can be a one-off process. However, since broadcast coverage cameras change viewpoint, field of view and sometimes position, camera parameters must be determined dynamically and unobtrusively.

For this purpose we have developed a completely automated calibration method that uses the pitch lines seen from each camera to determine the camera's position and pose (2).

Using an entirely optical based technique for camera calibration as opposed to techniques using any kind of mechanical sensor or measurement makes deployment across many cameras achievable, and potentially opens up the system for use even on archive material.

This method does require that cameras be framed so that a suitable number of pitch lines are visible. For wide shots, as are regularly used, this is often the case, but some close-up shots may be rejected because the calibration fails.

3D MODELLING AND RENDERING

A number of approaches have been investigated to reconstruct and represent dynamic scenes from multiple view video capture. Initial work in iview has evaluated previous reconstruction approaches for application in outdoor sports scenes such as football. Reviews of previous geometric and image-based representations can be found in (3,9,8). To reconstruct the players we assume that a segmentation or matting of the foreground (players) from the background (pitch) is available. This can be done by background subtraction (difference matting) or on colour-base (chroma-keying against the green).

In this work we focus on the representation of foreground objects (players) to evaluate capture requirements, visual quality, robustness, storage and rendering costs. The representations evaluated in initial work are:

Billboards: The simplest form of geometric representation is a planar billboard or sprite texture mapped with the captured video, as used in the Piero system. Reconstruction is required to estimate the approximate location of the player on the pitch for billboard placement. Billboards allow high-quality synthesis of adjacent camera views using footage from a single camera view. Interpolation to transition between real camera views may result in visual artefacts due to incorrect player geometry. This leads to a requirement for estimating more accurate player geometry or correspondence between views. A possible advantage of billboards over more complex geometry is robustness to matting errors.

Visual Hull: The visual-hull (VH) represents the maximum volume occupied by an object given a set of silhouettes from multiple views (8). This approach has been widely used for robust reconstruction of objects and indoor studio scenes from multiple camera views where foreground object silhouettes are segmented via chroma-key or other matting techniques. 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. View-dependent texture mapping can be combined with visual-hull rendering to increase the realism by resampling from the real cameras nearest to the desired view. Robustness of the visual-hull reconstruction is dependent on accurate camera calibration and silhouette extraction. The visual-hull provides an approximation of the scene geometry without concavities which can lead to incorrect alignment of images between views resulting in visual artefacts.

Photo Hull and Stereo Refinement: Photo-hull represents the maximum volume which is photo-consistent with multiple view video capture (14). Photo-consistency and stereo correspondence can be used to refine the visual-hull scene approximation. This approach has been used in previous studio based reconstruction of people to achieve improved visual quality (1,15). Visual-hull refinement using stereo correspondence or photo-consistency gives a global representation which integrates information from all views into a single surface. In the presence of errors in calibration or silhouette extraction this approach may fail. An alternative is to develop local approaches which optimise the correspondence or geometry between adjacent views.

View-dependent Geometry: Previous work (10) has shown that in the presence of calibration error a single surface geometry may fail to correctly align all images resulting in visual artefacts (blur, misalignment). View-dependent geometry based on stereo correspondence or photo-consistency between views can be used to represent the surface which achieves the best alignment between adjacent views and hence improve visual quality. This requires additional storage either as pairwise stereo correspondence, view-dependent displacements of a single mesh or a depth map for each captured image. View-dependent visual-hull VDVH (5) which refines the visual-hull to represent view-dependent geometry has previously been introduced to give optimal correspondence between adjacent views. This approach was developed for real-time interactive rendering of people from multiple camera studio capture.

Table 1 summarises the important properties of each of these approaches to reconstruction and rendering dynamic scenes. This analysis assumes capture from 8 cameras with SD-PAL resolution images. The visual-hull representation and texture maps are assumed to be at the same resolution as the captured images.

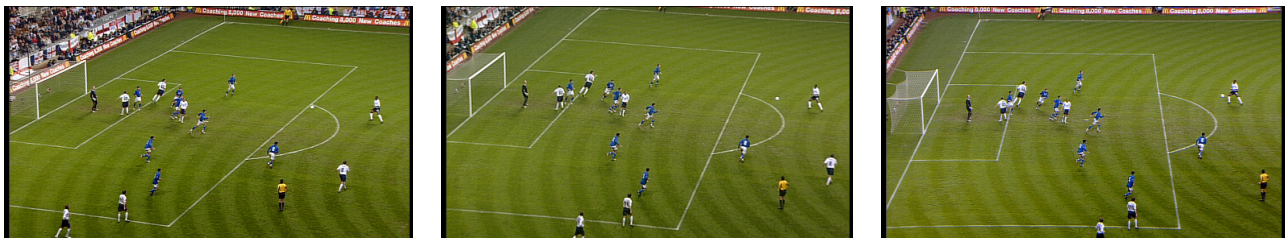
Representation	Novel view quality	Size (MB/s)	Rendering cost(tris.)	Robustness to Errors		
				Calibration	Matting	Reconstruction
Raw Video		250				
Billboards	low	60	1	high	low	low
Visual-hull	medium	30	2-20K	low	low	high
Photo-hull	medium	30	2-20K	low	medium	medium
VDVH (4)	medium	100	600K	low	low	medium
Stereo	high	100	200K	medium	medium	low

Table 1: Comparison of existing representations for novel view interpolation from multiple view video capture. The comparison assumes 8 cameras at SD-PAL resolution with players covering approximately 25% of each view.

RESULTS

In this section we compare results of applying the representations presented in the previous sections to multiple view capture of football. This comparison is qualitative to visually assess the relative merits of each approach. The approaches were previously developed and demonstrated for studio based reconstruction and have been applied directly to outdoor scene reconstruction without significant variations to the approach.

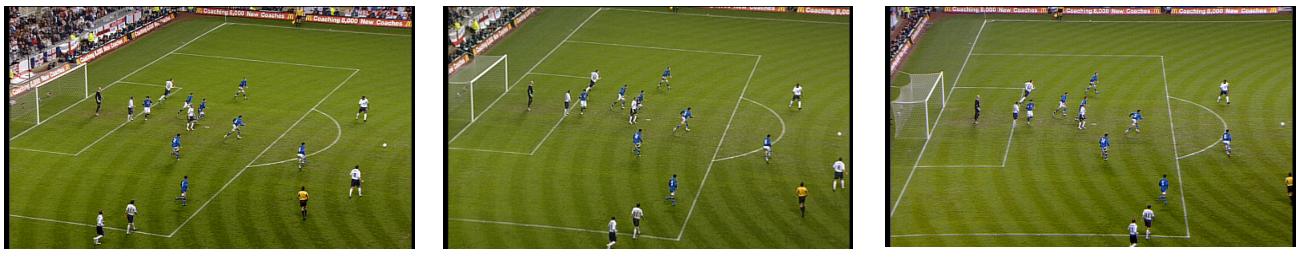
Figures 4 and 5 present a comparison of billboards, visual-hull and view-dependent visual-hull applied to novel view synthesis for frames from the camera set-up shown in Figure 2 (9 broadcast cameras). The initial comparison identifies limitations of all existing approaches in their application to football. Visual artefacts in the sequences result from a number of sources, primarily errors in: camera calibration; matting; player location; and incorrect geometry or correspondence. The billboard approach interpolates views by matting images onto planar polygons. The most significant artefacts observed in this approach are flicker due to errors in player localisation and ghosting due to incorrect correspondence between interpolated views which results from the simple geometry. Player localisation currently fails to reliably estimate the centroid of each player for groups of players in close proximity. The visual-hull and view-dependent visual-hull both currently exhibit similar performance, errors in matting and camera calibration resulting in missing limbs. The visual-hull geometry is sufficient to interpolate views, which significantly reduces the ghosting artefacts observed with planar billboards. The visual-hull implementation uses the player location to define the volume surrounding each player, errors in the localisation result in flicker.



(a) Video capture frames from multiple cameras at a single time instant

(b) Billboards (c) Visual-hull (d) View dependent visual-hull

Figure 4: Comparison of novel view synthesis using billboards, visual-hull and view-dependent visual-hull from a mid-viewpoint between not captured in the original cameras (frame 73).



(a) Video capture frames from multiple cameras at a single time instant

(b) Billboards (c) Visual-hull (d) View dependent visual-hull

Figure 5: Comparison of novel view synthesis using billboards, visual-hull and view-dependent visual-hull from a mid-viewpoint between not captured in the original cameras (frame 88).

This can be overcome using a global visual-hull representation as demonstrated for the second data set below (Figure 6). As the players are relatively small, errors in the visual-hull geometry which would result in incorrect alignment between views are not visible in these sequences.

Figure 6 shows results from the second camera set-up (narrow baseline configuration) as shown in Figure 3. Since the cameras are static and all looking into the penalty area of the pitch it is possible to compute a globally consistent model of that area. The shadow effects are added by the rendering system.

CONCLUSIONS

Initial results of applying the 3D reconstruction techniques developed for the use in studios to outdoor football scenes show the potential to visualise actions from novel viewpoints, even from above. This evaluation also identifies current limitations and challenges.

Simple rendering of the video on planar billboards gives high visual-quality allowing the synthesis of changes in viewpoint around the captured camera view. However, using billboards to transition between views results in significant visual artefacts due to incorrect geometry which causes misalignment of the images. This results in unacceptable visual artefacts such as ghosting and blur.

Visual-hull reconstruction from image silhouettes allows an approximate geometry to be extracted. View-dependent texture mapping from the captured views result in high-quality rendering. The visual-hull geometry is sufficient to allow high-quality transitions without noticeable misalignment in textures. Errors in silhouette extraction for any of the views result in gross errors such as missing limbs. Global calibration error also results in erosion of the visual-hull geometry from the true surface. The analysis performed demonstrates both the advantages and limitations of the global visual-hull reconstruction approach in outdoor sports scenes.

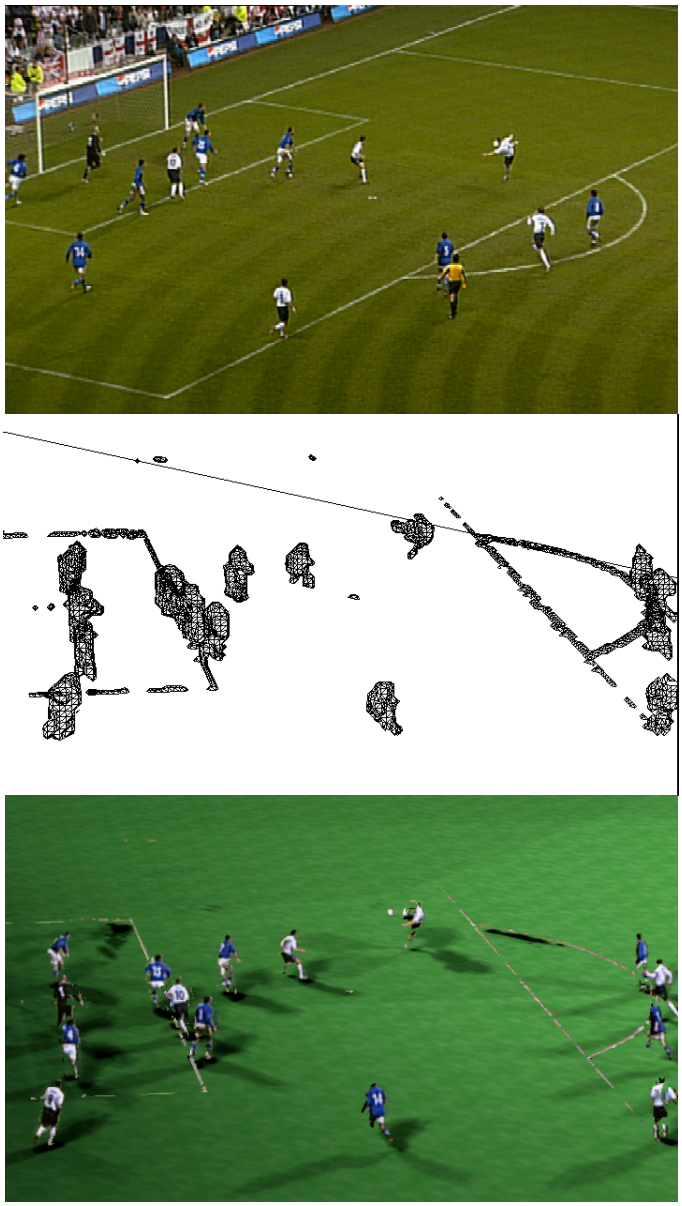


Figure 6 – 3D model using visual hull from second data set (top) rendered in wireframe (middle) and textured (bottom) mode.

The view-dependent visual hull achieves a similar visual-quality to visual-hull for view-interpolation. The principal limitation of both the VH and VDVH approaches is their sensitivity to errors in matting and camera calibration which cause artefacts such as missing limbs in the sequences tested. The initial comparative evaluation of existing novel view-synthesis techniques on outdoor football scenes indicates that none of the existing approaches achieve visually acceptable results for broadcast production. Future research within iview will address the limitations of these approaches to achieve robust free-viewpoint synthesis of broadcast quality view.

ACKNOWLEDGEMENTS

This research was supported by the DTI Technology Programme project 'iview: Free-viewpoint Video for Interactive Entertainment Production' TP/3/DSM/6/I/15515 and EPSRC Grant EP/D033926/1.

REFERENCES

1. The iview project, http://www.bbc.co.uk/rd/projects/iview
2. O. Grau, M. Prior-Jones and G.A. Thomas. 3D modelling and rendering of studio and sport scenes for TV applications. WIAMIS 2005, April 13-15 2005, Montreux, Switzerland.
3. C. Buehler, M. Bosse, L. McMillan, S. Gortler, and M. Cohen. Unstructured Lumigraph Rendering. In Proc. ACM SIGGRAPH, pages 425—432, 2001.
4. S.E. Chen and L. Williams. View Interpolation for Image Synthesis. In Proc. ACM SIGGRAPH, 1993.
5. G. Miller, A. Hilton, and J. Starck. Interactive Free-viewpoint Video. In IEE Conf. on Visual Media Production, 2005.
6. R. Pajarola, M. Sainz, and Y. Meng. DMESH: Fast Depth-Image Meshing and Warping. International Journal of Image and Graphics, 4(4):1—29, 2004.
7. J. Shade, S. Gortler, L.-W. He, and R. Szeliski. Layered Depth Images. In Proc. ACM SIGGRAPH, 1998.
8. H.-Y. Shum, S.B. Kang, and S.-C. Chan. Survey of image-based representations and compression techniques. IEEE Transactions on Circuits and Systems for Video Technology, 13(11), 2003.
9. G. Slabaugh, B. Culbertson, and T. Malzbender. A survey of methods for volumetric scene reconstruction from photographs. In Volume Graphics, 2001.
10. J. Starck and A. Hilton. Virtual view synthesis of people from multiple view video. Graphical Models, 67(6):600—620, 2005.
11. J. Starck, G. Miller, and A. Hilton. Video-Based Character Animation. In ACM SIGGRAPH/Eurographics Symposium on Computer Animation, pages 49—58, 2005.
12. O. Grau, M. Price and G.A. Thomas. Use of 3-D techniques for Virtual Production. In Proc. Of SPIE conference on 'Videometrics and Optical Methods for 3D Shape Measurement', 22-23 Jan.2001, San Jose, USA.
13. The Piero Project, http://www.bbc.co.uk/rd/projects/virtual/piero/index.shtml
14. Kutulakos,K.and Seitz,S. A Theory of Shape by Space Carving, International Journal of Computer Vision, 38(3):199-218,2000.
15. Starck, J. and Hilton, A. Model-based Multiple View Reconstruction of People. IEEE International Conference on Computer Vision, pages 915-922, 2003.