*Research White Paper*

*WHP 149*

*July 2007*

# A Robust Free-Viewpoint Video System for Sport Scenes

**O. Grau, G.A. Thomas, A. Hilton, J. Kilner, J. Starck**

*BRITISH BROADCASTING CORPORATION*

White Paper WHP 149

# A Robust Free-Viewpoint Video System for Sport Scenes

O. Grau, G.A. Thomas, A. Hilton, J. Kilner, J. Starck

## Abstract

This contribution describes robust methods to provide a free-viewpoint video visualisation of sport scenes using a multi-camera set-up. This allows generation of novel views of actions from any angle and is of interest for visualisation in TV productions. The system utilises 3D reconstruction techniques previously developed for studio use. This paper discusses some experiences found while applying these techniques for an uncontrolled outdoor environment and addresses robustness issues. This includes segmentation, camera calibration and 3D reconstruction. A number of different 3D representations, including billboards, visual hulls and view-dependent geometry are evaluated for the purpose.

This document was originally published in Proceeding of 3DTV conference 2007, Kos, Greece.

**Additional key words:** Sport replay, 3D reconstruction, image-based rendering

White Papers are distributed freely on request.

Authorisation of the Head of Research is required for publication.

# A ROBUST FREE-VIEWPOINT VIDEO SYSTEM FOR SPORT SCENES

*O. Grau, G. A. Thomas*

BBC Research, Tadworth, Surrey, UK.
Oliver.Grau @rd.bbc.co.uk

*A. Hilton, J.Kilner, J.Starck*

University of Surrey, UK.
A.Hilton @surrey.ac.uk

**ABSTRACT**

This contribution describes robust methods to provide a free-viewpoint video visualisation of sport scenes using a multi-camera set-up. This allows generation of novel views of actions from any angle and is of interest for visualisation in TV productions. The system utilises 3D reconstruction techniques previously developed for studio use. This paper discusses some experiences found while applying these techniques for an uncontrolled outdoor environment and addresses robustness issues. This includes segmentation, camera calibration and 3D reconstruction. A number of different 3D representations, including billboards, visual hulls and view-dependent geometry are evaluated for the purpose.

## 1. INTRODUCTION

In sport most interesting incidents tend to be over very quickly. A system that allows a replay from any angle adds a lot of value to the production of sport coverage. 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 exploration of the spatial scene information, which is usually important for understanding the event.

The work presented in this paper is part of the DTI-funded collaborative project iview [1], whose goal is to develop a system that allows the capture and interactive free-viewpoint replay of live sport events, as depicted in Fig. 1. The proposed system uses the input from multiple 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 many cameras, 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 visualise the scene from any angle.

Systems that capture the action with a number of cameras and provide a free-viewpoint functionality were first developed for the studio, for example [2, 3, 4, 5]. For use in an outdoor environment, very little work has been done using multiple cameras. Most approaches use just one camera, e.g. [6]. The work presented here addresses the problems in such an uncontrolled environment.

**Fig. 1**. Image of a football game from a broadcast camera.

The main difference between the scene as depicted in Fig. 1 and those addressed by previous projects is that the environment is not as well controlled as a studio. Even if cameras are mounted in fixed positions there are situations where the cameras are moving relative to the objects of interest, due to wind or because the entire stand of the stadium is moving under the weight of the audience. Furthermore the size of the objects in the images is usually smaller than in the studio, because a large area of the pitch has to be covered. Due to these factors, poor segmentation or inaccurate camera calibrations have an increased impact on the visual quality of the system.

The rest of this paper is structured as follows: The next section gives a brief overview of the system components. Section 3 then gives some details of the implemented processing modules and section 4 describes the replay. The paper finishes with some results and conclusions.

## 2. OVERVIEW

Fig. 2 gives an overview of the proposed system. The capture uses a time synchronised, calibrated multi-camera system. The minimal number of cameras is about four, but for good quality results a higher number is required. We are considering different configurations using broadcast coverage cameras and additional cameras. For more details on these

configurations and integration into a broadcast environment see [7].

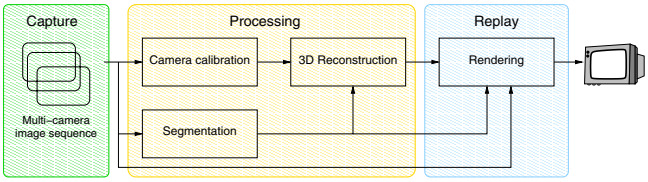


**Fig. 2**. Overview of the free-viewpoint system.

The processing module computes a 3D model of the scene. This is done using segmentation of objects from the background and 3D reconstruction. The next section describes some details of this processing.

The replay module renders the captured scene in realtime using the computed 3D model and the original camera images deploying view-dependent texture mapping [8].

The entire system can potentially operate in real-time. At the current stage the processing is done offline. That means the images are stored and the processing is run at a later stage. The replay module is designed to work at interactive rates.

## 3. PROCESSING

### 3.1. Camera Calibration

Most studio-based capture systems assume that the cameras are mounted statically and a calibration can be done once before the system is used. However in an outdoor environment the cameras are not necessarily mounted absolutely rigidly. Depending on the location the cameras might move slightly either caused by wind or vibrations of a big audience.

We therefore use a line-based approach for the calibration of camera parameters against the pitch lines [9]. This approach is very fast (can be computed in real-time on a PC) and robust giving online updates of camera parameters for moving cameras.

### 3.2. Segmentation

The segmentation separates the action, i.e. the players of a game from the background. Possible methods are difference- or chroma-keying against the green pitch. We investigated the latter option because it also works for moving cameras. A particular problem of a broadcast environment is that the pictures of the broadcast cameras are usually compressed (typically M-JPG). We evaluated two known techniques for chroma-keying for our application: Fast green subtraction in RGB colour space, and keying in HSV colour space. In addition to that we developed and tested a k-nearest neighbour approach.

*Fast Green*: This method is often implemented in commercial chroma-keyer and is based on the difference between the green channel intensity value for a given pixel and the maximum of the red and blue channel values:

$$d_{fg} = g - max(r, b) \tag{1}$$

The segmentation $S_{fg}$ is computed using threshold $\sigma_{fg}$:

$$S(x,y) = \begin{cases} 0 & , d_{fg} > \sigma_{fg} \\ 255 & , \text{otherwise} \end{cases} \tag{2}$$

*HSV*: This method is based on the distance of a pixel $I$ in HSV colour space to a background colour $P$. The segmentation $S_{HSV}$ is then computed using a threshold as described for the 'fast green' method in equation 2.

*K-nearest neighbour classifier*: This classifier is controlled by a simple GUI: The user clicks on positions in an image that represent background. The RGB colour values of that pixel are stored as a prototype $P_i = I$ into a list. All pixels in the image that are within a radius $r_1$ of the colour prototype are then marked as background as well. The user continues to choose background pixels until the resulting segmentation is satisfying.

The segmentation $S_{k-nearest}$ is computed by finding the nearest colour prototype $P_{best}$ from the list. With the distance $d$ of the pixel RGB values $I$:

$$d = Distance_{inRGB}(P_{best}, I) \tag{3}$$

In order to get continuous values a soft key can be obtained using a second radius $r_2$:

$$S'_{k-nearest} = \begin{cases} 0 & ,d <= r_1 \\ \frac{255(d-r_1)}{r_2-r_1} & , r_1 < d <= r_2 \\ 255 & , \text{otherwise} \end{cases} \tag{4}$$

Fig. 3 (left) shows the image pixels of Fig. 1 in RGB colour space. The pitch pixels are distributed in an elongated ellipsoid. Fig. 3 (right) shows 16 colour prototypes in RGB that approximate this distribution.

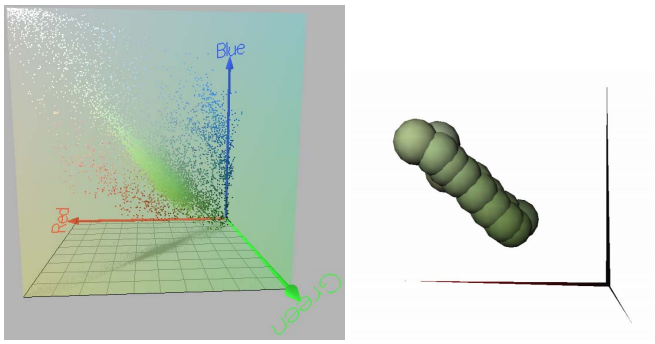


**Fig. 3**. RGB colour histogram (left) and selected colour prototypes (right).

### 3.3. 3D Reconstruction

Free-viewpoint rendering in sports production ideally requires a visual quality comparable to the source video together with reconstruction from sparse viewpoints and video-rate playback. Possible approaches include: billboards [10], visual hull [11], and the view-dependent visual hull [12].

Billboarding uses a single polygon placed co-incident with the object that it represents. This polygon is then rotated around an axis or point (typically the vertical axis) so that it retains its original position, but is constantly facing the virtual camera. An image of the original object is then applied to the polygon as a texture map. This technique can often give good results with very little overhead in reconstruction or rendering as large-scale parallax effects are handled by the relative positioning of the billboards, while the lack of small-scale parallax is often not noticed. However, the approach is limited to distant views and does not facilitate smooth transitions between views.

The visual hull (VH) [11] derives scene geometry that is consistent with a set of image silhouettes. Reconstruction projects the silhouettes of the foreground objects in each image using the calibration data for the relevant camera. It then finds the intersections of these projections. Each intersection defines the largest possible surface that could produce the silhouettes in the original images. This approach has been widely used for object reconstruction assuming accurate camera calibration and silhouettes.

In sports such as football which requires capture over a relatively large area with uncontrolled illumination the visual hull accuracy is reduced due to errors in camera calibration and matting together with image quantisation. A conservative visual hull approach taking into account these errors allows robust visual hull reconstruction in sports. The resulting multiple view images can then be textured onto the visual hull surface for view-dependent rendering of the players. This approach achieves robust reconstruction but does not accurately align overlapping images due to errors in geometry resulting in blur or doubling of features.

Refinement of an initial robust visual hull estimate provides an approach to high-quality reconstruction in the presence of global calibration errors [5]. Stereo correspondence between wide-baseline camera views constrained by the initial surface estimate allows refinement of the surface to locally align the multiple view images. This produces high-quality free-viewpoint rendering in the presence of global errors. Application of this approach to sports has been achieved by refinement of the view-dependent visual hull (VDVH) [12] using stereo correspondence to interpolate between captured views. The VDVH provides an exact sampling of the VH surface as a depth map from a specific camera viewpoint. This is then refined for pairs of views using a graph-cut to optimise stereo-correspondence and boundary constraints. VDVH is less sensitive to global errors due to camera calibration or matting, and unlike the VH it provides locally correct aligned textures. A comparative evaluation of free-viewpoint video in football using billboards, visual hull and view-dependent visual hull is presented in [13].

## 4. REPLAY

The replay module uses the 3D models of the scene together with the original camera images to produce a novel view of the scene. The camera images are applied using view-dependent texture mapping.

For high quality rendering three cameras are used and blended together. Cameras closer to the synthetic viewpoint get a higher weight. One option to achieve this is to use a simple argument based on the angle between virtual camera, real camera and the scene interest point. For the rendering we developed an OpenGL-based module that uses the view-dependant texture mapping, as described before.

In addition to the 'foreground action' a simple planar polygonal model of the pitch is inserted into the virtual scene model. The texture for this virtual pitch is computed by using a perspective projection of all available camera images into this polygon and combining these using a median filter.

## 5. RESULTS

The quality of calibration and segmentation are very important for the overall quality of the system. The line-based calibration produces an average residual error of around 1 pel and a maximal error of appr. 2 pel.

Fig. 4 & 5 show enlarged results of the football image in Fig. 1.

**Fig. 4**. Test image (left), results of fast greensegmentation (right).

The 'fast green' method (Fig. 4 right) emphasises compression artefacts, but is computationally very fast. The 'HSV' keyer (Fig. 5 left) produces slightly better detailed results.

The k-nearest neighbour classifier (Fig. 5 right) produces the best results since it can be interactively 'trained' to approximate the colour distribution of the background quite well.

Fig. 6 gives an example of a synthetic view (the goal keepers view). The scene was captured using 16 SD cameras mounted at approximately 20 m height. The novel view is generated using a visual hull reconstruction.

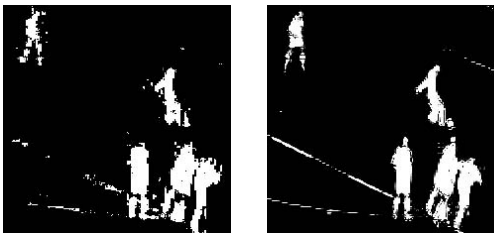

**Fig. 5**. Results of segmentation, HSV (left) k-nearest neighbour (right).

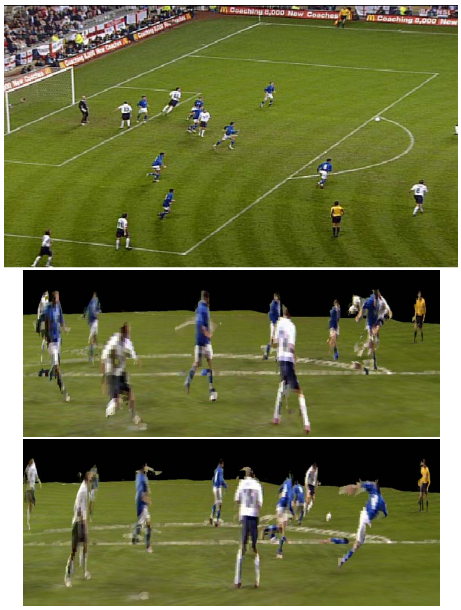

**Fig. 6**. Original camera image (top). Novel views (middle,bottom) from the goal keeper's position.

## 6. CONCLUSIONS

A system that provides a free-viewpoint video functionality of sports scenes was discussed. The system builds upon previous work done for a studio-based system. This paper discussed some of the experiences found while applying these techniques in an uncontrolled outdoor environment and addresses some of the robustness issues found.

First results show the potential of the new approach for action replay and strategy analysis of sport scenes. The visual hull technique seems to provide a robust platform for the 3D reconstruction. Current work focuses on improving the quality of the computed 3D models by improving the quality of the camera calibration and more robust 3D reconstruction algorithms.

## 7. ACKNOWLEDGEMENTS

This work has been funded by the UK DTI and EPSRC.

## 8. REFERENCES

[1] "iview," http://www.bbc.co.uk/rd/projects/iview.

[2] P. Rander, P.J. Narayanan, and T. Kanade, "Virtualized reality: Constructing time-varying virtual worlds from real events," in *Proceedings of IEEE Visualization '97*, October 1997, pp. 277–283.

[3] J. Carranza, C. Theobalt, M Magnor, and H.-P. Seidel, "Free-viewpoint video of human actors," *ACM Trans. on Computer Graphics*, vol. 22, no. 3, July 2003.

[4] Oliver Grau, Tim Pullen, and Graham A. Thomas, "A combined studio production system for 3-d capturing of live action and immersive actor feedback," *IEEE Tr. on CSVT*, vol. 14, no. 3, pp. 370–380, March 2004.

[5] J. Starck and A. Hilton, "Virtual view synthesis of people from multiple view video sequences," *Graphical Models*, vol. 67, no. 6, pp. 600–620, November 2005.

[6] N. Inamoto and H. Saito, "Fly through view video generation of soccer scene.," in *IWEC*, 2002, pp. 109–116.

[7] Oliver Grau and et al., "A free-viewpoint system for visualisation of sport scenes," in *Conference Proc. of International Broadcasting Convention*, Sept. 2006.

[8] Paul E. Debevec, George Borshukov, and Yizhou Yu, "Efficient view-dependent image-based rendering with projective texture-mapping," in *Proc. of 9th Eurographics Rendering Workshop*, Vienna, Austria, June 1998.

[9] Graham A. Thomas, "Real-time camera pose estimation for augmenting sports scenes," in *Proc. of 3rd European Conf. on Visual Media Production (CVMP2006)*, London, UK, November 2006, pp. 10–19.

[10] T. Koyama, I. Kitahara, and Y. Ohta, "Live mixed-reality 3d video in soccer stadium," *The 2nd IEEE and ACM International Symposium on Mixed and Augmented R eality*, pp. 178–186, 2003.

[11] R. Szeliski, "Rapid octree construction from image sequences," *CVGIP: Image Understanding*, vol. 58, no. 1, pp. 23–32, 1993.

[12] G. Miller and A. Hilton, "Exact view-dependent visual hulls," in *Proceedings of the 18th International Conference on Pattern Recog nition (ICPR)*, 2006.

[13] J.J.M. Kilner, J. Starck, and A Hilton, "A comparative study of free-viewpoint video techniques for sports eve nts," in *Proc. 3rd European Conference on Visual Media Production*, November 2006.