3D-TV Production from Conventional Cameras for Sports Broadcast

Adrian Hilton, *Member, IEEE,* Jean-Yves Guillemaut, *Member, IEEE,* Joe Kilner, *Member, IEEE,* Oliver Grau, *Member, IEEE,* and Graham Thomas, *Member, IEEE*

***Abstract*—3DTV production of live sports events presents a challenging problem involving conflicting requirements of maintaining broadcast stereo picture quality with practical problems in developing robust systems for cost effective deployment. In this paper we propose an alternative approach to stereo production in sports events using the conventional monocular broadcast cameras for 3D reconstruction of the event and subsequent stereo rendering. This approach has the potential advantage over stereo camera rigs of recovering full scene depth, allowing inter-ocular distance and convergence to be adapted according to the requirements of the target display and enabling stereo coverage from both existing and 'virtual' camera positions without additional cameras. A prototype system is presented with results of sports TV production trials for rendering of stereo and free-viewpoint video sequences of soccer and rugby.**

***Index Terms*—3DTV, 3D video, free-viewpoint video, multiple view reconstruction, camera calibration, image segmentation**

I. Introduction

Recent box-office success of stereo 3D movies has led to an increased demand for the production and delivery of 3DTV content with events such as the 2010 soccer world-cup being used to promote 3DTV services. Sports broadcast provides a context in which the viewing experience may be enhanced by stereo 3D. Currently stereo broadcast production of live sports events is achieved by using multiple stereo camera rigs alongside the conventional 2D cameras. In the 2010 world-cup seven additional stereo camera rigs were used in each stadium to provide 3DTV coverage. Production using dedicated stereo camera rigs significantly increases costs and introduces technical problems in maintaing accurate alignment of zooming stereo camera pairs. The resulting stereo footage also has a fixed inter-ocular distance and convergence at the time of capture. This prevents stereo adjustment in post-production for retargeting to displays of different size to maintain a consistent depth perception. Changing the inter-ocular distance in stereo camera views requires knowledge of the scene depth which is challenging to reconstruct from the closely spaced stereo camera pair.

In this paper we present an alternative approach to stereo 3DTV production based on multiple widely spaced monocular cameras such as those used in a conventional 2D soccer broadcast. The approach directly reconstructs the scene geometry from the camera views. Recent advances in multiple view reconstruction are exploited to provide a 3D scene proxy for subsequent stereo rendering. This allows full control of stereo rendering parameters in post-production avoiding any distortion between camera views due to camera zoom lens distortion or misalignment. Stereo inter-ocular distance and convergence are controlled in post-production allowing retargeting to different displays. The challenge in this approach is to maintain the visual quality of conventional 2D broadcast whilst allowing stereo rendering.

A prototype system for stereo 3DTV and free-viewpoint sports broadcast production is presented which uses the existing monocular match cameras. Production trials for soccer and rugby demonstrate a visual quality comparable to conventional 2D production with full control over viewpoint and stereo rendering in post-production. If the stereo pair is rendered from the viewpoint of a broadcast camera the 2D video can be augmented with 3D depth information from the 3D scene proxy allowing stereo rendering without loss of visual quality. This system demonstrates the potential of this approach for use in 3DTV production without a requirement for additional stereo camera rigs.

To date most multiple view video systems have been developed for studio applications with a fixed capture volume, controlled illumination and backgrounds. Live outdoor events such as sports present a number of additional challenges for both acquisition and processing. Multiple view capture systems in sports such as soccer must cover the action taking place over an entire pitch with video acquisition at sufficient resolution for 3D scene analysis and production of desired stereo or free-viewpoint virtual camera views. The system presented in this paper is based on use of the live broadcast cameras as the primary source of multiple view video. In a conventional broadcast for events such as premier league soccer these cameras are manually operated to follow the game play zooming in on events as they occur. Advances are presented in real-time through the lens camera calibration to estimate both the camera pose, focus and lens distortion from the pitch lines. A 3D scene proxy is then reconstructed at each frame starting with a volumetric reconstruction followed by a view-dependent refinement using information from multiple views. Production trials for both international soccer and rugby matches provide a qualitative evaluation of both stereo and free-viewpoint rendering from the conventional 2D broadcast match camera input. Results demonstrate stereo 3D and free-viewpoint rendering with a visual quality comparable to the captured monocular broadcast video.

A.Hilton, J.-Y.Guillemaut, J.Kilner: Centre for Vision, Speech and Signal Processing,University of Surrey, UK *a.hilton,j.guillemaut,j.kilner@surrey.ac.uk*

O.Grau, G.Thomas: BBC Research & Development, UK *oliver.grau,graham.thomas@bbc.co.uk*

II. BACKGROUND

A. Methodologies for Stereo and Free-viewpoint Video

Two principal methodologies have been investigated to rendering novel views of scenes captured from two or more camera viewpoints: interpolation and 3D reconstruction.

Interpolation: View interpolation directly estimates the scene appearance from novel viewpoints without explicitly reconstructing the 3D scene structure as an intermediate step [1], [2], [3]. This avoids the requirement for explicit 3D reconstruction but is in general limited to rendering viewpoints between the camera views. Interpolation has the advantage of circumventing inaccuracies in explicit reconstruction due to errors in camera calibration. The quality of rendered views is dependent on the accuracy of correspondences used to align multiple view observations. Extrapolation of novel views has also been investigated based on the colour consistency of observations from multiple views without explicit reconstruction [4]. A comprehensive survey of image-based rendering techniques for novel view synthesis is given in [5].

Reconstruction: Reconstruction of the 3D scene structure from multiple view images is commonly used as a basis for rendering novel views. Given multiple views of a dynamic scene such as a moving person a number of approaches have been used for reconstruction: visual-hull; photo-hull; stereo; and global shape optimisation. Visual-hull reconstruction intersects silhouette cones from multiple views [6], [7], [8], [9], [10] to reconstruct the maximal volume occupied by the scene objects. This requires prior segmentation of the foreground scene objects, such as a person, from the background. The photo-hull [11] is the maximal photo consistent volume between multiple views. An advantage of the photo-hull is that it does not require prior segmentation of the foreground. Stereo reconstruction from multiple views has been used to reconstruct dynamic scenes of moving people [12], [13]. Dense correspondence from stereo ensures reconstruction of surfaces which align the multiple view images reducing artefacts in rendering of novel views. However, stereo correspondence requires local variation in appearance across the scene surface and is ambiguous in regions of uniform appearance. To overcome this limitation research has investigated the combination of volumetric and stereo reconstruction in a global optimisation framework to ensure robust reconstruction in areas of uniform appearance and accurate alignment of images from multiple views [14], [15]. A comparison of approaches for reconstruction of static scenes from multiple views is presented in [16].

B. Multiple View Studio Reconstruction

Over the past decade there has been extensive research in multiple camera systems for reconstruction and representations of dynamic scenes. Following the pioneering work of Kanade et al.[12] introducing Virtualized Reality^{TM} there has been extensive research on acquisition of performances to allow replay with interactive control of a virtual camera viewpoint or *free-viewpoint video*. This system used 51 cameras over a 5 meter hemisphere to capture an actors performance. Reconstruction is performed by fusion of stereo surface reconstruction from multiple pairs of views. Novel viewpoints are then rendered by texture mapping the reconstructed surface. Other multiple camera studio systems with small numbers of cameras (6—12) have used the visual-hull [17], [18] and photo-hull [19]. Real-time free-viewpoint video with interactive viewpoint control has also been demonstrated [17], [20].

Recent advances have achieved offline production of free-viewpoint video with a visual quality comparable to the captured video. Zitnick et al.[13] presented high-quality video-based rendering using integrated stereo reconstruction and matting with a 1D array of 8 cameras over a 30° arc. Results demonstrate video-quality rendering comparable to the captured video for novel views along the 30° arc between cameras. High-quality rendering for all-round 360° views has also been demonstrated for reconstruction from widely spaced views (8 cameras with 30-45° between views) using global surface optimisation techniques which integrate silhouette constraints with wide-baseline stereo [14], [15], [21]. This approach refines an initial visual-hull reconstruction to obtain a surface which gives accurate alignment between widely spaced views.

C. Multiple View Reconstruction in Sports

Initial attempts have been made to transfer studio-based reconstruction methodologies to acquisition and reconstruction of outdoor events. The Virtualized Reality^{TM} technology [22] was used in the EyeVision[1]system to produce virtual camera sweeps as action replays for Super Bowl XXXV in 2001. Thirty motorised camera heads slaved to a single manually controlled camera were used to produce sweep shots with visible jumps between viewpoints.

More recently a number of groups have investigated volumetric [23] and image-based interpolation techniques [24], [25], [26] for rendering novel views in sports. Grau et al. [23] used a texture mapped visual-hull reconstruction from 15 camera of a soccer pitch to render novel views of the players. Interpolation of novel views between the real cameras without explicit 3D reconstruction has also been investigate in the context of sports. [24], [25], [26]. Inamoto and Saito [25] allow free-viewpoint video synthesis in soccer by segmenting the observed camera images into three layers: dynamic foreground (players); pitch; and background (stadium). Morphing is achieved by interpolation along the corresponding intervals of the epipolar line for the foreground layer. A layered representation for the spatio-temporal correspondence and occlusion of objects for pairs of views is proposed in [24] and applied to soccer view interpolation.

Germann et al. [27] recently proposed *articulated billboards* as an intermediate 3D proxy for rendering novel viewpoints of players in sports matches. This approach only requires two views in which the players skeletal pose is aligned in order to interpolate intermediate views based on a billboard representation of each body part. Currently manual interaction

[1]EyeVision www.ri.cmu.edu/events/sb35/tksuperbowl.html

is required to pose the articulated billboard limiting the use to free-viewpoint rendering of players in a single static frame or short sequences.

LiberoVision[2]recently introduced a commercial system for interpolation between pairs of match camera views in soccer broadcast. This system has the advantage of only using the existing broadcast cameras. The Piero[3] system developed by BBC R&D allows annotation of the broadcast video footage together with limit change in viewpoint using player billboards to extrapolate views around a single camera.

The system presented in this paper aims to allow 3D proxy reconstruction for stereo and free-viewpoint rendering of live action from informative viewpoints which add to the broadcast coverage of a sporting event. The system allows rendering of viewpoints on the pitch such as the referees or goal keepers view of events using the broadcast match cameras together with additional auxiliary cameras to increase coverage if available. This system has introduced automatic methods for online calibration, segmentation, reconstruction and rendering of stereo and free-viewpoint video for sports broadcast production.

III. Specification of Requirements for Sports TV

There are three critical issues for use of stereo and free-viewpoint video in sports TV broadcasts: visual quality; timing; and cost. In this section we identify the requirements and constraints for use of free-viewpoint video in sports TV production.

A. Visual Quality for Broadcast Production

The hardest technical constraint for stereo and free-viewpoint video of novel views in TV sports production is visual quality. Broadcast video quality equivalent to the live footage from the broadcast cameras is ideally required to be acceptable to the viewing public. In stereo production artefacts caused by a misalignment or distortion of views from stereo camera pairs may result in loss of depth perception and discomfort to viewers. The production of stereo views in post-production avoids the artefacts in stereo capture due to mismatching of zooming cameras but may introduce additional distortion due to errors in reconstruction of the 3D scene proxy. It is therefore important that the 3D proxy used for stereo rendering is sufficiently accurate to render the scene without visual artefacts. High-definition (HD) cameras are now widely used for acquisition at live sports events together with increasing use of HD transmission to the viewer. Stereo rendering needs to achieve HD quality for rendering of full-screen shots.

B. Production Requirements on Timing

The time taken to produce stereo or free-viewpoint video is critical to the potential uses in broadcast. From a production standpoint stereo video would ideally be available at video broadcast rate ($<$ 40ms/frame) with 100% reliability on visual quality allowing the sports producer to select stereo video streams for 3DTV broadcast as with conventional 2D broadcast. In practice due to both algorithm reliability and computational delay there are four critical time points where stereo and free-viewpoint video could be exploited in production.

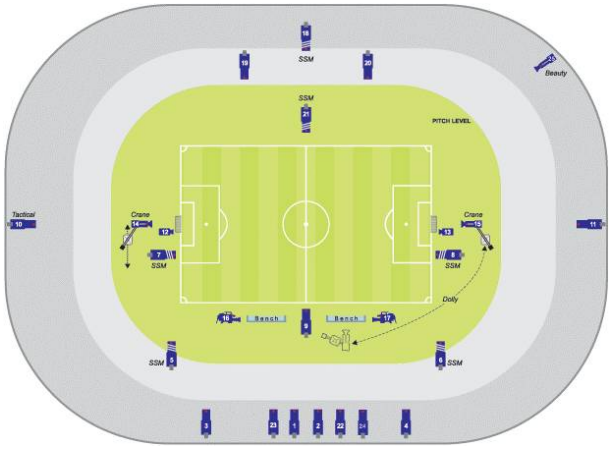


Fig. 1. Typical stadium broadcast camera layout for a major sporting event. Out of 26 cameras 1-4,10,11,19,20 provide potentially useful views, 5-8,18,21 are high-speed cameras and the remainder are at pitch level providing insufficient coverage for calibration or reconstruction.

Live Action: Video-rate capture, reconstruction and rendering for live stereo 3DTV broadcast with only a few frames delay.

Action Replay: Within seconds of an event happening (e.g. a player is fouled), and a novel view is offered in place of conventional *instant* replays.

Action Review: Within a few minutes, such that a novel view can be presented during half time or immediately after a match finishes.

Match Analysis: After many minutes, such that a novel view sequence made available for use as part of a post-match analysis programme (which may be later that day or week).

C. Acquisition Requirements

Production of sports events such as soccer for live broadcast typically use 12-18 match cameras at key locations around the stadium. In the 2006 FIFA World Cup[4] 26 HD cameras were used for coverage at each stadium as illustrated in Figure 1. The main broadcast cameras are typically located one side and on the ends of the stadium to avoid disorienting the viewer with switches to reverse views. All cameras are manually controlled by individual operators to cover both the action on the pitch and the crowd. This leads to the problem that even with 15-20 broadcast cameras only a small number will be covering the same area during normal play. Figure 2 shows a typical set of shots from match cameras for a penalty event during a cup-final match.

For stereo production in the 2010 world-cup a set of seven manually operated stereo camera pairs were located at the

[2]LiberoVision www.liberovision.com

[3]Piero www.bbc.co.uk/rd/projects/virtual/piero/

[4]http://www.fifa.com/worldcup/archive/germany2006/news/newsid=13449.html

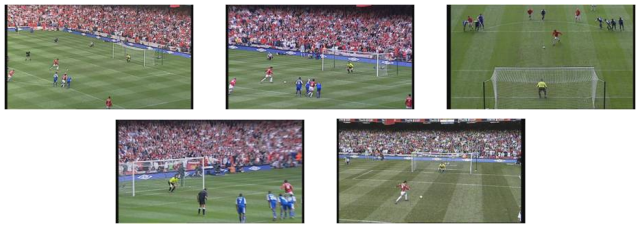


Fig. 2. Broadcast match camera views for a penalty event during a cup-final. Of the 15 match cameras only the views shown were of the penalty.

stadia in addition to the conventional broadcast cameras. Video-rate processing of the stereo streams is employed to correct for distortions of left and right camera views due to lens differences which occur during zooming. Direct stereo capture and processing enables live broadcast of 3DTV content with a fixed inter-ocular distance. The use of stereo camera rigs in addition to the conventional broadcast cameras considerably increased the production cost which is only viable for high-profile events.

IV. A Stereo 3D Video System for Sports TV

This section presents a system for stereo and free-viewpoint video in TV production of sports events. The system has been developed to utilise footage from both the manually operated monocular match cameras and fixed auxiliary cameras if available to ensure full stadium coverage. Automatic camera calibration from the pitch markings has been developed to allow combination of footage from multiple camera views, including the moving and zooming match cameras. In this section we review both the acquisition system and algorithms developed to facilitate broadcast quality stereo 3D production.

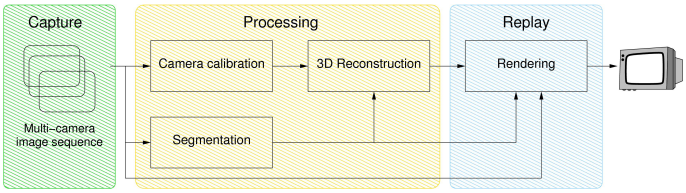


Fig. 3. Overview of the 3DTV video system.

A. System Overview

An overview of the 3DTV production system is shown in Figure 3. Capture is performed using time synchronised acquisition from both auxiliary and match cameras. Synchronisation using genlock is a standard process in conventional broadcast acquisition. Uncompressed camera footage is stored directly to disk for offline processing. Automatic calibration of all cameras is performed from the pitch lines of the captured footage. This avoids the need for prior camera calibration and allows the use of footage from match cameras. The calibration is capable of real-time operation for use during live match footage. Calibration estimates the extrinsic and intrinsic parameters of each camera including lens distortion. Matting of foreground (players) from the background (pitch) is performed using chroma and difference key matting. This allows the approximate segmentation of the foreground players for subsequent processing to produce a 3D scene proxy for stereo rendering and free-viewpoint video. A robust 3D scene reconstruction and segmentation refinement is then performed taking into account errors in calibration (1-2 pixel rms) and initial segmentation (2-3 pixels). View-dependent rendering is then performed to render stereo 3D and free-viewpoint video.

B. Video-rate Calibration of Live Broadcast Footage

One way in which camera calibration data can be derived is by performing an initial off-line calibration of the position of the camera mounting using a theodolite or range-finder, and mounting sensors on the camera and the lens to measure the pan, tilt, and zoom. However, it is often the case in international sports that the cameras are controlled by a host broadcaster and only access to the match camera feeds is available. A more attractive way of deriving calibration data is by analysis of the camera image sequence. The lines on a sports pitch are usually in known positions, and these can be used to compute the camera pose. In some sports, such as soccer, the layout of some pitch markings (such as those around the goal) are fully specified, but the overall dimensions vary between grounds.

For free-viewpoint video in sports TV production we have developed a real-time (50-60Hz) camera pose estimation system for the live broadcast cameras[28], [29]. The online calibration estimates the camera position, orientation, focal length and lens distortion from the match footage using pitch markings. Camera calibration is computed to minimise the reprojection error of observed pitch lines. Calibration is based on a multi-hypothesis line tracking approach using edge points closest to the predicted line position. This provides robustness to the appearance of other nearby edge points. The method includes an automatic initialisation process which takes about one second to evaluate. The stages of the calibration process are as follows:

Initial Estimation of Camera Position: As broadcast cameras are commonly mounted in a fixed location on a pan and tilt head an initial estimate of the camera position is obtained from multiple images with a wide range of camera orientations. The camera position, orientation and field-of-view is estimated with the position constrained to a common value for all images. This initial calibration from images over a wide range of orientations significantly reduces the ambiguity between distance of the camera from reference features and the focal length. Figure 4 shows an example of camera positions computed first from 42 individual images, and then with all images used simultaneously to compute a common position. The position is computed by considering all images simultaneously and lies roughly where these lines of uncertainty cross. Repeating the process with different sets of images showed that the position could be estimated with a consistency of about 0.3m (about 0.4% of the distance from the camera to the centre of the pitch), giving an indication of the accuracy which is use to initialise online calibration.

Initialisation: Before the tracker can be run at full video rate, it is necessary to initialise it by determining roughly what its values of pan, tilt and field-of-view are. This process needs to be carried out when the tracker is first started, and also whenever it loses track (for example if the camera briefly zooms in tightly to an area with no lines). A Hough transform is used to quickly establish how well the image matches the set of lines that would be expected to be visible from a given pose. An exhaustive search process is used to establish the pose which gives the best line matches to the observed image. For each pre-determined camera position, we search over the full range of plausible values of pan, tilt, and field-of-view, calculating the match value by summing the values in the bins in the Hough transform that correspond to the line positions that would be expected. Figure 5 presents examples of the Hough space corresponding to the single frame shown.

Tracking: The tracking process uses the pose estimate from the previous image, and searches a window of the image centred on each predicted line position for points likely to correspond to pitch lines. A simple line detection filter uses knowledge of the predicted line width and orientation to produce a measure of the extent to which each pixel looks like it may be at the centre of a pitch line. Pixels having a filter output above a given threshold are hypothesised to be candidates for lying on the line, and those closest to the predicted line position are used. An iterative process is used to adjust the camera pan, tilt and focal length in order to minimise the distance in the image between the detected line points and the projection into the camera image of the corresponding lines in the model. The approach can also estimate lens distortion, although reliable values can only be computed when one or more long lines are visible that do not pass close to the image centre, and any curvature in the pitch is known. An alternative approach to determining lens distortion is to assume that there is a fixed relationship between distortion and focal length, and use multiple images to solve for the values of a few coefficients relating distortion to focal length, for example when computing the initial camera position or as a separate off-line lens calibration process. Figure 6 shows the estimated camera angles for a 20s sequence (in the absence of any lens distortion correction). The second derivative of the pan angle gives an indication of the level of noise, as the true movement is inherently smooth. The spikes are caused by lines coming into or out of view, mainly due to small differences between the assumed and true positions of the lines on the pitch. Evaluation of the real-time tracking on longer sections of sports footage [28] shows that the noise error in the estimated pan angle is approximately 0.02° which typically corresponds to about 1 pixel.

C. Foreground Segmentation

For the segmentation of players colour-based methods, like chroma-keying against the green of soccer and rugby pitches have been considered. However, the colour of grass varies significantly on pitches. This is due to inhomogeneous illumination in the uncontrolled environment and anisotropic effects in the grass caused by the process of 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 and a motion compensated difference key. After segmentation we compute the foreground/background colours for 'mixed pixels' with a method similar to that described by Hillman in [30].

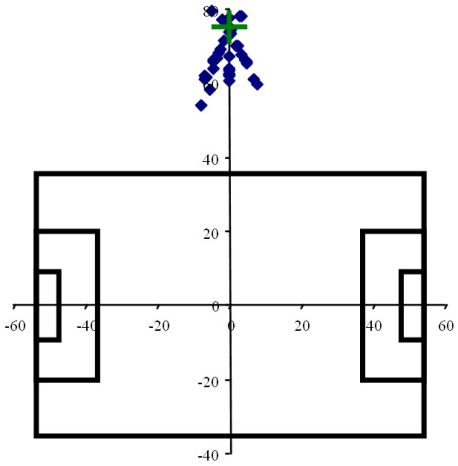


Fig. 4. Camera positions estimated from pitch lines in the image, using individual images (blue diamonds) and simultaneously with all images (green cross)

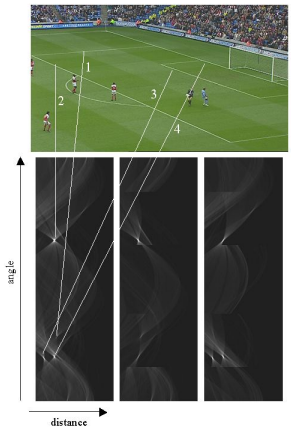


Fig. 5. Calibration initialisation using Hough space: (top) original image, (bottom) Hough space for full image(left), top or left half of image (centre) and bottom or right half(right)

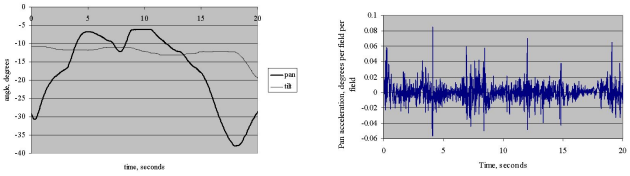


Fig. 6. Calibration parameters for a 20s sequence (left) Pan and tilt angles (right) second derivative of pan

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 in RGB space r_1 of the colour prototype are then marked as background. The user continues to select background pixels until the resulting segmentation is satisfactory. The segmentation $S_{k-nearest}$ is computed by finding the nearest colour prototype P_{best} from the list with the minimum RGB colour distance d of the pixel values I:

$$d = argmin_I\{d_{rgb}(P_{best}, I)\} \tag{1}$$

The segmentation is then given by:

$$S_{k-nearest} = \begin{cases} 0 & , d < r_1 \\ 1 & , \text{otherwise} \end{cases} \tag{2}$$

In order to get continuous values a soft key can be obtained using a second radius r_2. See [31] for details.

Motion compensated difference keying: Difference keying is often used as a simple segmentation technique. It is based on the difference in colour space between a pixel I of the image and the corresponding pixel I_{bg} in the background plate. The background plate can be created by either taking a picture of the scene without any foreground objects or if this is not possible the background plate can be generated by applying a temporal median filter over a sequence to remove moving foreground objects. The difference between I and I_{bg} can be computed in any colour space. We used the difference in RGB space here:

$$\delta = d_{rgb}(I, I_{bg}) \tag{3}$$

The segmentation S_{diff} is computed as a binarisation with threshold σ:

$$S_{diff} = \begin{cases} 0 & , \delta < \sigma \\ 1 & , \text{otherwise} \end{cases} \tag{4}$$

Difference keying assumes correspondences between image pixel I and I_{bg} in the background plate and therefore requires static cameras. However, under known nodal (pan,tilt) 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 (section IV-B).A clear plate of foreground objects is created by applying a temporal median filter on the contributing patches. The colour distance δ defined in equation 4 is computed by projecting each pixel I into the spherical map using the camera parameters of the image.

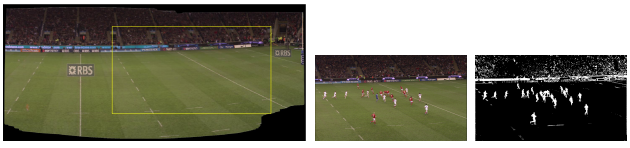


Fig. 7. Difference keying. The camera image (middle) is compared against a spherical panorama of the scene (left) giving the key (right).

Aperture correction in broadcast cameras Broadcast cameras have a control known as aperture correction or sometimes *detail*. The aperture correction is used to *sharpen* an image to emphasise high-frequency image components and is therefore a high-boost filter. Figure 8 shows an example of a broadcast image. The image was taken during a rugby match with a Sony HDC-1500 high definition camera.

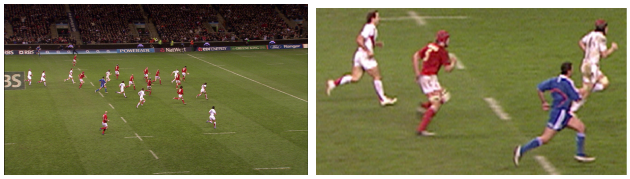


Fig. 8. Image of a sport scene from a broadcast camera (left) and detail (right).

In sport productions it is quite common to add a lot of *detail*. The effect can be seen in the detail picture (figure 8 bottom) as over- and undershoots of the image signal (visible as black haloes of white shirts against the background). In broadcast this feature is intended to give a better perceived contrast of objects. For segmentation the effect is a challenge since significant colour changes take place at the edge of objects of up to 3-5 pixels and can therefore be a significant problem.

The effect of the aperture correction can be compensated by a symmetric low-pass filter applied to the luminance channel of the broadcast image. We propose to compute the segmentation on the compensated image to improve the robustness and quality of the key.

(a) raw broadcast image (b) after aperture correction

Fig. 9. Detail of difference key computed on broadcast image before and after aperture correction for close-up from Figure 8 (right)

Figure 9 shows results of the difference keying applied to the original broadcast image (shown in Figure 9 (a) and after compensation of the aperture correction (Figure 9 (b)). In contrast to the global colour-based methods the pitch lines are suppressed except in areas with shadows. As expected the segmentation in the compensated image is more precisely aligned to object edges. This is clearly visible in the right image of figure 9. A more detailed analysis and description of the implemented compensation filter can be found in [32].

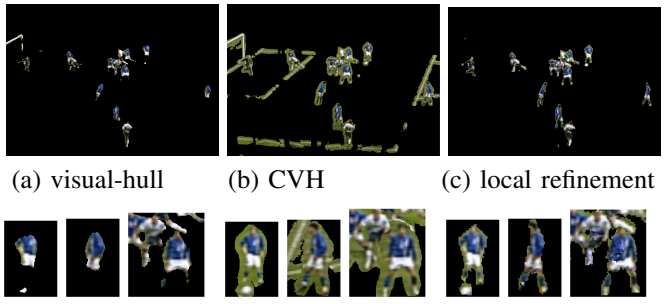
D. 3D Player Reconstruction and Segmentation

Stereo 3D video production for outdoor sports scenes captured over large areas must be robust to errors in the online camera calibration and natural scene matting. Typically for moving match cameras and chroma-key or difference matting errors are of the order 1-3 pixels. In typical camera footage the foreground players are relatively small 10-20pixel width

with arms and legs of the order 3-6 pixels. Direct multiple view reconstruction from erroneous foreground mattes can result in gross errors such as missing arms and legs. Due to the wide-baseline between cameras direct stereo matching between adjacent views is also problematic. In this section we present algorithms developed for robust rendering of novel views from multiple cameras in the presence of matting and calibration errors. The approach comprises two stages: conservative visual-hull reconstruction to recover a coarse scene approximation which encloses the true scene surfaces; and view-dependent optimisation to simultaneously refine the surface reconstruction and foreground segmentation to estimate a scene approximation which aligns images across multiple adjacent wide-baseline views and accurately segments the foreground player boundary. This approach achieves high-quality rendering of novel views in the presence of calibration and matting errors.

1) Conservative Visual-Hull: The conservative visual-hull (CVH) [33] is a volumetric approximation of the scene from multiple view image silhouettes up to a maximum error in the camera calibration and matting. The CVH is a global multiple view reconstruction which encloses the true scene surface. In the presence of camera calibration errors there is no single global scene reconstruction which corresponds to the observations from all views. A conservative visual hull with an n-pixel tolerance is obtained by dilating the image silhouettes by n pixels prior to silhouette intersection. Typically in this work $n = 3$ gives an upper-bound on the combined calibration and image segmentation errors. This ensures that the true scene surface projects to inside the dilated silhouettes for all views given errors in calibration and matting.

Figure 10 presents examples of novel view rendering of a soccer match for the visual-hull and conservative visual-hull. The visual-hull demonstrates the problem of global reconstruction from multiple views in the presence of calibration and matting errors. A number of players have missing arms or legs. The CVH in Figure 10(b) reconstructs a surface which enclosed the true foreground scene objects including narrow limbs. However, significant visual artifacts occur with the CVH rendering as the surface is over extended at the boundaries and does not accurately align adjacent camera views for rendering. In our system the CVH provides an initial global scene reconstruction which is then locally refined for accurate image alignment and boundary extraction.



(a) visual-hull (b) CVH (c) local refinement

Fig. 10. Novel view rendering for scene reconstructions in the presence of camera calibration and matting errors.

2) View-dependent 3D Reconstruction and Segmentation Refinement: The CVH provides a robust initial estimate of a surface which encloses the true scene surfaces. This surface is then locally refined with respect to a specific camera viewpoint to obtain a surface approximation which aligns the adjacent images and accurately segments the foreground boundaries. The CVH surface is an initial approximation which provides constraints to enable wide-baseline stereo matching between adjacent views. Stereo correspondence is constrained to lie inside the CVH reducing the likelihood of false matches. This approach was previously introduced for wide-baseline reconstruction in multiple camera studios[14] and has recently been extended for refinement in outdoor sports scenes [34], [33]. The critical advance required for high-quality rendering in sports production is the simultaneous refinement of both the initial 2D image segmentation and initial CVH surface reconstruction. Refinement in a view-dependent framework is robust to errors in the global camera calibration where there is no global reconstruction which is consistent with all camera views. Local view-dependent refinement combines information from multiple views together with priors on background appearance to achieve robust segmentation and improvements in the surface reconstruction. Simultaneous matting and reconstruction from multiple views of sports scenes was introduced in [34]. This approach has been extended to incorporate prior information on background and foreground appearance and improved optimisation of multiple view image cues.

The problem of simultaneous segmentation and reconstruction from multiple views consists in partitioning the image into its constituent foreground and background layers and assignising depth estimates at each layer's pixels. More formally this defines a labelling problem where we seek the mappings $l : \mathscr{P} \rightarrow \mathscr{L}$ and $d : \mathscr{P} \rightarrow \mathscr{D}$, which respectively assign a layer label l_p and a depth label d_p to every pixel p in the reference image. $\mathscr{P}$ denotes the set of pixels in the reference image; $\mathscr{L}$ and $\mathscr{D}$ are discrete sets of labels representing the different layer and depth hypotheses. $\mathscr{L} = \{l_1, \ldots, l_{|\mathscr{L}|}\}$ may consist of one background layer and one foreground layer (classic segmentation problem) or of multiple foreground and background layers. In this paper we assume multiple foreground layers corresponding to players at different depths and a single background layer. The set of depth labels $\mathscr{D} = \{d_1, \ldots, d_{|\mathscr{D}|-1}, \mathscr{U}\}$ is formed of depth values d_i obtained by sampling the optical rays together with an unknown label $\mathscr{U}$ used to account for occlusions. Occlusions are common and can be severe when the number of cameras is small, especially in the background where large areas are often visible only in a single camera.

The labelling problem is formulated in a Bayesian framework [34] and extended to incorporate multiple visual cues [35]. The maximum likelihood solution leads to the minimisation of an energy function:

$$E(l,d) = \lambda_{\text{colour}} E_{\text{colour}}(l) + \lambda_{\text{contrast}} E_{\text{contrast}}(l) + \lambda_{\text{match}} E_{\text{match}}(d) + \lambda_{\text{smooth}} E_{\text{smooth}}(l,d), \quad (5)$$

where $E_{\text{colour}}(l)$ and $E_{\text{match}}(l,d)$ are likelihood terms for image foreground and background layer assignments based on colour

Fig. 11. An input image and its adaptively attenuated contrast.

models and for depth assignments based on stereo matching scores. $E_{\text{contrast}}(l)$ and $E_{\text{smooth}}(l,d)$ are contrast and smoothness priors on the labelling. The parameters λ_{colour}, λ_{match}, $\lambda_{\text{contrast}}$ and λ_{smooth} control the contribution of each term.

The colour energy is defined in terms of the probability $P(I_p|l=l_i)$ that a pixel p belongs to a foreground or background layer l_i as:

$$E_{\text{colour}}(l) = \sum_{p\in\mathscr{P}} -\log P(I_p|l_p). \tag{6}$$

This probability is evaluated according to a linear combination of a local per-pixel model $P_{\text{l}}(I_p|l_p=l_i)$ and a global Gaussian mixture colour model $P_{\text{g}}(I_p|l_p=l_i)$ defined as follows:

$$P(I_p|l_p=l_i) = wP_{\text{g}}(I_p|l_p=l_i) + (1-w)P_{\text{l}}(I_p|l_p=l_i), \tag{7}$$

where w is a real value between 0 and 1 controlling the contributions of the two models. A dual colour model combining global and local components allows for dynamic changes in the background. The local model is only applicable to background layers which are static; in the case of foreground layers, this term is ignored. The local component of the colour model for a static layer l_i is represented by a single Gaussian distribution for each pixel p:

$$P_{\text{l}}(I_p|l_p=l_i) = N(I_p|\mu_{ip},\Sigma_{ip}), \tag{8}$$

where the parameters μ_{ip} and Σ_{ip} represent the mean and the covariance matrix of the Gaussian distribution at pixel p. To define this model, a reference background image is constructed a priori from the image sequence by mosaicing known background pixels and computing the local per pixel mean and variance (similar to the approach in Section IV-C). The global component of the colour model is represented by the Gaussian Mixture Model (GMM)

$$P_{\text{g}}(I_p|l_p=l_i) = \sum_{k=1}^{K_i} w_{ik}N(I_p|\mu_{ik},\Sigma_{ik}), \tag{9}$$

where N is the normal distribution and the parameters w_{ik}, μ_{ik} and Σ_{ik} represent the weight, the mean and the covariance matrix of the k^{th} component for layer l_i. K_i is the number of components of the mixture model for layer l_i. This model is learnt from a single key-frame per camera where foreground has been manually segmented from the background.

The contrast term encourages layer discontinuities to occur at high contrast locations. This naturally encourages low contrast regions to coalesce into layers and favours discontinuities to follow strong edges. This term is defined as

$$E_{\text{contrast}}(l) = \sum_{(p,q)\in\mathscr{N}} e_{\text{contrast}}(p,q,l_p,l_q), \text{ with} \tag{10}$$

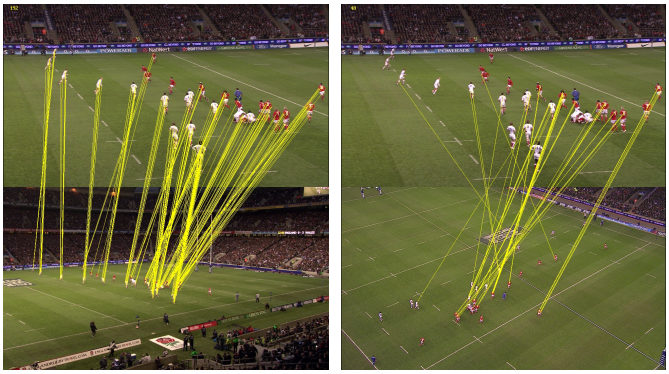


Fig. 12. Matched sparse features for two camera pairs.

$$e_{\text{contrast}}(p,q,l_p,l_q) = \begin{cases} 0 & \text{if } l_p = l_q, \\ \exp\left(-\beta C(I_p,I_q)\right) & \text{otherwise.} \end{cases} \tag{11}$$

$\mathscr{N}$ denotes the set of interacting pairs of pixels in $\mathscr{P}$ (a 4-connected neighbourhood is assumed) and $||\cdot||$ is the L_2 norm. $C(\cdot,\cdot)$ represents the squared colour distance between neighbouring pixels, and β is a parameter weighting the distance function. Although various distance functions are possible for $C(\cdot,\cdot)$, we use the attenuated contrast [36]

$$C(I_p,I_q) = \frac{||I_p - I_q||^2}{1+\left(\frac{||B_p-B_q||}{K}\right)^2 \exp\left(-\frac{z(p,q)^2}{\sigma_z}\right)}, \tag{12}$$

where $z(p,q) = \max(||I_p - B_p||, ||I_q - B_q||)$. B_p is the background colour at pixel p; it is provided by the local component of the colour model previously defined. β, K and σ_z are parameters which are set to the standard values suggested in [36]. This formulation uses background information to adaptively normalise the contrast, thereby encouraging layer discontinuities to fall on foreground edges (see Figure 11).

The matching energy combines sparse and dense correspondence between camera views:

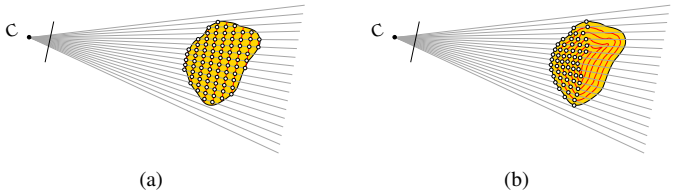
$$E_{\text{match}}(d) = E_{\text{dense}}(d) + E_{\text{sparse}}(d). \tag{13}$$

The dense matching score is defined as

$$E_{\text{dense}}(d) = \sum_{p\in\mathscr{P}} e_{\text{dense}}(p,d_p), \text{ with} \tag{14}$$

$$e_{\text{dense}}(p,d_p) = \begin{cases} S(P(p,d_p)) & \text{if } d_p \neq \mathscr{U}, \\ S_{\mathscr{U}} & \text{if } d_p = \mathscr{U}. \end{cases} \tag{15}$$

$P(p,d_p)$ denotes the coordinates of the 3D point along the optical ray passing through pixel p and located at a distance d_p from the reference camera. The function $S(\cdot)$ measures the similarity of the reference camera with the auxiliary cameras in which the hypothesised point $P(p,d_p)$ is visible. For weakly textured scenes such as the ones considered in this paper, standard normalised cross correlation similarity measures are inadequate. A more appropriate choice in this case is an error tolerant photo-consistency measure similar to [37]. This computes photo-consistency over extended regions of radius r_{tol} rather than single pixels, and thereby compensates for calibration errors or non-uniform image sampling. The



(a) (b)

Fig. 13. Comparison of locally constant depth prior (a) with visual hull iso-surface based depth prior (b).

photo-consistency score between the reference image and the auxiliary camera i is defined as

$$\text{photo}_i(X) = \max_{(q-\pi_i(X))^2<r_{\text{tol}}} \frac{(I_p - I_q^i)^2}{\sigma_i^2}, \tag{16}$$

where σ_i^2 normalises the photo-consistency measure for each auxiliary camera i and the function $\pi_i(X)$ projects the hypothesised 3D point X into the image plane of camera i. A robust combination rule is defined as the sum of the k most photo-consistent pairs denoted by $\mathscr{B}_k$

$$S(X) = \sum_{i\in\mathscr{B}_k} \text{photo}_i(X). \tag{17}$$

The sparse matching score is defined as

$$E_{\text{sparse}}(d) = \sum_{p\in\mathscr{P}} e_{\text{sparse}}(p,d_p), \text{ with} \tag{18}$$

$$e_{\text{sparse}}(p,d_p) = \begin{cases} 0 & \text{if } \mathscr{F}(p) = \emptyset \text{ or } d_p \in \mathscr{F}(p), \\ \infty & \text{otherwise.} \end{cases} \tag{19}$$

$\mathscr{F}(p)$ denotes the set of depth labels located within a distance T from a sparse constraints at pixel p. This forces the reconstructed surface to pass nearby existing sparse 3D correspondences. Because of calibration errors, we do not require the reconstruction to match exactly the sparse constraints, but allow a tolerance controlled by the parameter T. We use affine-covariant features [38], [39] which are known to be robust to changes in viewpoint and illumination. In this paper, we used the Hessian-affine feature detector. Features are represented using the SIFT descriptor and matched based on a nearest neighbour strategy. Robust matching is ensured by restricting the search to areas within a tolerance distance from the epipolar lines. The left-right spatial consistency (reciprocity) constraint is enforced together with temporal consistency which requires corresponding features between camera views to be in correspondence temporally with the previous or the next frame (see Figure 12 for some examples of correspondences found between two adjacent views).

The smoothness term encourages the depth labels to vary smoothly within each layer. This is useful in situations where matching constraints are weak (poor photoconsistency or a low number of sparse constraints) and insufficient to produce an accurate reconstruction without the support from neighbouring pixels. It is defined as

$$E_{\text{smooth}}(l,d) = \sum_{(p,q)\in\mathscr{N}} e_{\text{smooth}}(l_p,d_p,l_q,d_q), \text{ with} \tag{20}$$

$$e_{\text{smooth}}(l_p,d_p,l_q,d_q) = \tag{21}$$
$$\begin{cases} \min(|d_p-d_q|,d_{\max}) & \text{if } l_p = l_q \text{ and } d_p,d_q \neq \mathscr{U}, \\ 0 & \text{if } l_p = l_q \text{ and } d_p,d_q = \mathscr{U}, \\ d_{\max} & \text{otherwise.} \end{cases}$$

Discontinuities between layers are assigned a constant smoothness penalty $d_{\max}$, while within each layer the penalty is defined as a truncated linear distance. Such a distance is discontinuity preserving as it does not over-penalise large discontinuities within a layer; this is known to be superior to simpler non-discontinuity functions (see [40], [41]). This term also encourages unknown features to coalesce within each layer. The choice of shape prior is crucial. A commonly used prior is to assume locally constant depth (see Figure 13(a)). In this case, a label (l_p,d_p) corresponds to the point from layer l_p and located at a distance d_p from the reference camera centre along the ray emanating from pixel p. Although this yields good quality results when supported by strong matching cues, this results in bias towards flat figure models which do not give good alignment between views. An alternative approach which we use here is to place samples along the iso-surfaces of the visual hull (see Figure 13(b)), which results in a reconstructed surface biased towards the visual hull iso-surfaces. We call this prior the iso-surface prior. In this case, a label (l_p,d_p) corresponds to the first point of intersection between the ray emanating from pixel p and the d_p-th iso-surface in the interior of the visual hull's connected component corresponding to layer l_p. To account for calibration and matting error, we use the error tolerant visual hull proposed in [33]. Unlike the fronto-parallel prior, the iso-surface prior is view-independent and results in reconstructions more realistic and likely to coincide in the absence of strong matching cues. It can be noted that the choice of a fronto-parallel or an iso-surface prior affects the correspondence between labels and the 3D points they represent, however it does not change the formulation in Equation 21 since the set of depth values remain an ordered set of discrete values.

Optimisation of the energy defined by Equation 5 is known to be NP-hard. However, an approximate solution can be computed using the expansion move algorithm based on graph-cuts [40]. The expansion move algorithm proceeds by cycling through the set of labels $\alpha = (l_\alpha,d_\alpha)$ in $\mathscr{L}\times\mathscr{D}$ and performing an α-expansion iteration for each label until the energy cannot be decreased (see [40]). An α-expansion iteration is a change of labelling such that each pixel p either retains its current value or takes the new label α. Each α-expansion iteration can be solved exactly by performing a single graph-cut using the min-cut/max-flow algorithm [42]. After convergence of the algorithm, the result obtained is guaranteed to be a strong local optimum [40]. The α-expansion algorithm was initialised with the visual hull estimate; convergence has been found to be insensitive to the choice of initialisation. In practice, convergence is usually achieved in 3 or 4 cycles of iterations over the label set. We improve computation and memory efficiency by dynamically reusing the flow at each iteration of the min-cut/max-flow algorithm [43]. This usually results in a speed-up of an order of two.

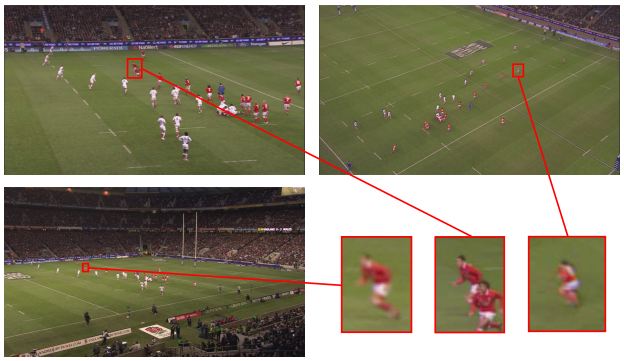


Fig. 14. Two moving broadcast camera views (top) and a locked-off camera view (bottom-left) from a rugby match.

E. 3D Sports TV Production Trials

Production trials of the system have been conducted for soccer and rugby. In both cases free-viewpoint video sequences were generated for events of editorial value identified by sports producers. Sports events were captured using both the moving broadcast cameras and a small number (4–6) of additional fixed cameras to give coverage of the entire pitch area. Typically in a high-profile soccer or rugby match with 12-18 cameras only 6-8 cameras are viewing the events of interest for free-viewpoint production. All cameras in the production trials were captured in uncompressed HD-SDI 4:2:2 format for subsequent processing.

1) 3D Scene Reconstruction and Segmentation: Figure 14 shows camera views from an international rugby match for two match cameras and one static camera with a typical wide-baseline and difference in framing between views. The close-ups of one player show the difference in resolution of the players between camera views. This illustrates the variation in viewpoint and scale for a set of multiple view cameras. Players are at a relatively small scale in the static cameras due to the requirement to cover a wide area of the pitch. In the broadcast camera views players are at a larger scale but the scale varies between views and there is a high-degree of motion blur due to camera movement.

A comparison of results for segmentation algorithms for soccer and rugby is presented in Figure 15. Chroma-key and difference key are 2D image segmentation techniques which show errors in segmentation as both additional foreground clutter and areas of the foreground (arms/legs) which are incorrectly classified as background. Background cut [36] is another 2D image segmentation technique which improves the results by combining local and global models; this increases robustness and adds tolerance to non-static background elements, however it yields inaccurate results in ambiguous areas. In contrast the multiple view segmentation and reconstruction refinement algorithm presented in section IV-D2 produces a clean foreground segmentation with correct classification of the foreground objects even in ambiguous situations where the foreground and background have similar colour ie lines or muddy parts of the players legs. This is due to the combination of information from multiple views to overcome single view visual ambiguities. The refined segmentation allows improvements in the reconstruction quality.

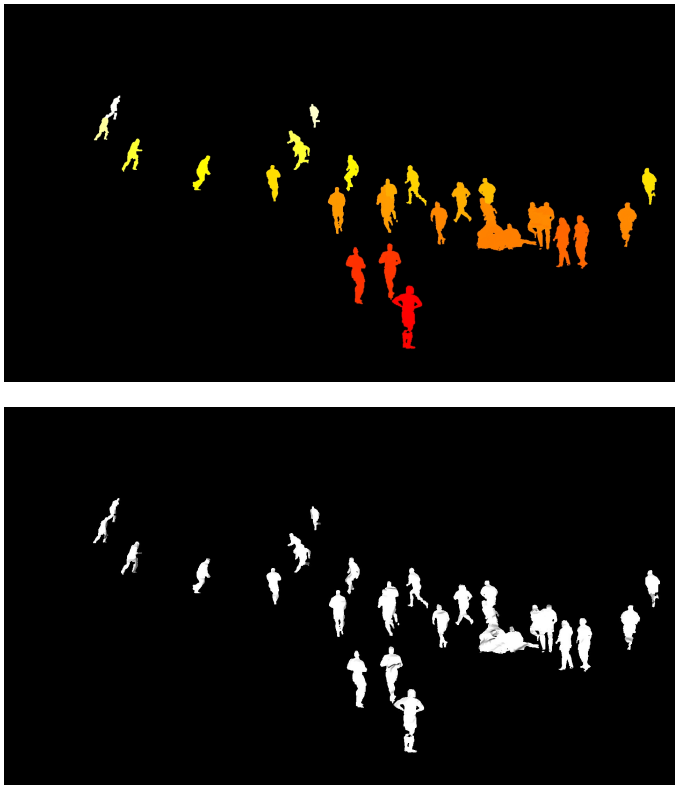


Fig. 16. A layered depth image representation (top) and the corresponding view-dependent mesh (bottom).

The proposed technique estimates a layered-depth representation. This produces a richer segmentation of the image into multiple layers in contrast with a standard foreground/background segmentation and incorporates depth information at each pixel of the foreground layers (background layers, which are static, can be more conveniently represented using a pre-defined model and do not require automatic depth estimation). An example of layered depth representation is shown in Figure 16. It defines a view-dependent 2.5D foreground representation which can be easily converted into a regular mesh with vertices defined by image pixel locations. Vertex connectivity is decided based on the layer segmentation and thresholding of the angle separating the line segment connecting 3D surface points defined by pairs of neighbouring pixels and the optical ray passing through the midpoint of the pixel pair. This allows pixels belonging to different layers or located at a depth discontinuity to be correctly converted into separate mesh components. This view-dependent representations is used for stereo rendering or free-viewpoint videos as discussed in the following sections.

To evaluate reconstruction quality, the proposed technique was compared to three standard techniques: (i) conventional visual hull, (ii) conservative visual hull (with 2 pixel tolerance), and (iii) stereo refinement of the conservative visual hull with no colour, contrast or smoothness term. Results are shown in Figure 17. As expected the visual hull produces large truncations in the presence of calibration and segmentation errors. These truncations have been eliminated in

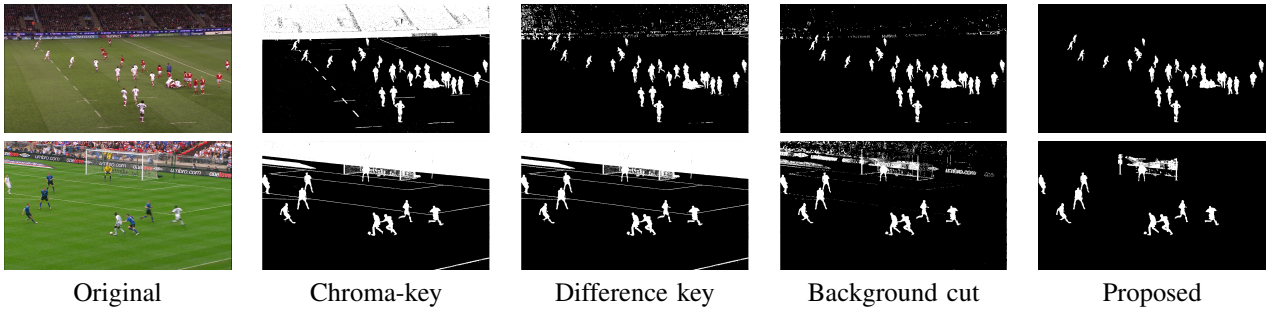


Fig. 15. Example of segmentation results on rugby (top) and soccer (bottom) data (see attached video for full sequence).

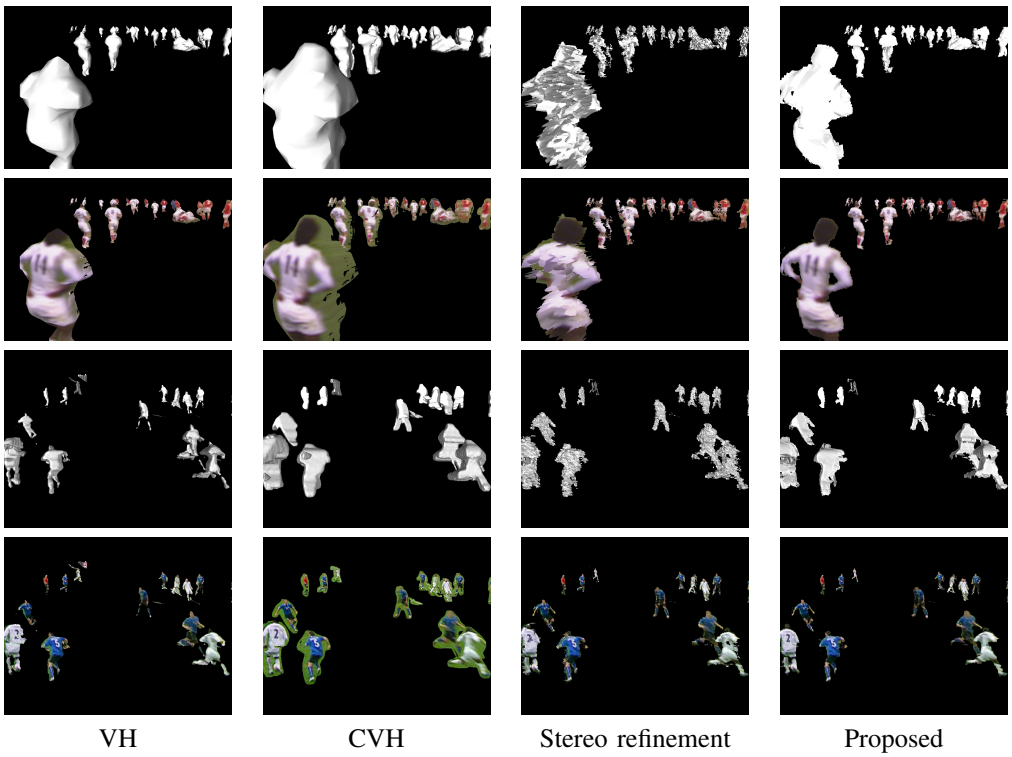


Fig. 17. Example of reconstruction results on rugby (top) and soccer (bottom) data (see attached video for full sequence).

the conservative visual hull but have been replaced by some protrusions and phantom volumes. Stereo refinement of the conservative visual hull results in a very noisy reconstruction; this illustrates the weakness of the available photo-consistency information. In contrast, the proposed technique yields a smooth reconstruction with accurate player boundaries and the elimination of phantom volumes. This accurately aligns wide baseline views based on stereo matches and gives a smooth surface approximation based on iso-surfaces of the visual-hull shape prior in regions of uniform appearance which commonly occur on player shirts or due to views sampled at significantly different resolutions. This is suitable for stereoscopic content production and high quality free-viewpoint rendering.

2) Stereo Rendering: The proposed method is well-suited for stereoscopic content production because of the flexibility it provides for automatic content production from multiple standard (monocular) input cameras. The layered depth representation produced by the proposed technique is compatible with existing 3D video representations such as Video plus Depth (VD) representations used for a single video stream or their multi-stream extensions such as Multiview Video plus Depth (MVD) and the Layered Depth Video (LDV). More recent research in developing 3D video standards is focused on extending the applicability to 3D displays with varying specifications and providing control over the rendering parameters [44], [45]. These new standards present many similarities with the proposed technique and it is expected that they will facilitate future integration of our technique into commercial 3DTV devices. Two different 3DTV formats are supported here:

A stereoscopic image pair can be generated by using one original camera image and rendering a second view. The second view is rendered by using the original image as a texture source for the background scene. This procedure preserves shadows. Unrevealed background areas need to be filled. For this purpose we use the clean background plates computed during the foreground segmentation process, described in section IV-C. This information does not contain proper shadows, but provides better information than using other camera views,

since these have other problems, for example differently colour balanced cameras and anisotropic effects of the scene.

A layered depth representation (LDV) can be generated by using the original camera image and adding the scene depth information derived by the 3D reconstruction. The static scene components, like the ground and the stadium are generated manually in a separate process. The simplest implantation of a LDV representation is using just the camera image and a pixel-wise depth map. In addition one (or more) maps with occluded data can be added. We use the background plates derived from the segmentation process as described above for this purpose.

For the scope of this paper, a rendering platform has been developed in order to demonstrate the stereoscopic content production capabilities of the system. The rendering platform is implemented using OpenGL and provides full control of the inter-ocular distance as well as the convergence distance. It also allows free-viewpoint video rendering which will be discussed in the next section. In this section, we concentrate on synthesising stereoscopic output from input cameras locations. Foreground layers are modelled in the form of view-dependent meshes extracted from the layered depth representation. To provide robustness to segmentation errors, we use meshes obtained from the considered input camera as well as the nearest adjacent cameras. Such a local representation is adequate for stereoscopic rendering without suffering from the artefacts observed with global reconstruction methods that were described in the previous section. The stadium backgrounds are manually modelled from the input images and textured using the background plate images that were automatically generated from input video data (Section IV-C). Left and right views are synthesised by texture mapping the input image and background plate onto the foreground and background layers. Use of more sophisticated view-dependent texture mapping techniques which combine multiple views is not necessary here given that the rendering viewpoints are located near an input camera.

Modelling of the foreground and background layers using separate texture maps reduces data transfer requirement as it decouples the dynamic foreground elements (which require update at each frame) from the static background elements (which require less frequent update). In addition, this attenuates the effect of segmentation errors which can deteriorate the final video quality; in particular use of a background plate guarantees that no foreground texture is accidentally mapped onto the background layers. A drawback of this approach however is that shadows and background motion are lost in the background plate generation process which is based on temporal filtering. While the elimination of the background motion is usually not too problematic, the absence of shadows tends to produce unrealistic sequences where players do not seem to connect to the background. In order to improve the degree of realism of the synthesised images, the scene is augmented with virtual shadows. Soft shadows are generated by using a virtual light source with a shadow mapping algorithm to cast shadows generated by players onto the ground.

Adequate definition of the left and right camera parameters is crucial as those control the 3D effect experienced by the viewer. Two key parameters in a stereoscopic system are the inter-ocular distance, ie the distance separating left and right cameras' optical centres, and the convergence distance, ie the distance at which 3D points have zero parallax. The inter-ocular distance controls the amplitude of the depth effect. The convergence distance controls the location of the scene with respect to the 3D display in viewer space. Points located at the convergence distance have zero parallax and will therefore appear to be located at screen depth in the viewer space. In order to avoid discomfort to the viewer, these parameters should be defined so as not to severely break the accommodation/convergence relationship to which the eyes are used to [46]. In our implementation, left and right views are located on either side of the monocular principal camera view; alternatively the principal camera view could have been used to define either the left or right view, thereby reducing the novel view generation to a single view instead of two.

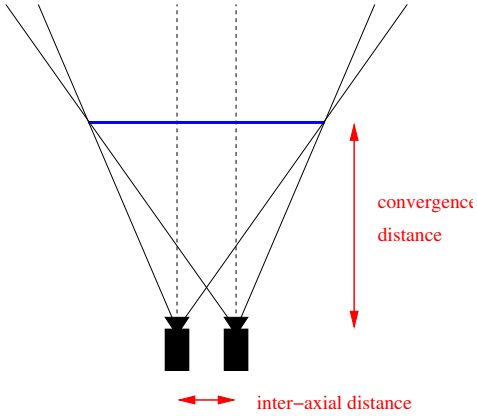


Fig. 18. Stereo camera configuration with inter-ocular distance and convergence distance parameters.

The proposed stereoscopic rendering approach has been tested with rugby and soccer data (see attached video for full sequences). Results in the case of the rugby trial are shown in Figure 19. To illustrate the flexibility of the system, the same frame has been rendered with different user-defined inter-ocular and convergence distances. Adjustment of these parameters can be performed in real-time. This provides a flexible and intuitive interface for stereoscopic content production which can be used to optimise viewer comfort, retarget a sequence to a specific display device or facilitate the creation of stereoscopic visual effects. Experiments conducted suggest that the videos produced using this technique convey a realistic sense of depth and are comfortable to view.

3) Free-viewpoint Rendering: Free-viewpoint rendering results for rugby and soccer trials are shown in Figures 20 and 21. In both cases the virtual camera sequences were generated for production trials of specific camera views which give added value to the match analysis. In the case of the rugby sequence shots show specific game plays and for soccer the shot was specified to view the offside line in a contentious incident. Free-viewpoint camera moves can either take place by freezing the action at a single frame or whilst the action is taking place according to the production requirements. All sequences

Fig. 19. Examples of synthesised stereoscopic images illustrating adjustment of inter-ocular and convergence distances. Each column of images corresponds to a different inter-ocular distance setting; the distance starts from 0 (left column) and increases as we move along the columns. Each row of images represents a different convergence distance setting; the top row places all the scene behind screen depth while the bottom row places the player wearing the white shirt number 10 at screen depth. Images are displayed in optimised red-cyan anaglyph format. Please see attachment for full videos.

were generated with automatic calibration, 2D segmentation, reconstruction and refinement. View-dependent rendering is performed using the view-dependent geometry to render images from the adjacent views. The stadium backgrounds are manually modelled using either images from the captured sequences as in Figure 20(a) or a synthetic appearance Figure 20(c). Rugby and soccer present different challenges for free-viewpoint production, rugby is particularly challenging as the players are distributed across the field and groups of players form rucks and malls where individual players come into contact and cannot be isolated. The approach developed does not make any prior assumptions on player shape allowing high-quality free-viewpoint rendering of both isolated and tightly packed groups of players.

Free-viewpoint rendering with the proposed approach achieves an image quality comparable to that of the input image sequences as demonstrated in the closeup of Figure 20(b). Degradation in image quality will occur if there are no real cameras which see a part of the scene or there are insufficient views for reconstruction. The proposed approach is robust to the wide-baseline moving camera views at different resolutions which occur in broadcast coverage. The free-viewpoint rendering system takes advantage of the manually operated broadcast cameras which generally frame the play to give higher player resolution than the static auxiliary cameras. The system can operate from the match cameras only but this limits coverage and virtual camera viewpoints to sections of the play where there are sufficient views. Addition of a small number of auxiliary cameras adds to the production cost but ensures complete coverage of the game play and increased range of views for free-viewpoint production. The correct trade-off between coverage and cost will be determined by the production requirements for a specific sport or event. Production trials have demonstrated free-viewpoint shots which add value to the commentary and are of a quality suitable for broadcast.

(a) Rugby virtual camera sequences at 20 frame intervals

(b) Virtual camera closeup

(c) Rugby virtual camera sequences with a virtual stadium model

Fig. 20. Free-viewpoint video rendering of rugby to show pitch level views for commentary

Fig. 21. Free-viewpoint video rendering of soccer to show an offside incident

V. Conclusions

Stereo and free-viewpoint video production for 3DTV sports broadcast presents significant challenges to achieve a visual quality comparable to captured video with minimal delay from the manually controlled moving and zooming match cameras. Capture of stadium sports such as soccer and rugby requires acquisition over a large area with relatively uncontrolled conditions. In this paper we have presented a system for stereo 3DTV production from conventional monocular match cameras used for 2D broadcast production. This represents an alternative to the use of additional stereo camera rigs and avoids the problems associated with zoom lens matching and correction required in live production. The approach presented allows the rendering of stereo views without mismatches between camera views and with full control of inter-ocular distance and convergence in post-production. This allows stereo rendering for different display sizes or transmission of image+depth for stereo rendering at the point of display.

This system advances previous studio based 3D reconstruction from multiple static camera views to wide-baseline moving broadcast cameras covering a large area. Reconstruction is robust to relatively large calibration errors and uncontrolled scene illumination and backgrounds. Production trials of the system have been conducted on soccer and rugby to generate stereo 3D and free-viewpoint video sequences. The system allows automatic reconstruction and stereo rendering from the match cameras with a visual quality comparable to captured video.

A number of open-problems remain to achieve widespread deployment in broadcast production: calibration and segmentation of close-up and pitch level camera views where pitch lines are not visible; rendering quality of close-up shots which are limited by the available camera resolution; validated accuracy of stereo and free-viewpoint rendering for match decisions (offside); temporal coherence of rendering and representation for moving scenes; video-rate production of stereo views for live 3DTV broadcast; and interfaces for rapid free-viewpoint or stereo shot production.

Acknowledgements: This work was supported by TSB Technology Programme and EPSRC on projects *iview: Free-viewpoint video for entertainment content production* and *i3Dlive: Interactive methods for live action media* and EU IST FW7 Project *3D4YOU - Content Generation and Delivery for 3D Television*.

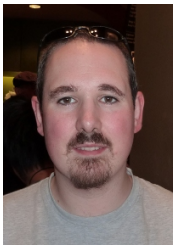
References

[1] S. Chen and L. Williams, "View Interpolation for Image Synthesis," in *Proc. ACM SIGGRAPH*, 1993.
[2] S. Seitz and C. Dyer, "View morphing: Synthesizing 3D metamorphosis using image transforms," *Proc. ACM SIGGRAPH*, pp. 21—30, 1996.
[3] D. Kim, D. Min, and K. Sohn, "A stereoscopic video generation method using stereoscopic display characterization and motion analysis," *IEEE Transactions on Broadcasting*, vol. 54, no. 2, pp. 188–197, 2008.
[4] M. Irani, T. Hassner, and P. Anandan, "What Does the Scene Look Like from a Scene Point ?" in *European Conference on Computer Vision*, 2002.
[5] H.-Y. Shum, S. Kang, and S.-C. Chan, "Survey of image-based representations and compression techniques," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 11, 2003.
[6] A. Laurentini, "The visual hull concept for silhouette based image understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 2, pp. 150–162, 1994.
[7] Moezzi,S. and Katkere,A. and Kuramura,D.Y. and Jain,R., "Reality Modeling and Visualization from Multiple Video Sequences," *IEEE Computer Graphics and Applications*, pp. 58–63, November 1996.
[8] W. Matusik, C. Buehler, R. Raskar, S. Gortler, and L. McMillan, "Image-based visual hulls," *Proceedings of ACM SIGGRAPH*, pp. 369–374, 2000.
[9] F. Franco and E. Boyer, "Exact Polyhedral Visual Hulls," in *British Machine Vision Conference*, 2003, pp. 329–338.
[10] G. Miller, A. Hilton, and J. Starck, "Interactive Free-viewpoint Video," in *IEE European Conf. on Visual Media Production*, 2005, pp. 50—59.
[11] S. Seitz and C. Dyer, "Photorealistic scene reconstruction by voxel coloring," *International Journal of Computer Vision*, vol. 35, no. 2, pp. 1–23, 1997.
[12] T. Kanade and P. Rander, "Virtualized reality: Constructing virtual worlds from real scenes," *IEEE MultiMedia*, vol. 4, no. 2, pp. 34—47, 1997.
[13] C. Zitnick, S. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High-quality video view interpolation using a layered representation," in *Proc. ACM SIGGRAPH*, 2004, pp. 600—608.
[14] J. Starck and A. Hilton, "Virtual view synthesis of people from multiple view video," *Graphical Models*, vol. 67, no. 6, pp. 600—620, 2005.
[15] G. Miller, J. Starck, and A. Hilton, "Projective Surface Refinement for Free-Viewpoint Video," in *IET European Conference on Visual Media Production*, 2006, pp. 153–162.
[16] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A Comparison and Evaluation of Multi-View Stereo Reconstruction Algorithms," in *Conference on Computer Vision and Pattern Recognition*, 2006, pp. 519—528.
[17] W. Matusik, C. Buehler, R. Raskar, S. Gortler, and L. McMillan, "Image-based visual hulls," *Proceedings of ACM SIGGRAPH*, pp. 369–374, 2000.
[18] S. Moezzi, L. Tai, and P. Gerard, "Virtual view generation for 3d digital video," *IEEE Multimedia*, vol. 4, no. 1, pp. 18–25, 1997.
[19] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade, "Three-dimensional scene flow," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 3, 2005.
[20] F. Franco, C. Menier, E. Boyer, and B. Raffin, "A Distributed Approach for Real-Time 3D Modeling," in *CVPR Workshop on Real-Time 3D Sensors and their Applications*, 2004.
[21] J. Starck and A. Hilton, "Surface Capture for Performance-Based Animation," *IEEE Computer Graphics and Applications*, vol. 27, no. 3, pp. 21—31, 2007.

[22] T. Kanade, P. Rander, and P. Narayanan, "Virtualized reality: Constructing virtual worlds from real scenes," *IEEE Multimedia*, vol. 4, no. 1, pp. 34–47, 1997.
[23] O. Grau, M. Prior-Jones, and G. Thomas, "3d modelling and rendering of studio and sport scenes for tv applications," in *Proc. of WIAMIS*.
[24] K. Connor and I. Reid, "A Multiple View Layered Representation for Dynamic Novel View Synthesis," in *British Machine Vision Conference*, 2003.
[25] N. Inamoto and H. Saito, "Virtual Viewpoint Replay for a Soccer Match by View Interpolation From Multiple Cameras," *IEEE Trans.Multimedia*, vol. 9, no. 6, pp. 1155—1166, 2007.
[26] K. Kimura and H. Saito, "Player viewpoint video synthesis using multiple cameras," in *IEE European Conference on Visual Media Production*, 2005, pp. 112—121.
[27] M. Germann, A. Hornung, R. Keiser, R. Siegler, S. Wurmlin, and M. Gross, "Articulated Billboards for Video-based Rendering," *Computer Graphics Forum*, vol. 29, no. 2, pp. 585—594, 2010.
[28] G. Thomas, "Real-time Camera Pose Estimation for Augmenting Sports Scenes," in *European Conference on Visual Media Production*, 2006, pp. 10—19.
[29] ——, "Real-Time Camera Tracking using Sports Pitch Markings," *Journal of Real Time Image Processing (Available as BBC R&D White Paper 168. http://www.bbc.co.uk/rd/publications/whitepaper168.shtml)*, vol. 2, no. 2—3, pp. 117–132, 2007.
[30] P.Hillman, J.Hannah, and D.Renshaw, "Foreground/background segmentation of motion picture images and image sequences," *IEE Transactions on Vision, Image and Signal Processing*, vol. 142(4), pp. 387–397, August 2005.
[31] O. Grau, G. Thomas, A. Hilton, J. Kilner, and J. Starck, "A robust free-viewpoint video system for sport scenes," in *Proceeding of 3DTV conference 2007*, Kos, Greece, 2007.
[32] O.Grau and J.Easterbrook, "Effects of camera aperture correction on keying of broadcast video," in *Proc. of the 5rd European Conference on Visual Media Production (CVMP)*, 2008.
[33] J. Kilner, J. Starck, A. Hilton, J. Guillemaut, and O. Grau, "Dual Mode Deformable Models for Free-Viewpoint Video of Outdoor Sports Events," in *IEEE Int.Conf. on 3D Imaging and Modeling* , 2007.
[34] J. Guillemaut, A. Hilton, J. Starck, J. Kilner, and O. Grau, "A Baysian Framework for Simultaneous Reconstruction and Matting," in *IEEE Int.Conf. on 3D Imaging and Modeling* , 2007.
[35] J.-Y. Guillemaut, J. Kilner, and A. Hilton, "Robust Graph-Cut Scene Segmentation and Reconstruction for Free-Viewpoint Video of Complex Dynamic Scenes," in *IEEE Int.Conf. on Computer Vision, ICCV* , 2009.
[36] J. Sun, W. Zhang, X. Tang, and H.-Y. Shum, "Background cut," vol. 3954, 2006, pp. 628–641.
[37] K. Kutulakos, "Approximate N-view stereo," in *ECCV*, vol. I, 2000, pp. 67–83.
[38] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, "A comparison of affine region detectors," *IJCV*, vol. 65, no. 1-2, pp. 43–72, 2005.
[39] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *PAMI*, vol. 27, no. 10, pp. 1615–1630, 2005.
[40] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *PAMI*, vol. 23, no. 11, pp. 1222–1239, 2001.
[41] V. Kolmogorov and R. Zabih, "What energy function can be minimized via graph cuts?" *PAMI*, vol. 26, no. 2, pp. 147–159, 2004.
[42] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *PAMI*, vol. 26, no. 9, pp. 1124–1137, 2004.
[43] K. Alahari, P. Kohli, and P. Torr, "Reduce, reuse & recycle: Efficiently solving multi-label MRFs," in *CVPR*, 2008.
[44] A. Smolic, K. Mueller, P. Merkle, and A. Vetro, "Development of a new mpeg standard for advanced 3d video applications," in *Proc International Symp on Image and Signal Processing and Analysis*, 2009, pp. 400 –407.
[45] A. Vetro, A. M. Tourapis, K. Muller, and T. Chen, "3D-TV content storage and transmission," *IEEE Transactions on Broadcasting*, 2011, in press.
[46] L. Lipton, *Foundations of the Stereoscopic Cinema*. Van Nostrand Reinhold Company, 1982.

Adrian Hilton (BSc(hons), DPhil, CEng) is Professor of Computer Vision and Graphics at the University of Surrey, UK. His research interest is robust computer vision to model and understand real world scenes. Contributions include technologies for the first hand-held 3D scanner, modelling of people from images and 3D video for games, broadcast and film. He currently leads research investigating the use of computer vision for applications in entertainment content production, visual interaction and clinical analysis.

Jean-Yves Guillemaut received an MEng degree from the Ecole Centrale de Nantes, France, in 2001, and a PhD degree from the University of Surrey, U.K., in 2005. He is currently a Research Fellow in the Centre for Vision, Speech and Signal Processing, University of Surrey, U.K. His research interests includes free-viewpoint video and 3D TV, image/video-based scene reconstruction and rendering, image/video segmentation and matting, camera calibration, and active appearance models for face recognition.



Joe Kilner graduated from Churchill College, Cambridge in 1999. Until 2006 he worked in the games and the financial industries developing a wide range of software, from user-interface frameworks to real-time animation engines. In 2010 he completed a PhD on the 3D reconstruction and analysis of sporting events at the University of Surrey, where he is currently employed as a Research Fellow in the Centre for Vision, Speech and Signal Processing. His research interests include 3D reconstruction, representation and rendering, and 3D motion analysis.

Oliver Grau received a Diploma (Master) and a PhD from the University of Hanover, Germany. From 1991-2000 he worked as a research scientist at the University of Hanover and was involved in several national and international projects, in the field of industrial image processing and 3D scene reconstruction for computer graphics applications. In 2000 he joined the BBC Research & Development Department in the UK. He was working on a number of national and international projects on 3D scene reconstruction and visualization. His research interests are in new innovative tools for visual media production using image processing, computer vision and computer graphic techniques and he published a number of research papers and patents on this topic. Dr. Grau was and is active as reviewer for scientific journals, research bodies like EPSRC, EC-FP7 and as a programme committee member of several international conferences. Further he was the initiator and chair of CVMP, the European Conference on Visual Media Production in London.

Graham Thomas joined the BBC Research Department at Kingswood Warren in 1983. His PhD included the development of motion estimation methods for standards conversion, which led to an Emmy award and a Queens Award. Since 1995 he has been leading a team of engineers developing 3D image processing and graphics techniques for TV production, and is currently the Section Lead for Production Magic at BBC R&D. His work has led to many commercial products including the free-d camera tracking system, Chromatte retroreflective chromakey cloth, and the Piero sports graphics system. Graham has led or worked in many UK and European collaborative projects, has written many papers, and holds over 20 patents. He is a chartered engineer and member of the IET.