FascinatE Production, Audio and Networking

Author(s) and company: G.A. Thomas (BBC), O. Schreer (HHI), B. Shirley (UOS), J. Spille(DTO), H. Kropp (DTO), J.M. Batke (DTO), S. Abeling (DTO), F. Keiler (DTO), R. Oldfield (UOS), O.A. Niamut (TNO), A. Kochale (DTO), J. Ruiz Hidalgo (UPC), J-F. Macq (ALU), G. Kienast (JRS).

Document status: Final

Confidentiality: Public

Introduction

At IBC 2011 the FascinatE project presented three papers in the conference section, focusing on the production side of FascinatE, the format agnostic audio aspects of the project and the network and terminal end. This document contains all three documents reproduced here with permission from the IBC.

COMBINING PANORAMIC IMAGE AND 3D AUDIO CAPTURE WITH CONVENTIONAL COVERAGE FOR IMMERSIVE AND INTERACTIVE CONTENT PRODUCTION

G.A. Thomas[1], O. Schreer[2], B. Shirley[3], J. Spille[4]

[1]BBC R&D, UK; [2]Fraunhofer HHI, Germany; [3]University of Salford, UK; [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.

FORMAT-AGNOSTIC APPROACH FOR 3D AUDIO

H. Kropp[1], J. Spille[1], J.M. Batke[1], S. Abeling[1], F. Keiler[1],
R. Oldfield[2], and B. Shirley[2]

[1]Technicolor, Research & Innovation, Germany
[2]Acoustics Research Centre, UK

Abstract

In the market exists a large variety of media devices, reaching from mobile handsets equipped with headphones up to an ultra-high resolution display connected with a large loudspeaker setup. This makes it difficult for the broadcast industry to provide all of these devices with appropriate data at once. In the EU-funded FascinatE project a system is being developed that considers the individual requirements of a particular end-user device and allows a user to interactively navigate in an audiovisual scene. This paper focus on the latest audio related developments in capturing and replaying audio objects and the entire sound field with respect to the selected view on any loudspeaker setup. First results of a UK Premier League football match show practical aspects of the spatial audio recording and its playback on a 3D loudspeaker setup that can be used for small event rooms.

ADVANCED AUDIOVISUAL RENDERING, GESTURE-BASED INTERACTION AND DISTRIBUTED DELIVERY FOR IMMERSIVE AND INTERACTIVE MEDIA SERVICES

O.A. Niamut[1], A. Kochale[2], J. Ruiz Hidalgo[3], J-F. Macq[4], G. Kienast[5]

[1]TNO, NL; [2]Deutsche Thomson OHG, DE; [3]Universitat Politècnica de Catalunya, ES; [4]Alcatel-Lucent, BE; [5]Joanneum Research, AT.

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, delivery and rendering system capable of supporting both these trends is being developed by a consortium of European organisations including partners from the broadcast, film, telecoms and academic sectors, in the EU-funded FascinatE project. This paper reports on the latest project developments in the delivery network and end-user device domains, including advanced audiovisual rendering, computer analysis and scripting, content-aware distributed delivery and gesture-based interaction. The paper includes an overview of existing immersive media services and concludes with initial service concept descriptions and their market potential.

COMBINING PANORAMIC IMAGE AND 3D AUDIO CAPTURE WITH CONVENTIONAL COVERAGE FOR IMMERSIVE AND INTERACTIVE CONTENT PRODUCTION

G.A. Thomas[1], O. Schreer[2], B. Shirley[3], J. Spille[4]

[1]BBC R&D, UK; [2]Fraunhofer HHI, Germany; [3]University of Salford, UK;
[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].

Figure 1 - OmniCam

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 it 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.

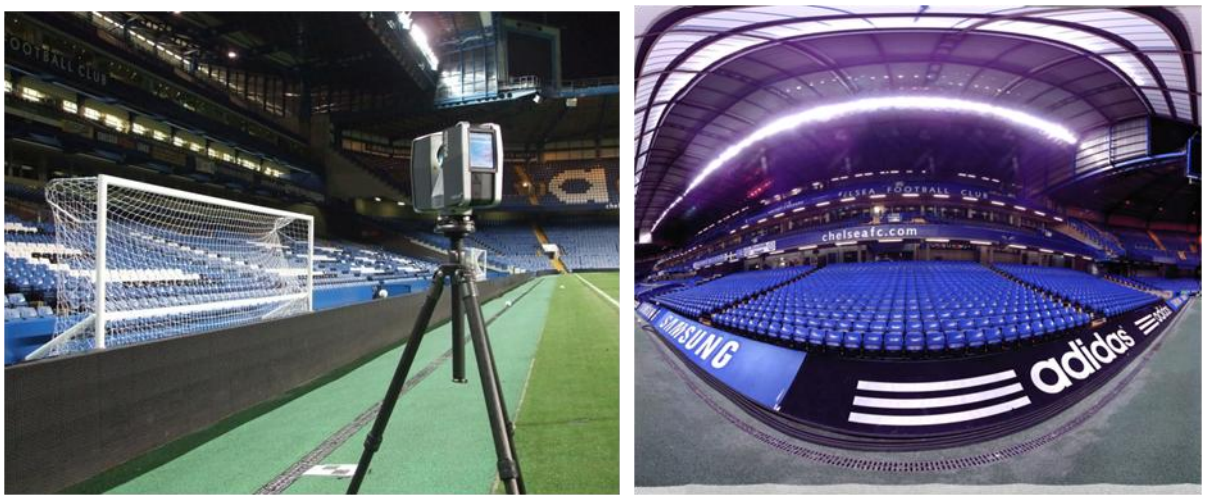


Figure 2 - 3D laser scanner (left), captured planar view (right)

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.

Figure 3 - Chelsea test shoot: camera platform (left), calibration of the OmniCam (middle), Soundfield® mics and stereo pair (right)

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.

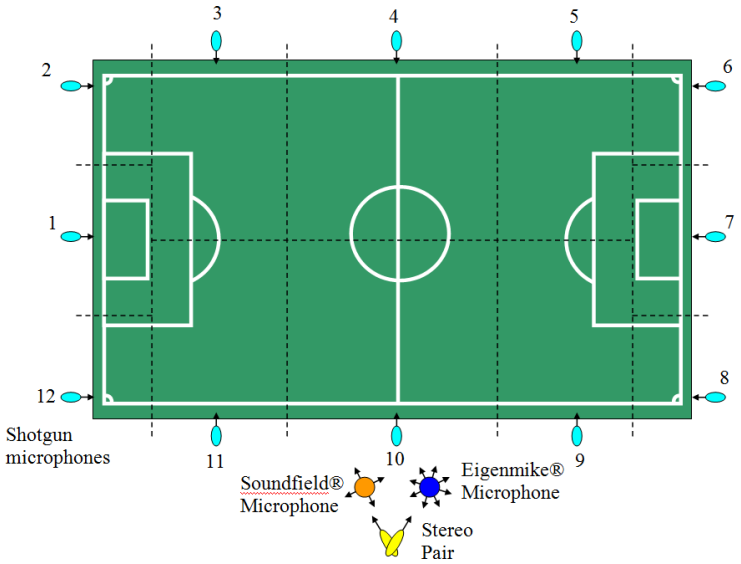


Figure 4 – Microphone positions

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).

Figure 5 - Stitched panoramic view

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.

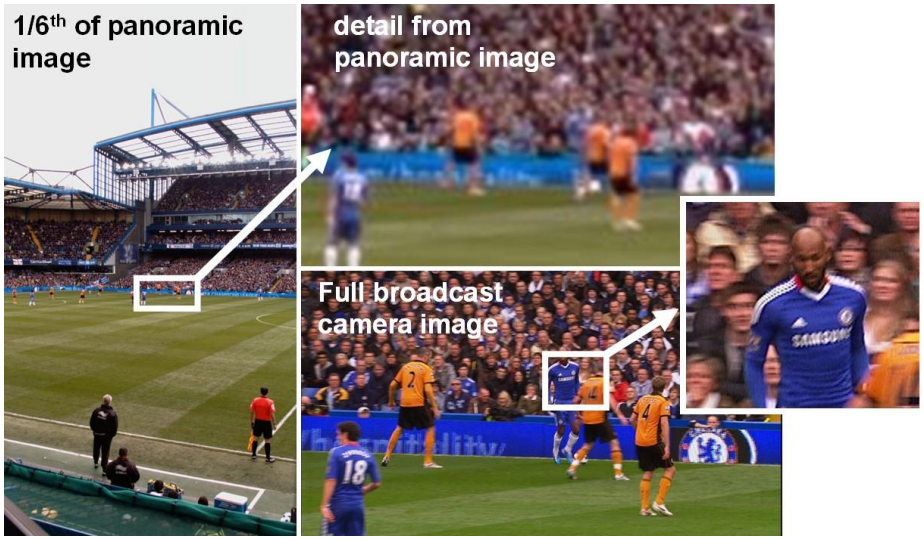


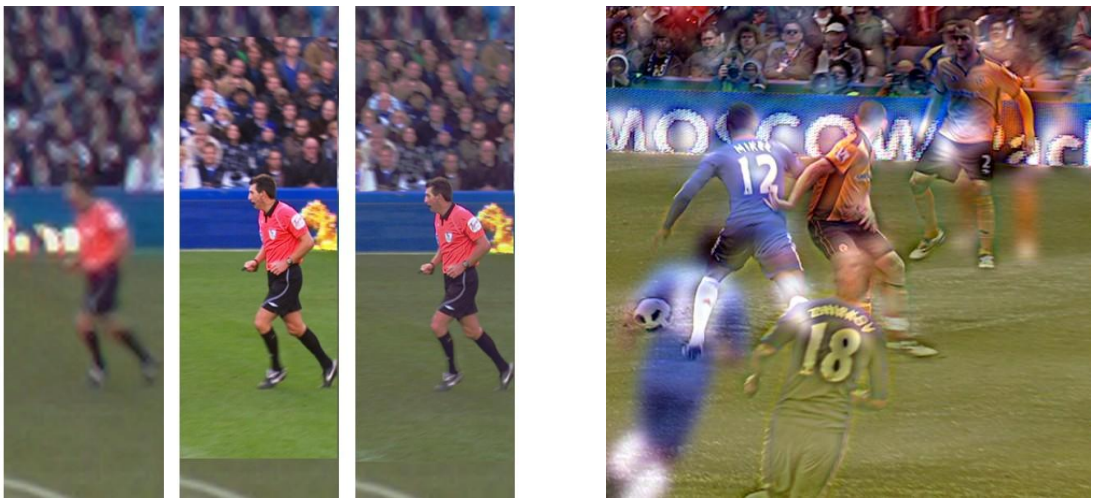
Figure 6 - Comparison between OmniCam and broadcast camera

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.



(a) (b) (c) (d)

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

1. Schäfer, R., Kauff, P. and Weissig, C. 2010. Ultra high resolution video production and display as basis of a format agnostic production system. Proceedings of IBC 2010.
2. Kropp, H. et al. 2011. Format-agnostic approach for 3D audio. Submitted to Proceedings of IBC 2011.

3. Niamut, O. et al. 2011. Advanced audiovisual rendering, gesture-based interaction and distributed delivery for immersive and interactive media services. Submitted to Proceedings of IBC 2011.
4. Faro 'Focus^{3D}' laser scanner. http://www.faro.com/focus/
5. The ARRI Alexa camera. http://www.arridigital.com/alexa
6. Dawes, R., Chandaria, J. and Thomas, G.A. 2009. Image-based Camera Tracking for Athletics. Proceedings of the IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB 2009), May 2009. Available as BBC R&D White Paper http://www.bbc.co.uk/rd/publications/whitepaper181.shtml
7. Gibb, A., Thomas G.A. 2010. Fusion of images using complementary filters and histogram equalisation. Conference on Visual Media Production (CVMP2010), November 2010.

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.

FORMAT-AGNOSTIC APPROACH FOR 3D AUDIO

H. Kropp[1], J. Spille[1], J.M. Batke[1], S. Abeling[1], F. Keiler[1],
R. Oldfield[2], and B. Shirley[2]

[1]Technicolor, Research & Innovation, Germany
[2]Acoustics Research Centre, UK

ABSTRACT

In the market exists a large variety of media devices, reaching from mobile handsets equipped with headphones up to an ultra-high resolution display connected with a large loudspeaker setup. This makes it difficult for the broadcast industry to provide all of these devices with appropriate data at once. In the EU-funded FascinatE project a system is being developed that considers the individual requirements of a particular end-user device and allows a user to interactively navigate in an audiovisual scene. This paper focus on the latest audio related developments in capturing and replaying audio objects and the entire sound field with respect to the selected view on any loudspeaker setup. First results of a UK Premier League football match show practical aspects of the spatial audio recording and its playback on a 3D loudspeaker setup that can be used for small event rooms.

INTRODUCTION

Visual 3D presentation in cinema has led to a new spectacular audience experience and to commercial success in the film industry. As opposed to video, the audio part in the film and broadcast sectors is still aimed primarily at conventional 2D playback systems. An introduction of spatial audio in a broader market would not only improve the audiovisual effects but would also stimulate commercial perspectives of the broadcast industry.

Spatial audio is one key feature of the European funded project FascinatE which stands for "Format-Agnostic SCript-based INterAcTive Experience". The project is primarily looking at broadcasting live events to give the viewer a more interactive and immersive experience. This is done by providing a format agnostic output stream that will be interactively modified and adapted to a viewer's particular kind of devices. Auditory spatial impression, e.g., of a football stadium, can be transported home, adapted to a viewpoint and replayed on loudspeaker scenarios from simple stereo to more complex 3D setups.

Therefore sophisticated 3D audio technologies are required to capture the spatial sound field information, to transmit and to map them onto various end user devices. FascinatE combines two kinds of spatial wave field decompositions which are known as Higher-Order Ambisonics and Wave Field Synthesis. Higher-Order Ambisonics describes the entire sound field with high spatial resolution at a specific location, whereas Wave Field Synthesis reconstructs a sound field out of a set of recorded sound sources together with their spatial coordinates.

This paper will explain differences between Higher-Order Ambisonics and Wave Field Synthesis concerning their respective key challenges of capturing and the rendering of audio scenes for a spatial loudspeaker setup. It will be shown how a user's viewpoint influences the spatial orientation of the sound field and how one audio sequence can be

mapped onto different loudspeaker setups. Finally, we will summarise the main differences and benefits of FascinatE's sound field description.

This is one of three papers being submitted covering various aspects of the FascinatE project. This paper focuses on a format-agnostic approach for 3D audio, whilst the others focus on developments at the production side [4], and the delivery network and end-user terminal [9]. The project would like to offer some demos relating to these papers on a New Technology Campus booth.

FASCINATE PROJECT

The FascinatE Project [1] is based on a 'format-agnostic approach' which enables the adaptation to different end-user devices that can vary from modern displays in connection with large loudspeaker setups down to mobile devices combined with headphones, rather than being limited to specific audio or video formats. 'Format-agnostic' means that specific parameters like image resolution or accuracy of the sound field description can be chosen according to the requirements of the particular production and the specific end-user devices. To control such parameters, more interactivity is required on both the production and end-user side. The need for improved interactivity by the user has a direct influence on FascinatE's architecture.

FascinatE Architecture

FascinatE is a network-based approach leading to an architecture that can be separated into the three main parts; the production side, a network, and the end-user terminal (s. Figure 1). All network related connections are controlled by specific FascinatE rendering and scripting nodes. The rendering nodes enable the adaptation of individual data and band-width requirements, whereas the scripting nodes are responsible for controlling the rendering nodes by transferring and analysing interactive control sequences from the production or end-user side. More details about the specific network nodes can be found in [3,9].

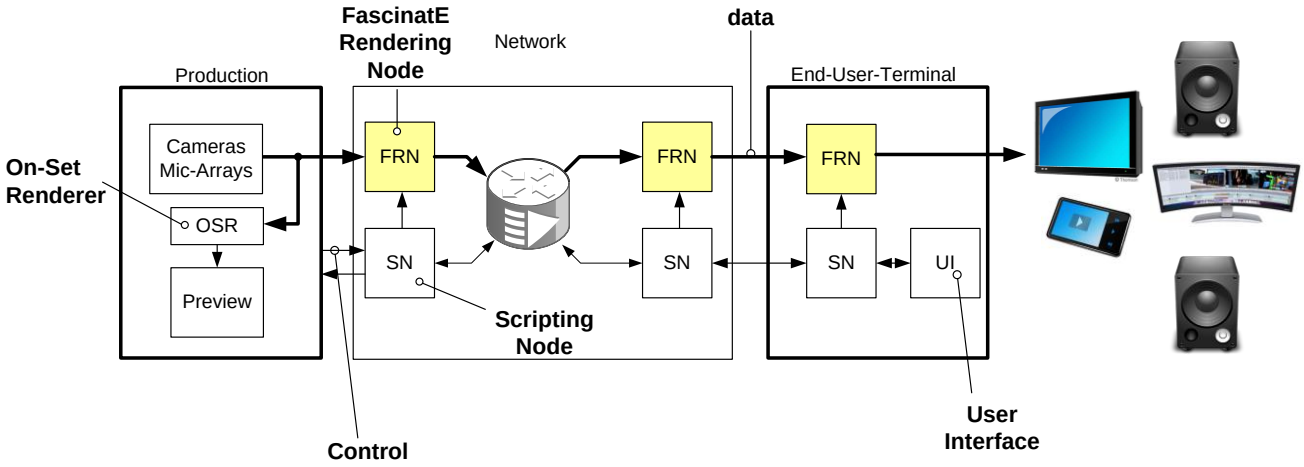


Figure 1 – Overview about FascinatE's architecture with production, network- and end-user related blocks

At the production side, this architecture requires the development of new video and audio capture systems as well as systems to control the shot framing options presented to the viewer [1]. In FascinatE a 2D panoramic view is used that provides not only high resolution but also different view points. The greatest challenges concerning video capturing and rendering is described in more detail in [4].

Here, the focus is on the challenges for FascinatE's audio related tasks, which are sound field capturing and rendering to specific loudspeaker setups for example in a small event room. In the case of audio we have to deal with spatial effects in a 3D sound field so that spatial Higher-Order Ambisonics (HOA) and Wave Field Synthesis (WFS) are facilitated technologies. The following chapter introduces the main principles of a sound field description based on HOA or WFS.

SOUND FIELD DESCRIPTIONS

The spatial audio reproduction system for FascinatE aims at to replay the sound field accompanying a visual scene. Sound fields describe the air pressure distribution within a region of space generated by different sound sources, i.e., $p(\boldsymbol{x},t)$ depends on spatial position (described by vector $\boldsymbol{x}$) and time t. A sound field can also be formulated as an acoustic wave field $g(\boldsymbol{x},\omega)$ at the same position $\boldsymbol{x}$ but in the frequency domain with radian frequency ω [5]. If $g(\boldsymbol{x},\omega)$ is inserted into the Helmholtz equation and expressed in spherical coordinates, the general solutions are based on a separation in a radial and an angular dependent part. With respect to different frequencies, the solutions are a linear combination of complex modes [5] containing spherical Bessel functions for the radial part and spherical harmonics for the angular part. These general solutions describe every point within a sphere, so that it becomes very difficult to generate a sound field with stimulated loudspeakers, if the sound sources are located within the sphere.

However, a simplification is based on the Kirchhoff-Helmholtz equation. It says that in a source-free volume, the sound field can be fully described by the sound pressure and its derivative on the surface enclosing the volume [2] which is more efficient than using the whole volume. Practically, this means that a sound field can be reproduced by several loudspeakers distributed on this surface.

Wave Field Synthesis (WFS)

This effect is used in wave field synthesis to (re)create a sound field in a volume enclosed by loudspeakers (also called secondary sources). The theory of wave field synthesis is essentially an application of the Huygen's Effect which states that each point on a wave front can be considered as the starting point of a new point source. If an infinite number of these point sources are distributed along the wave front it can be completely recreated. However, WFS attempts to reconstruct any sound field using arrays of a finite number of loudspeakers. Conceptually, if these loudspeakers are fed with different amplitudes and delays the sound field within the enclosed spaces can be recreated.

Principally one can derive two driving functions for the WFS system corresponding to the rendering of point sources and plane waves. For FascinatE, point sources can be used to render audio objects (sound sources) that can be recorded and positioned in the space, and plane waves can be used for diffuse, ambient sound or for audio objects a long way away.

WFS, therefore can be used to accurately recreate a sound field if each of the audio objects in the sound field is recorded as a single audio source with its corresponding spatial coordinates, the ambiance can also be recreated if recorded using ambient microphones such as a Soundfield® microphone and can be reproduced with each component direction of the sound field can be rendered as a plane wave from the corresponding direction.

A great advantage of WFS reproduction is that the spatial impression and localisation of sound sources is almost independent of listener position which means that WFS lends

itself to applications where there is a large audience as then everyone can have a similar listening experience. This is different in the case of Ambisonics.

Higher-Order Ambisonics (HOA)

Ambisonics, which was developed in the early 1970's [6,7], is a method employing a mathematical approximation of the sound field. It assumes that loudspeakers are far enough from a listeners' region, so that instead of point sources driving spherical waves, plane waves are considered.

In principle, Ambisonics uses the spherical harmonics on a spherical surface to model the superposition of plane waves from different directions as a linear combination of orthonormal basis vectors. The components in each basis vector are the spherical harmonics of one specific direction. Due to their vector related description, coefficients for different directions are combined in form of matrices. In Ambisonics one has to distinguish between encoder and decoder matrices as given in Figure 2 .

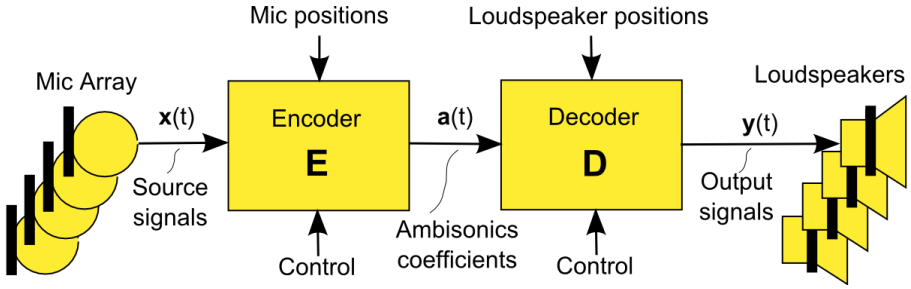


Figure 2 – The Ambisonics encoder and decoder operations

For each sample time, an encoder matrix $\boldsymbol{E}$ maps all sound source signals, e.g., recorded by microphones, from different directions $\boldsymbol{x}$ into a single vector where its components represents the Ambisonics coefficients. Mathematically this process is a simple matrix-vector calculation:

$$\boldsymbol{a} = \boldsymbol{E} \cdot \boldsymbol{x} .$$

So at each sample time we derive a vector $\boldsymbol{a}$ containing all recorded source information encoded in its vector components. This is a format-agnostic approach because the Ambisonics representation is space invariant. This means it can be mapped onto any arbitrary loudspeaker setup, where the positions of the loudspeakers and the sound sources can be different. Therefore, the decoder matrix $\boldsymbol{D}$ performs the inverse operation to map the Ambisonics coefficients onto concrete loudspeakers:

$$\boldsymbol{y} = \boldsymbol{D} \cdot \boldsymbol{a}$$

In general, the accuracy of the Ambisonics sound field reproduction depends on the number of coefficients; the more coefficients the better the accuracy. This is very similar to signal sampling where its accuracy improves with the number of digits. But here, we still have a spatial sampling, where the spatial distance between possible sampling points have to be considered [8]. A drawback of the early versions of Ambisonics is its limitation to 16 coefficients. Furthermore, it is based only on the angular dependencies in the acoustic wave field $g(\boldsymbol{x},\omega)$, whereas the radial part is neglected. Thus, the accuracy of the sound field is restricted to a small reference area, often called "sweet spot".

Higher-Order Ambisonics circumvents these drawbacks for the following reasons:

- It is no longer restricted to a limited number of coefficients. Modern HOA implementations are built from vectors with hundreds of coefficients.
- Using spherical Bessel functions allows the consideration of radial dependencies [8].
- Using point sources allows a more realistic modelling of the sound field [8].

Nevertheless, improvements in HOA and WFS require an additional overhead for an implementation, especially if both methods are combined. The following chapter describes these aspects in relation to the FascinatE project.

IMPLEMENTATION ASPECTS

The aim of FascinatE's audio related tasks is to reproduce the entire sound field at the end-user side which matches the sound field recorded on the production side. The challenging task is to achieve this on any loudspeaker setup with respect to the interactively controlled view at the end-user terminal. This requires as much information as possible about the audio objects and the sound fields. To perform individual modifications by the end-user, for example zooming in a specific region of a football stadium, audio objects and sound fields have to be captured and transmitted separately. At the reproduction side all these audio signals have to be adequately composed and adapted to the existing loudspeaker setup. More details about the recording and reproduction techniques based on WFS and HOA are given in the following paragraphs.

Audio Recording

A large part of the recording process for WFS is in the determining of both the position and content of audio objects. This allows the dynamic tracking of single sound objects as it exists in today first party video gaming scenarios, which give a dynamic impression of the sound image. In addition the ambience sound, for example from the crowd in a football game, has to be captured without position data. Ideally audio objects can be captured with close up microphones as it is done with the referee voice at ruby matches in the UK today. However it is not possible to use close up microphones for all sound objects, due to some regulations or comb filter effects [12]. Therefore individual solutions have been developed in the FascinatE project which are described in the results.

Concerning the ambience sound, it will not be possible in a real scenario to capture and track all sound sources, e.g. from all football supporters, at once. Therefore the ambience will be recorded as Higher-Order Ambisonics format. First experiences were made using a Soundfield® microphone with 4 capsules and an Eigenmike® microphone with 32 capsules. Both microphones provide more directional information, than conventional stereo or surround recordings. The Soundfield® microphone directly provides an early versioned Ambisonics format, whereas the Eigenmike® signals can be used for a HOA representation with up to 25 coefficients. However, the latter signals require additional post processing steps, like a specific filtering to consider radial dependencies [8].

Audio Reproduction

At each terminal a specific FascinatE Rendering Node exists that contain two main blocks as depicted in Figure 3; a scene composer and a so called presenter.

The scene composer is used to select and compose single audio objects or recorded sound fields according to the user request. However, it is not an ordinary mixing of different audio channels. It allows the user for example rotation or fading of specific objects if they are described in WFS format. Due to the fact that HOA encoding and decoding is based on matrices, a HOA representation can also be used to perform a linear operation like rotating the entire sound field according to the dynamically selected view.

The final presenter receives the selected and composed audio objects as well as the adapted sound field signals. Currently, all audio objects are Ambisonics encoded in the presenter under consideration of the composed positions and before they are mixed with the HOA encoded sound field signals.

The composer is controlled by production script parameters, like the default view direction and user requests, while the presenter is controlled by the device parameters, which are the loudspeaker positions in case of audio (s. Figure 3).

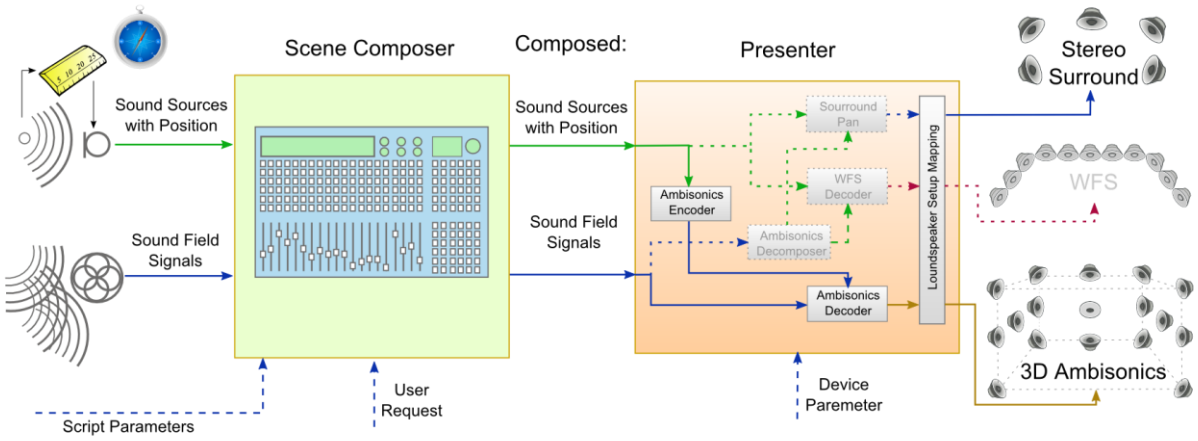


Figure 3 – Mixing of different audio objects in FascinatE

In future implementations it is planned to implement among others a wave field synthesis (WFS) decoder to be able to feed a WFS loudspeaker setup as well (s. dotted lines and gray boxes in Figure 3).

RESULTS

Experimental Loudspeaker Mapping

In principle, FascinatE allows the mapping onto typical loudspeaker setups (stereo and surround), an appropriate WFS setup, or a 3D setup that considers height loudspeakers for HOA. In the case of HOA based scenarios we have tested several decoders that depend on a specific loudspeaker setup. At the beginning, we used pure HOA ‘mode matching’ decoders [3]. This HOA decoder type can have more than 16 coefficients but its method is restricted to plane waves which lead to a poor directional localisation.

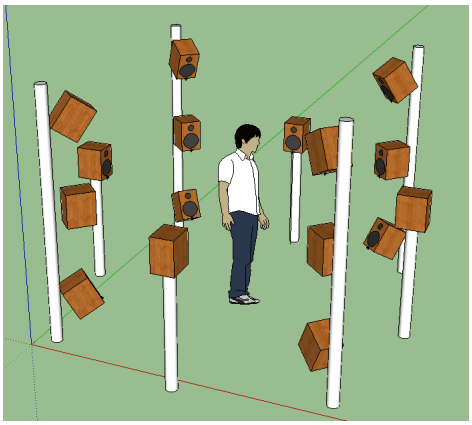


Figure 4 – Loudspeaker setup with 16 loudspeakers were used in the FascinatE project

An improvement for the directional perceiving was derived, by the use of additional panning functions [3]. Informal tests were performed with a setup of 16 loudspeakers (s. Figure 4). We get satisfying results for this setup in combination with an accuracy of 25 Ambisonics coefficients. The loudspeaker positioning in Figure 4 was chosen due to

practical considerations, having four columns with three loudspeakers each and additional loudspeakers between these columns.

Real-Life Example: Football Game

A comprehensive test shoot was performed on the in October 2010 at the UK Premier League match between Chelsea FC and Wolverhampton Wanderers in London. Some specific audio visual scenes were recorded and post processed by members of the FascinatE team.

In order that the audio capture was appropriate for WFS rendering it was vital that the on-pitch sound sources were separated from the ambient crowd noise so they could be rendered as point sources and plane wave respectively. The microphone layout used for the test capture was as shown in Figure 5. The shotgun microphones around the pitch are intended to capture only the on-pitch sounds and are therefore chosen to be as directional as possible but microphone limitations and the high level and close proximity of the crowd means that unwanted noise is still picked up. Consequently for a standard broadcast, the level of each microphone channel is only raised by the sound engineer when the action on the pitch is within the vicinity of that microphone. For FascinatE, different users will choose different views and will consequently require a different audio mix so this process cannot be used. An automatic approach [10] can therefore be used which looks for key audio events such as ball kicks and whistle blows on the pitch and activates microphone channels only when these events are detected. The approximate location of the audio events can be determined by time delay estimation algorithms [11]. These audio objects can then be extracted from the microphone signals and positioned in space in the WFS renderer. The ambient sound from the crowd is recorded with the Soundfield® microphone, stereo pair or the Eigenmike® and can be rendered as plane waves from different directions.

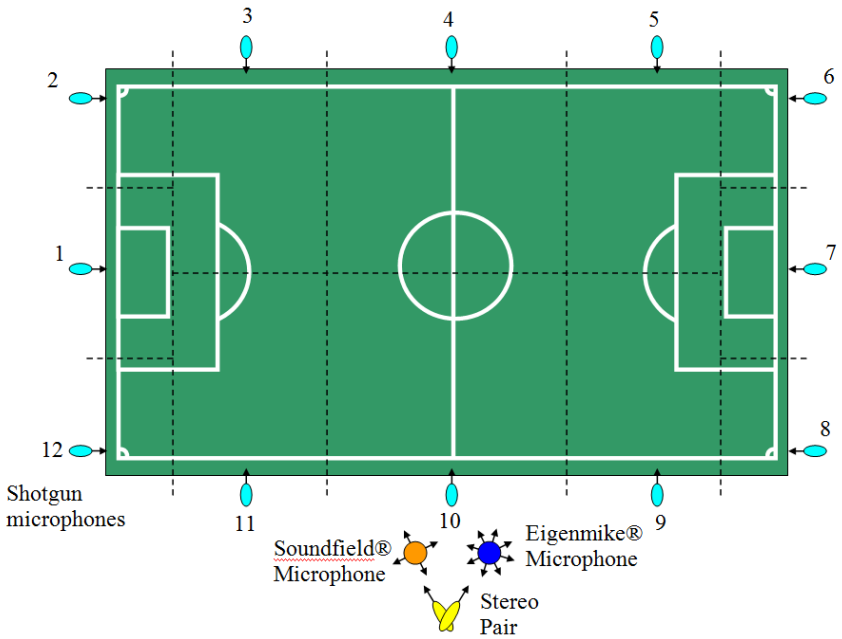


Figure 5 – Microphone setup used to record the audio from a football match.

Playback tests of this content were also performed on the above the loudspeaker setup from Figure 4 . Additional post processing steps allow us to following the interactively selected view point and results in the feeling of “being there”.

CONCLUSIONS

This paper describes the audio related challenges of a format agnostic approach by the use of combined sound field description techniques based on Wave Field Description (WFS) and Higher-Order Ambisonics (HOA). We made first real-life experiences with a spatial sound recording and a playback environment leading to a nearly realistic impression of the recorded sound scene. The proposed method enables sound scene descriptions rather than being restricted to loudspeaker channel based formats. The consideration of additional meta data improves the interactivity of a user.

REFERENCES

1. FascinatE Project. Official FascinatE Website. http://www.fascinate-project.eu. 2010.

2. P. Nelson and S. Elliot. Active Control of Sound. New York: Academic Press. 1992.

3. J.-M. Batke, et al. Using VBAP-Derived panning functions for 3DAmbisonics Decoding. Int. Symp. on Ambisonics and Spherical Acoustics. May 6-7. Paris.

4. G. A. Thomas. et al. Combining panoramic image and 3D audio capture with conventional coverage for immersive and interactive content reproduction. Submitted to Proceedings of IBC 2011. Amsterdam. Netherlands. Sept. 8-13 2011.

5. E. G.Williams. Fourier Acoustics. Academic Press. 1999.

6. D. H. Cooper and T. Shiga. Discrete-Matrix Multichannel Stereo. J. Audio Eng. Soc.. 20:346–360. 1972.

7. Michael A. Gerzon. Ambisonics in Multichannel Broadcasting and Video. J. Audio Eng. Soc.. 33 (11):859–871. 1985.

8. J.-M. Batke, et al. Recording Spatial Audio Signals for Interactive Broadcast Systems. Forum Acusticum. Danmark. Aarlborg. June 27 – July 1 2011.

9. O. Niamut, et al. 2011. Advanced audiovisual rendering, gesture-based interaction and distributed delivery for immersive and interactive media services. Submitted to Proceedings of IBC 2011. Amsterdam. Netherlands. Sept. 8-13 2011.

10. R.G. Oldfield and B. G. Shirley. Automatic mixing and tracking of on-pitch football action for television broadcasts. 130th Conv. Audio Eng Soc, London, UK, May 2011.

11. C. H. Knapp and G. C. Carter. The generalized correlation method for estimation of time delay. IEEE Trans. Acoust., Speech, Signal Process. ASSP-24, 320–327. 1976.

12. Hinata et al. Live Production of 22.2 Multichannel Sound for Sports Programs. AES 40th International Conference, Tokyo, Japan, 2010 October 8–10

ACKNOWLEDGEMENTS

We would like to thank our colleagues for their contributions to this work. We would also like to thank the FascinatE project for permission to publish this paper. 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]. We would also like to thank SIS Live and the Chelsea FC for their assistance with the test shoot.

ADVANCED AUDIOVISUAL RENDERING, GESTURE-BASED INTERACTION AND DISTRIBUTED DELIVERY FOR IMMERSIVE AND INTERACTIVE MEDIA SERVICES

O.A. Niamut[1], A. Kochale[2], J. Ruiz Hidalgo[3], J-F. Macq[4], G. Kienast[5]

[1]TNO, NL; [2]Deutsche Thomson OHG, DE; [3]Universitat Politècnica de Catalunya, ES; [4]Alcatel-Lucent, BE; [5]Joanneum Research, AT.

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, delivery and rendering system capable of supporting both these trends is being developed by a consortium of European organisations including partners from the broadcast, film, telecoms and academic sectors, in the EU-funded FascinatE project. This paper reports on the latest project developments in the delivery network and end-user device domains, including advanced audiovisual rendering, computer analysis and scripting, content-aware distributed delivery and gesture-based interaction. The paper includes an overview of existing immersive media services and concludes with initial service concept descriptions and their market potential.

INTRODUCTION

New kinds of ultra high resolution sensors and ultra large displays are generally considered to be a logical next step in providing a more immersive experience to end users. High resolution video immersive media services have been studied by NHK in their Super Hi-Vision 8k developments [1]. In the international organization CineGrid.org [2], 4k video for display in large theaters plays a central role. At Fraunhofer HHI, a 6k multi-camera system, called the OmniCam, and an associated panoramic projection system was recently developed [3]. However, the notion of immersive media with high resolution video, stereoscopic displays and large screen sizes seems contradictory to leveraging the user's ability to select and control content and have it available on personal devices.

Within the EU FP7 project FascinatE [4] a capture, production and delivery system capable of supporting interaction, such as pan/tilt/zoom (PTZ) navigation, with immersive media is being developed by a consortium of 11 European partners from the broadcast, film, telecoms and academic sectors. The FascinatE project aims to develop a system that allows end-users to interactively view and navigate around an ultra high resolution video panorama showing a live event, with the accompanying audio automatically changing to match the selected view. The output is adapted to the particular kind of device, ranging from a mobile handset to an immersive panoramic display. At the production side, an audio and video capture system is developed that delivers a so-called Layered Scene, i.e. a multi-resolution, multi-source representation of the audiovisual environment. In addition, scripting systems are employed to control the shot framing options presented to the viewer. Intelligent networks with processing components are used to repurpose the content to suit different device types and framing selections, and user terminals supporting innovative gesture-based interaction methods allow viewers to control and display the

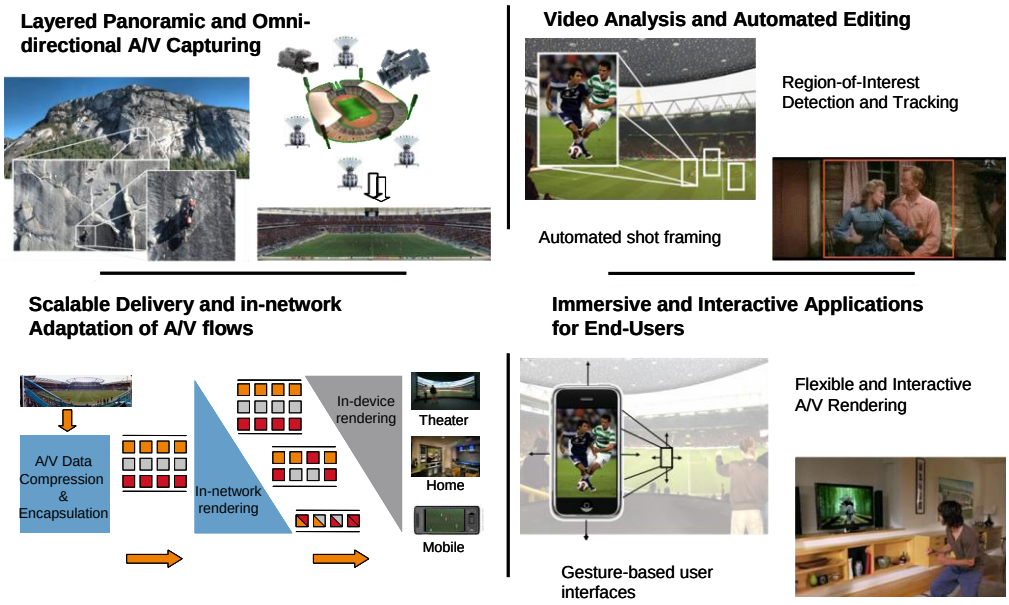


Figure 1 – Key innovation areas in FascinatE.

content suited to their needs. FascinatE considers four key innovation areas (Figure 1), from which we identify five technological developments, referred to as technical attributes, that enable FascinatE immersive and interactive media services:

1. *Layered Scene Production*, where audiovisual scenes are captured with clusters consisting of multiple cameras and microphones;
2. *Metadata and Scripting*, providing knowledge to steer further processing and adaptation of the content within the network and on the terminal;
3. *Scalable Delivery and In-Network Audio/Video Adaptation*, leveraging efficient delivery and media-aware network-based processing required for supporting low-end terminals;
4. *Flexible and Interactive Audio/Video Rendering*, adapting the content to the end-user terminal with the associated screen and speaker set-ups;
5. *Gesture Based User Interaction*, enabling natural end-user navigation.

In this paper we focus on the latter four aspects, where we limit rendering to video only. Both the production aspects, such as capturing the scene and reproducing it for a given viewing direction and field-of-view, as well as the audio aspects, such as 3D audio capture and rendering, are detailed in two additional IBC papers from the FascinatE project. Also, in an earlier paper at the IBC2010 conference [5] the initial goals and challenges for the project and the inherent format-agnostic approach that the project takes, were outlined.

FascinatE considers three main use cases, each with its associated target end device and screen type (Figure 2); in the theatre case, the captured content is transmitted to and displayed on a large panoramic screen, enabling multiple viewers to simultaneously see the content and interact with it. In contrast, in the home viewing situation a limited number of viewers consumes the content via a large TV screen and interacts using gestures, e.g. by selecting players to follow when watching a sports game and zooming in on interesting events. Lastly, in the mobile use case, users can employ their individual devices, such as smartphones and tablets, to personalize their views at e.g. live concerts.

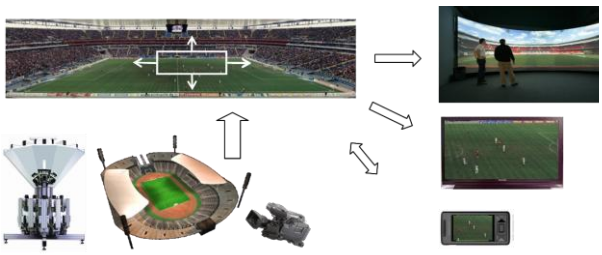


Figure 2 – Main use cases in FascinatE.

METADATA AND SCRIPTING

Figure 3 – Example metadata: ROIs from a developed real-time GPU-based person tracker on panoramic image stream.

In the FascinatE system various types of metadata need to be managed. Based on the FascinatE system architecture a study of the metadata flow in the system and potential metadata formats has been performed. For some types of metadata there are obvious candidate formats (e.g. MPEG-7, MPEG-21), which cover the FascinatE requirements. These include A/V analysis metadata, rights/licensing and device/network capabilities. For sensor parameters, calibration metadata and user profiles, at least one format exists that covers the requirements good enough. The gaps can be closed by defining extensions for the candidate formats. Finally, for other types of metadata, such as knowledge about domain & scene, production rules, visual grammar, user interactions, script templates and scripts, no obvious candidate format could be identified. For those, an application-specific format, or a comprehensive extension of an existing format, will be defined within the project.

The FascinatE Scripting Engines are the components that take decisions about what is visible and audible at each playout device and prepare the audiovisual content streams for display. Such components are referred to as *Virtual Directors*. There are two main types of scripting engines: The Production Scripting Engines (PSE) are responsible for real-time decision making to select content/camera views. Decisions are influenced by e.g. content relevance, visual grammar, privacy and licensing rules, terminal capabilities. The output of a PSE will be a production script (P Script). The Delivery Scripting Engines (DSE) take care of the format-agnostic preparation of content streams and generate delivery scripts (D Script).

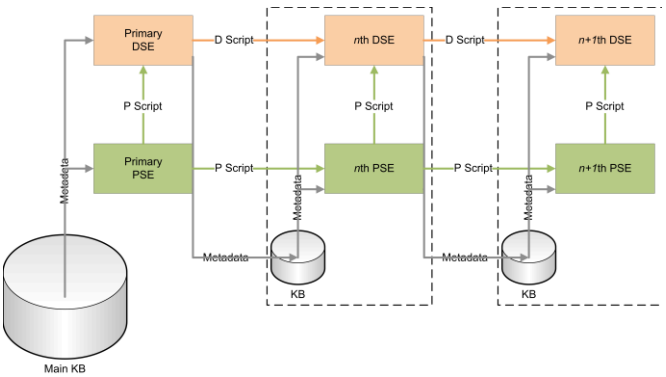


Figure 4 – Cascade of Scripting Engines.

The scripting architecture will be shaped as a cascade (Figure 4) where each stage consists of triples of a PSE, a DSE and a local Knowledge Base (KB) for metadata storage. Edit decisions will be more and more restricted down the cascade and only necessary metadata is passed on. This allows handling content differently for various end user groups and/or terminal types. The cascading architecture will allow flexible adaptation to the complexity of the aforementioned use cases.

The key requirements of the scripting process are as follows; decisions have to be made with a constant, maximum decision making delay; synchronisation of audio and video must be maintained despite separate processing in the workflow; a continuous real-time stream of low level cues must be provided by the content analysis process; a formalised description of rules must be available that drives the decision process, including modelling user preferences and aesthetic rules for different genres. After evaluating different approaches for the PSE decision making process, a rule-based approach using the Complex Event Processing engine JBoss Drools [6] was chosen.

DELIVERY NETWORK

The common denominator of the use cases described earlier is the need for the network to ingest the whole set of A/V data produced to support these immersive and personalized applications. This translates into very demanding bandwidth requirements. As an example, the live delivery of the Layered Scene format would require an uncompressed data rate of 16Gbps. In situations where the full layered scene is to be received by the terminal, say in the case of a theatre with large-scale immersive rendering conditions, the delivery merely requires massive end-to-end bandwidth provisioning. But FascinatE also aims at delivering immersive video services to terminal devices with lower bandwidth access or less processing horsepower. In particular, a high-end home set-up capable of processing the full layered scene for interactive rendering (as described in the next section), but with typical residential network access, may be unable to receive 16Gbps of the full layered scene. In such situations however, a high-quality interactive video experience can still be offered, provided that some forms of in-network filtering are put in place and deliver, at any point in time, only the portions of the layered scene that are required to be rendered by the terminal. In order to support immersive and interactive media consumption to a large range of terminals in a scalable way, the project has focused so far on some particular delivery mechanisms.

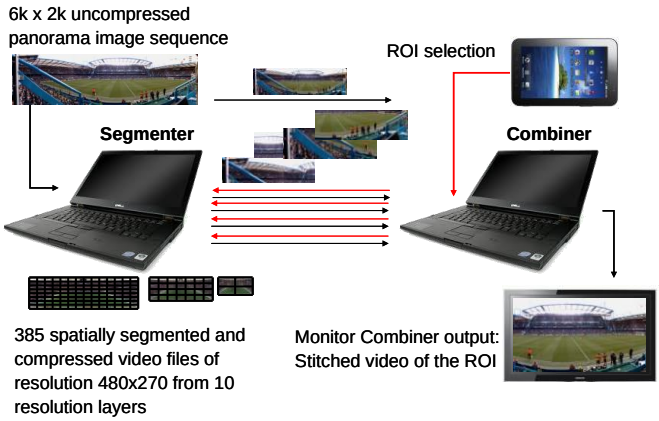


Figure 5 – HAS-based tiled streaming prototype.

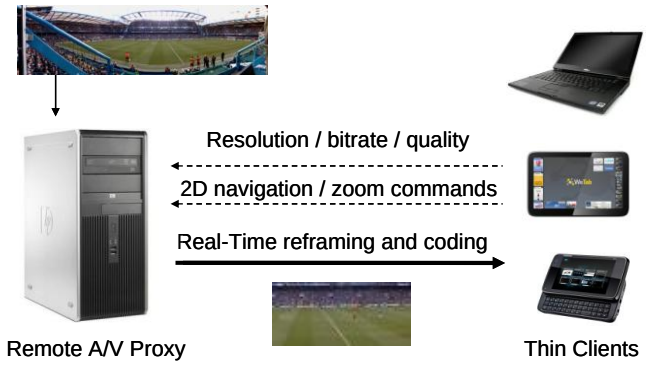


Figure 6 – Remote A/V proxy prototype.

For supporting a flexible transport of the A/V data, a tiled streaming mechanism is employed to package the A/V data at the network ingest point, using various schemes for temporal and spatial segmentation of video panoramas. First results focus on how the obtained segments can be efficiently transported under constrained bandwidth resources. In [7], we show the benefits of video spatial segmentation (tiling) in two ways. First we analyse the trade-off between the bandwidth gain of the tile filtering against the compression overhead of the tiling process itself. Then we show how a tiled streaming system can accommodate the network latency in case of low-delay interactivity. This is done by transmitting tiles in the neighbourhood of the client's current view, with an adapted rate-distortion level. In [8], we describe an implementation of tiled streaming based on adaptive HTTP streaming (HAS), enabling PTZ navigation and region of interest (ROI) selection (Figure 5). Finally in case of low-powered devices, such as mobile phones or tablets, one of the FascinatE goals is to introduce media proxies, capable of performing some or all rendering functions on behalf of the end-client. To assess the feasibility of our approach, prototypes have been built to support real-time navigation within a rectangular panorama for a thin client device, while all the required cropping and rescaling operations are performed at the network-side, before being delivered ready-to-display towards the terminal (Figure 6).

FLEXIBLE AND INTERACTIVE VIDEO RENDERING

Free view based on 3D models still fails to create high quality images comparable to today's HDTV programs. A logical step is to make use of available camera technology and graphical processing power to render images with higher resolution while allowing the consumer to select individually a favoured perspective. This increases the immersive experience by added detail and enhanced interactivity. Today's content rendering in media terminals is understood as decoding and formatting to present at connected displays and loudspeakers. The format such as framing for the displayed view is already defined by the production and the selected business model.

The FascinatE project has specified the Layered Scene as a generic data model to represent multiple layers of audio visual information formed by clusters of cameras and microphones, the creation of virtual cameras having freedom of perspective selection can be supported. The projection of such a scene selection on a display of the end users terminal will be achieved by scalable FascinatE Rendering Nodes (FRN), (Figure 7).

The scalability of the rendering process is reached by cascading multiple FRNs along the work chain for production, delivery and involved terminals. They are divided into rendering operations for device-independent compositions and dependent-presentation processes. This relates to requests from a user to look into a specific area of the panorama (composition) and showing that on a specific display (presentation). The configuration of this scene rendering is done based on scripts derived from the original generic scene representation. They also describe ROIs for tracked objects or predefined views within the panorama. Virtual camera navigation in a cylindrical panorama and optional available overlaid perspectives of shot cameras require powerful system architectures of the end user devices. The FascinatE clients ensure scalability and low latency of content presentation. Profiles to describe functions and levels to structure system parameters are required to organize a scalable terminal infrastructure and have been described in [9].

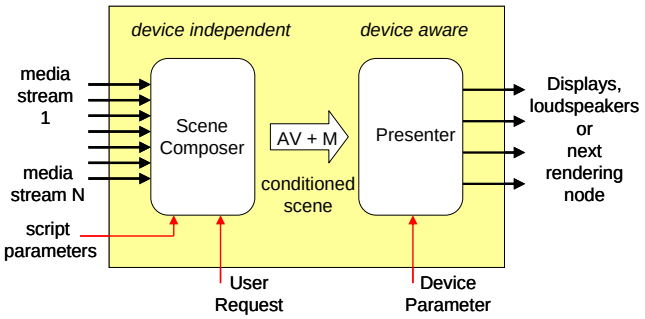


Figure 7 – The FascinatE Rendering Node.

For the FRN prototype rendering, three categories were specified and implemented; the live video layer processes the panorama and the optional shot frames. A ROI layer renders live video related markers and object indicators. The Graphical User Interface (GUI) layer finally produces graphical elements for information, navigation, logos or object lists. In an example of a rendered image (Figure 8) two ROIs markers are placed over a video sequence of panorama content. Additionally, navigation markers, logos and a panorama overview are shown.

Figure 8 – Rendered image of FascinatE Prototype

GESTURE-BASED USER INTERACTION

The FascinatE project is working in providing seamless user interaction with the system by detecting and recognizing user gestures. There is a global tendency to replace external devices, such as remote controls, keyboards or mice, with device-less gesture recognition solutions applied to many applications related to the interaction between users and machines. In FascinatE, the objective is to obtain these device-less, but also marker-less, gesture recognition systems that allow users to interact as naturally as possible, providing a truly immersive experience. Therefore, a user of the FascinatE system will be able to interact with it from the couch without the need of any external device on their hands. The gestures allow the user to perform simple interactions, such as selecting different channels on their TVs, to more innovative interactions such as automatically following players in a football match or navigating through the high resolution panoramic views of the scene. A home setup consisting of a depth sensor attached to TV set is proposed in order to detect and classify user gestures. The depth sensor can be either a time-of-flight or a more recent Microsoft Kinect sensor and provides the necessary 2.5D information for the gesture recognition system.

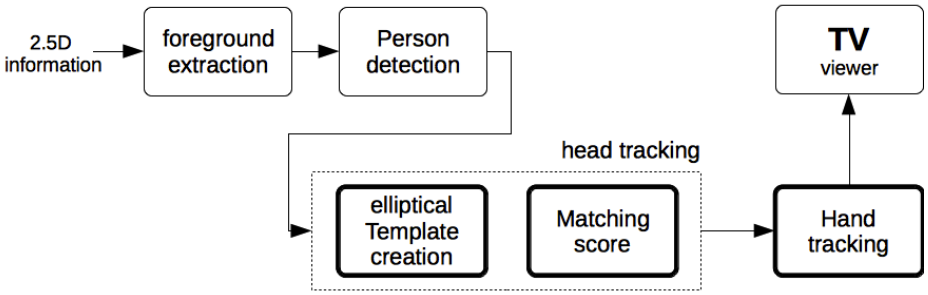


Figure 9 – Block diagram of the proposed gesture recognition system and visualization on TV set.

In order to interpret user gestures, head and hands are tracked by exploiting the 2.5D information [10]. The gesture recognition system consists of several blocks (Figure 9). First of all, foreground extraction and person detection is performed in the raw data. With that information, a head tracking algorithm locates the head of the user within the scene by an elliptical fitting and a best matching score. In a second step, a virtual 3D bounding box is attached to the head position, in such a way that hands lie in the box when moved before the body. An estimate of the position of the hand(s) is obtained after segmenting and grouping the 3D points in the bounding box.

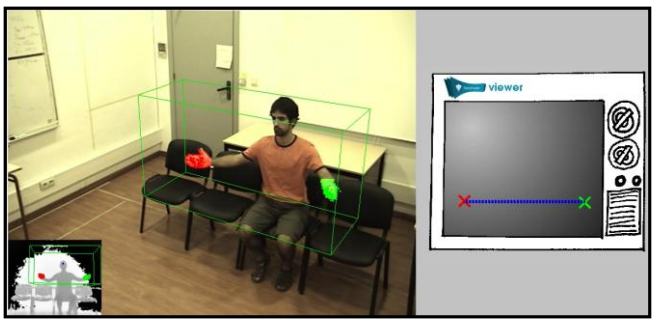


Figure 10 – 3D virtual bounding box (green) with tracked hands (red,green) and user feedback on a TV set.

The gesture recognition system has been integrated in a real-time system controlling the panorama image rendered by the FRN described above. Figure 10 shows a practical example of how the gesture recognition works. On the left side, both tracked hands (red and green) inside the virtual 3D bounding box (green) are overlapped in a user home setup. The right side of Figure 10 shows a possible user feed-back on a TV set where the user can visualize the relative position of his/her hands on top of the TV content. In the integrated system the user is able to control de FRN to navigate and zoom in and out of the panorama image performing several pre-defined device-less gestures.

MARKET OVERVIEW AND SERVICE POTENTIAL

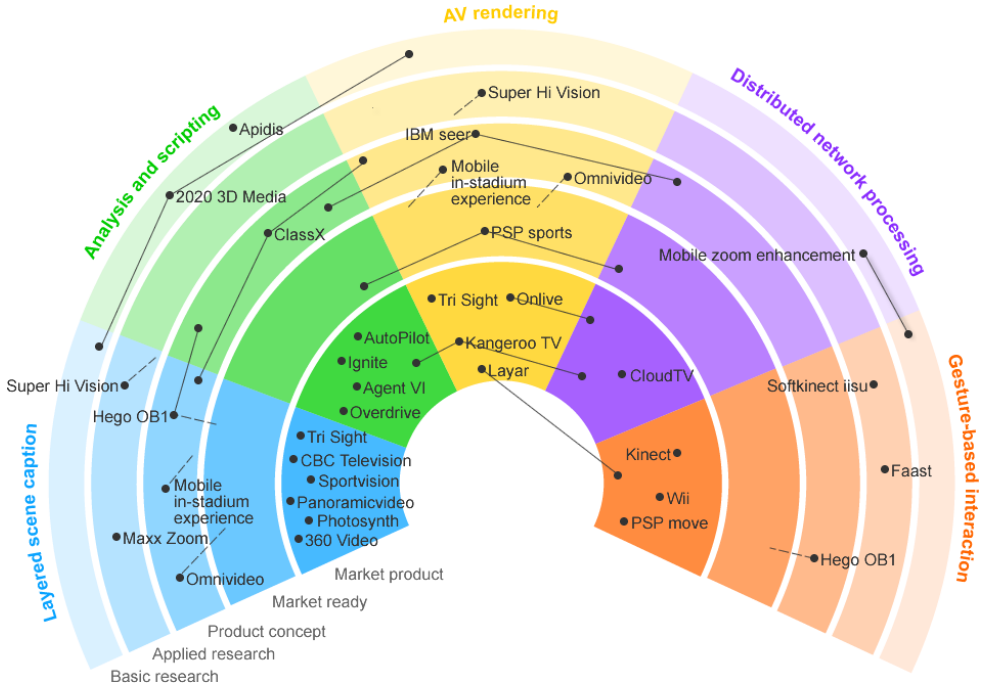


Figure 11 – Technology radar for the FascinatE technical attributes.

The five technical attributes, including the four technologies described in this paper, form the basis for future FascinatE services. In order to asses the service potential, we have investigated the presence of similar technology and services in the current market. A total of 36 service and technology examples in relevant markets were identified through desktop research, via interviews with relevant stakeholders and from a project-wide questionnaire. These services were then classified, a.o. based on the maturity and offered service value, and visualized in the form of a technology radar (Figure 11). From this analysis we can make the following observations. First, R&D developments in the area of delivery networks and gesture-based interaction related to FascinatE are relatively scarce. Current R&D projects mostly consider advanced and interconnected Content Delivery Networks, but network-based processing only to a limited extent. For gesture-based interaction, the Kinect has had a significant impact on developments and we expect more in that area in the near future from the gaming market. However, other types of advanced interactions are limited. In contrast, there is a more pronounced presence of technologies and services related to layered scene capture, scripting and audiovisual rendering, although the existing technologies mainly focus on single layer scenes, e.g. a stitched panorama. A more detailed analysis of the market overview can be found in [11].

The information in Figure 11 allows us to better scope and elaborate upon the three main use cases described in the first section. Within the consortium, we have established a first set of so-called service concepts; a description of potential services, including information on the value proposition, the intended customer segments and relationships, and the channels through which the services are distributed and consumed. This set includes the following five concepts:

1. *iDirector*, providing the home viewer with director-like functionality during live events and enabling the viewer to orchestrate the different views that are available from the layered scene capture;
2. *Immersive Experience*, providing an immersive experience of live events to viewers, by presenting the audiovisual information on a panoramic screen and a 3D-audio setup, such that the viewer experiences the feeling of being physically present at the event;

3. *Mobile Magnifier*, providing users the option to, while watching a live event, select a specific view, which can then be played out on a mobile device;
4. *Cost Efficient Event Reporting*, enabling a production team to direct the capture of an event, supported by automated selection of areas of interest;
5. *Omni Security Cam*, enabling a surveillance team to minimize the human monitoring task by automatically selecting the content that requires immediate action.

FUTURE WORK

Further study in the FascinatE project will focus on developing the technical attributes and integrating them into the overall FascinatE system. For scripting, the definition of rule sets for different scenarios will continue. For the FascinatE Delivery Network, a reference architecture will be specified and the selected delivery and proxying mechanisms will be integrated. For the FascinatE Rendering Node, the applicability of multi-core architectures will be assessed and the identified bottlenecks to pass video elements fast enough to processing units will be tackled. For gesture-based interaction, new gestures will be investigated in order to allow the user to have a further control of the FascinatE system, i.e. allowing the user to point zones on the screen to automatically select regions of interest and navigate through menus or manage the sound of the system.

REFERENCES

1. M. Maeda, Y. Shishikui, F. Suginoshita, Y. Takiguchi, T. Nakatogawa, M. Kanazawa, K. Mitani, K. Hamasaki, M. Iwaki and Y. Nojiri. “Steps Toward the Practical Use of Super Hi-Vision”. NAB2006 Proceedings, Las Vegas, USA, April 2006.
2. Cinegrid, http://www.cinegrid.org Visited: May 12, 2011.
3. Fraunhofer HHI, http://www.hhi.fraunhofer.de/en/departments/image-processing/applications/omnicam Visited: May 12, 2011.
4. FascinatE, http://www.fascinate-project.eu Visited: May 12, 2011.
5. Ralf Schäfer, Peter Kauff, and Christian Weissig. “Ultra high resolution video production and display as basis of a format agnostic production system”, IBC2010 Proceedings, Amsterdam, Netherlands, September 2010.
6. JBoss Drools – The Business Logic integration Platform. http://www.jboss.org/drools Visited: May 12, 2011.
7. P. Rondao Alface, J.-F. Macq, N. Verzijp, “Evaluation of Bandwidth Performance for Interactive Spherical Video”, WoMAN’11 Proceedings, Barcelona, Spain, July 2011.
8. O.A. Niamut, M.J. Prins, R. van Brandenburg, A. Havekes “Spatial Tiling And Streaming In An Immersive Media Delivery Network”, EuroITV2011 Adjunct Proceedings, Lisbon, Portugal, June 2011.
9. M. Borsum, J. Spille, A. Kochale, E. Önnevall, G. Zoric, J. Ruiz. “AV Renderer Specification and Basic Characterisation of Audience Interaction”, FP7 FascinatE Deliverable D5.1.1, July 2010. Available at http://www.fascinate-project.com/wp-content/uploads/2010/09/Fascinate-D5.1.1-RendererSpecification-AudienceInteraction.pdf
10. X. Suau, J.R. Casas and J. Ruiz-Hidalgo, “Real-Time Head and Hand Tracking based on 2.5D data”, ICME2011 Proceedings, Barcelona, Spain, July 2011.
11. O.A. Niamut, T.T. Bachet, A.J.P. Limonard, “High-Resolution Video, More Is More?”, EuroITV2011 Proceedings, Lisbon, Portugal, June 2011.

ACKNOWLEDGMENTS

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.