Towards Scalable And Interactive Delivery of Immersive Media

O.A. Niamut[1], J.-F. Macq[2], M.J. Prins[1], R. Van Brandenburg[1], N.Verzijp[2], P. R. Alface[2]

[1]TNO, Delft, The Netherlands; [2]Alcatel-Lucent Bell Labs, Antwerp, Belgium

E-mail: [1]{omar.niamut, martin.prins, ray.vanbrandenburg}@tno.nl, [2]{jean-francois.macq, nico.verzijp, patrice.rondao_alface}@alcatel-lucent.com

Abstract: **Within the EU FP7 project FascinatE, a capture, production and delivery system capable of allowing end-users to interactively view and navigate around an ultra-high resolution video panorama showing a live event is being developed. Within this system, network-based processing is used to repurpose the audiovisual content to suit delivery towards different device types and user selection of regions of interest. In this paper we report on the ongoing developments of the FascinatE delivery network functionality. We present the delivery network architecture and its constituent functional components. The content segmentation procedures at the ingest of audiovisual data are considered and two delivery mechanisms are discussed.**

Keywords: immersive media, high-resolution video, content aware networking, adaptive streaming

1 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 visual experience to end users. This notion of immersive media with ultra-high definition TV (UHDTV) and displays, highlighted by the NHK work on 8K Super Hi-Vision video [1] and the Fraunhofer HHI 6K OMNICAM system [2] seems contradictory with the explosive growth of device diversity. That is, having the content available on an increasing number of mobile devices, such as smartphones and tablets, each with its own characteristics, facilitates the user in selecting and controlling content. In contrast, UHDTV still assumes a more or less passive behaviour on the end user's side. The relevance of this contradiction for future ICT research is recognised in the NEM Strategic Research Agenda [3], as it refers to technologies for transport, coding and rendering, e.g. content-centric networks, spatial and ultra-high resolution video and video over the device continuum, as being vital to immersive media reproduction.

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 Representation (LSR), i.e. a multi-resolution, multi-source representation of the audiovisual environment [5]. In addition, content analysis and 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 content suited to their needs.

This paper focuses on the FascinatE delivery network. This network needs to ingest the whole set of audiovisual (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 in an LSR consisting of an OMNICAM and three HD image sequences would require an uncompressed data rate of more than 16 Gbps. In situations where the full LSR is to be received by an end-user terminal, say in the case of a theatre with large-scale immersive rendering conditions, the delivery requires massive end-to-end bandwidth provisioning, even when using mezzanine or broadcast video compression. 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 LSR for interactive rendering, but with typical residential network access, may be unable to receive the data rate of the complete LSR. 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 functionality on behalf of the end-client.

Corresponding author: Omar Niamut, TNO, Brassersplein 2, 2612CT Delft, +31 651916242, omar.niamut@tno.nl

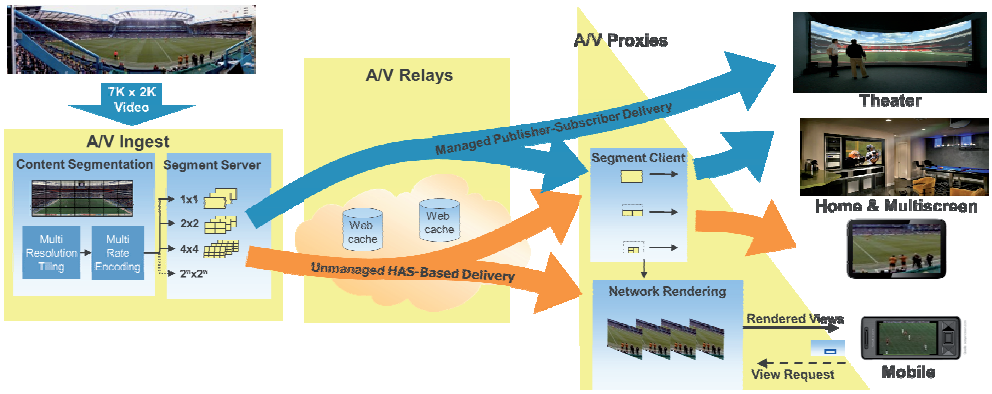

Figure 1 – FascinatE network functionality and delivery mechanisms.

The paper is organized as follows; we first describe the FascinatE network architecture and its constituent functional components in section 2. Then, in section 3, we discuss the ingest of A/V data and content segmentation. The FascinatE delivery network includes two delivery mechanisms, which are described in sections 4 and 5, respectively. Finally, in section 6, we discuss the future work planned in the project.

2 THE FASCINATE DELIVERY NETWORK

FascinatE considers three main use cases, each with its associated target end-device and screen type. First, in the theatre or public screen case, the captured content is transmitted to and displayed on a large panoramic screen, enabling multiple viewers to simultaneously see the content. In contrast, in home viewing situations 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, for mobile usage, users can employ their individual devices, such as smartphones and tablets, to personalize their views.

In the first case, that is large displays for public viewing, dedicated optical networks such as Cinegrid [6] are already employed for uncompressed UHDTV transmissions. In contrast, for current and near-future home viewing situations based on IPTV or DOCSIS cable networks, we expect bandwidths ranging between 20-100 Mbps; not enough to transmit the full LSR, even in compressed form. Furthermore, in the case of mobile broadband networks, bandwidths of up to 20 Mbps are foreseen. Hence, the role of the FascinatE delivery network is different for each of these cases, as shown in Figure 1. In this paper, we concentrate on a delivery network architecture with functional components that facilitate the transmission of the LSR to devices for home viewing, such as Connected TVs and set-top boxes, and to mobile devices such as smart phones and tablets.

2.1 Related Work

A network-based approach for interaction with immersive media was recently demonstrated by KDDI [7]. The demonstrated prototype allows a user to zoom into a region of interest (ROI) on a mobile device. The ROI parameters are sent to a network proxy, which then crops the transmitted video to reduce the overall video bandwidth. That is, since the user is looking at a specific ROI, only that spatial part of the video can be transmitted without loss of resolution. PTZ interaction with video was previously studied by Mavlankar and Khiem. In [8] a video coding approach is described which allows for extracting ROIs directly from the coded bit stream. In [9] a tiled streaming approach is presented. Within FascinatE, the merits of these approaches are studied and incorporated into the delivery network architecture.

2.2 Delivery Network Architecture

Figure 1 also shows the three high-level active delivery components in the FascinatE delivery network. These components provide the following functionality at specific stages of the delivery, namely, ingest, storage and forwarding, and rendering.

- A/V Ingest: receives as input the full LSR and performs initial view rendering applicable for all or a large fraction of the end-users. This rendering is performed by an instance of the FascinatE Rendering Node (FRN). Furthermore, content is prepared for the actual delivery by a content segmentation operation, resulting in FascinatE media delivery units that we refer to as segments. This operation is described in section 3.
- A/V Proxy: at the other end of the network, this block is responsible for ensuring that the A/V segments required by one user or a local set of end-users are delivered and reassembled according to their interactivity requests. The proxy can also perform in-network A/V processing using an FRN instance to adapt to personalized requests and/or personalized delivery conditions, such as access bandwidth and device capabilities.

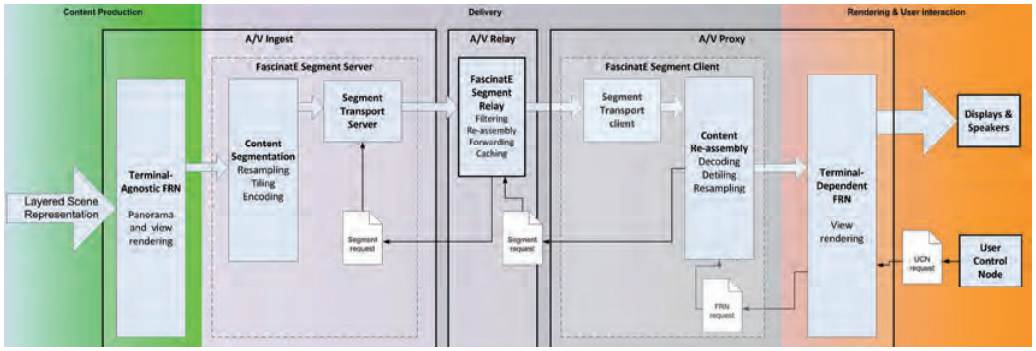

Figure 2 – FascinatE delivery network architecture.

- A/V Relay: in between these two network demarcation points, the transport of A/V segments needs to act as an end-to-end filter that accommodates the network capabilities as well as the aggregated requests of the deployed A/V proxies. This can be ensured by intermediate transport nodes, that can aggregate, cache and/or relay segment requests at the transport protocol control level, and also serve as demarcation points between delivery modes for the downstream A/V flows.

Figure 2 shows a detailed version of the FascinatE delivery network architecture, including the aforementioned functional components and the neighbouring content production and user interaction domains.

2.3 Delivery Mechanisms

The actual transport of A/V segments takes place between Segment Transport Servers and Clients. Two specific delivery mechanisms are developed between the A/V Ingest and the A/V Proxy, catering for different usage scenarios and network deployments. On the one hand, the tiled HTTP Adaptive Streaming (HAS) mechanism is suitable for web-based over-the-top delivery, in e.g. CDN or cloud video deployments. This mechanism is described in section 4. On the other hand, the PUB/SUB mechanism fits the requirements of managed delivery networks such as IPTV over xDSL and cable. This mechanism is described in section 5.

3 CONTENT SEGMENTATION

A key realisation in developing the delivery network architecture is the fact that inside the delivery network, i.e. between A/V ingest and A/V proxy, the adaptive delivery of parts of the content based on the viewing behaviour of the client (or the user) can be supported by spatially segmenting the A/V data into tiles that relate to a specific spatial region of a video frame. In most cases, tiles are grouped for a certain time period, in which case they are called segments. The particular grouping can be dependent on the transport protocol used, but globally, the FascinatE delivery network is aimed at delivery of tiled and segmented content. Regular H.264 video coding can be employed. Trade-offs in the encoding of spatially segmented content are reported in [9].

Content segmentation is required to recast the LSR content into segments that are suitable for network encapsulation and further transport functions. The general concept behind spatial segmentation is to spatially partition each video frame into rectangular pieces called tiles. All frames representing a single area of the video are taken together, encoded and stored as a new independent video stream, or spatial segment. The result is a large number of video files, each representing a specific area of the original video file. Encoding each spatial segment as an independent video stream allows an A/V Proxy to only request a subset of segments, based on the ROI selected by the user for which it performs the spatial recombination. Upon reception of the individual spatial segments, the A/V Proxy can then recombine them with a content reassembling operation and pass the result to the end-user device. In certain cases, a user may also want to see an overview of the entire video. In order to do so, the A/V Proxy would need to receive all spatial segments, resulting in enormous bandwidth requirements. Furthermore, it would need to downscale the resulting video, e.g. in order to be able to present it on a small smartphone display. To solve these inefficiencies we create multiple resolution scales. Each scale is a collection of spatial segments that together encompasses the entire video. However, each scale does so in a different resolution. For example, the top scale might consist of 144 (12x12 tiling) segments, each of resolution 640x360, together encompassing the original OMNICAM resolution of 7680x4320, while the bottom scale might only consist of 4 segments, with a combined resolution of 1280x720. It is even possible to create a scale consisting of only a single segment, with a 640x360 resolution, still showing the entire video but at a much lower quality.

The resulting spatial segmentation system provides an efficient method -in terms of required bandwidth- for receiving parts of an ultra-high resolution video. By only having to receive those areas of a video in which a user is interested, combined with support for a wide variety of display sizes and resolutions, it is possible to exploit next-generation ultra-high resolution camera systems with current generation delivery networks. Note that the combined usage of spatial segments and multiple resolution scales may lead to a significant number of video files.

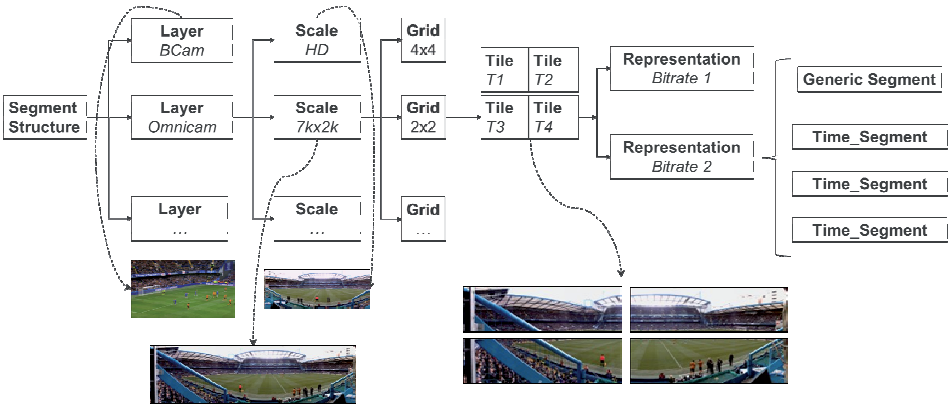

Figure 3 – Content segmentation hierarchy.

Figure 3 shows the hierarchy of segmented content and some example segments. This shows that, starting from an LSR, video content is segmented at the following levels:

- For every layer in the LSR, one or more resolution scales are created;
- For every scale, one or more MxN tiling grids are created, leading to a set of tiled video streams per scale;
- For every tile, one or more representations at different quality settings are created;
- For every representation, the associated tiled video stream can be temporally segmented.

4 WEB-BASED DELIVERY

Web-based or over-the-top delivery refers to open Internet video transport as provisioned by overlay networks such as a Content Delivery Network (CDN) or a cloud video platform. HTTP adaptive streaming (HAS) is emerging as a popular transport protocol for video streaming format. For the interactive online video services considered in FascinatE, we aim at extending HAS with spatial segmentation, so that interaction with high resolution video can be supported by a CDN provider operator, without the need for a complete overhaul of its current CDN deployment. In particular, the prototype described in section 4.3 supports a companion screen scenario, where the final rendering stage is performed on e.g. a Connected TV, whereas the interactive control is done on a thin client, e.g. a tablet device. The developed prototype allows one to freely navigate into the 7K x 2K OMNICAM video using a second screen, as shown in Figure 4.

4.1 HTTP Adaptive Streaming

HTTP adaptive streaming has recently emerged as a standard for video delivery over best-effort networks. HAS enables the delivery of (live) video by means of the HTTP protocol, by providing the video in segments that are independently requested by the client from a web server. A video is temporally split in several video segments, which in itself are standalone video files. These segment files can be delivered separately. When recombined they provide a seamless video stream. A video can be provided in several representations: alternative versions of the same content that differ in resolution, the number of audio channels and/or different bitrate. All representations are temporally aligned such that segments of different representations can be interchanged. An ISO/IEC HAS standard has recently been created by MPEG and is referred to as DASH (Dynamic Adaptive Streaming over HTTP [10]).

Figure 4 - Prototype allowing for second-screen pan/tilt/zoom interaction with the panoramic video.

4.2 HAS and Tiled Streaming

The aforementioned HAS solutions focus on temporal-segmentation. HTTP adaptive streaming can however also be combined with the spatial content segmentation procedure described in section 3. Each video tile is individually encoded and temporally segmented according to common HAS solutions. An advantage of using HAS for the delivery of spatial tiles is that the inherent time-segmentation makes it relatively easy to resynchronize different spatial tiles. That is, all HAS tiles are temporally aligned such that segments from different tiles can be easily recombined to create the reassembled picture. As long as the time segmentation process makes

sure that time-segments between different spatial tiles have exactly the same length, the relative position of a frame within a time segment can be used as a measure for the position of that frame within the overall timeline.

In HAS solutions such as MPEG-DASH, a *manifest file* is used to describe the structure of the segmented content. This manifest is referred to as a Media Presentation Description (MPD). The MPD includes all information that a HTTP client needs to retrieve the media segments corresponding to a media session, such as the Media Presentation, alternative representations of the media, specific groupings of media and segment and media information, e.g. segment length, resolution, audio and video codecs and the container format. The MPD as defined in MPEG DASH can be readily extended with resolution scale and spatial tiling information.

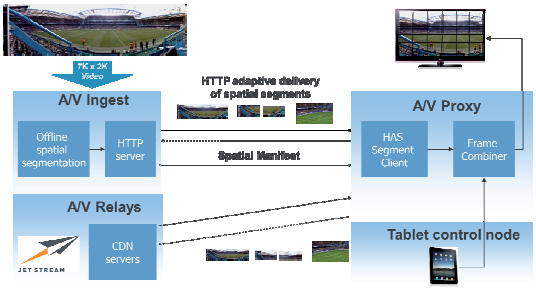

Figure 5 – Tiled HAS proof of concept.

4.3 Proof of Concept

A proof of concept prototype of the tiled HAS delivery mechanism has been developed as part of the overall FascinatE delivery network testbed. This proof of concept is illustrated in Figure 5. It supports HAS-based delivery of H.264 encoded content segments that allows the system to be deployed on current CDNs and cloud infrastructures. This was achieved by developing an integrated set functions for the A/V Ingest and A/V Proxy components based on the tiled HAS mechanism, and by using an actual CDN as the A/V Relay. The main functions at the A/V Ingest, Relay and Proxy are the following:

- At the A/V ingest: the main component is the tiled HAS server that hosts the segmented LSR content.
- At the A/V relay: the main component is a live CDN delivery server, for scalable and distributed delivery of segments.
- At the A/V proxy: the main components are the tiled HAS segment client which requests the segments, and the frame combiner which performs the content reassembly function and adapts the reassembled view to the target device.

Further functionalities incorporated in the prototype are trick play, picture-in-picture through multi-ROI rendering and predetermined ROI selection. Also, additional layers from the LSR, e.g. from broadcast cameras can be made available on the companion screen.

5 MANAGED DELIVERY

Managed delivery refers to a video transport platform that is fully under control of a service provider. This includes controlling how content is represented and transported over the complete end-to-end delivery path, from ingest till client device, as well as some policies regarding the management of network resources and performances. A typical example in managed IPTV services is the allocation of linear TV channels to provisioned multicast trees. For the interactive TV services tackled in FascinatE, we aim at extending the optimized transport approach of managed delivery, so that low-delay interaction and high-quality video can be supported by a network operator at a reasonable cost in terms of resources consumed.

5.1 Publisher/Subscriber mechanism

One of the main challenges for the FascinatE delivery network is that the requirements on the type of transport technology seem contradictory depending on which end of the network one looks at:

- At the delivery network ingress where the whole LSR is ingested, the network elements are responsible for pushing the content through the network, agnostic to the actual user requests.
- At the terminal side, user-specific portions of the layered scene may be requested. Therefore, if the capabilities of the end-to-end network and of the terminal cannot support a plain transmission of the full LSR, the terminal has to send some requests upstream to pull those parts of the LSR which are required for rendering.

To cover these two requirements, we propose to use a message-queue mechanism, which specifies “publisher” and “subscriber” functions that can work asynchronously at each end of the network. This approach fits well for a deployment in a managed network, such as next generation IPTV systems, that would be required to support a large number of end-devices with various bandwidth and processing capabilities. In this Publisher/Subscriber (PUB/SUB) mechanism, the tiled-based representation naturally leads to assigning each publisher to a given spatial tile. The published data is transported over a combination of unicast and multicast channels, organized according to the multi-resolution hierarchy of spatial segments described in section 3.

In addition to a better control of the transport channels, a managed network context also opens the possibility to put more processing functions into the network. In particular, our managed solution supports an end-to-end scenario, where the final rendering stage is also performed in the network, so as to support thin clients. In this case, the thin client does not need to directly subscribe to the segmented data, but directly receives a pre-rendered video stream. This requires the network to include in the A/V Proxy rendering functions are responsible for making the received segments ready for delivery to the end-device. Such a Video Proxy prototype (shown in Figure 6) has been developed so as to allow any thin client device to freely navigate into the 7K x 2K OMNICAM videos.

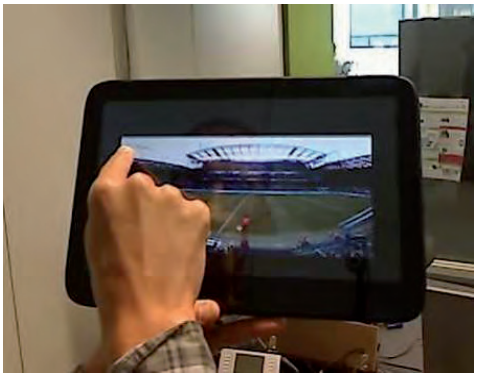

Figure 6 - Continuous interactive video rendering on a Thin Tablet Client.

The end-device only has to send its pan-tilt-zoom navigation commands (e.g. from a touch-based user interface) to the proxy and receives back the requested sequence of views, fully pre-rendered by the network and delivered at a resolution and bandwidth that match the device capabilities. With this approach, high-resolution video content can be watched interactively in a natural manner, even on a low-power and small-display device.

5.2 Proof of Concept

A proof of concept prototype of the PUB/SUB delivery mechanism has been developed as part of the overall FascinatE delivery network testbed. The current set-up is illustrated in Figure 7. It supports live in-network rendering (A/V Proxy) that can serve multiple Video Thin Clients. Two rendering mode are supported. A 2D mode consists in continuously reframing and rescaling the panorama according to the stream of user navigation commands. A 3D mode, relying on GPU acceleration, compensates in addition the geometrical distortion of the cylindrical representation of the OMNICAM content.

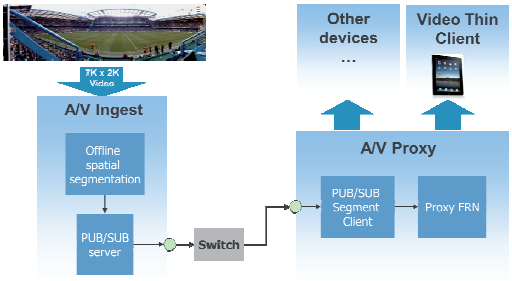

Figure 7 – PUB/Sub proof of concept.

The role of the PUB/SUB mechanism is to connect the full hierarchy of tiled content at the A/V Ingest till the A/V Proxy where only a subset of this content is required on behalf of end clients. The delivery of video tiles is optimized so as to guarantee that each A/V Proxy receives the required subset of LSR data and minimize the bandwidth usage between ingest and proxy. This work is based on our previous work [11] where this joint optimization of video coding and tile selection was studied.

6 FUTURE WORK

In the remainder of the FascinatE project, the delivery network architecture and proof of concept implementations will evolve so as to support live delivery of ultra-high and interactive video services, in manner that can scale to many client with heterogeneous access bandwidth and end-device processing power. Further studies will evaluate the scalability performance of the proposed approaches. Additionally, other FascinatE components will be supported and incorporated, such as scripted view rendering, interactive audio rendering and gesture-based interaction.

Acknowledgment

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. Thanks to SIS Live and the Premier League for their assistance and permission to use the football images. The authors would like to thank the project partners for the insightful discussions and collaboration. Also thanks to CDN provider Jet-Stream for making available their cutting-edge CDN testing environment.

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] R. Schäfer, P. Kauff, and C. Weissig. "Ultra-high resolution video production and display as basis of a format agnostic production system", IBC2010 Proceedings, Amsterdam, Netherlands, September 2010.

[3] NEM Strategic Research Agenda - Position Paper on Future Research Directions, 2nd edition, September 2011.

[4] O. Schreer, G. Thomas, O.A. Niamut, J-F. Macq, A. Kochale, J-M. Batke, J. Ruiz Hidalgo, R. Oldfield, B. Shirley, G. Thallinger. "Format-agnostic Approach for Production, Delivery and Rendering of Immersive Media", NEM Summit 2011, Torino, Italy, 27th September, 2011.

[5] G.A. Thomas, O. Schreer, B. Shirley, J. Spille. "Combining panoramic image and 3D audio capture with conventional coverage for immersive and interactive content production", IBC 2011, Amsterdam, The Netherlands, 11th September, 2011.

[6] P. Grosso, L. Herr, N. Ohta, P. Hearty and C. de Laat. "Super high definition media over optical networks", Future Generation Computer Systems, Volume 27, Issue 7, Pages 881-990, July 2011.

[7] KDDI R&D Labs, Three Screen Service Platform. http://www.youtube.com/watch?v=urjQjR5VK_Q. Visited: May 10th, 2012.

[8] Mavlankar, A., "Peer-to-Peer Video Streaming with Interactive Regionof- Interest", Ph.D. Dissertation, Department of Electrical Engineering Stanford University, April 2010

[9] Khiem, N., Ravindra, G., Carlier, A., and Ooi., W. 2010. Supporting zoomable video streams with dynamic region-of-interest cropping. In Proceedings of the first annual ACM SIGMM conference on Multimedia systems (MMSys '10). ACM, New York, NY, USA, 259-270.

[10] T. Stockhammer, "Dynamic Adaptive Streaming over HTTP - Standards and Design Principles", MMSys'11, February 23–25, 2011, San Jose, California, USA.

[11] P. R. Alface, J.-F. Macq, and N. Verzijp, "Interactive Omnidirectional Video Delivery: A Bandwidth-Effective Approach", Bell Labs Technical Journal 16(4): 135-147, 2012.