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

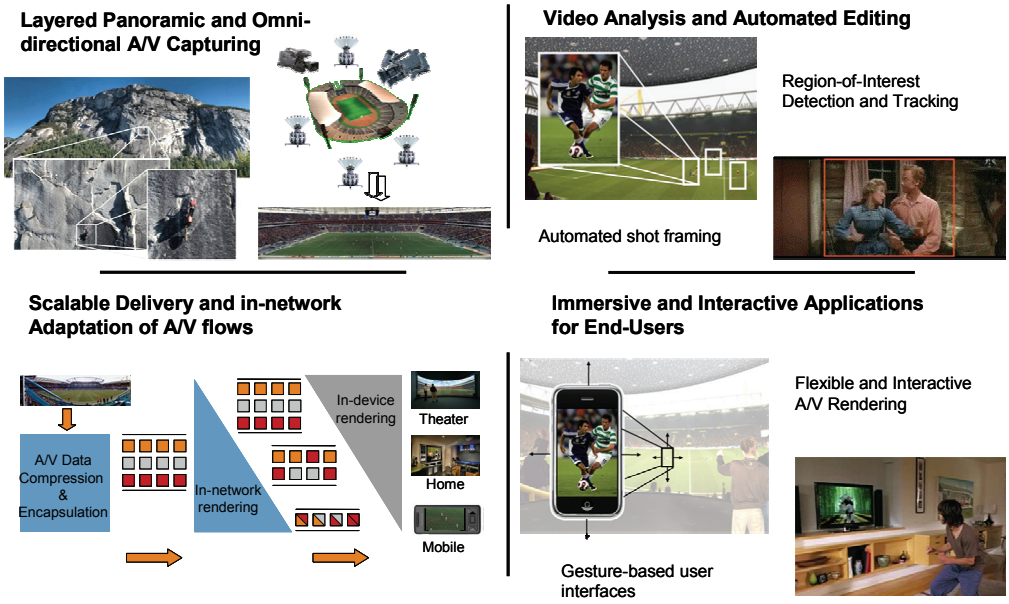

Figure 1 – Key innovation areas in FascinatE.

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.

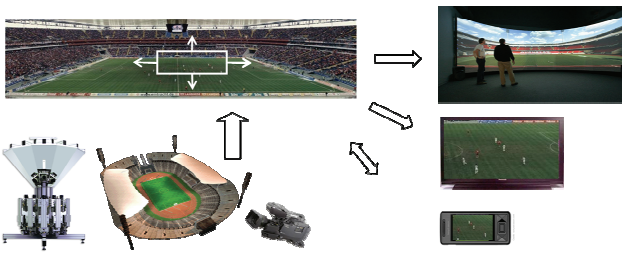

Figure 2 – Main use cases in FascinatE.

METADATA AND SCRIPTING

Figure 3 – Automated tracking results are used for scripting.

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, 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 (Figure 3). 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).

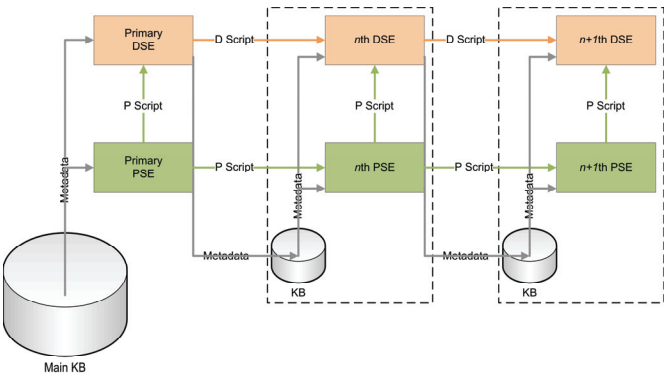

Figure 4 – Cascade of Scripting Engines.

The scripting architecture is 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 allows a 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 audio-video data produced to support these immersive and personalized applications. This typically 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 (Figure 5). In particular, a high-end home set-up capable of processing the full layered scene for interactive rendering, with typical residential network access, may be unable to receive the very high data rate of the complete layered scene. In such situations, a high-quality interactive video experience can still be offered, provided that some forms of in-network filtering are put in place and deliver only the portions of the layered scene that are required to be rendered by the terminal. 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.

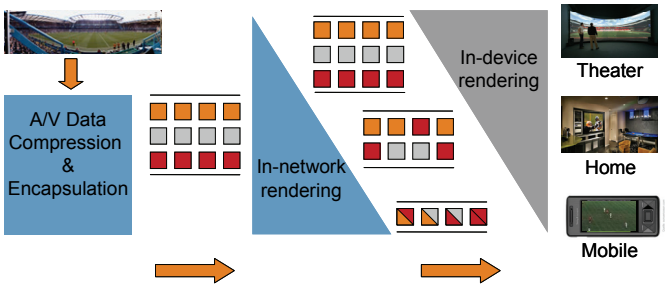

Figure 5 - Overview of network delivery scenarios.

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. 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 [7], we investigate the optimal sizing of rectangular panorama tiling for interactive navigation in high-resolution spherical video under varying bandwidth and delay constraints. In [8], we describe an implementation of tiled streaming based on adaptive HTTP streaming, enabling PTZ navigation. In order to assess the feasibility of A/V proxies, 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).

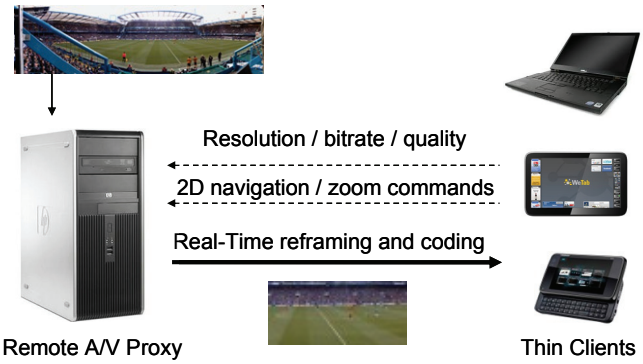

Figure 6 – Remote A/V proxy prototype.

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 regions of interests (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].

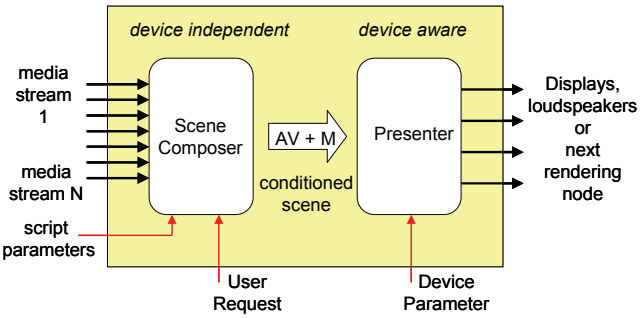

Figure 7 – The FascinatE Rendering Node.

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.

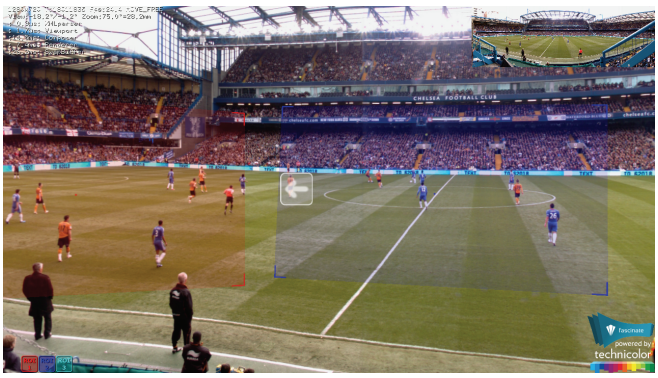

Figure 8 – Rendered image of FascinatE Prototype

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 and gesture recognition technologies are being widely applied to many applications related to the interaction between users and machines. In FascinatE, the objective is to obtain 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.

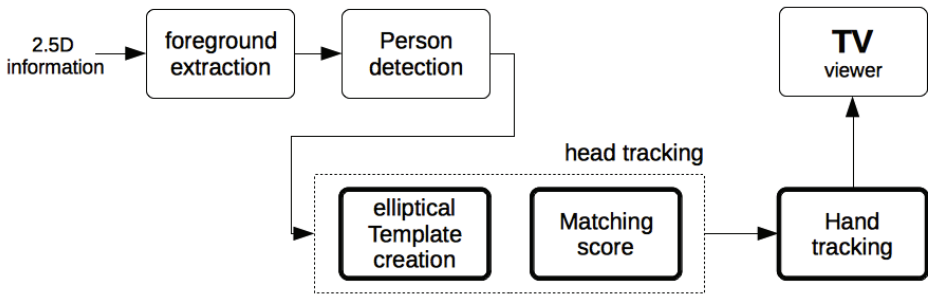

Figure 9 – Block diagram of the proposed gesture recognition system and visualization on TV set.

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 block (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 matching score to find the best possible match of the head position. 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 10, an example of both tracked hands (red and green) inside the virtual 3D bounding box (green) are overlapped in the 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. Eventually, the user could be able to point zones on the screen, navigate through menus or perform gestures to control some functionalities of FascinatE's TV-based home system.

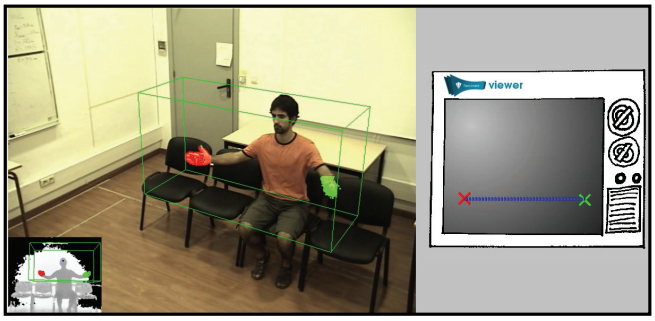

Figure 10 – 3D virtual bounding box (green) with tracked hands (red, green) and user feedback on a TV set.

MARKET OVERVIEW AND SERVICE POTENTIAL

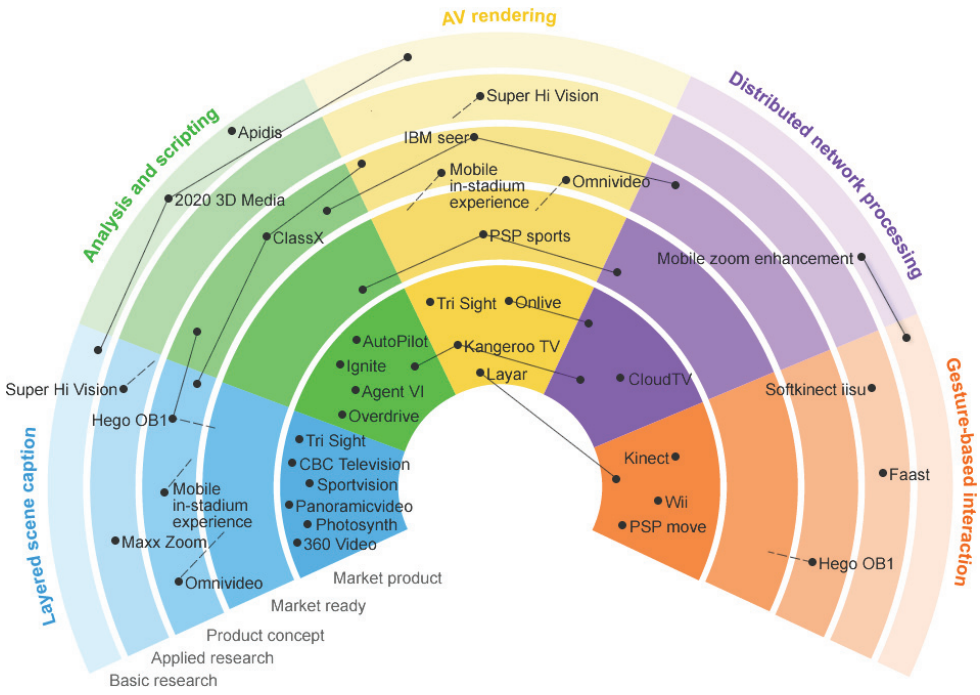

Figure 11 – Technology radar for the FascinatE technical attributes.

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”, to appear in 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”, to appear in 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?”, to appear in 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.