BBC

Research & Development

White Paper

WHP 284

September 2014

Immersive Live Event Experiences - Interactive UHDTV on Mobile Devices

O.A. Niamut[1], G.A. Thomas[2], E. Thomas[1], R. van Brandenburg[1], L. D'Acunto[1], R. Gregory-Clarke[2]

[1]TNO, NL; [2]BBC R&D, UK

BRITISH BROADCASTING CORPORATION

White Paper WHP 284

Immersive Live Event Experiences – Interactive UHDTV on Mobile Devices

O.A. Niamut[1], G.A. Thomas[2], E. Thomas[1], R. van Brandenburg[1], L. D'Acunto[1], R. Gregory-Clarke[2]

[1]TNO, NL; [2]BBC R&D, UK

Abstract

This paper reports on the latest developments around tiled streaming. As an extension of HTTP adaptive streaming, it retains all the benefits of this streaming technology, while adding the possibility of interaction when consuming UHDTV on mobile devices. In particular, we discuss the underlying principles and aspects, such as multi-layer resolution scaling, spatial segmentation and adaptive streaming. Then we present insights from a number of technology validation tests and demonstrations, such as a live dance performance in Manchester, May 2013; a training tool for professional skiers employing tiled streaming in Schladming, host of the Alpine Skiing World Championship 2013; and tests incorporating 'augmented reality'-style overlays in an athletics stadium in preparation for a trial at the 2014 Commonwealth Games. Finally, we report on the status of ongoing standardization efforts in the MPEG-DASH ad-hoc group, where tiled streaming is considered as a new feature, referred to as Spatial Relationship Description.

This document was originally published at IBC 2014, Amsterdam, September 2014, and in *The Best of IET and IBC, 2014, Vol. 6, pp. 38-43.*

Additional key words: Venue Explorer, graphical overlays, FascinatE project

White Papers are distributed freely on request.

Authorisation of the Chief Scientist or General Manager is required for publication.

White Paper WHP 284

Immersive Live Event Experiences – Interactive UHDTV on Mobile Devices

O.A. Niamut[1], G.A. Thomas[2], E. Thomas[1], R. van Brandenburg[1], L. D'Acunto[1], R. Gregory-Clarke[2]

[1]TNO, NL; [2]BBC R&D, UK

1 Introduction

The momentum behind UHDTV is quickly gathering and UHDTV displays are entering the market at an increasing pace. And whereas the production of UHD content is lagging behind, online content providers such as Netflix, Amazon and YouTube have announced their plans for releasing series and films in initial UHD formats such as 4K. With UHD format recommendations and requirements emerging from ITU [1] and DVB[1], with the specification of a 4K Blu-ray format[2], and with several live 4K trials over existing broadcast[3] and CDN infrastructure[4], the future for UHD, starting with 4K, looks bright. For mobile devices, the expectations for and benefits of UHD formats are less clear. That is, 4K UHD tablets and smartphones featuring limited screen sizes make for less-then-ideal candidates for displaying e.g. 4K UHD video. And with an 8K UHDTV standard emerging, the discrepancy between native content resolution and screen rendering resolution remains. While it is possible to enable regular UHDTV experiences on tablets and smartphones, this experience can be enriched by (i) allowing end-users to freely extract a region-of-interest and navigate around the ultra-high resolution video, and (ii) add scalable 'augmented reality'-style overlays to the video. Such an approach requires efficient delivery and media-aware network-based processing in order to support mobile terminals and bandwidth limitations in the access networks.

An emerging technology, referred to as tiled streaming, enables such interaction with streaming video, in such a way that end-users can enjoy the full UHD resolution, even if their device is not capable of rendering and displaying the video in its entirety. As an extension of HTTP adaptive streaming, it retains all the benefits of this streaming technology, while adding the possibility of interaction when consuming UHDTV on mobile devices. This paper reports on the latest developments around tiled streaming. In particular, we (i) discuss the underlying principles and aspects, such as multi-layer resolution scaling, spatial segmentation and adaptive streaming, and (ii) present insights from a number of technology validation tests and demonstrations, such as a live dance performance in Manchester, May 2013; a training tool for professional skiers employing tiled streaming in Schladming, host of the Alpine Skiing World Championship 2013; and tests incorporating 'augmented reality'-style overlays in an athletics stadium in preparation for a trial at the 2014 Commonwealth Games. We further report on the status of ongoing standardization efforts in the MPEG-DASH ad-hoc group, where tiled streaming is considered as a new feature, referred to as Spatial Relationship Description.

[1] https://www.dvb.org/news/uhdtv--new-evidence-and-new-questions-for-dvb
[2] http://www.hollywoodreporter.com/behind-screen/ces-as-ultra-hd-train-669587
[3] http://www.broadbandtvnews.com/2014/05/27/4k-smash-for-french-open/
[4] http://www.iptv-news.com/2014/05/vienna-state-opera-streams-in-4k-with-elemental/

2 Related work

With recent capturing systems for panoramic and omnidirectional UHD video, new types of media experiences are possible where end-users have the possibility to freely choose their viewing direction and zooming level. Many different examples of such interactive video delivery have been demonstrated or deployed. In the entertainment sector, companies like Immersive Media[5] and Mativision[6] offer web streaming and mobile app solutions to cover events with a 360-degree video camera. However, such solutions rely on streaming or downloading a complete spherical panorama to end-user devices, where the final rendering of the interactive viewport takes place. In an alternative approach, KDDI[7] has shown a solution where all rendering takes place on the server side. Here, a low-powered and low-resolution mobile phone sends a spatial request to the server, requiring the server to reframe and rescale the content accordingly before compression and streaming to the end-user device. Tiled streaming has emerged as a scalable and bandwidth-efficient approach to interactive UHD. Initially proposed by Mavlankar [2] and further developed in [3-5], interactive UHD is enabled by a multi-layer tiling approach where the video is split into multiple independently-encoded tiles which are stitched on the end-user device.

In addition to interactive video delivery, panoramic and UHD video also provide a good opportunity for adding hotspots and overlays. That is, due to the static nature of the background, additional and interactive overlay graphics can be positioned in the world reference frame rather than having to account for camera motion. Such hotspots were initially incorporated into static panoramic systems, such as Quicktime VR, introduced in 1994. More recently, systems using video panoramas have been developed for use in sports broadcasting, where real-time data relating to player tracking can be overlaid on the scene prior to selecting a window for broadcast[8]. Approaches for adding interactive overlay graphics to conventional web video at the client side are starting to appear[9], but these are not yet generally being applied to interactive delivery of panoramic and UHD video.

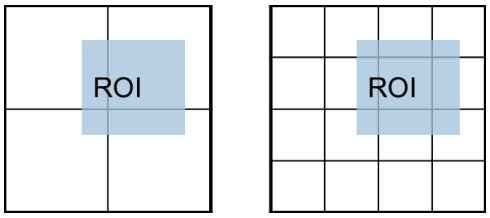

Figure 1: Example of 2x2 (left) and 4x4 (right) tiling grids. Depending on the tile size, a single ROI overlaps multiple tiles, and thus requires multiple tiles for reconstruction.

3 Tiled streaming for zoomable video

Zoomable video allows users to selectively zoom and pan into regions of interest (ROI) within the video. Such interaction typically requires dynamic cropping of ROIs in the source video, as well as unicast streaming of the cropped ROIs. The concept of tiled streaming addresses the limitations of today's networks when streaming high resolution content, as well as allowing new applications in video streaming such as interactive panning and zooming. Tiled video provides a better user experience than with predefined or dynamically cropped regions of interest. Moreover, it provides better image quality for the selected region than by simply enlarging pixel dimensions. Finally, the adaptive version of tiled streaming offers a new adaptation dimension after bandwidth, resolution and quality, i.e. the possibility at a given bandwidth to choose between full-frame video in low quality and a spatial area in higher quality.

[5] http://immersivemedia.com/
[6] http://www.mativision.com/
[7] http://www.engadget.com/2010/11/29/kddi-develops-a-zoom-enhance-system-for-hd-movie-streaming-on-sm/
[8] http://www.stats.com/pdfs/SportVU_SonyDAV.pdf`
[9] https://popcorn.webmaker.org/

3.1 Basic Concepts of Tiling

A tiled video can be obtained from a single video file or stream in by partitioning each individual video frame into independently-encoded videos. Tiles are thus defined as a spatial segmentation of the video content into a regular grid of independent videos that can each be encoded and decoded separately. We denote the tiling scheme by *MxN* where *M* is the number of columns and *N* is the number of rows of a regular grid of tiles. See Figure 1 for two examples of regular tiling grids. The tiling and subsequent separate encoding of tiled videos leads to a reduced compression performance, due to reduced exploitation of spatial correlation in the original video frame being limited to tile boundaries. This compression performance loss can be reduced by using multiple resolution layers. Each additional layer originates from a lower resolution version of the original video frame, tiled into a grid with fewer tiles. If the tiling is small enough (such as thumbnails), the bitrate overhead of using another resolution layer is affordable. This multi-resolution tiling increases the quality of user-defined zooming factors on tiles. Once a user zooms into a region of the content, the system will provide the highest resolution tiles that are included in the requested region.

3.2 Optimizing Performance on Low-Powered Devices With Overlapping Tiling

Figure 2 provides an example of an overlapping tiled grid, as used in [6]. Such a tiling is beneficial for mobile devices that are equipped with a single hardware decoder only. In this case, the total number of tiles required to reconstruct the requested ROI can be reduced to one, while the ROI size and total number of tiles remain approximately the same. The effectiveness of overlapping tiles in reducing the number of tiles required to reconstruct a given ROI is determined by the overlapping factor, which gives the relative overlap (per planar direction) of a particular tile in relation to its size. Choosing a larger overlapping factor results in larger overlapping areas, and thus in fewer tiles being required to reconstruct a given ROI. The downside of this overlap is that these redundant pixels result in a larger amount of data to be stored on the server side. Also, overlapping tiles result in heterogeneous tile sizes.

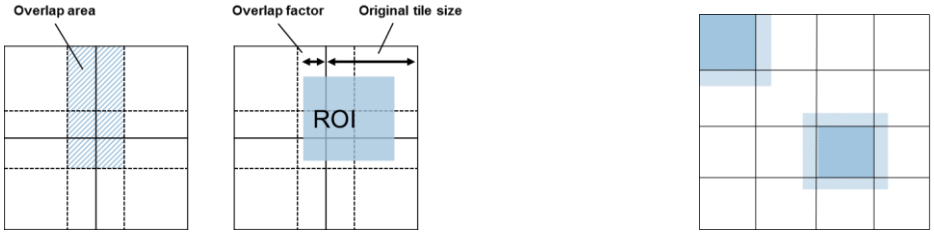

Figure 2: (left) Tiling scheme with overlapping tiles (2x2 grid) and an overlap factor of one third. (right) Overlapped tiling scheme with an overlap factor of one fourth.

3.3 Tiled Adaptive Streaming

As shown in **Error! Reference source not found.**, spatial segmentation can be complemented with the temporal segmentation of HTTP Adaptive Streaming (HAS). The scalability properties of HAS enable zoomable video to be available to a large number of users thanks to efficient bandwidth utilisation, cacheability and simpler inter-tile synchronization. Tiled streaming can be integrated within HAS by having each video tile individually encoded and then temporally segmented according to any of the common HAS solutions (e.g. MPEG DASH [7] or Apple HLS[10]). This leads to a form of tiled adaptive streaming, where all tiles are temporally aligned such that segments from different tiles can be recombined to create the reassembled picture. An advantage of using HAS for the delivery of spatial tiles is that the inherent time-segmentation makes it relatively easy to resynchronise different spatial tiles when recombining tiles into a single picture or frame. 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. For example,

[10] HTTP Live Streaming, see http://tools.ietf.org/html/draft-pantos-http-live-streaming-13

frame number *n* within time segment *s* of tile *A* is synchronised with frame number *n* within time segment *s* of tile *B*. On the client side, timestamps provided by the segment container can be used to ensure perfect synchronisation between the segments that make up the final viewport to be rendered on the screen of the end user.

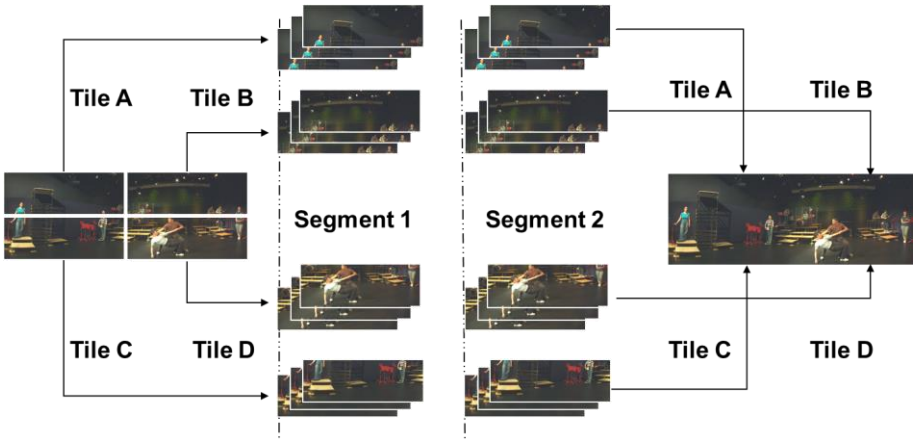

Figure 3: With tiled HAS, a video is tiled in a certain grid. Each tile is encoded and segmented using HAS segments. In this example, the grid is 2 by 2.

4 Applications of tiled streaming technology

In this section we present insights from a number of recent and planned technology validation tests and demonstrations, such as (i) the FascinatE system, used during a live dance performance in Manchester, May 2013; (ii) the iCaCoT system, used as a training tool for professional skiers in Schladming, February to April 2014; and (iii) preparations for tests including 'augmented reality'-style overlays in an athletics stadium for the 2014 Commonwealth Games.

4.1 FascinatE System

Tiled streaming was employed for the distribution of panoramic video sequences in the context of the European-funded project FascinatE[11]. A live system was demonstrated during a dance performance in Manchester, May 2013, see Figure 4 and 5. Here, a multi-camera audio-visual scene representation, including both panoramic 6K and regular 1080p video content, was spatially tiled and temporally segmented. For tiling, a dyadic tiling approach was used, ranging from 1x1 to 8x8 tiling grids, resulting in a multi-resolution set of panoramic video layers, with the original resolution of 6976 by 1920 pixels as a base layer.

Figure 4: Panoramic 6K image of the FascinatE live demonstration, based around the performance of 'Deeper than all roses', a composition from Stephen Davismoon, featuring rock band Bears?Bears! and live performance artists Joseph Lau and Shona Roberts.

[11] Format-Agnostic SCript-based INterAcTive Experience, see http://www.fascinate-project.eu/

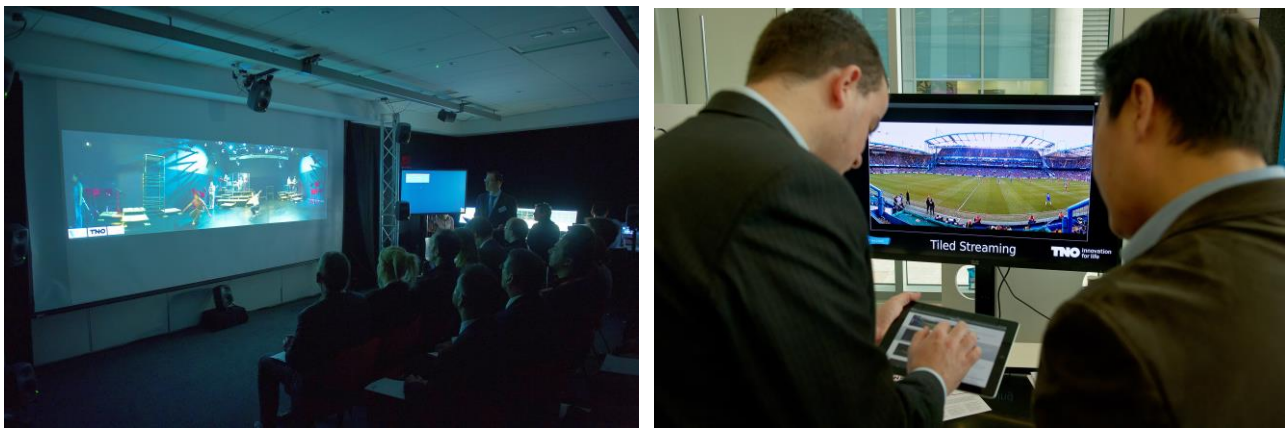

Figure 5: The live performance was shown in a separate room, delivered via tiled streaming. The presenter could navigate through video panorama using a tablet. Additional footage from earlier recordings was also shown on tablets.

4.2 iCaCoT System

Tiled streaming was further incorporated into the iCaCoT[12] training application, in the context of the European-funded project EXPERIMEDIA[13]. Here, the goal was to provide ski coaches in Schladming, host of the Alpine Skiing World Championship 2013, with a tablet application through which they could provide their trainees with real-time feedback. This was achieved using a combination of a set of static high-resolution cameras along the ski-slope, tiled streaming and advanced trick play and drawing features, see Figure 6. In collaboration with Schladming2030, the Austrian venue partner, we performed several experiments with a live 4K tiled streaming system. A significant challenge was to cope with the particular conditions of the experiment location, as the material had to be protected from extreme weather conditions in terms of temperature, humidity, and so on. As the on-site connectivity to the open Internet ruled out any off-site video processing such as tiling and encoding, the overall system had to be stand-alone and compact.

For tiling of 4K video at 24 fps in real-time, a software-based approach was not a viable solution. Therefore, a hardware-accelerated pipeline based on the Intel Media SDK[14] was used. This pipeline, running on a single machine, is able to decode the original panorama, to produce the tiles and to independently encode them in H.264 at 24 fps. In practice, the pipeline can handle more than 10 tiles at roughly 1920x1080 resolution in parallel.

Figure 6: Interactive camera-based coaching and training using tiled streaming for video navigation. Several application screenshots, with the third screenshot showing user-drawn markers helping to provide the skier useful feedback.

[12] Interactive CAmera-based COaching and Training, see http://www.experimedia.eu/2014/02/20/icacot/
[13] Experiments in live social and networked media experiences, see http://www.experimedia.eu/
[14] Intel Media Software Development Kit, see http://software.intel.com/en-us/vcsource/tools/media-sdk-clients

4.3 Commonwealth Games 2014 system

These experiments will aim at verifying that navigation around a high resolution video using tiled streaming will contribute to a higher sense of interaction and engagement amongst users, particularly in the case of large-scale, event-based programming. This form of content often contains several distinct regions of interest, and a number of different things that a user may want to look at in more detail. Hence, we hypothesise that the user experience could be further enhanced through the use of overlaid, interactive graphics which provide extra information about the scene.

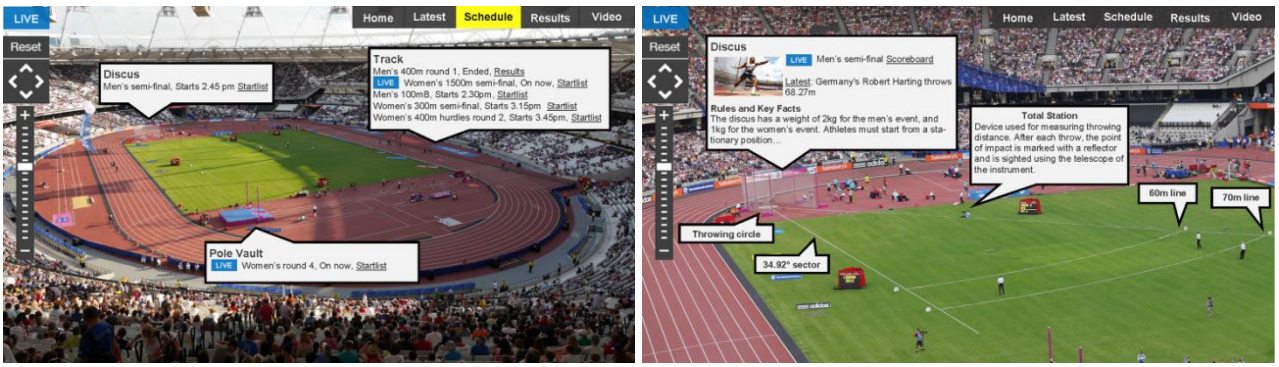

Figure 7: A possible user interface for the Venue Explorer system

An example of such an application is an athletics event, which typically features several different track and field events occurring simultaneously. As well as being able to pan and zoom around the scene as they wish, the user can also be presented with data which offers more detail about what they are looking at. This could include the locations and times of the sports taking place on the day in question, the names of the athletes that are visible, the current height of the high jump bar, and so forth. The intention is to provide the type of rich data that a user would typically be interested in anyway, but that they would ordinarily have found either from burnt-in graphics provided by a broadcaster, or else a self-initiated search. Presenting this information as optional overlays in this way allows the user to access the level of detail they want, when they want it, without having to leave the application. We plan to test interactive overlays on panoramic video as part of a closed trial of a prototype system, known as the Venue Explorer, at the Commonwealth Games in Glasgow in July-August 2014. The plan is to capture a wide-angle view of the athletics stadium using a 4k camera, encode this as a set of overlapping tiles, and stream these to an HTML5-based client using MPEG-DASH. Data relating to sports events (including live updates) will be used to render interactive overlays in the client, according to options selected by the viewer. Initial work on the system has been using video captured at the London Anniversary Games in summer 2013, as shown in Figure 7.

5 Standardisation of tiled streaming

MPEG-DASH (ISO/IEC 23009) is the adaptive streaming technology as standardised by MPEG. After having published a first edition of the standard [7], MPEG experts are now aiming at extending the original scope of adaptive streaming to new use cases. Tiled streaming use cases and their relevancy were presented to MPEG-DASH working group during the 104th MPEG meeting in April 2013. There, MPEG-DASH experts acknowledged the usefulness of such use cases and agreed on starting a so-called Core Experiment. The goal of this Core Experiment was to steer the group effort towards a technical solution enabling the tiled streaming use cases. The discussions within the Core Experiment reached a consensus among the group at the 107th MPEG meeting last January. Consequently, MPEG has initiated the publication process of this new feature which should end in the course of 2015.

Conceptually this new feature, called Spatial Relationship Description (SRD), allows an author of the MPEG-DASH media presentation description (MPD) to describe how the various tiles are spatially related with each other. This description handles both intra and inter layer relationships. Thus far, the standard only allowed for an *AdaptationSet* to define perceptually-equivalent content. Therefore, describing different tiles under the same *AdaptationSet* would violate this rule. With the

SRD feature, the concept of a tile, as described so far in this paper, is mapped onto the *AdaptationSet* element of the MPD. It is also important to note that this new mapping decouples the tile concept from a particular video. To offer backwards compatibility, it remains possible to benefit from the regular properties of *AdapationSets*, namely the availability of several *Representations* in different bitrates, codecs, resolutions, and so on. In practice, the new MPEG-DASH SRD feature will specify a set of parameters in order to describe tiles with respect to a common reference space. These parameters are (*x,y*), respectively horizontal and vertical positions of the tile, (*w,h*). respectively width and height of the tile, and (*W,H*), respectively width and height of the reference space. All these values are expressed in an arbitrary unit as chosen by the MPD author.

6 Future work

In subsequent developments, we aim to improve the live tiling process, such that it can handle a range of UHD formats, including higher resolutions and frame rates. Incorporation of multi-sensor systems will also be considered. In the near future, we aim to produce live 4K footage and deliver this via tiled streaming to end-users in a large-scale user trial. This allows us to determine the impact of live 4K tiled streaming on a regular production environment and to measure the effects of tiled streaming on bandwidth and latency in a regular content delivery setting. Furthermore, the ways in which users interact with panoramic UHD video and overlays will be studied in a series of user trials, which will inform future developments. It is also planned to evaluate the use of interactive panoramic video with overlays in other application scenarios.

7 References

1. ITU-R Recommendation BT.2020, “Parameter values for ultra-high definition television systems for production and international programme exchange”. August 2012.
2. Mavlankar A., P.Agrawal, D.Pang, S. Halawa, N-M Cheung, B.Girod. “An Interactive Region-Of-Interest Video Streaming System For Online Lecture Viewing”. Special Session on Advanced Interactive Multimedia Streaming, Proc. of 18th International Packet Video Workshop (PV). December 2010, Hong Kong.
3. Khiem N., G. Ravindra, A. Carlier and W.T Ooi, “Supporting Zoomable Video Streams via Dynamic Region-of-Interest Cropping”, in Proc. of 1st ACM Multimedia Systems Conf. pp. 259-270, 22-23 February 2010, Scottsdale, Arizona.
4. Brandenburg R. van, Niamut O., Prins M., Stokking H., “Spatial Segmentation For Immersive Media Delivery”. In Proc. of 15th Int. Conf. on Intelligence in Next Generation Networks (ICIN). pp. 151-156, 4-7 October 2011, Berlin, Germany.
5. Quax P., P.Issaris, W.Vanmontfort, W.Lamotte. “Evaluation of Distribution of Panoramic Video Sequences in the eXplorative Television Project”. NOSSDAV’12, 7-8 June 2012, Toronto, Canada.
6. Pang D., S. Halawa, N-M Cheung, B.Girod. “ClassX Mobile: Region-of-Interest Video Streaming to Mobile Devices with Multi-Touch Interaction”. In Proc. of MM’11, November 28–December 1 2011, Scottsdale, Arizona, USA.
7. MPEG, “ISO/IEC DIS 23009-1 Information Technology -- Dynamic adaptive streaming over HTTP (DASH) -- Part 1: Media presentation description and segment formats”, January 5, 2012.

8 ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Union’s Seventh Framework Programme (FP7/2007-2013) under grant agreements no. 248138 and no. 287966.