Format-agnostic Approach for Production, Delivery and Rendering of Immersive Media

O. Schreer[1], G. Thomas[2], O.A. Niamut[3], J-F. Macq[4], A. Kochale[5], J-M. Batke[5], J. Ruiz Hidalgo[6], R. Oldfield[7], B. Shirley[7], G. Thallinger[8]

[1]Fraunhofer Heinrich-Hertz Institute, Berlin, Germany; [2]BBC R&D, London, UK; [3]TNO, Delft, The Netherlands; [4]Alcatel-Lucent, Antwerp, Belgium; [5]Image Processing Lab, Technicolor, Hannover, Germany; [6]UPC, Barcelona, Spain; [7]University of Salford, Manchester, UK; [8]JOANNEUM RESEARCH, Graz, Austria

E-mail: [1]Oliver.Schreer@hhi.fraunhofer.de, [2]Graham.Thomas@bbc.co.uk, [3]omar.niamut@tno.nl, [4]jean-francois.macq@alcatel-lucent.com, [5]axel.kochale@technicolor.com, [5]Jan-Mark.Batke@technicolor.com, [6]j.ruiz@upc.edu, [7]R.Oldfield@salford.ac.uk, [7]B.G.Shirley@salford.ac.uk, [8]georg.thallinger@joanneum.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 the EU-funded FascinatE project. In this paper, different aspects of the format agnostic approach are discussed which we believe can be a promising concept for future media production and consumption. The different parts of the complete multimedia production and delivery process are revisited demonstrating the requirements and the potential of such an advanced concept.**

Keywords: Ultra-high definition, panoramic imaging, spatial audio, media aware networking, gesture-based interaction.

1 INTRODUCTION

It is an often-expressed view that the broadcast 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 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 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 or wave field synthesis.

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

Corresponding author: Oliver Schreer, Fraunhofer Heinrich-Hertz Institute, Einsteinufer 37, 10587 Berlin, Germany, Phone: +49 30 31002-620, E-mail: Oliver.Schreer@hhi.fraunhofer.de

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. 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 innovation areas of the project. In section 3 and 4, aspects of the acquisition and production as well as networking and delivery are presented. Section 5 discusses the role of spatial audio and in section 6, the rendering approach as well as the gesture interface are presented. A conclusion ends the paper.

2 INNOVATION AREAS

FascinatE aims to improve upon the current immersive media developments by focussing on our key innovation areas, shown in Figure 1.

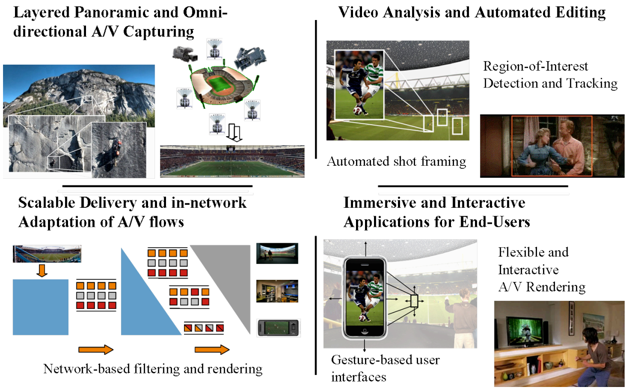
Layered Panoramic and Omni-directional A/V Capturing
Video Analysis and Automated Editing
Region-of-Interest Detection and Tracking
Automated shot framing
Scalable Delivery and in-network Adaptation of A/V flows
Immersive and Interactive Applications for End-Users
Flexible and Interactive A/V Rendering
Network-based filtering and rendering
Gesture-based user interfaces

Figure 1: Key innovation areas in FascinatE

From these areas, we five technologies are identified that the project partners are actively pursuing, referred to as technical attributes:

1. *Layered Scene Production*, where audiovisual scenes are captured in multiple resolutions, frame rates and dynamic ranges with A/V sensor 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, based on production knowledge and metadata;

3. *Scalable Delivery and In-Network Audio/Video Adaptation*, enabling the efficient delivery and media-aware network-based processing that is required for the support of low-end terminals and bandwidth limitations in the access networks;

4. *Flexible and Interactive Audio/Video Rendering*, adapting the content to the end-user terminals with their associated screen and speaker set-ups;

5. *Gesture Based User Interaction*, enabling natural end-user navigation based on simple and intuitive gestures.

FascinatE further considers three main use cases with associated devices and screens (see Figure 2). In the theatre case, the captured content is transmitted to and displayed on a large panoramic screen, with the associated 3D audio being presented through a multi-loudspeaker set-up. This enables multiple viewers to simultaneously see the content and interact with it. 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 smart phones and tablets, to personalize their views at e.g. live concerts.

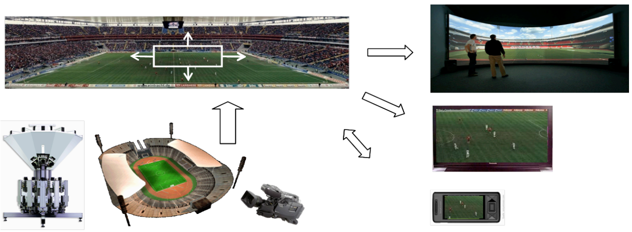

Figure 2: FascinatE main use cases and terminals

The expected impact of the technological developments in the aforementioned five areas has been investigated by introducing three technology dominance scenarios. These describe what part of the TV broadcasting value chain will be primarily influenced by the technical attributes.

In the production-centric scenario, the focus is on innovations in the production domain, assuming a current state-of-the-art delivery network and excluding any FascinatE functionality in the network and terminal. The distribution of FascinatE content is tailored to a specific delivery format and one or more rendered views in the form of TV channels and/or media streams are presented to the user. End-user interaction is limited to switching between channels and selecting streams. The degree of interaction allowed for current end-users is determined at the production side by the number of views made available.

In the terminal-centric scenario, an idealistic delivery network is assumed which allows for distribution of the full layered scene, with the terminal receiving and rendering all the captured A/V streams. Production scripts are sent towards the terminal, containing production-side knowledge that specifies the required processing steps. Significant computational load is moved to the terminal side for the processing of production-generated scripts in response to user commands.

In the network-centric scenario, the delivery network contains processing functionality for (partly) rendering the layered representation of the A/V scene and processing the accompanying production and delivery scripts. Within the delivery network, the layered scene will be rendered to a format tailored to the access network and the requesting or targeted terminal. The script processing is also located in the delivery network and receives the interaction commands from the terminal side and the production script from the production side. Based on these inputs it can control the rendering function to provide the right view in the appropriate format to the terminal.

3 ACQUISITION AND PRODUCTION

The format agnostic nature on the production side requires a large variety of audio-visual sensors. In order to allow the succeeding modules to access the content in the most efficient way, a layered scene description has been developed. This description contains all the available data, but also the relationship between them in the form of metadata. In terms of visual information, this contains either geometrical information (i.e. which view of the scene), the spatial and temporal resolution or dynamic range. The metadata for audio sensors contains the geometrical information i.e. position and orientation, number of capsules, capsules arrangement and transport format (mono, stereo, surround, ...).

A set of different visual sensors are currently part of the FascinatE system. An ultra-high resolution omni-directional camera consisting of 6 individual HD cameras mounted in a special mirror rig is used to capture a 180° panoramic view of the scene at a resolution of 7K x 2K pixels (see Figure 3, left). In addition to that, a number of HD broadcast cameras are available allowing pan/tilt and zoom into special parts of the scene. Cameras with different dynamic range or frame rate are envisaged as well. The audio scene is captured by a number of shot gun microphones, stereo microphones, soundfield microphones and an Eigenmike® [2] (see Figure 3, right).

The captured raw content needs to be further processed according to the layered scene description. Hence a complete stitched panoramic view is produced out of the individual HD cameras of the omni-directional camera (see Figure 4). A registration of the different views (e.g. panoramic and zoom in view) is performed to offer the user additional information of the scene at higher resolution, better dynamic range or even higher frame rate. The set of audio signals is processed to achieve a soundfield description, which can then be used to drive a large variety of sound systems in a format agnostic manner.

The complete acquisition and production pipeline has been investigated and tested in a first test shoot at a UK premier league football match in October 2010.

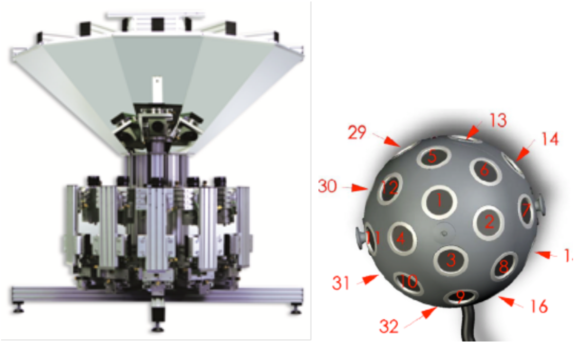

Figure 3: Omni-directional camera by Fraunhofer HHI (left), Eigenmike® (right)

This showed that the format agnostic concept for immersive media is well suited for sports events. A variety of other live events are also considered such as music festivals, pop concerts or big theatre events like opera or musicals.

4 ROLE OF NETWORK SUPPORTING IMMERSIVE MEDIA

The common denominator of the use cases described before is the need for the network to ingest the whole set of A/V data produced to support immersive and personalized applications. This typically translates into very demanding bandwidth requirements. As an example, the live delivery of the immersive A/V material currently used within the project would require an uncompressed data rate of 16 Gbps. 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 power. In particular, a high-end home set-up capable of processing the full layered scene for interactive rendering, but with typical residential network access, may be unable to receive the data rate of the complete 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.

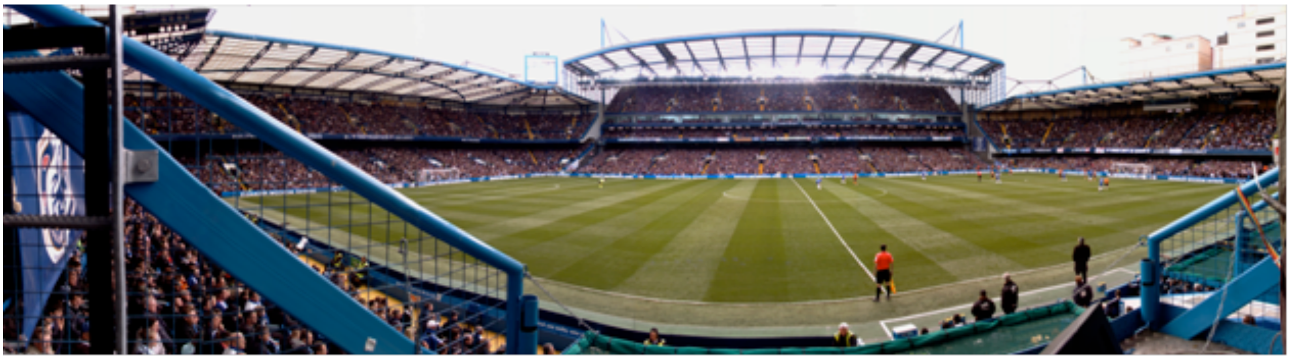

Figure 4: Panoramic view

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 (if not all) rendering functionality on behalf of the end-client.

4.1 Delivery mechanisms

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, at the A/V Ingest, Transport and Proxy (see Figure 5). 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. Current results focus on how the obtained segments can be efficiently transported and filtered, e.g. under constrained bandwidth resources. In [3], the optimal sizing of rectangular panorama tiling is investigated for interactive navigation in high-resolution spherical video under varying bandwidth and delay constraints. In [4], an implementation of tiled streaming is described based on adaptive HTTP streaming, enabling pan/tilt/zoom interaction.

In order to assess the feasibility of A/V proxies, prototypes have been built to support real-time navigation (panning and zooming) 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.

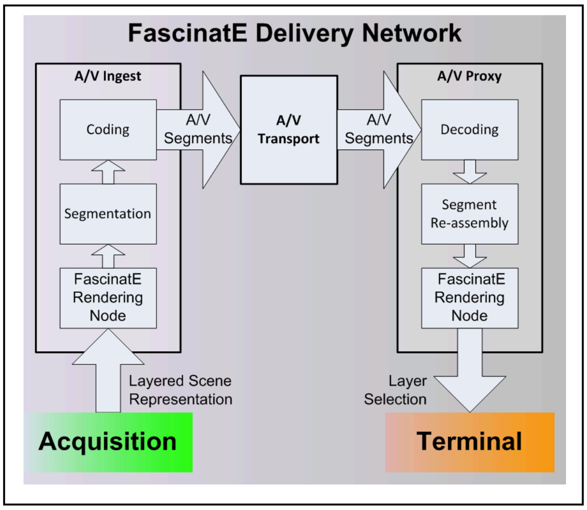

Figure 5: FascinatE Delivery Mechanisms and Network.

4.2 Types of networks

Several classes of delivery networks and technologies are relevant for supporting the FascinatE services. The project aims at defining a reference delivery architecture that can take a variety of forms depending on selected evolutionary paths. We give hereafter examples of current delivery networks that can evolve to support the interactive and immersive video services considered within FascinatE [5]:

IPTV. It is characterized by being entirely deployed over an operator's walled-garden delivery network. This allows the use of native IP multicast for transmitting live streams, which provides the most efficient use of bandwidth resource within an IP-based delivery network. Moreover, this IPTV context offers the possibility to implement mechanisms closer to the end-user so as to improve QoS or interactivity. Examples include Retransmission and Fast Channel Change mechanisms that can be implemented in the edge or access parts of the network. From a service prospective, IPTV has been paving the road towards more interactivity within the TV experience.

CDN. Today the mechanisms and the resources provided by Content Delivery Networks (CDN) represent the main forces that make the open Internet support large scale delivery of on-demand video content. Some CDNs also support live video delivery with an infrastructure that can scale to a high and geographically-spread demand. However, being generally built as a distributed overlay infrastructure, CDN-based delivery typically leads to long end-to-end delay and response time, in comparison to the interactivity requirements targeted in FascinatE. Still, some form of CDN technology seems very appropriate in order to deploy Fascinate content on a large, or even global, geographical scale. Such a CDN infrastructure would play a role analogue to a contribution network as its main role would be to push the entire (or a pre-defined portion of the) FascinatE layered scene closer to the delivery edge. From there on, more flexible mechanisms would need to be put in place to support interactivity and content request on a per user/user-groups basis.

HBB. One of the more recent trends in broadcast networks is that they can be complemented by IP connectivity, leading to hybrid broadcast broadband (HBB) platforms providing a very similar experience to IPTV for the end-user. FascinatE will investigate how hybrid delivery can benefit from combining the flexibility of web-based delivery with the transport efficiency of broadcast, in order to support the FascinatE requirements;

5 FORMAT AGNOSTIC AUDIO

Today's audio formats include stereo (2.0) and surround (5.1) formats as well as extensions of surround sound (7.1, 9.1, and others). The content is played back on the respective loudspeaker setups at the consumer side. Thus, audio content needs to be delivered in a variety of formats. Stereo and surround sound systems have in common that they are restricted to 2-dimensional playback. This means that the acoustical image contains width and depth. Loudspeaker setups for 3-dimensional playback also including height information are seldom in use. Examples include the proposal for the NHK ultra high definition TV with the 22.2 format [6], the 2+2+2 arrangement of Dabringhaus [7], the Auro-3D approach of van Baelen [8], and the 10.2 setup of Holman. All of the mentioned audio formats require a distinct loudspeaker setup with specified loudspeaker positions for playback of the different channel signals. A clear trend of future audio formats is an increasing number of channels. Typically each channel carries a signal dedicated to a specific loudspeaker position (e. g. stereo: first channel = left loudspeaker, second channel = right loudspeaker). A further increase of the channel number will not be a solution, since a high number of loudspeaker signals with specified positions have limited practical value. As consequence, we are considering a new approach: Instead of using the loudspeaker signals as content representation a soundfield description is used, which can be transmitted using several audio channels. Two major soundfield description technologies are established, the "Wave Field Synthesis (WFS)" and the "Higher Order Ambisonics (HOA)" approach [9].

5.1 Sound Field Descriptions

The spatial audio reproduction system for FascinatE aims to replay the soundfield accompanying a visual scene. To drive such a system the description of the desired soundfield is necessary. The soundfield description in turn determines the necessary production and editing steps. Using a WFS system the soundfield is synthesised from several audio objects that are positioned in a sound scene. The acoustical environment is auralised by additional objects that create reverberation [10]. For the sound scene creation, basic parameters like source positions and acoustical parameters need to be known. The next section describes the basic concepts underlying WFS. Contrary to WFS the HOA approach does not use a parametric description of audio objects, but uses a description of the entire soundfield at a certain position instead [11]. In section 5.3, the theoretical background of HOA is briefly described and the necessary steps for signal processing are outlined.

5.2 Wave Field Synthesis

Wave field synthesis uses a sound field description that is based on the Kirchhoff-Helmholtz equation which states that if the pressure and particle velocity at an infinite number of points on the surface of a volume are known, then the pressure at any point in that volume can be determined. This means that the soundfield in a room can be controlled using sources (loudspeaker) on the surface of the room. 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. A combination of plane wave sources can be used to accurately recreate the original soundfield. Because the soundfield is correct over the volume each person in the listening space will get the correct spatial impression no matter where they are. Using this description the audio can also be down mixed to less complex systems such as stereo, allowing format agnostic rendering.

5.3 Higher-Order Ambisonics

Ambisonics is a soundfield description method employing a mathematical approximation of the soundfield in one reference point, often referred to as "sweet spot". The basic idea of Ambisonics is to describe a soundfield using the coefficients of a spatial Fourier transform [9]. In FascinatE this Ambisonics representation is being treated as an intermediate format. The advantage of this representation is a scalable representation of a soundfield with respect to the spatial resolution. In other words, if a high spatial resolution is not required, the coefficients of the highest order may be simply omitted. Moreover, the HOA representation is agnostic to the loudspeaker setup since the transmitted signals do not refer to specific loudspeaker positions, but describe time varying coefficients of a soundfield. This in turn leads to some specific points that require special attention. On the recording side, microphone arrays can be used to obtain the coefficients of an Ambisonics representation.

6 RENDERING AND INTERACTIVITY

6.1 Format agnostic rendering

Free view based on 3D models still fails to create high quality images comparable to today's High Definition (HD) programs. One logical step is to make use of available camera technology and graphical processing power to render images with higher resolution while allowing the content 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 limited by formats selected by the production and the business model. This results in a potential mismatch to display and loudspeaker setups used for content playback. To outpace this format limitation, the FascinatE project has specified a generic data model to represent multiple layers of audio visual information formed by so called clusters of cameras and microphones. Such a cluster captures a panorama with additional HD and high dynamic range (HDR) camera shots. This captured scene supports the creation of virtual cameras having freedom of perspective selection. The projection of such a scene selection on a display of the end users terminal will be achieved by scalable FascinatE Rendering Nodes (FRN) (see Figure 6).

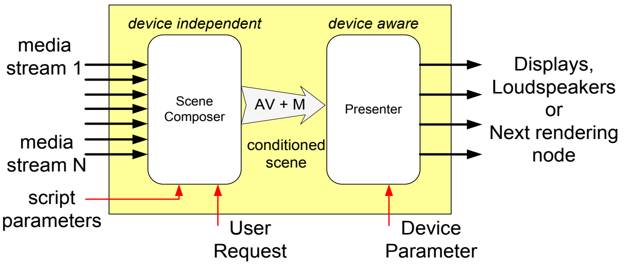

Figure 6: FascinatE Rendering Node (FRN)

Scalability 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 regions of interests (ROIs) for tracked objects or predefined views within the panorama or from other cameras out of the cluster. Virtual camera navigation in a cylindrical panorama and optional available overlaid perspectives of shot cameras (ROIs) 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 [12]. Further study in the FascinatE project aims to provide assessments of applicability of multi core architectures and tackling identified bottlenecks to pass video elements fast enough to the processing units. For the FRN prototype, rendering 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.

6.2 Gesture interaction

Gesture recognition technologies are being widely applied to many applications related to the interaction between users and machines. There is a global tendency to replace external devices, such as remote controls, keyboards or mice, with device-less gesture recognition solutions. Indeed, the objective is to obtain device-less, but also marker-less, gesture recognition systems that allow users interact as natural as possible, providing a truly immersive experience. The FascinatE project is working in providing seamless user interaction with the system by detecting and recognizing user gestures. 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 high resolution panoramic views of the scene. In order to interpret user gestures, head and hands are tracked by exploiting the 2.5D information [13]. 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. In a second step, a three dimensional virtual 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. On the left image of Figure 7, an example of both tracked hands (read and green) inside the virtual 3D bounding box (green) are overlapped in the user home setup. The gesture recognition system performs in real time and enables many interesting applications. The right side of Figure 7 shows a possible user feedback on a TV set where the user can visualize the relative position of his/her hands on top of the TV content. A communication channel using TCP/IP has been created between the FascinatE Rendering Node (FRN) and the gesture recognition system. Each time a gesture is recognized it is transmitted to the FRN as an XML message. At the moment, only navigation and zooming functionality are implemented, but eventually, the user could be able to point zones on the screen, navigate through menus or follow regions of interest in the FascinatE's TV-based home system.

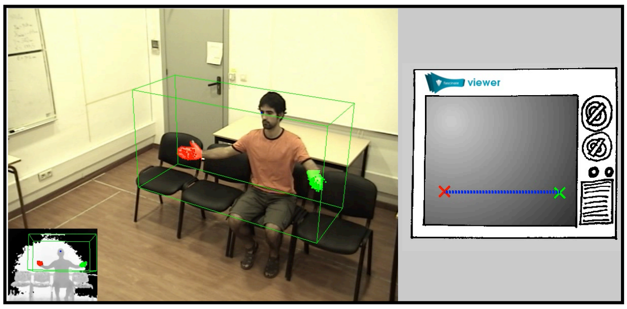

Figure 7: 3D virtual bounding box with tracked hands (red, green) and user feedback on a TV set

7 CONCLUSION

We presented a format agnostic approach for future immersive multimedia production, delivery and consumption based on recent research results of the FP7 EU-funded project FascinatE. Several parts of the complete processing chain have been discussed in order to identify the necessary requirements or even to show first solutions towards a future multimedia framework.

Acknowledgement

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 248138.

References

[1] R. Schäfer, P. Kauff, C. Weissig, "Ultra high resolution video production and display as basis of a format agnostic production system", Proceedings of IBC 2010.

[2] http://www.mhacoustics.com/mh_acoustics/Eigenmike_microphone_array.html

[3] P. Rondao Alface, J.-F. Macq, N. Verzijp, "Evaluation of Bandwidth Performance for Interactive Spherical Video", to appear in Proceedings of IEEE Workshop on Multimedia-Aware Networks (WoMAN'11), July 2011.

[4] O.A. Niamut, M.J. Prins, R. van Brandenburg, A. Havekes "Spatial Tiling And Streaming In An Immersive Media Delivery Network", to appear in Adjunct Proceedings of EuroITV 2011, June 2011.

[5] A. Havekes, O.A Niamut, M.J. Prins, J-F Macq, P. Rondao Alface, N. Verzijp, "Capabilities of current and Next-Generation delivery networks involved in FascinatE services", FP7 FascinatE Deliverable D4.2.1, February 2011. http://www.fascinate-project.com/?page_id=690

[6] K. Hamasaki, T. Nishiguchi, R. Okumaura, and Y. Nakayama. Wide listening area with exceptional spatial sound quality of a 22.2 multichannel sound system. In Audio Engineering Society Preprints, Vienna, Austria, May 2007.

[7] MDG. Mdg-musikproduktion dabringhaus und grimm. http://www.mdg.de.

[8] Galaxy Studios Group. Homepage auro-3d. http://auro-3d.com, visited 2010-07-16. URL http://auro-3d.com.

[9] S. Spors and J. Ahrens. A comparison of wave field synthesis and higher-order ambisonics with respect to physical properties and spatial sampling. In 125th AES Convention, San Fransisco, USA, 2008.

[10] Diemer de Vries. AES Monograph - Wave Field Synthesis. Audio Engineering Society Inc., 2009.

[11] M. A. Poletti. Three-dimensional surround sound systems based on spherical harmonics. J. Audio Eng. Soc., 53(11):1004–1025, Nov. 2005.

[12] M. Borsum, J. Spille, A. Kochale, E. Önnevall, G. Zoric, J. Ruiz "FascinatE deliverable D5.1.1: AV Renderer Specification and Basic Characterisation of Audience Interaction"

[13] X. Suau, J.R. Casas and J. Ruiz-Hidalgo, "Real-Time Head and Hand Tracking based on 2.5D data", In International Conference on Multi-media and Expo (ICME), 2011.