Free-viewpoint Video for TV Sport Production

A.Hilton,J.-Y.Guillemaut,J.Kilner, O.Grau and G.Thomas

Abstract Free-viewpoint video in sports TV production presents a challenging problem involving the conflicting requirements of broadcast picture quality with video-rate generation of novel views, together with practical problems in developing robust systems for cost effective deployment at live events. 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 football must cover the action taking place over an entire pitch with video acquisition at sufficient resolution for analysis and production of desired virtual camera views. In this chapter we identify the requirements for broadcast production of free-viewpoint video in sports and review the state-of-the-art. We present the *iview* free-viewpoint video system (www.bbc.co.uk/rd/iview) which enables production of novel camera views of sports action for use in match commentary, for example the referees viewpoint. Automatic online calibration, segmentation and reconstruction is performed to allow rendering of novel viewpoints from the moving match cameras. Results are reported of production trials in football (soccer) and rugby which demonstrate free-viewpoint video with a visual quality comparable to the broadcast footage.

Adrian Hilton
University of Surrey, UK, e-mail: a.hilton@surrey.ac.uk

Jean-Yves Guillemaut
University of Surrey, UK, e-mail: j.guillemaut@surrey.ac.uk

Joe Kilner
University of Surrey, UK, e-mail: j.kilner@surrey.ac.uk

Oliver Grau
BBC R&D, UK, e-mail: oliver.grau@bbc.co.uk

Graham Thomas
BBC R&D, UK, e-mail: graham.thomas@bbc.co.uk

1 Introduction

Multiple view video systems are increasingly used in sports broadcasts for production of novel *free-viewpoint* camera views and annotation of the game play. This chapter identifies the requirements for free-viewpoint video in sports TV broadcast production, reviews previous work in relation to the requirements and presents a system designed to meet these requirements. Free-viewpoint video in sports TV production is a challenging problem involving the conflicting requirements of broadcast picture quality with video-rate generation. Further practical challenges exist in developing systems which can be cost effectively deployed at live events. 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 football must cover the action taking place over an entire pitch with video acquisition at sufficient resolution for analysis and production of desired virtual camera views. The *iview* system presented in this chapter 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 football 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 marks. Free-viewpoint video is then produced starting with a volumetric reconstruction followed by a view-dependent refinement using information from multiple views. Results are presented from production trials at international soccer and rugby sports events which show high-quality virtual camera views. This article is intended to be of benefit to the research community in defining broadcast industry requirements and identifying open problems for future research.

There is a demand from the broadcast industry for more flexible production technologies at live events such as sports. Currently cameras are operated manually with physical restrictions on viewpoint and sampling rate. The ability to place *virtual* cameras at any location around or on the pitch is highly attractive to broadcasters greatly increasing flexibility in production and enabling novel delivery formats such as mobile TV. Examples of the type of physically impossible camera views which could be desirable are the goal keepers view, a player tracking camera or even a ball camera. Virtual Replay (www.bbc.co.uk/virtualreplay) is an example of an existing technology using synthetic graphics and manual match annotation derived from broadcast footage of premier league soccer which allows viewers to select a virtual camera view. Manual annotation is labour intensive and the resulting virtual replay using generic player models and approximate movements is far from realistic. Ultimately the challenge is to achieve the flexibility of synthetic graphics with the visual-quality of broadcast TV in an automated approach.

This article presents the *iview* system (www.bbc.co.uk/rd/iview) which is being developed to address the requirements of broadcast production. The system primarily utilises footage from match cameras used for live TV broadcast. Match cameras provide wide-baseline views and are manually controlled to cover the play.

Typically major sporting events will have 12-18 manually operated high-definition cameras in the stadium. However, only a fraction of these will be viewing specific events of interest for production of free-viewpoint renders to enhance the commentary, other cameras are used for coverage of the pitch, crowd, coaches and close-up shots of players. Robust algorithms are required for both recovery of camera calibration from the broadcast footage and wide-baseline correspondence between views for reconstruction or view interpolation. At major sports events direct access to match cameras is limited as they are often operated by a third-party. Therefore *iview* uses a method for real-time camera calibration from the match footage [46]. Player segmentation is performed using chroma-key and difference matting techniques independently for each camera view [15]. Automatic calibration and player segmentation for moving broadcast cameras results in errors of the order of 1-3 pixels which is often comparable to the size of players arms and legs in the broadcast footage. Robust reconstruction and rendering of novel viewpoints is achieved in the *iview* system by an initial conservative visual-hull reconstruction followed by a view-dependent refinement. View-dependent refinement simultaneously refines the player reconstruction and segmentation exploiting visual cues from multiple camera views. This achieves free-viewpoint rendering with pixel accurate alignment of neighbouring views to render novel views with a visual quality approaching that of the source video. Results of the current *iview* system in production trials are presented. Limitations of the existing approaches are identified with repsect to the requirements of broadcast production. This leads to conclusions on future research advances required to achieve free-viewpoint technology which is exploited by the broadcast industry for use in production.

2 Background

Over the past decade there has been significant interest in the interpolation of novel viewpoints from multiple camera views. Recently research in this field for novel viewpoint interpolation from multiple view image sequences has been refered to as *video-based rendering*, *free-viewpoint video* or *3DTV*. In this section we review the principal methodologies for synthesis of novel views with a focus on the application to view-interpolation in sports. We first review the state-of-the-art in studio based systems where multiple cameras are used in an environment with controlled illumination and/or backgrounds. Studio systems have been the primary focus of both research and film production to capture a performance in a limited volume and allow playback of novel view sequences as free-viewpoint video. The extension of multiple camera systems to sports TV production requires capture of a much larger volumes with relatively uncontrolled illumination and backgrounds. Here we review the current state-of-the-art in sports production of live events such as football and identify the limitations of existing approaches in the context of TV production.

2.1 Methodologies for Free-viewpoint Video

Three principal methodologies have been investigated to rendering novel views of scenes captured from two or more camera viewpoints: interpolation; reconstruction; and tracking.

Interpolation: View interpolation directly estimates the scene appearance from novel viewpoints without explicitly reconstructing the 3D scene structure as an intermediate step [6, 40]. Interpolation between camera views requires the 2D-to-2D image correspondence. Given the correspondence image morphing and blending techniques can be used to interpolate intermediate views [40]. Commonly multiple view epi-polar and tri-focal constraints on visual geometry are used in the estimation of correspondence to reduce complexity and ambiguities which implicitly exploit the 3D scene structure. Both sparse feature based and dense stereo techniques have been used to estimate correspondence for view interpolation. View interpolation avoids the requirement for explicit 3D reconstruction but is in general limited to rendering viewpoints between the camera views. This 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 correspondence 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 [24]. A comprehensive survey of image-based rendering techniques for novel view synthesis if given in [41]. Recent research [45] has demonstrated perceptually convincing rendering of novel viewpoints for dynamic scenes by purely image-based interpolation between real camera views.

Reconstruction: Reconstruction of the 3D scene structure from multiple view images is commonly used as a basis for rendering novel views. The intermediate reconstruction of scene structure provides a geometric proxy which can be used to combine observations of scene appearance from multiple views in order to render images from novel viewpoints. The visual geometry of multiple views is now well understood [21]. 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 intersect silhouette cones from multiple views [29, 36, 30, 11, 32] 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 [38] is the maximal photo consistent volume between multiple views. Given the visual-hull as the maximum occupied volume the photo-hull is a sub-volume. An advantage of the photo-hull is that it does not require prior segmentation of the foreground. The photo-hull relies on the availability of a photo-consistency measure to distinguish different surface regions, in real scenes this may result in poor results. To overcome limitations of the visual and photo hull approaches a number of probabilistic approaches to volumetric reconstruction from multiple views have recently been introduced [3, 12]. A survey

of volumetric reconstruction approaches with further details can be found in [42]. Both the visual and photo hull are maximal approximations of the scene structure and therefore commonly result in significant visual artifacts in rendering novel views due to inaccurate geometry and resulting misalignment of the observed images. Stereo correspondence has been used for surface reconstruction from image pairs, a survey of approaches is given in [37]. Stereo reconstruction from multiple views has been used to reconstruct dynamic scenes of moving people [25, 49]. Dense correspondence from stereo ensures reconstruction of a surface which align the multiple view images reducing artifacts 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 [43, 33]. A comparison of approaches for reconstruction of static scenes from multiple views is presented in [39].

Tracking: An alternative approach to rendering novel viewpoints of a known scene such as a moving person is model-based 3D tracking as proposed for free-viewpoint video [5]. Model-based 3D tracking of human motion from multiple view video has receive considerable attention in the field of computer vision [34]. A generic 3D humanoid model approximating the surface shape and underlying skeletal structure is aligned with the image observations to estimate the pose at each frame. Typically an analysis-by-synthesis framework is used to optimise the pose estimate at each frame which minimises the re-projection error of the 3D humanoid model with the multiple view image observations. State-of-the-art multiple view 3D tracking of human motion [5, 9] achieves reconstruction of gross human pose but does not accurately reconstruct detailed movement such as hand rotations. Given the estimated pose sequence the model can be rendered from an arbitrary viewpoint. Shape reconstruction and observed appearance of the subject over time can be used to render realistic free-viewpoint video of the subjects performance. Model-based 3D tracking has the advantage over direct reconstruction of providing a compact structured representation. However, the shape reconstruction is limited by the constraints of the prior model. Existing approaches to model-based tracking can not cope with long hair or loose clothing such as skirts. Recent approaches [1, 48] have exploited models constructed from examples of the performer to more accurately represent body shape and incorporate loose clothing and long-hair. Current 3D human tracking systems from video are limited to individual people in a controlled studio environment with high-quality imagery. If these limitations are overcome this approach has the potential to provide an alternative for free-viewpoint rendering in more general scenes such as sports with multiple players and inter-player occlusion.

Given multiple camera views of a scene containing multiple moving people 3D scene reconstruction is commonly used to provide an explicit intermediate representation for rendering novel views. In subsequent sections we review the state-of-

the-art in both studio and outdoor reconstruction with respect to the requirements of free-viewpoint video.

2.2 Studio production of Free-viewpoint Video

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.[26] 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 system with small numbers of cameras (6—12) have used the visual-hull from multiple view silhouettes [31, 35] and photo-hull [47]. Real-time free-viewpoint video with interactive viewpoint control has also been demonstrated [13, 31]. An advantage of these volumetric approaches over stereo correspondence is the use of widely spaced views reducing the number of cameras required. However, due to visual ambiguity with a limited number of views both visual and photo-hull approaches are susceptible to phantom volumes and extraneous protrusions from the true scene surface. All of these approaches typically result in a reduction of visual quality relative to the camera image sequences in rendering novel views. This is due to the approximate scene geometry which results in misalignment of images between views.

A number of subsequent approaches have exploited temporal consistency in surface reconstruction resulting in improved visual quality by removing spurious artifacts. Volumetric scene flow [47] was introduced to enforce photo-metric consistency over time in surface reconstruction. Spatio-temporal volumetric surface reconstruction from multiple view silhouettes using level-sets to integrate information over time has also been investigated [14]. Incorporation of temporal information improves the visual quality by ensuring smooth variation in surface shape over time.

Recent advances have achieved offline production of free-viewpoint video with a visual quality comparable to the captured video. Zitnick et al.[49] 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 [43, 33, 44]. This approach refines an initial visual-hull reconstruction to obtain a surface which gives accurate alignment between widely spaced views. As an alternative to refinement of geometry reconstruction direct texture correspondence and interpolation using optic-flow tech-

niques has been used to correct misalignment due to errors in geometry and achieve high-quality rendering on relatively low-resolution models [10] This approach could be advantageous in the presence of camera calibration error as it avoids the requirement for a single surface reconstruction which is consistent between all views.

In film production rigs of high-resolution digital cameras have been used to produce high-quality virtual camera sweeps at a single time instant, as popularised by the *bullet-time* shots in the Matrix. These systems typically use very narrow baseline configurations (approximately 3° spacing) of hundreds of cameras. Semi-automatic tools are used in post-production for interpolation between camera views to render a continuous free-viewpoint camera move. This technology is now in widespread use for film and broadcast advertising with acquisition of both studio and outdoor scenes, such as dolphins jumping, to produce film quality sequences (timeslicefilms.com, digitalair.com). Virtual views are limited to interpolating along preplanned paths along which the cameras are positioned. This restricts the general use of current technologies used in film production for free-viewpoint video.

2.3 Free-viewpoint video in sports and outdoor scenes

Transfer of studio technology to sports events requires acquisition over a much larger area with uncontrolled conditions such as illumination and background. Studio systems typically have all cameras focused on a fixed capture volume together with constant illumination and background appearance. To date the majority of studio systems have used fixed cameras with similar focal lengths. Achieving broadcast quality free-viewpoint video of live events such as sports taking place over large areas presents a significant additional challenge.

Initial attempts have been made to transfer studio-based reconstruction methodologies to acquisition and reconstruction of outdoor events. The Virtualized Reality^{TM} technology [26] was used in the EyeVision system to produce virtual camera sweeps as action replays for Super Bowl XXXV in 2001 (www.ri.cmu.edu/events/sb35/tksuperbowl.html). In this system the pan, tilt and zoom of more than 30 cameras spaced around the stadium were slaved to a single manually controlled camera to capture the same event from different viewpoints. Switching between cameras was used in the Super Bowl television broadcast to produce sweep shots with visible jumps between viewpoints.

More recently a number of groups have investigated volumetric [18] and image-based interpolation techniques [7, 23, 28] for rendering novel views in sports. Grau et al. [18] used a texture mapped visual-hull reconstruction obtained from multiple view silhouettes to render novel views of the players. Their approach utilises a set of 15 static cameras position around one end of a football pitch. The reconstruction allows flexible production of free-viewpoint rendering from any virtual camera viewpoint. Rendered views are lower visual quality than the captured video due to inaccuracies in the visual hull geometry causing misalignment between views which results in blur and double exposure. This work forms the basis for the free-viewpoint rendering system described in further detail in section 4.

Interpolation of novel views between the real cameras without explicit 3D reconstruction has also been investigate in the context of sports. [7, 23, 28]. Inamoto and Saito [23] allow free-viewpoint video synthesis in football by segmenting the observed camera images into three layers: dynamic foreground (players); pitch; and background (stadium). Segmentation of dynamic foreground regions corresponding to players is performed based on motion and colour. Image-based novel view synthesis is performed by morphing of the foreground layer between adjacent pairs of views. Morphing is achieved by interpolation along the corresponding intervals of the epipolar line for foreground. Results from four static camera views of the penalty area in football demonstrate interpolation of intermediate views. The approach does not take into account inter-player occlusion which may result in visual artifacts. A layered representation for the spatio-temporal correspondence and occlusion of objects for pairs of views is proposed in [7] and applied to football view interpolation. A MAP formulation is presented using the EM algorithm to estimate the object layer segmentation and parameters for correspondence and occlusion. This approach implicitly represents the 3D scene structure using correspondence and does not require prior camera calibration to interpolate intermediate views. Results are presented for interpolation of views between a pair of fixed cameras covering the goal area in football. The layered representation allows view interpolation of intermediate views in the presence of significant inter-player occlusion with a visual quality comparable to the source video. Interpolation based approaches to novel view rendering in sports are limited to the generation of intermediate views between the match cameras. Potential advantages include avoiding the requirement for camera calibration or explicit 3D reconstruction and direct rendering of novel views from the camera images. Extrapolation of novel viewpoints away from the captured views is limited due to the lack of explicit geometry and may result in unrealistic image distortions.

All current systems for free-viewpoint video in sports use special rigs of auxiliary cameras to capture footage over a wide-area and result in a visual quality below that acceptable for TV broadcast. In addition due to the use of static cameras players are captured at relatively low resolution. One solution to this is the remote controlled camera slaved to a single operator used in the EyeVision system allowing close-ups of the critical action from multiple views. The use of auxiliary cameras in addition to the match cameras adds a significant cost to rigging which may prohibit deployment of free-viewpoint video at all but key sports events such as cup finals or the SuperBowl. The added-value in broadcast production depends on the trade-off between potential use and costs such as rigging additional cameras.

An attractive alternative would be to use the manually controlled match cameras used for live broadcast which primarily focus on the play action of interest. Use of match cameras would also avoid the prohibitive costs associated with rigging of additional cameras. Subsequent sections of this article specify the multiple camera capture requirements for sports and identify practical limitations in using either match cameras or special camera rigs. LiberoVision (www.liberovision.com) recently introduced a commercial system for interpolation between pairs of match camera views in football broadcast. This system has the advantage of only using the existing broadcast cameras. Current technology is limited to viewpoint inter-

polation at a single static frame and does not allow free-viewpoint rendering of the game play from novel views such as the referee viewpoint. The system is currently in use for half-time and post-match commentary of football. The Piero system (www.bbc.co.uk/rd/projects/virtual/piero/) developed by BBC R&D allows annotation of the broadcast video footage together with limit change in viewpoint using player billboards together with camera calibration to extrapolate views around a single camera. Red Bee Media commercialised the Piero system and have introduced free-viewpoints of static plays using manually posed player models to illustrate the action. Current commercial technologies are limited to view-interpolation at a single frame and do not allow novel camera views of the action.

The *iview* free-viewpoint video system presented in this chapter aims to allow 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. *iview* has introduced automatic methods for online calibration, segmentation, reconstruction and rendering of free-viewpoint video for sports broadcast production.

3 Specification of Requirements for TV Sports Production

There are three critical issues for use of 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.

3.1 Visual Quality for Broadcast Production

The hardest technical constraint for 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. Visual artifacts such as flicker and jitter are unacceptable. In addition, the visual resolution of the broadcast footage should be preserved without spatial blur or ghosting artifacts. The occurrence of visual artifacts due to reconstruction errors such as false volumes due to multiple player occlusion is also unacceptable. In broadcast production the visual quality is a primary consideration which must be satisfied for a new technology be widely used.

High-definition (HD) cameras are now widely used for acquisition at live sports events together with increasing use of HD transmission to the viewer. Free-viewpoint rendering needs to achieve HD quality for rendering of full-screen shots. For commentary free-viewpoint video rendering of novel views may also be used as inpicture inserts at a lower resolution to illustrate different views of the play requiring

a slightly lower resolution. However, the visual-quality at the displayed resolution should be equivalent to that of conventional TV broadcast footage.

3.2 Production Requirements on timing

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

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 that 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).

Aiming for the *replay* timeframe is an extreme challenge. Calibration, definition of viewpoint and rendering must all be done within a small number of seconds. Additionally, the production crew must make split second decisions about which replays to present before the match action continues, so must be 100% sure of the reliability of any view offered. In the case of review or analysis, time pressure is much more relaxed. Particularly for analysis, it may even be possible to prepare novel view sequences remotely, having had video sequences transmitted from the outside broadcast venue. Reliability levels required for non time-critical post-match analysis are significantly lower as it is possible to have an operator check sequence quality. The added value to sports broadcast production of free-viewpoint video decreases rapidly with the time-taken for shot production. Consequently the level of expenditure available for production systems will also decrease depending on the production time.

3.3 Acquisition Requirements

Production of sports events such as football for live broadcast typically use 12-18 match cameras at key locations around the stadium. This number may be increased at major sports events such as the SuperBowl or a cup final. In the 2006 FIFA World Cup 25 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

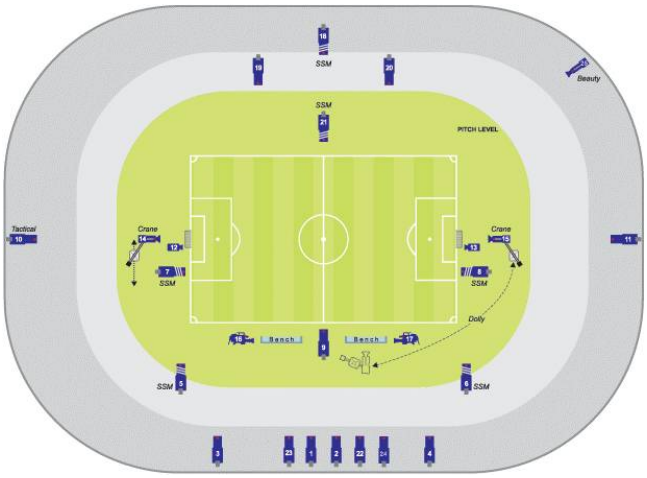


Fig. 1 Typical stadium broadcast camera layout for a major sporting event. Out of 25 cameras 1-4,10,11,19,20 provide potentially useful views, 5-8,18,21 are high-speed cameras and the remainder are at pitch level providing insufficient coverage for calibration or reconstruction.

stadium to avoid disorienting the viewer with switches to reverse views. Additional high-speed, overhead and reverse side camera views are typically used for action replays in commentary. 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. An exception to this is events such as penalties where more camera angles may be available on a particular player. In addition the framing varies between cameras from zoomed in shots of individual players with little background visible for calibration to wider shots of groups of players. Figure 2 shows a typical set of shots from match cameras for a penalty event during a cup-final match where there were 15 match cameras in total only 8 of which captured the event.

Ideally the broadcast match camera views would be used for free-viewpoint video production. However, as shown in Figure 2 only a small proportion of these will have overlapping fields of view even for a specific event in the game where normal play is suspended. In addition as the match cameras are individually operated it is necessary to calibrate the camera orientation and zoom from the broadcast footage. Calibration requires fixed features such as pitch markings to be visible in the shot which may not be the case for zoomed in shots of individual players or cameras at pitch level. Therefore, in practice it may be desirable to use auxiliary cameras in addition to the match cameras to ensure sufficient coverage for high-quality free-viewpoint video production. Placement of cameras may be restricted to the pre-defined camera positions due to access restrictions. The requirement for additional cameras introduces a significant cost for rigging and operation. This cost must be justified in terms of the additional value of the free-viewpoint video shots.

Fig. 2 Broadcast match camera views for a penalty event during a cup-final. Of the 15 match cameras only the 8 views shown were of the penalty.

In addition to the camera placement simultaneous access to the broadcast camera footage from multiple camera feeds is required. Commonly only a limited number of camera views of interest for a particular time instant are stored. These are typically stored in a compressed format for subsequent editing in post match commentary. Acquisition and storage of multiple camera views is required for subsequent free-viewpoint video rendering.

3.4 Production Requirements for use of free-viewpoint video

An important consideration in the development of free-viewpoint technology for sports is the end-use in production. To justify the additional cost and complexity of free-viewpoint video technology the content produced must add value to the programme. The added value to broadcasters and viewers must go beyond the short-term novelty of a special-effect such as the virtual camera fly-through shots produced from the EyeVision system at the SuperBowl. Added value requires the production of free-viewpoint camera shots which add to the experience of viewing the game and the associated match commentary. Examples of shots which might add value in a football (soccer) match include the referees view during a specific event, the goal-keepers view or a players viewpoint. Virtual cameras able to generate free-viewpoint video of such views during or shortly after a particular event are more likely to add-value and be exploited in the programme production.

4 A Free-viewpoint System for Sports TV Production

This section presents a system being developed for free-viewpoint video in TV production of sports events. The system is being developed to utilise footage from both the manually operated 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 free-viewpoint video production. The current prototype system has been used for production trials capturing test footage of live events for off-line processing to evaluate the performance with respect to the broadcast production requirements, as specified in section 3. Results of evaluation for both international soccer and rugby are presented.

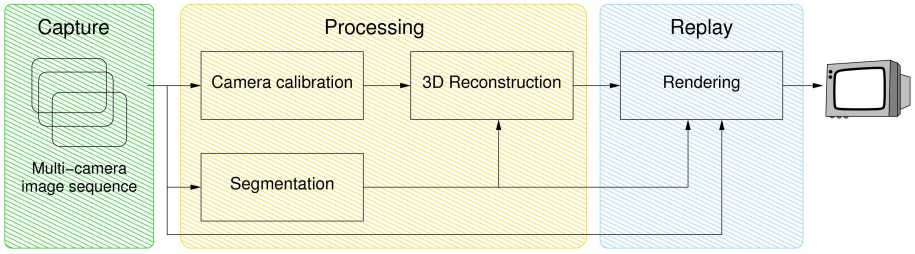


Fig. 3 Overview of the iview free-viewpoint video system.

4.1 System Overview

An overview of the free-viewpoint TV 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 direct 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 free-viewpoint video. Aperture correction is also applied to each video sequence to correct for the camera edge enhancement used in standard broadcast footage.

The calibration of moving match cameras achieves an rms error of 1-2 pixels. Likewise matting in relatively uncontrolled outdoor conditions with changing illumination achieves a segmentation within 1-2 pixels of the true foreground with the addition of background clutter for the crowd, hoardings and on-pitch advertising. The accumulation of errors from calibration and matting can cause gross errors in the reconstruction of the scene such as loss of limbs as their image size may also be of the order of a few pixels. Therefore, robust algorithms have been developed for scene reconstruction and free-viewpoint rendering in the presence of matting and calibration errors.

Initial reconstruction from multiple views is performed using a conservative visual-hull algorithm, taking into account the maximum error to ensure that the foreground is within the reconstructed volume. This allows robust reconstruction of a coarse approximation of the scene structure. However, scene rendering of novel views using the conservative visual hull surface results in visual artifacts due to errors in the geometry resulting in poor alignment of the foreground between views. Given the conservative visual hull as an initial reconstruction techniques have been developed for view-dependent refinement of both the surface reconstruction and foreground matting. Initialisation of the view-dependent reconstruction from a conservative approximation of the scene structure ensures that the resulting refined segmentation accurately separates the foreground and background in regions of visual ambiguity by integrating information from multiple views. View-dependent surface refinement enables robust reconstruction in the presence of camera calibration error for the rendering of high-quality novel viewpoints. The replay module renders the captured scene in realtime using the computed 3D model and the original camera images deploying view-dependent texture mapping [8]. The replay module is designed to work at interactive rates allowing free-viewpoint camera path planning for editorial shot production.

Currently the capture, calibration, matting and visual-hull together with interactive rendering of novel views can be performed in real-time. Reconstruction refinement is currently performed as an off-line process taking several minutes per frame. The capture, reconstruction and rendering components are integrated in a pipeline for use in live production.

4.2 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, this is costly and sometimes very difficult, for example if it is not possible to gain access to the camera mounts, or if cameras are mounted on non-rigid structures such as a crane. 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 are 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 football, the layout of some pitch markings (such as those around the goal) are fully specified, but the overall dimensions vary between grounds. Pitches are also often raised in the middle by $30-50cm$ for drainage. For soccer the Football Association specifies that the pitch length must be in the range 90-120m, and the width 45-90m; for international matches, less variation is allowed (length 100-110m and width 64-75m). It is thus necessary to obtain a measurement of the actual pitch.

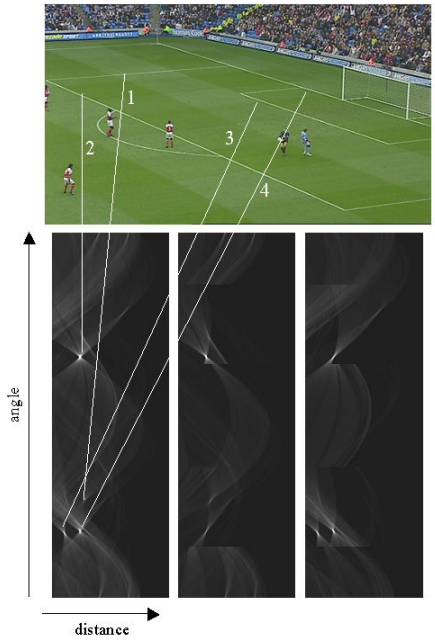


Fig. 4 Calibration initialisation using Hough space(a) Complete (b) top or left half (c) bottom or right half

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[46]. 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 line tracking using a multi-hypothesis approach based on 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 implemented in such a way that the process can be carried out in about one second. We also take advantage of the fact that TV cameras at outside broadcasts are often

mounted on fixed pan/tilt heads, so that their position remains roughly constant over time. Further details can be found in [46], the stages of the calibration 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. Evaluation of this approach found that that an accuracy of 0.3m for the initial camera position estimate was obtained with 10-20 images across the pitch. This estimate computed offline is used to initialise the camera tracking and can be automatically refined during tracking from the online calibration.

Initialisation: 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 4 presents examples of the Hough space corresponding to the single frame shown. The initialisation has been found to reliably give an initial estimate of the camera calibration on sports footage in less than 1s from an unknown orientation and focal-length [46]. Figure 6 shows some example images successfully used for initialisation of the calibration.

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 straight line is fitted through each set of points, and an iterative process is used to minimise the distance in the image between the ends of the observed line and the corresponding line in the model. Figure 5 shows the estimated camera angles for a 20s sequence. Evaluation of the real-time tracking on sports footage [46] shows that the noise error in the estimated pan angle is approximately 0.02° which corresponds to 1 image line.

4.3 Foreground Segmentation

For the segmentation of players colour-based methods, like 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 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 reconstruc-

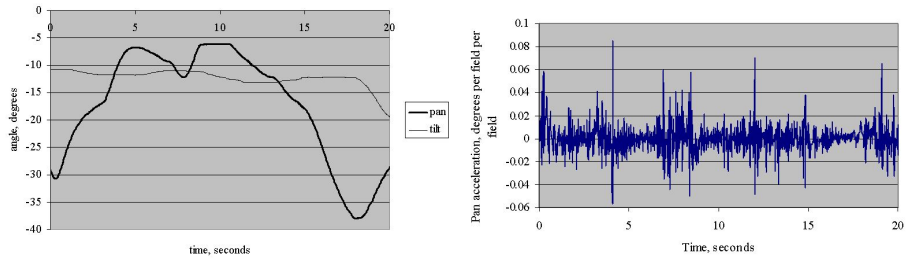


Fig. 5 Calibration parameters for a 20s sequence (left) Pan and tilt angles (right) second derivative of pan

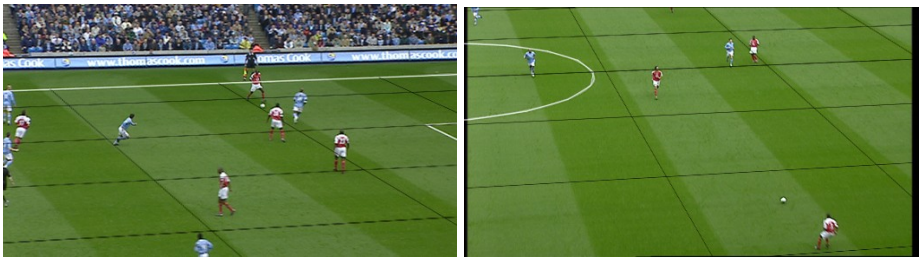


Fig. 6 Online calibration from pitch markings for two match cameras

tion. Therefore two improved methods have been implemented and tested: A global colour-based 'k-nearest-neighbour classifier' 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 [22].

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 RGB values I:

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

The segmentation is then given by:

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

In order to get continuous values a soft key can be obtained using a second radius r_2. See [16] 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:

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

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 4.2).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).

4.3.1 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.

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

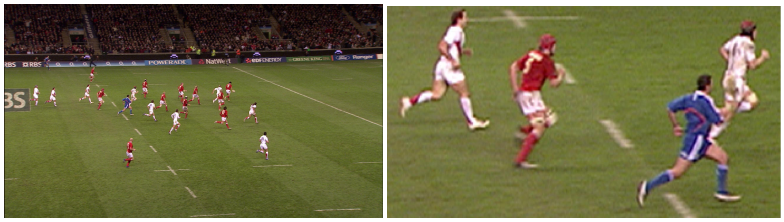


Fig. 8 Image of a sport scene from a broadcast camera (left) and detail (right).

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 simple linear 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.

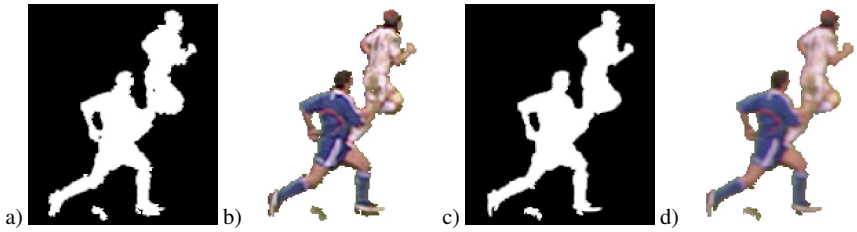


Fig. 9 Detail of difference key computed on broadcast image (a+b) and after compensation of aperture corection (c+d).

Figure 9 shows results of the difference keying applied to the original broadcast image (shown in Figure 9 a+b) and after compensation of the aperture correction (Figure 9 c+d). 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 [17].

4.4 Free-viewpoint video production from match cameras

Free-viewpoint 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.

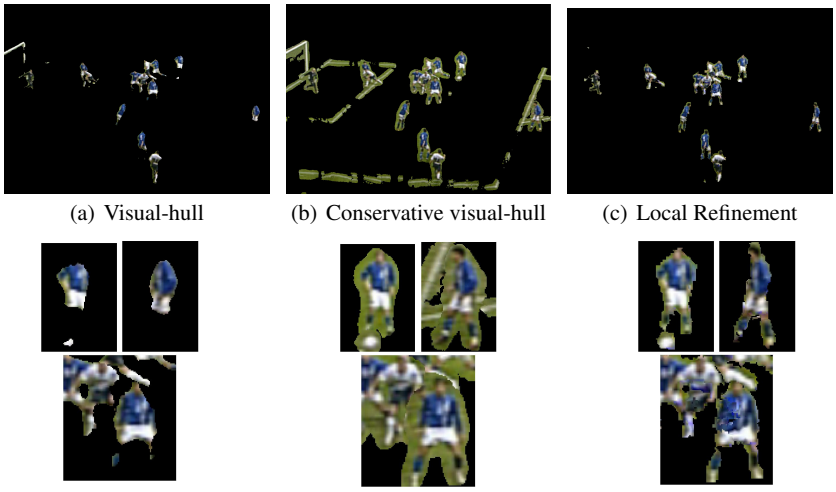
4.4.1 Conservative Visual-Hull

The conservative visual-hull (CVH) [27] 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 football 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.

4.4.2 Local View-dependent Visual-hull 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 a specific camera viewpoint to obtain a surface approximation which aligns the adjacent images and



(a) Visual-hull (b) Conservative visual-hull (c) Local Refinement

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

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[43] and has recently been extended for refinement in outdoor sports scenes [20, 27]. 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 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 [20]. 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 is formulated in a Bayesian framework [20]. The maximum likelihood solution leads to the minimisation of an energy function:

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

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

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

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

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 4.3). The probability of a pixel belonging to a foreground or background layer is evaluated according to a combination of local per-pixel and a global Gaussian mixture colour models.

The contrast energy term encourages changes in layer label to occur in regions of high image contrast and is based on the difference in image values for adjacent pixels p and q as:

$$E_{contrast}(l) = \sum_{(\mathbf{p},\mathbf{q})\in\mathscr{N}} \frac{\pi}{4} e_{contrast}(l,\mathbf{p},\mathbf{q}), \tag{7}$$

where

$$e_{contrast}(l,\mathbf{p},\mathbf{q}) = \begin{cases} 0 & if\ l_\mathbf{p} = l_\mathbf{q}, \\ \exp(-\beta C(\mathbf{I_p},\mathbf{I_q})) & otherwise. \end{cases} \tag{8}$$

$\mathscr{N}$ denotes the pixel neighborhood, and $C(\cdot,\cdot)$ is a squared colour distance between adjacent pixels.

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

$$E_{match}(l,d) = \sum_{\mathbf{p}\in\mathscr{P}} e_{dense}(l,d,\mathbf{p}) + \sum_{\mathbf{p}\in\mathscr{P}} e_{sparse}(l,d,\mathbf{p}) \tag{9}$$

where $e_{dense}(l,d,\mathbf{p})$ is the photo-consistency measure with respect to adjacent camera views at a depth d along the optical ray passing through pixel $\mathbf{p}$. A robust photo-consistency score is evaluated between the reference and adjacent camera. The sparse correspondence energy $e_{sparse}(l,d,\mathbf{p})$ is introduced to constrain the depth where feature correspondences have been detected between views using a Hessian-affine feature detector with SIFT descriptors which is robust to changes in viewpoint and illumination respectively.

The smoothness term constrains the local continuity of the reconstructed surface as:

$$E_{smooth}(l,d) = \sum_{(\mathbf{p},\mathbf{q})\in\mathscr{N}} w_{\mathbf{p},\mathbf{q}} e_{smooth}(l,d,\mathbf{p},\mathbf{q}), \tag{10}$$

where

$$e_{smooth}(l,d,\mathbf{p},\mathbf{q}) = \tag{11}$$
$$\begin{cases} \min(|d_{\mathbf{p}} - d_{\mathbf{q}}|, d_{max}) & if\ l_{\mathbf{p}} = l_{\mathbf{q}}\ and\ d_{\mathbf{p}}, d_{\mathbf{q}} \neq \mathscr{U}, \\ 0 & if\ l_{\mathbf{p}} = l_{\mathbf{q}}\ and\ d_{\mathbf{p}}, d_{\mathbf{q}} = \mathscr{U}, \\ d_{max} & otherwise. \end{cases}$$

Discontinuities between layers are assigned a constant smoothness penalty equal to d_{max}, while within a layer the penalty is defined as a truncated linear distance. This defines a discontinuity preserving function which does not over-penalise large discontinuities within a layer. Iso-surfaces of the conservative visual hull are sampled which results in a reconstructed surface which is biased towards the visual hull shape. This is useful in the case where other reconstruction cues are weak which commonly occurs in regions of uniform appearance such as the players shirts. Optimisation of the energy defined by 5 is performed using the α-expansion graph-cut [2] to obtain an efficient approximation to the global solution.

4.5 Free-viewpoint Rendering Results

Production trials of the iview system have been conducted for football 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 football 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. Figure 11 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 the complete 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 12. 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. In contrast the multiple view segmentation and reconstruction refinement algorithm presented in section 4.4.2 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

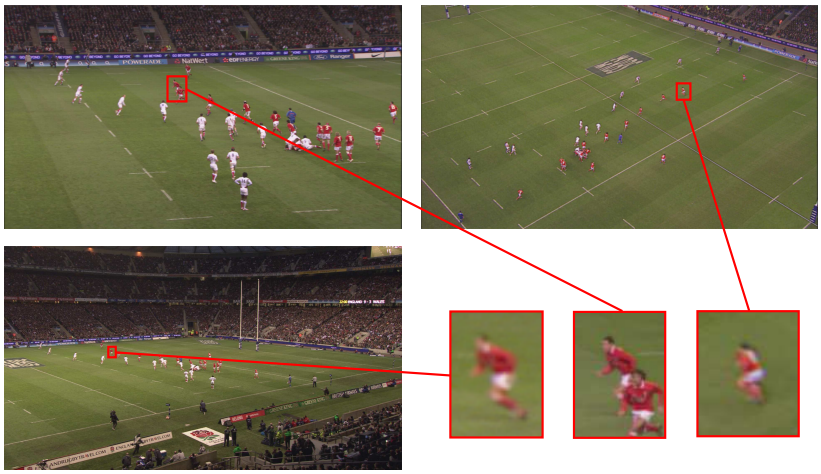


Fig. 11 Two moving broadcast camera views (top) and a locked-off camera view (bottom-left) from a rugby match.

overcome single view visual ambiguities. The refined segmentation allows improvements in the reconstruction quality and subsequent free-viewpoint rendering.

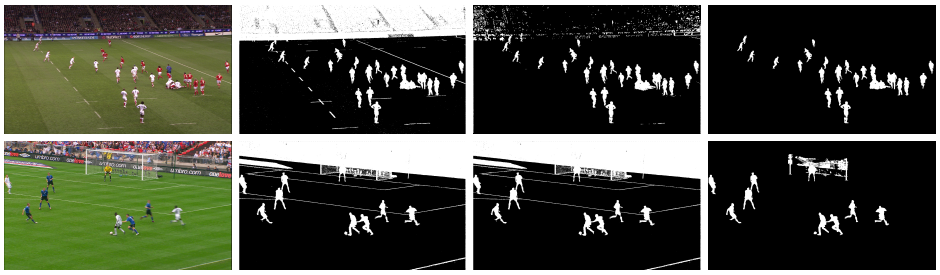


Fig. 12 Example of segmentation results on rugby (top) and soccer (bottom) data (see attached video for full sequence).

Free-viewpoint rendering results for rugby and soccer trials are shown in Figures 13 and 14. In both cases the virtual camera sequences were produced for production trials to generate specific camera views which give added value to the match commentary. 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 at a single frame or whilst the action is taking place according to the production requirements. All sequences were generated with automatic calibration, 2D segmentation, reconstruction and refinement. View-dependent rendering is performed using the view-dependent geometry to ren-

der images from the adjacent views with dynamic feathering over cameras rendered from different views[19]. The statdium backgrounds are manually modelled using either images from the captured sequences as in Figure 13(a) or a synthetic appearance Figure 13(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 can not be isolated. The approach developed in the iview project 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 13(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 iview 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 cameras. The current iview 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.

5 Conclusions

Free-viewpoint video in live sports broadcast production presents significant challenges to achieve a visual quality comparable to captured video with minimal delay from the manually controlled moving and zooming match cameras. In addition, capture of stadium sports such as soccer and rugby requires acquisition over a large area with relatively uncontrolled conditions. Advances over previous studio based free-viewpoint video are necessary to achieve robust reconstruction for wide-baseline camera views in the presence of calibration and segmentation errors for manually operated moving camera which capture the scene at different resolutions. The iview free-viewpoint video system presented in this chapter incorporates automatic calibration, segmentation and reconstruction from the broadcast cameras. Online calibration uses the pitch lines to estimate extrinsic and intrinsic parameters. Scene reconstruction is performed in two stages: a conservative visual-hull reconstruction provides a robust initial reconstruction in the presence of calibration and matting errors; and view-dependent refinement of reconstruction and segmentation has been introduced which combines cues from multiple views. This approach is

(a) Rugby virtual camera sequences at 20 frame intervals

(b) Virtual camera closeup

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

Fig. 13 Free-viewpoint video rendering of rugby to show pitch level views for commentary

robust to online calibration and 2D image segmentation errors enabling reconstruction from wide-baseline moving and zooming broadcast cameras. Production trials of the iview system have been conducted on soccer and rugby to generate novel camera sequences which add to the match coverage. The iview system allows automatic reconstruction and free-viewpoint rendering from the match cameras. Results demonstrate free-viewpoint rendering with a visual quality comparable to the captured video.

Fig. 14 Free-viewpoint video rendering of soccer to show an offside incident

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

Acknowledgements: This work was supported by TSB Technology Programme Project TP/3/DSM/6/l/15515 and EPSRC EP/D033926/1 iview: Free-viewpoint video for entertainment content production.

References

1. E.de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun. Performance Capture from Sparse Multi-view Video. *Proc. ACM SIGGRAPH*, 27(3), 2008.
2. Y. Boykov, O. Veksle, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(11):1222—1239, 2001.
3. A. Broadhurst, T.W. Drummond, and R. Cipolla. A probabilistic framework for space carving. In *International Conf. on Computer Vision*, pages 388—393, 2001.
4. C. Buehler, M. Bosse, L. McMillan, S. Gortler, and M. Cohen. Unstructured Lumigraph Rendering. In *Proc. ACM SIGGRAPH*, pages 425—432, 2001.
5. J. Carranza, C. Theobalt, M. Magnor, and H.-.P. Seidel. Free-viewpoint video of human actors. In *Proc. ACM SIGGRAPH*, pages 565—577, 2003.
6. S.E. Chen and L. Williams. View Interpolation for Image Synthesis. In *Proc. ACM SIGGRAPH*, 1993.
7. K. Connor and I. Reid. A Multiple View Layered Representation for Dynamic Novel View Synthesis. In *British Machine Vision Conference*, 2003.
8. P. Debevec, Y. Yu, and G. Borshukov. Efficient view-dependent image-based rendering with projective texture-mapping. *9th Eurographics Rendering Workshop*, pages 105–116, 1998.
9. J. Deutscher, A. Davidson, and I. Reid. Automatic partitioning of high-dimensional search spaces associated with articulated body motion capture. In *Conference on Computer Vision and Pattern Recognition*, pages 669—676, 2001.
10. M. Eisemann, B. Decker, M.A. Magnor, P. Bekaert, E. Aguiar, N. Ahmed, C. Theobalt, and A. Sellent. Floating Textures. *Comput.Graph.Forum*, 27(2):409—418, 2006.
11. F.S. Franco and E. Boyer. Exact Polyhedral Visual Hulls. In *British Machine Vision Conference*, pages 329–338, 2003.
12. F.S. Franco and E. Boyer. Fusion of Multi-View Silhouette Cues Using a Space Occupacy Grid. In *International Conf. on Computer Vision*, pages 1747—1753, 2005.
13. F.S. 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.
14. B. Goldluecke and M. Magnor. Space-Time Isosurface Evolution for Temporally Coherent 3D Reconstruction. In *Conference on Computer Vision and Pattern Recognition*, pages S–E, Washington, D.C., USA, July 2004. IEEE Computer Society, IEEE Computer Society.
15. O. Grau, A. Hilton, J. Kilner, G. Miller, T. Sargeant, and J. Starck. A Free-Viewpoint Video System for Visualisation of Sports Scenes. *SMPTE Motion Imaging Journal*, May/June, 2007.
16. O. Grau, G.A. 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.
17. Oliver Grau and Jim Easterbrook. Effects of camera aperture correction on keying of broadcast video. In *Proc. of the 5rd European Conference on Visual Media Production (CVMP)*, 2008.
18. Oliver Grau, Michael Prior-Jones, and Graham Thomas. 3d modelling and rendering of studio and sport scenes for tv applications. In *Proc. of WIAMIS*.
19. J-Y. Guillemaut, J. Kilner, J. Starck, and A. Hilton. Dynamic Feathering: Minimising Blending Artefacts in View Dependent Rendering. In *IET European Conference on Visual Media Production*, pages 1—8, 2007.
20. J.Y. Guillemaut, A. Hilton, J. Starck, J.J. Kilner, and O. Grau. A Baysian Framework for Simultaneous Reconstruction and Matting . In *IEEE Int.Conf. on 3D Imaging and Modeling* , 2007.
21. R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.
22. Peter Hillman, John Hannah, and David Renshaw. Foreground/background segmentation of motion picture images and image sequences. *IEE Transactions on Vision, Image and Signal Processing*, 142(4):387–397, August 2005.
23. N. Inamoto and H. Saito. Virtual Viewpoint Replay for a Soccer Match by View Interpolation From Multiple Cameras. *IEEE Trans.Multimedia*, 9(6):1155—1166, 2007.

24. M. Irani, T. Hassner, and P. Anandan. What Does the Scene Look Like from a Scene Point ? In *European Conference on Computer Vision*, 2002.
25. T. Kanade and P. Rander. Virtualized reality: Constructing virtual worlds from real scenes. *IEEE MultiMedia*, 4(2):34—47, 1997.
26. T. Kanade, P.W. Rander, and P.J. Narayanan. Virtualized reality: Constructing virtual worlds from real scenes. *IEEE Multimedia*, 4(1):34–47, 1997.
27. J.J. Kilner, J. Starck, A. Hilton, J.Y. 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.
28. K. Kimura and H. Saito. Player viewpoint video synthesis using multiple cameras. In *IEE European Conference on Visual Media Production*, pages 112—121, 2005.
29. A. Laurentini. The visual hull concept for silhouette based image understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(2):150–162, 1994.
30. W. Matusik, C. Buehler, R. Raskar, S.J. Gortler, and L. McMillan. Image-based visual hulls. *Proceedings of ACM SIGGRAPH*, pages 369–374, 2000.
31. W. Matusik, C. Buehler, R. Raskar, S.J. Gortler, and L. McMillan. Image-based visual hulls. *Proceedings of ACM SIGGRAPH*, pages 369–374, 2000.
32. G. Miller, A. Hilton, and J. Starck. Interactive Free-viewpoint Video. In *IEE European Conf. on Visual Media Production*, pages 50—59, 2005.
33. G. Miller, J.R. Starck, and A. Hilton. Projective Surface Refinement for Free-Viewpoint Video. In *IET European Conference on Visual Media Production*, pages 153–162, 2006.
34. T. Moeslund, A. Hilton, and V. Kruger. A Survey of Advances in Vision-Based Human Motion Capture and Analysis. *Computer Vision and Image Understanding*, 104(2-3):90—127, 2006.
35. S. Moezzi, L.C. Tai, and P. Gerard. Virtual view generation for 3d digital video. *IEEE Multimedia*, 4(1):18–25, 1997.
36. Saied Moezzi, Arun Katkere, Don Y.Kuramura and Ramesh Jain. Reality Modeling and Visualization from Multiple Video Sequences. *IEEE Computer Graphics and Applications*, pages 58–63, November 1996.
37. D. Scarstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1):7—42, 2002.
38. C.M. Seitz and C.R. Dyer. Photorealistic scene reconstruction by voxel coloring. *International Journal of Computer Vision*, 35(2):1–23, 1999.
39. S.M. 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*, pages 519—528, 2006.
40. S.M. Seitz and C.R. Dyer. View morphing: Synthesizing 3D metamorphosis using image transforms. *Proc. ACM SIGGRAPH*, pages 21—30, 1996.
41. H.-Y. Shum, S.B. Kang, and S.-C. Chan. Survey of image-based representations and compression techniques. *IEEE Transactions on Circuits and Systems for Video Technology*, 13(11), 2003.
42. G. Slabaugh, B. Culbertson, and T. Malzbender. A survey of methods for volumetric scene reconstruction from photographs. In *Volume Graphics*, 2001.
43. J. Starck and A. Hilton. Virtual view synthesis of people from multiple view video. *Graphical Models*, 67(6):600—620, 2005.
44. J. Starck and A. Hilton. Surface Capture for Performance-Based Animation. *IEEE Computer Graphics and Applications*, 27(3):21—31, 2007.
45. T. Stich, C. Linz, C. Wallraven, D. Cunningham, and M. Magnor. Perception-motivated Interpolation of Image Sequences. *In Proc. ACM APGV*, pages 97—106, 2008.
46. G.A. Thomas. Real-time Camera Pose Estimation for Augmenting Sports Scenes. In *European Conference on Visual Media Production*, pages 10—19, 2006.
47. S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade. Three-dimensional scene flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(3), 2005.
48. D. Vlasic, I. Baran, W. Matusik, and J. Popović. Articulated mesh animation from multi-view silhouettes. In *Proc. ACM SIGGRAPH*, pages 1–9, 2008.

49. C.L. Zitnick, S.B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. High-quality video view interpolation using a layered representation. In *Proc. ACM SIGGRAPH*, pages 600—608, 2004.