BBC

Research White Paper

WHP 180

November 2009

Stereoscopic 3D sports content without stereo rigs

Oliver Grau, Vinoba Vinayagamoorthy

BRITISH BROADCASTING CORPORATION

BBC Research White Paper WHP 180

Stereoscopic 3D sports content without stereo rigs

Oliver Grau, Vinoba Vinayagamoorthy

Abstract

This contribution describes an alternative approach to generate stereoscopic content of sports scenes from regular broadcast cameras without the need for special stereo rigs. This is achieved by using 3D reconstruction previously developed for applications in post-match analysis. The reconstruction method requires at least 4-5 cameras and computes the 3D information automatically.

Two different target formats for the delivery of stereoscopic 3DTV are discussed: The display independent LDV format and conventional binocular stereo. The results viewed on two different displays demonstrate the potential of the method as an alternative production method for stereoscopic 3D content.

This document was originally published in Proceedings of IBC 2009.

Additional key words: Stereoscopic content production, 3D reconstruction

STEREOSCOPIC 3D SPORTS CONTENT WITHOUT STEREO RIGS

Oliver Grau, Vinoba Vinayagamoorthy

BBC R&D, UK

ABSTRACT

This contribution describes an alternative approach to generate stereoscopic content of sports scenes from regular broadcast cameras without the need for special stereo rigs. This is achieved by using 3D reconstruction previously developed for applications in post-match analysis. The reconstruction method requires at least 4-5 cameras and computes the 3D information automatically.

Two different target formats for the delivery of stereoscopic 3DTV are discussed: The display independent LDV format and conventional binocular stereo. The results viewed on two different displays demonstrate the potential of the method as an alternative production method for stereoscopic 3D content.

INTRODUCTION

Due to the successful re-introduction of stereoscopic 3D (S3D) in the cinemas there is recently also a growing interest in broadcast industry in S3D content. In particular the coverage of live sports events like football or rugby is a focus of interest. Unfortunately the production of stereo content increases costs significantly. To add S3D to the regular broadcast each camera position would need to be equipped with a pair of cameras in a special stereo rig. That would increase the costs beyond simply doubling the camera budget, since it requires special skills to set-up, to operate and additional broadcast infrastructure, e.g. in OB vans.

At the current state S3D content is captured with special stereo rigs, which are either built up using miniature cameras or through the use of a special mirror rig, so that the content can be captured with an inter-ocular distance of approximately 6.5 cm, which represents the average eye distance of the population. Special camera systems, like boxed super-zoom-lenses or high-speed cameras that are commonly used for the coverage of sports content cannot be easily used in a stereo rig. For this and other reasons, it is very likely that S3D productions will only be able to share some of the conventional broadcast equipment and will be operated mostly alongside the conventional broadcast coverage.

In this contribution we present an alternative production method, which derives 3D stereoscopic content from the normal broadcast coverage cameras, plus optional locked-off cameras. The approach makes use of multi-camera 3D reconstruction techniques that have been previously developed for post-match analysis of sports scenes [1+2]. These techniques perform an automatic camera calibration and segmentation of the foreground action. From this information a 3D model of the foreground action is computed. This data is then converted into a 3DTV format. The *3D4You* project [3] investigates formats that are independent from the 3D display as it stores depth information alongside the image, allowing the generation of either stereoscopic or multi-view for different kinds of 3D displays. Alternatively, the data can be stored as a conventional S3D image pair for the left and right eye with a fixed inter-ocular distance.

The rest of this paper is structured as follows. The next section gives a very brief overview of our approach, followed by a more detailed description of the processing modules. The section following that discusses 3DTV delivery formats and aspects of the method that has been developed to convert the 3D action into these formats. The paper finishes with a description of first results and some conclusions.

OVERVIEW OF THE SYSYEM

As depicted in Figure 1 the system has three component blocks:

1. The capture system is fully integrated into the OB[1] infra-structure and performs a live ingest of the multi-camera video streams. For our tests we implemented the capture system with standard IT server hardware equipped with HD-SDI frame grabber cards.

[1] outside broadcast

2. The processing block automatically computes 3D information from the multi-camera streams.
3. The format conversion converts the image- and 3D-data into a 3DTV delivery format.

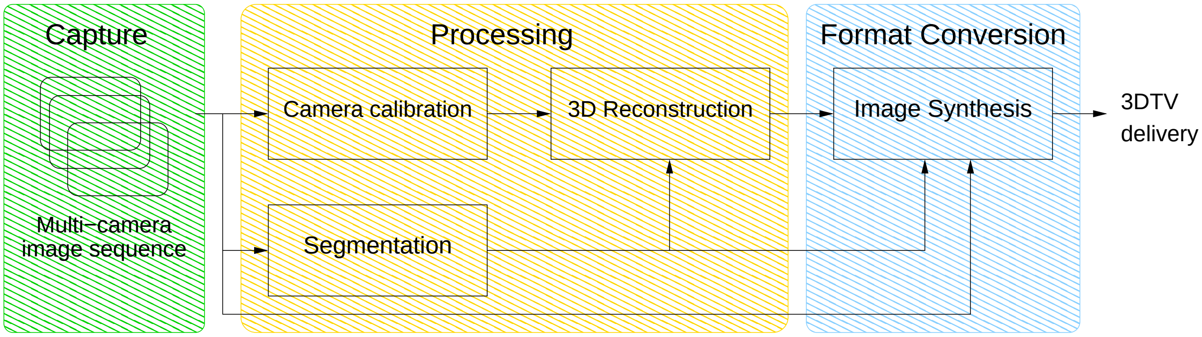

Figure 1 System overview

In the current experimental implementation the video streams are stored to disk while the processing and format conversion are run offline to produce the stereoscopic 3DTV content.

PROCESSING MODULES

Camera calibration

The camera calibration determines the parameters of the broadcast cameras. Our approach does this automatically from the images without any additional special hardware (like pan-tilt heads). The method detects the pitch lines of a sports ground, the dimensions of which need to be known. From this information the camera position, orientation and internal parameters (focal length and lens distortions) are determined. A detailed description of the calibration method can be found in [4].

Foreground Segmentation

The foreground segmentation separates the foreground action from the background, i.e. the pitch and other background objects like stands.
For the segmentation of players, colour-based methods such as chroma-keying against the green of football and rugby pitches have been considered. However, the colour of grass varies significantly on pitches. This is due to uneven illumination and anisotropic effects in the grass caused by the process of lawn-mowing in alternating directions. Under these conditions chroma-key gives a segmentation that is too noisy to achieve a high-quality visual scene reconstruction. Therefore two improved methods have been implemented and tested: a global colour-based ‘k-nearest-neighbour classifier' classifier (KNN) and a motion compensated difference keyer. Details of the KNN classifier can be found in [2].

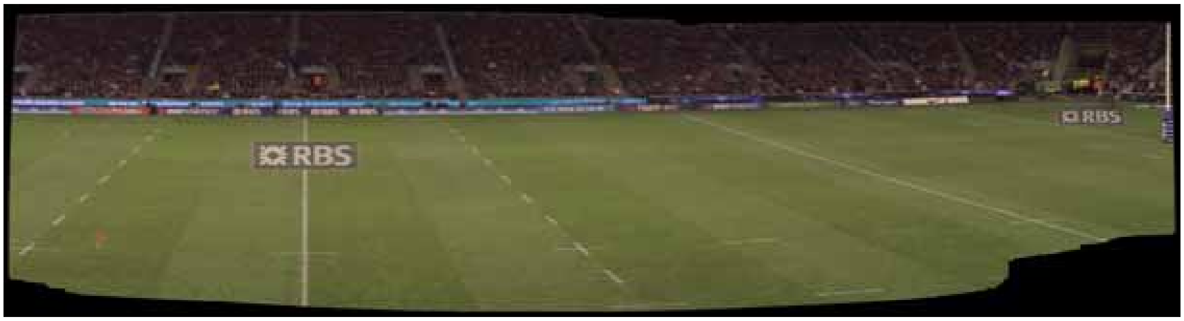

Figure 2 Spherical background plate of a rugby pitch

The difference-keyer derives a key based on the difference in colour of a pixel in a camera image to a model of the background that is stored as a background plate. This method is usually applied to static cameras. However, under known nodal movement of the camera, a background plate can be constructed by piecewise projection of the camera images into a spherical map. This transformation is derived from the camera parameters, as computed in the camera calibration. A plate clear of foreground objects is created by applying a temporal median filter to the contributing patches. An example background plate is shown in Figure 2. The actual difference key is then derived from the difference of a pixel in the current camera image to its reference in the spherical map, taking into account the camera movement and zoom. The latter parameters are known from the camera calibration.

The difference keyer is superior in most cases to a global colour-based approach. It is even able to separate between foreground and mostly stationary but cluttered backgrounds and is used for the results presented in this paper. The colour-based approach on the other hand is relatively simple to implement, but limited mainly to the segmentation of players against the pitch. Figure 3 shows an example difference key for a moving camera.

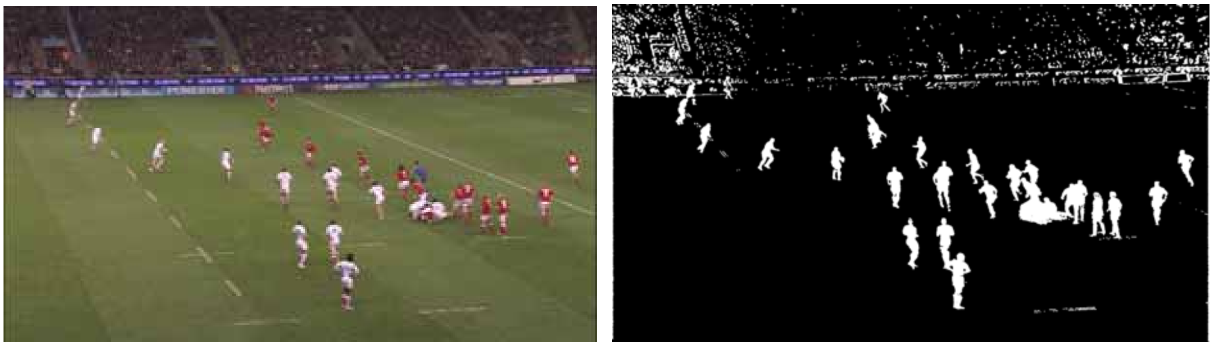

Figure 3: Camera image (left) and difference key (right).

3D Reconstruction of foreground action

The 3D reconstruction uses a visual hull or shape-from-silhouette approach to compute 3D models of the scene action (for more details see [2]). For the rugby scene shown in Figure 3, this is done for an area of 50 x 50 x 3 m (width x depth x height) with a volumetric resolution of 512 x 512 x 16. This only gives a relatively coarse 3D surface model, but the image rendering makes use of the keys generated by the segmentation and uses them as an alpha channel to get a better shape of the boundaries. Figure 4 shows a 3D model overlaid onto the original camera image.

A limitation of using only 3 m height is that the ball will not be represented when it is kicked higher than this. Using a higher volumetric area is not without problems. One problem is the increase in the computational effort required. Another problem is that the only areas of the active volume that can be reconstructed are those seen by at least one camera. Since the 'higher' areas are only marginally of interest during most of a game these areas would not been covered by many cameras. The approach taken to tackle this in the iview project was to model the ball in a separate pass: an operator sets the active area manually to roughly where the ball is. Although this is acceptable for post-match analysis, it would be desirable to automate this approach for S3D coverage.

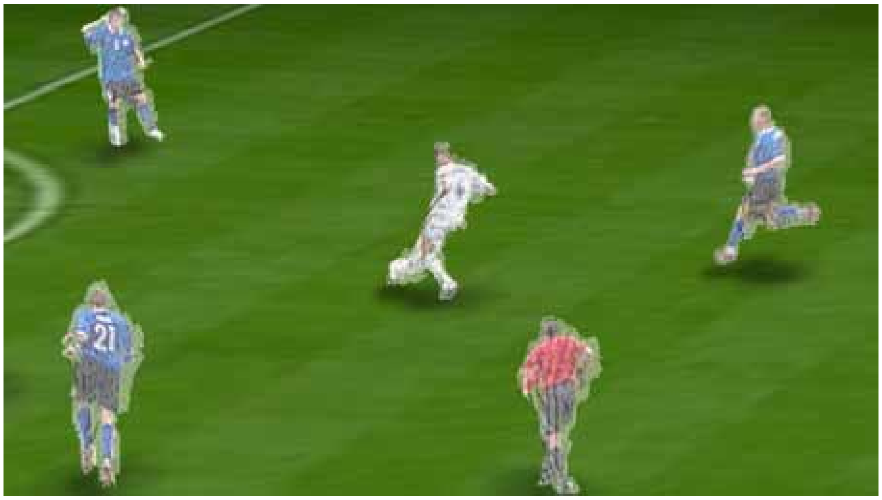

Figure 4: Wire-frame overlaid 3D model

3D Reconstruction of background objects

The generation of stereoscopic images also requires the 3D geometry of background objects. The pitch can be approximated by a planar polygon as its dimensions are known, as a pre-requisite for the camera calibration step. In comparison to the foreground action, the stadium is further away and for our initial tests we approximated it with very few polygons. This can be done with a CAD or post-production modelling tool. If more detailed models are needed then these have to be aligned to the images. This can be achieved with image-based modelling tools, which is a subject of our current investigations.

The modelling of the background objects can be done offline in preparation of a game and needs to be done only once per location.

CONVERSION TO STEREOSCOPIC FORMAT

After reconstruction of foreground action and background, these objects are made available as 3D descriptions in form of a polygonal surface mesh. This scene description is then the basis for the conversions into a 3DTV format.

There is currently no regular 3DTV service available and moreover there is, as of yet, no standard format. A likely candidate is a side-by-side aligned stereo image pair (binocular stereo). One implementation could be to scale the left and right image 50% horizontally and merge them into a new image which is of the same size as the original ones. This will reduce the horizontal resolution of the images, but demands no further changes in the rest of the transmission chain other than a display capable of handling this format on the viewer side. The viewer cannot change the depth scaling at his end as the inter-ocular distance is fixed with this format.

Alternative approaches for 3DTV delivery are investigated in the 3D4You project [3], with the goal to define a display-independent delivery format. One option is to use one video stream plus depth information (stored like an alpha-channel). An extension of this 'video-plus-depth' is 3D-Layered Depth Video (LDV), which is constructed by adding additional occlusion layers. The end product is one central camera view with all video and depth data from multi-view capturing mapped during post-production. This format is very compact and efficient in terms of data compression and low-complexity rendering on a 3D display.

The conversion of the 3D data acquired by our approach into the LDV or conventional binocular stereo formats is achieved by synthesising the required information from the reconstructed 3D scene description.

Conversion to LDV

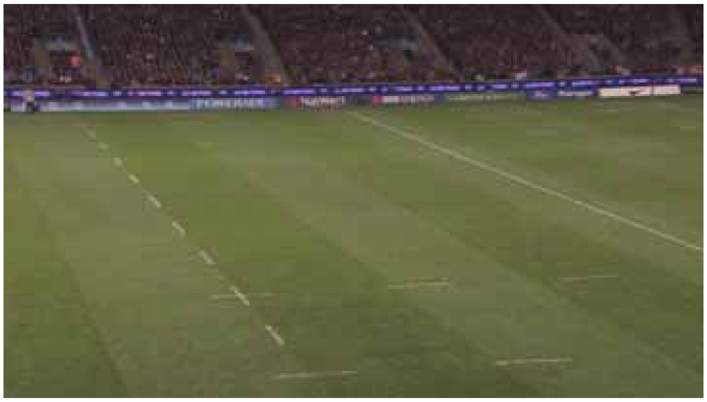

Figure 5: Background layer

To generate the LDV format the original camera images are augmented by a depth channel and by a background layer. The resulting data was viewed on a Philips WOWvx display [5]. The depth information was generated by rendering the 3D scene description with a scan-line renderer. Our implementation is based on OpenGL. The depth information is retrieved by reading out the z-buffer.

The depth values in LDV (and image-plus-depth) are mapped to Byte-Integers. The depth range is therefore limited to a minimum and maximum depth value in a particular scene. The depth values are converted to disparity values within a range of 0 and 255, where a value of 0 corresponds to objects located at the maximum disparity behind the screen while 255 corresponded to objects closest to the observer [5]. A disparity value of 128 corresponded to objects on the screen. The disparity values of objects such as the goal posts in Figure 6 were constructed with disparity values between 128 and 255 in order to project them as foreground objects to the observer.

The background layer is an image from the same camera angle as the camera image, but without the foreground action. One option to generate this layer is to fill in the obscured background from information taken from other cameras. One problem found with this procedure is that the areas might look slightly different in colour, due to different colour characterises of the cameras (mismatched colour balance) or anisotropic effects - see remarks about segmentation above. Furthermore, the occluded area might not been visible in other cameras with the same degree of detail.

Another approach to fill un-revealed areas is to use background images generated from the spherical background plate as used in the difference keying technique described above. Although the generated images are slightly less detailed and miss shadows, the results of this approach look quite promising. Figure 5 shows the generated background layer for the camera angle of the scene depicted in Figure 3.

The resulting 2D, depth and background layer sub-images are merged into one image in the required format to provide a 1920 x 1080 image for each frame to be played on a 42" Philips WOWvx display [5].

Conversion to binocular stereo

For binocular stereo the original, monoscopic camera view is used as the left view and the image for the right view is synthesised by rendering from a camera viewpoint that is laterally offset by the intro-ocular distance. An advantage of this approach is that the intraocular distance can be freely varied. This is of particular interest since the optimal distance depends on the screen size and viewing distance. Screens in the cinema require usually a different inter-ocular distance than for domestic TV viewing.

RESULTS

Three short sequences, with a total length of about a minute, taken from a pre-recorded rugby game have been used to test the approach. A number of broadcast cameras from that game were captured and four camera positions were then used. A further 6 additional locked-off cameras were recorded to complement the broadcast cameras. From this set of 10 cameras, a 3D model of the action was computed.

The 3D data set was then converted into image-plus-depth, LDV and binocular stereo. Figure 6 shows an example frame from that data set from different camera angles. The image-plus-depth and LDV were displayed on an auto-stereoscopic Philips WOWvx display. The depth augmentation works very well, resulting in a good visual quality. The lack of detail in the background objects (stadium) is not disturbing.

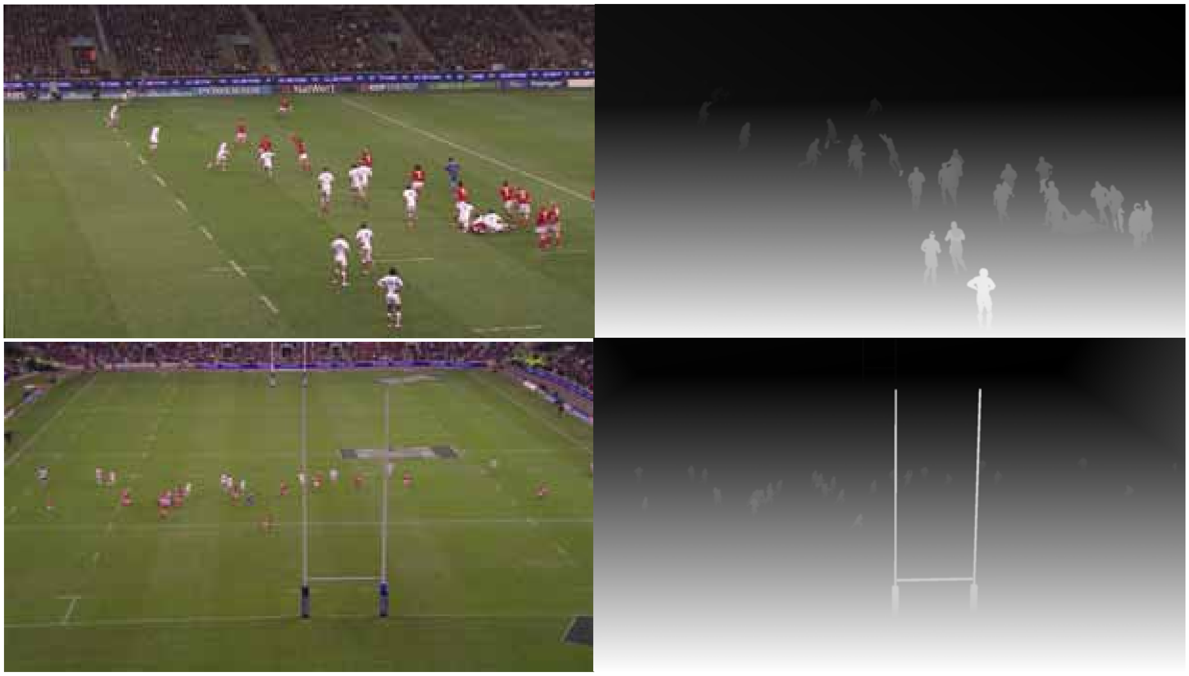

Figure 6 Images from a rugby game (left) and synthesised depth (right)

The binocular data was viewed on a 19 inch Trimon ZM-M190 display using polarisation glasses. Compared to the Philips display the image resolution is higher and artefacts are more clearly visible. In particular it would be desirable to have finer detail in the background model as it can be seen that parts are too flat.

CONCLUSIONS

This paper described an alternative approach to generate stereoscopic content of sports scenes from regular broadcast cameras. This is achieved by using 3D reconstruction previously developed for applications in post-match analysis and synthesis of the target 3DTV format. The reconstruction method needs at least 4-5 cameras to generate the 3D information automatically. This number of cameras is usually available in the coverage of high-end games. In order to improve the robustness and quality of the system additional locked-off cameras can be used.

The availability of a 3D description of the scene has the major advantage that different target formats can be handled. That includes the display-independent delivery formats discussed, like LDV, or the ability to render binocular stereo sequences with different, fixed inter-ocular distances for different end-terminals.

Further research is currently being carried out to make the methods more robust so that the processing would run fully automated. Although the experimental implementation of the system runs currently offline it is believed that the algorithms used could be optimised and run in real-time.

ACKNOWLEDGEMENTS

This research was supported by the TSB project iview and FP7 3D4You.

REFERENCES

1. The iview project, http://www.bbc.co.uk/rd/projects/iview

2. O. Grau and G.A. Thomas and A. Hilton and J. Kilner and J. Starck, A Robust Free-Viewpoint Video System for Sport Scenes. Proceeding of 3DTV conference 2007, Kos, 2007.

3. IST 3D4You project, http://www.3d4you.eu

4. G.A. Thomas. Real-time camera tracking using sports pitch markings, Journal of Real Time Image Processing, Vol. 2, No. 2-3 , pp. 117-132, November 2007.

5. 3D Interface Specifications of the WOWvx display, White paper Philips 3D Solutions, 15 February 2008.