Combining Panoramic Image and 3D Audio Capture with Conventional Coverage for Immersive and Interactive Content Production

Graham Thomas (BBC), Oliver Schreer (HHI), Ben Shirley (Salford Uni), Jens Spille (Technicolor)
Combining Panoramic Image and 3D Audio Capture with Conventional Coverage for Immersive and Interactive Content Production

Graham Thomas (BBC), Oliver Schreer (HHI), Ben Shirley (Salford Uni), Jens Spille (Technicolor)

Abstract

The media industry is currently being pulled in the often-opposing directions of increased realism (high resolution, stereoscopic, large screen) and personalisation (selection and control of content, availability on many devices). A capture, production and delivery system capable of supporting both these trends is being developed by a consortium of European organisations in the EU-funded FascinatE project. This paper reports on the latest developments and presents results obtained from a test shoot at a UK Premier League football match. These include the use of imagery from broadcast cameras to add detail to key areas of the panoramic scene, and the automated generation of spatial audio to match the selected view. The paper explains how a 3D laser scan of the scene can help register the cameras and microphones into a common reference frame.

This document was originally published in the proceedings of IBC2011. The slides shown during the presentation are included as an appendix.

Additional key words: Ambisonics, image registration
COMBINING PANORAMIC IMAGE AND 3D AUDIO CAPTURE WITH CONVENTIONAL COVERAGE FOR IMMERSIVE AND INTERACTIVE CONTENT PRODUCTION

G.A. Thomas\textsuperscript{1}, O. Schreer\textsuperscript{2}, B. Shirley\textsuperscript{3}, J. Spille\textsuperscript{4}

\textsuperscript{1}BBC R&D, UK; \textsuperscript{2}Fraunhofer HHI, Germany; \textsuperscript{3}University of Salford, UK; \textsuperscript{4}Technicolor, Germany

ABSTRACT

The media industry is currently being pulled in the often-opposing directions of increased realism (high resolution, stereoscopic, large screen) and personalisation (selection and control of content, availability on many devices). A capture, production and delivery system capable of supporting both these trends is being developed by a consortium of European organisations in the EU-funded FascinatE project. This paper reports on the latest developments and presents results obtained from a test shoot at a UK Premier League football match. These include the use of imagery from broadcast cameras to add detail to key areas of the panoramic scene, and the automated generation of spatial audio to match the selected view. The paper explains how a 3D laser scan of the scene can help register the cameras and microphones into a common reference frame.

INTRODUCTION

It is an often-expressed view that the TV industry should adopt a common video production format, which would not only be unified across the world, but also support a wide range of applications. Traditionally, the shot selection, framing and audio mix is designed to support the particular ‘story’ that the director is aiming to tell, and will have been produced with a particular reproduction system in mind (e.g. widescreen HD with 5.1 surround sound). Although some provisions are sometimes made to allow repurposing for other devices (such as maintaining a 4:3 ‘safe area’ within a 16:9 frame), such content is not ideal for supporting extreme variations in viewing device, e.g. from mobile phones to ultra-high-resolution immersive projection systems with 3D audio support. Audiences increasingly expect to be able to control their experience, for example by selecting one of several suggested areas of interest, or even by freely exploring the scene themselves. Traditionally-produced content offers very limited support for such functionality. Whilst such a degree of freedom may not be appropriate for all kinds of content, it has the potential to add useful interactivity to any kind of programme where there is no obvious single ‘best’ shot that will satisfy all viewers.

An approach to overcoming the limitations of current production systems to help meet these requirements is the so-called ‘format agnostic’ approach [1]. The main idea of this is to develop a completely new production system, which does not use fixed numbers of frames, lines and pixels, or even geometry. Such an approach requires a paradigm shift in video production, towards capturing a format-agnostic representation of the whole scene from a given viewpoint, rather than the view selected by a cameraman based on assumptions about the viewer’s screen size, loudspeaker set-up and interests.

The ideal format-agnostic representation of a scene would involve capturing a very wide angle view of the scene from each camera position, sampled at a sufficiently high
resolution that any desired shot framing and resolution could be obtained. However, this is not only impractical, but would be wasteful, as less interesting areas of the scene would be captured at the same high resolution as the key areas of interest. This leads to the concept of a ‘layered’ scene representation, where several cameras with different spatial resolutions and fields-of-view can be used to represent the view of the scene from a given viewpoint. The views from these cameras can be considered as providing a ‘base’ layer panoramic image, with ‘enhancement layers’ from one or more cameras more tightly-framed on key areas of interest. Other kinds of camera, such as high frame-rate or high dynamic range, could add further layers in relevant areas. This ‘layered’ concept can be extended to audio capture, by using a range of microphone types to allow capture of the ambient sound field, enhanced by the use of additional microphones to capture localised sound sources at locations of interest. This allows an audio mix to be produced to match any required shot framing, in a way that can support reproduction systems ranging from mono, through 5.1, to higher-order Ambisonics (HOA) or wave field synthesis (WFS).

This paper presents some of the latest results of the EU-funded ‘FascinatE’ project, which is developing a capture, delivery and reproduction system to evaluate the concepts outlined above. The project addresses several different levels of interactivity: at simplest, the production tools developed could be used to allow local or specialist broadcasters to customize and tailor coverage of live events for a specific audience. In this scenario, the users’ experience will not be interactive although will be improved by being tailored to their locality and interests (for example, by showing a sporting event in a manner designed for supporters of a particular team). At the other extreme, all captured content could be delivered to the user. This would allow them to switch between a number of shot sequences selected by the director, optimised locally for their particular screen size. Users could even construct and define their own shot selection and framing, with matching audio that they could further customise, for example by adding various commentary channels.

The following section describes the approach being taken to scene capture for both audio and video, and how a 3D laser scan of the scene can be used to register all sources in a common reference frame. This is followed by a report on a test capture carried out at a Premier League football match in October 2010, illustrating the first practical application of the ideas to acquire a data set to support the work of the project. Two specific aspects of production using the layered scene are then discussed: the use of conventional HD broadcast cameras to provide additional detail in key areas, and the rendering of the captured audio to match the chosen view of the scene. Further details of the way in which the project is handling audio may be found in [2], and a discussion of the approach being taken to the delivery network and end-user terminal is given in [3].

**SCENE CAPTURE**

**Video**

Building on the concept of a layered scene representation, the approach taken by the FascinatE project is to make use of any available video feeds from conventional broadcast cameras, and capture additional very-wide-angle images from one or more locations co-sited with these cameras. The wide-angle capture makes use of an ultra-high resolution omni-directional camera - the so called OmniCam (see Figure 1). With this system a full 180° panoramic view can be captured resulting in a total resolution of 7k x 2k pixels. Details of the system can be found at [1].
Due to the high resolution of the captured image, fast-moving objects in the foreground become blurred, due to the current relatively low capturing frame rate of 30 fps. Hence, in the next revision of this system, a new camera [5] will be used which overcomes this limitation. This new camera operates at 50/60 fps and moreover, is equipped with a high-quality sensor with high dynamic range, low noise, and brilliant image quality, especially for difficult lighting situations. The use of a high dynamic range camera is particularly important for panoramic imaging applications, as the field-of-view is very likely to encompass both very bright areas (such as sky) and very dark areas (such as shadows).

Audio
The FascinatE project presents a number of interesting challenges for audio capture; firstly the format-agnostic approach of the system requires all audio to be captured in such a way that they be rendered across the full range of current reproduction systems, and secondly, that the audio is in a form that can be rendered to take into account the interactive control that the user will have.

Audio reproduction formats represented in the FascinatE project include stereo, HRTF generated binaural reproduction, 5.1, 7.1 surround systems, higher-order ambisonics and wavefield synthesis.

A particular challenge for audio is posed by the necessity within FascinatE to match the sound of the event to the visual effect of zooming into the picture. Although in reality the user is zooming into a 2D video, the visual effect in some cases will be that the user's position travels past objects that will move to the sides and behind the viewing position as they move out of shot. For this reason FascinatE audio must have a depth dimension that has to be mapped to the panoramic 2D video scene. For example, if while watching a football match the user zooms past the ball position to a region of interest at the opposite side of the pitch, their expectation is likely to be that the sound of the ball being kicked will move behind their new viewpoint.

To allow audio to be reproduced to match the visual appearance of the scene it is necessary not just to capture a sound field from the camera position, but instead to capture 'audio objects' with appropriate coordinate positions so that they may be rendered to any point around the user. The capture mechanism to allow this feature is very much dependent on the particular situation of the recording. For some events close microphone techniques at audio sources can be used to accurately generate audio objects that can be manipulated in response to user control. For other events, such as the football match described below, the situation is considerably more challenging. Further details of the way in which the project is handling audio may be found in [2].
**3D scan**

In order to register the different sensors of the FascinatE system in a common coordinate system, a 3D laser scanner [4] is used. This scanner allows an accurate 3D scan of a large environment such as a football stadium, including recognition of special markers. This allows the correct measurement of 3D positions of all the different sensors, such as microphones and cameras. The scanner not only provides a 3D ‘point cloud’ representing the scene, but also a colour image. In Figure 2 (left), the 3D scanner is shown and on the right, the planar view of the captured colour image is presented.

![Figure 2 - 3D laser scanner (left), captured planar view (right)](image)

In addition to directly measuring the locations of the various cameras, the 3D scan data can be used to help estimate the pan, tilt and field-of-view of the broadcast cameras, by providing an accurate depth map of features visible in the background. Computer vision techniques can then be used to identify features in the broadcast camera images and thus track the camera movement [6], for example by matching them with features visible in the OmniCam.

**TEST SHOOT**

On 23rd October 2010, the FascinatE consortium carried out the first test shoot at a live event: the UK Premier League football match Chelsea vs. Wolverhampton Wanderers, at Stamford Bridge, London. The aim of this shoot was to get a complete set of audiovisual material in order to research and develop the new concepts of format agnostic production. Therefore the omnidirectional high-resolution camera system [1], the new high-dynamic range camera [5], an Eigenmike® and two Soundfield® mics were brought to London and installed on different camera platforms in the stadium (see Figure 3). Thanks to close cooperation between BBC and their outside broadcast supplier, the consortium was able to get the recordings of four broadcast cameras, twelve shotgun microphones and several stereo microphones located around the pitch.

Various practical issues had to be overcome during the test shoot. In particular, whilst rigging the omni-directional camera system, care had to be taken in locating it so that the views of spectators were not obstructed. Rain also posed another potential problem, as any drops of water on the mirrors or upward-facing cameras would impair the panoramic image. Luckily, the weather remained dry. After the match, a complete 3D laser scan of the stadium was captured. In this way, it was possible to accurately register all the camera and microphone positions as required for matching of visual and sound events.
It was impossible to attach microphones to the players or referee, and even techniques such as microphone arrays for localising and capturing audio sources were impractical owing to the limitations imposed by the event.

Out of necessity the FascinatE project therefore took advantage of existing recording equipment used at the stadium: 12 shotgun microphones spaced around the pitch (for on-pitch sound) and added several sound field microphones – Soundfield® microphones at either end of the half-way line and a single 32 capsule Eigenmike® situated close to the camera position (Figure 4). Using these microphones a scenario has been developed whereby areas of the pitch determined by microphone placement have been defined as static audio objects that may be either active or inactive depending on automatic assessment of key audio events. This combination then allows the user to dynamically change their viewing direction and apparent location with appropriate panning effects being applied to sound sources.

The audio and video content contained all the 90 minutes of the match of which about 10 minutes (occupying about 1 Tbyte) was selected for distribution to the consortium members. From the selected clips of the omnidirectional camera, a fully stitched panorama has been produced and made available (see Figure 5).
MERGING OF BROADCAST CAMERAS INTO PANORAMIC IMAGE

As discussed in the introduction, one aim of the FascinatE project is to evaluate the ‘layered scene’ concept. One aspect of this is the use of the broadcast cameras to provide higher resolution to key areas of the panoramic scene. To evaluate the potential gain from this approach, tests were conducted with some of the images from the test shoot.

The OmniCam horizontal resolution is approximately 7K pixels, which covers 180 degrees – an equivalent resolution to an HD camera with a horizontal field-of-view of approximately 50 degrees. The main camera covering a football match typically has a horizontal field-of-view of around 30 degrees, although close-up cameras can go as tight as 5 degrees or less. To get the equivalent resolution of such a tight zoom from a 180-degree camera would require a horizontal image resolution of approximately 70K pixels. Using a broadcast camera to enhance resolution in areas of interest thus has the potential to increase the resolution by around a factor of 10 in each direction – well beyond what a practical omnidirectional camera could achieve. Figure 6 shows a comparison of the resolutions.

Some initial experiments have been conducted to assess the challenges in forming a composite image from broadcast and OmniCam images [7]. An issue to be overcome is mismatches in the brightness and colorimetry of the cameras. One approach that has been investigated is the use of histogram matching: the RGB histogram of each image is evaluated in the area of overlap, and a lookup table is computed to re-map the colours of one image to make the two colour histograms match. Figure 7 (a)-(c) shows a small part of the OmniCam image, into which a section from the broadcast camera has been overlaid. The colour mismatch is clearly visible in the central image, particularly on the
grass. The colour histogram equalisation that has been applied in the right-hand image has virtually eliminated any obvious colour difference.

An alternative approach, which avoids having to correct for any level shifts, is to take the high frequency components from the broadcast camera image, and the low frequency components from the OmniCam. This guarantees that flat areas of colour will match exactly. The approach could be extended to use an adaptive filtering strategy, to ensure that detail could instead be taken from the OmniCam where this happened to give more high frequency energy (e.g. in areas of the background that suffered from motion blur in the moving broadcast camera).

Initial experiments with this approach have shown that its success depends critically on the accurate alignment of the images. In this test shoot, there was a distance of around 3m between the OmniCam and the broadcast camera capturing the close-ups, and this resulted in significant parallax differences between the images. Figure 7 (d) shows an example of a part of a composite image, where the low spatial frequencies have been taken from the OmniCam and the high spatial frequencies from the broadcast camera. The images were aligned to match the two players near the centre. Although the detail layer from the broadcast camera correctly enhances the appearance of these players, the background and players at other depths show significant misalignment. Whilst it would be possible to apply some disparity compensation to the processing, it is clear that there would be significant areas of the scene that were only visible in one of the two cameras. In this situation, it is unrealistic to expect to be able to produce a perfect merged image, and instead we are aiming to identify the best approach to producing a visually-acceptable transition between the cameras, so that a virtual zoom could be produced, starting on a wide shot from the OmniCam, and ending up with the close-up from the broadcast camera. This would meet the requirement for a user to be able to seamlessly move from viewing a wide shot to a region-of-interest covered by a broadcast camera.

Figure 7 - Comparison between approaches for merging images: OmniCam image (a), direct overlay of broadcast camera image over the central part (b), overlay after colour histogram matching (c) and taking high frequencies from broadcast camera (d).
GENERATION OF AUDIO TO MATCH VIDEO

One principle of FascinatE is to transmit as much information as possible to the terminal in its original format, rather than transcoding from one format into another. Therefore audio objects and sound field recordings are transmitted separately. This allows the user to interact with the content independently, for example selecting audio objects like the TV commentator and rotating the sound field depending on the viewing direction. At the terminal the sound field signal will be decoded and the audio objects will be placed at the appropriate locations, before being passed to the reproduction system.

Audio objects will be used for dedicated sound events like a ball kicks and a referee whistle blow; a position will be added to other sources such as the TV commentator. However it will not possible to capture and track 45000 football supporters at once. Therefore the ambience will be recorded as higher order Ambisonics format.

It seems likely that a shift in user expectations may occur when the user becomes an active participant in defining the scene rather than a passive viewer. In the current football broadcast scenario the panning of a camera has no corresponding panning effect on the reproduced audio from the event. In shifting to a viewer-defined scene however the situation is closer to a first-person video gaming scenario where every pan is accompanied by a corresponding shift in the audio scene. Listening tests have been devised and pilot tests carried out within the project in order to assess this possible paradigm shift in user expectation. A representation of user-controlled FascinatE scene manipulation has been developed giving users control of camera panning within the test shoot panorama. Two scenarios have been presented initially: a static scene with no rotation (the current broadcast norm) and dynamic pan response with both audio objects and rendered ambience rotated according to the user’s defined view. Early results from the pilot study, which involved 5 participants, indicate a likely user preference towards the active participant scenario where the entire sound field, including the audio objects, rotates with the view of the scene. Qualitative evidence from participants in the pilot study suggests that movement of audio objects on the football pitch derived from pitch-side shotgun microphones has a greater subjective effect than rotating the crowd ambience recorded by surround microphones. A full set of tests is planned to determine the optimal audio rendering protocols for the FascinatE system.

CONCLUSION

This paper has outlined the principles of a format-agnostic production system, to support ‘virtual re-shooting’ of events under the control of either the production team or end users, to suit different devices and user preferences. The concept of a layered scene representation has been introduced, to tailor the resolution of the captured scene to match both the areas of interest and the capabilities of practical production hardware. The first results from an experiment to test these ideas in the context of a football match have been presented.

REFERENCES


ACKNOWLEDGEMENTS

The authors would like to thank their colleagues in the FascinatE project for their contributions to the work reported here. They would also like to thank SIS Live for their assistance with the test shoot, and the Premier League for permission to use the football images. The research leading to these results has received funding from the European Union’s Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 248138.
Combining Panoramic Image and 3D Audio Capture with Conventional Coverage for Immersive and Interactive Content Production

Graham Thomas  BBC R&D, UK
Oliver Schreer  Fraunhofer HHI, Germany
Ben Shirley  University of Salford, UK
Jens Spille  Technicolor, Germany

Some challenges for future media production & delivery

Users expect the possibility of interacting with or customising their experience

Increasing diversity of end-user devices
Some new opportunities for domestic immersive video – “Surround Video”

Project wide-angle view beyond TV around the walls – we used a conventional projector and a hemispherical mirror

Viewer concentrates on main image but gets sense of immersion from projected image, which aligns with main display

Surround Video test
The FascinatE project concept

- Pan/tilt/zoom broadcast cameras
- Fixed omni-directional camera
- OmniCam
- Alexa Broadcast camera

Calibrate & merge into single reference frame

Automated shot framing / scripting

Compress & encapsulate

Layered format-agnostic representation

Broadband Net for Immersive Media

Local renderer

Proxy renderer for small devices

Interaction

Multiple mic types

FASCINATE test shoot - cameras

OmniCam

Alexa

Broadcast camera
FASCINATE test shoot - cameras

FASCINATE test shoot - microphones

Soundfield®

Stereo pair

Gun mics
FASCINATE test shoot – 3D scan

Panoramic & broadcast cameras compared

7K x 2K panorama

Detail from panoramic image

Full broadcast camera image
Use of broadcast camera to enhance Omnicam

Region of Omnicam image

Overlay shows problems with photometric matching and parallax
Use of broadcast camera to enhance Omnicam

Histogram matching can largely solve photometric issues

Comparison with Omnicam shows that parallax issues remain
Use of broadcast camera to enhance Omnicam

Switch with area-of-interest aligned
FascinatE Audio

- Object based approach
  - Audio objects
    - Close miked and tracked sources
    - Derived from multiple microphones
  - Ambient sound field component
    - Ambisonics Microphones
      - Eigenmike
      - SoundField microphone
    - Stereo Microphones

Audio Rendering Preferences

- Audio Rendering
  - Shift in viewer expectation?
    - Passive Viewer vs. Active Participant

- Pilot Test
  - Audio objects placed in 3D model
  - Ambient sound field from Eigenmike®
  - User can pan around scene from fixed position
  - Audio rendering options
    - Static
    - Rotating sound field and audio objects
  - Results
    - Preference towards rotating sound field
Conclusion

- The FascinatE project is investigating the practicality of a format-agnostic production system
- Shown that reframing and immersive scene composition can benefit from the availability of multiple sources – demonstrated with both audio and video
- More work needed to determine the best way of using video from different kinds of cameras (resolution, parallax, colour matching)
- Initial audio experiments have shown that ability to create audio to match a selected view is preferred over a simple static audio field
- Next steps: Further development of capture and reproduction systems and experiments with other types of content

Thanks to all the FascinatE partners

Come and see us on booth 8.G44 in the Future Zone!

graham.thomas@bbc.co.uk  www.bbc.co.uk/rd

www.fascinate-project.eu