MPEG DASH SRD - Spatial Relationship Description

Omar A. Niamut
TNO
Anna van Buerenplein 1
2595 DA, Den Haag, Netherlands
omar.niamut@tno.nl

Emmanuel Thomas
TNO
Anna van Buerenplein 1
2595 DA, Den Haag, Netherlands
emmanuel.thomas@tno.nl

Lucia D'Acunto
TNO
Anna van Buerenplein 1
2595 DA, Den Haag, Netherlands
lucia.dacunto@tno.nl

Cyril Concolato
LTCI, CNRS, Télécom ParisTech
Université Paris-Saclay
75013, Paris, France
concolato@telecom-paristech.fr

Franck Denoual
CANON RESEARCH CENTRE
Rue de la Touche Lambert
35510 Cesson-Sévigné, France
franck.denoual@crf.canon.fr

Seong Yong Lim
ETRI
218 Gajeong-ro, Yuseong-gu,
Daejeon, 34129, KOREA
seoji@etri.re.kr

ABSTRACT
This paper presents the Spatial Representation Description (SRD) feature of the second amendment of MPEG DASH standard part 1, 23009-1:2014 [1]. SRD is an approach for streaming only spatial sub-parts of a video to display devices, in combination with the form of adaptive multi-rate streaming that is intrinsically supported by MPEG DASH. The SRD feature extends the Media Presentation Description (MPD) of MPEG DASH by describing spatial relationships between associated pieces of video content. This enables the DASH client to select and retrieve only those video streams at those resolutions that are relevant to the user experience. The paper describes the design principles behind SRD, the different possibilities it enables and examples of how SRD was used in different experiments on interactive streaming of ultra-high resolution video.

CCS Concepts
• Information systems→Multimedia streaming

Keywords
"video"; "mobile video"; "ultra-high definition"; "streaming"; "tiled streaming"; "standards"; "MPEG-DASH"

1. INTRODUCTION
The advent of ultra-high resolution video (e.g. 4K, 8K and beyond) in combination with an increasing heterogeneity of display devices (e.g. UHD TV sets, tablet devices, smartphones, smartwatches) introduces opportunities for new usages in video streaming; for instance, interactive pan and zoom features. However, the streaming of high resolution video over today's networks raises problems due to bandwidth restrictions in the access and home networks. Also, the video decoders of mobile display devices may be unable to handle ultra-high resolutions, given their often limited hardware capabilities. A solution to both issues is found in streaming only spatial sub-parts of a video to the display device, in combination with the form of adaptive multi-rate streaming that is intrinsically supported by MPEG DASH. This solution, often referred to as tiled streaming, enables streaming of parts of a video in its native resolution, with lower bandwidth requirements on access and home networks. It has been incorporated in the MPEG DASH standard, as the Spatial Representation Description (SRD) feature of the second amendment of MPEG DASH standard part 1, 23009-1:2014 [1]. The feature extends the Media Presentation Description (MPD) of MPEG DASH by describing spatial relationships between associated pieces of video content. This enables the DASH client to select and retrieve only those video streams at those resolutions that are relevant to the user experience.

This paper presents the design principles behind SRD, the different possibilities it enables and examples of how SRD was used in different experiments on interactive streaming of ultra-high resolution video. The paper is structured as follows. In section 2, we consider related work on zoomable video systems and the streaming of spatial sub-parts of a video. Section 3 provides the underlying concepts behind tiled streaming. The design principles, definitions and examples of the MPEG-DASH SRD feature are discussed in section 4. In section 5, we look at several experiment examples. Finally, in section 6, we provide conclusions and an outlook towards future deployments.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
MMSys'16, May 10-13, 2016, Klagenfurt, Austria
© 2016 ACM. ISBN 978-1-4503-4297-1/16/05…$15.00
DOI: http://dx.doi.org/10.1145/2910017.2910606

2. RELATED WORK
With recent capturing systems for UHD video, new types of media experiences are possible where end users have the possibility to choose their viewing direction and zooming level. Different examples of such interactive region-of-interest (ROI) video streaming have been demonstrated or deployed. Interactive ROI video streaming was initially explored in depth by [2, 3]. The authors developed various methods in the context of ClassX, an interactive ROI streaming system for online lecture viewing, selecting tiled streaming as the best compromise between bandwidth, storage, processing and device requirements. Tiled streaming, explained in more detail in section 3 relies on a tiling of video into independent video streams. The ClassX system [4] allows for capture and interactive streaming of online lectures. To reduce tile switching delay, [5] studied a crowd-driven ROI prediction scheme to pre-fetch future selected regions. This scheme exploited user viewing statistics collected at the server to make ROI predictions. The experiments showed that crowd-driven prefetching can substantially reduce average ROI switching delays compared to a system without prefetching.

Some of the first comparisons between regular encoding and coding for tiled streaming were investigated by [6]. In particular, they compared regular monolithic streaming with tiled streaming. Their results indicated that a monolithic stream with proper choice of parameters achieves better bandwidth efficiency than tiled streams. The research was later extended with studies of user

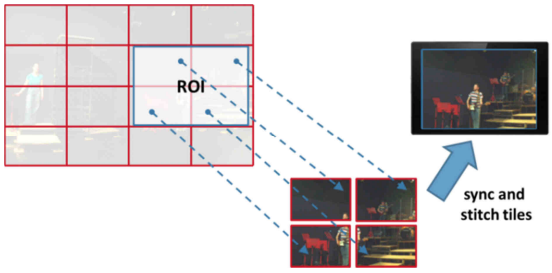

Figure 1. Tiling refers to a spatial partitioning of a video.

access patterns in [7]. A zoomable video system was further explored by [8]. There, the focus was on enabling low-delay interaction with high resolution and high-quality video, with constraints on the available bandwidth and processing capabilities as encountered in current network technologies and devices. The authors studied bandwidth requirements for tiled streaming as well as the performance of media containers. The authors also noted that, since rapid seeking is required for tiled streaming, the choice for a particular media container has an impact on seeking performance. Several codec and container implementations only support seeking to the nearest I-frame only. This results in increased switching delays, as the client waits until all decoded tile frames can be synchronised. The zoomable video systems discussed in this section are all based on a form of tiled streaming, but none of them use standard adaptive streaming systems such as MPEG-DASH. In the next section, we describe the underlying concepts behind this approach to interactive region-of-interest (ROI) video streaming.

3. TILED STREAMING

Tiling refers to a spatial partitioning of a video where tiles correspond to independently decodable video streams [9]. A tiled video can be obtained from a single video by partitioning each frame into multiple frames of smaller resolution and by aggregating the smaller frames coming from the same partition/region of the input frame into a new video. Here, tiles are defined as a spatial segmentation of the video content into a regular grid of independent videos. In particular, the partitioning into tiles is temporally consistent. 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. Classic video compression corresponds to a 1x1 tiling. While this approach nicely removes all dependencies between tiles, the drawbacks of this approach are that synchronisation between tiles (temporally as well as in terms of global encoding rate) must be ensured and that a ROI might require more than one tile to be accessed for reconstructing the view, as depicted in Figure 1. Moreover, the compression performance is also reduced as the existing redundancy between tiles can no longer be exploited. Therefore, tile-based compression for random access into ultra-high resolution video has different objectives than classic video compression. While a good compression performance for the total video is still desirable at the server side, the objective is also to maximise the quality of a ROI reconstructed by an interactive client and minimise the total bitrate of the tiles it needs to decode.

Tiling is typically performed within a multi-resolution scheme, where lower resolutions can help to reduce the compression loss of tiled content at higher resolutions. If the lower resolution tiling is small enough (such as thumbnails), the bitrate overhead of using another resolution layer is affordable. Moreover, multi-resolution tiling allows for increasing the quality of user defined

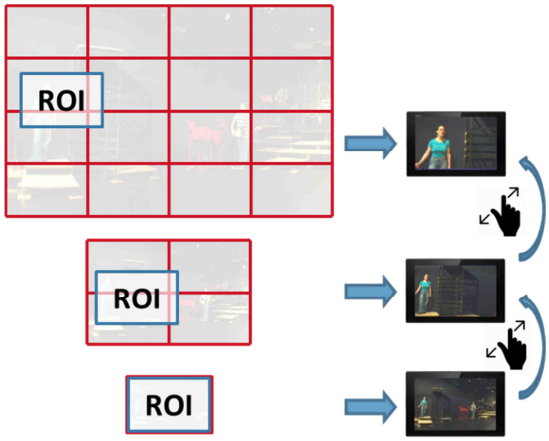

Figure 2. Multiple resolution layers are created from the original source video to increase the quality of user-defined zooming factors on tiles.

zooming factors on tiles. Once the user zooms into a region of the content, the server will provide the highest resolution tiles that are jointly constitute the requested region. Because of transmission latency or possible data loss, the device reconstructing the ROI may end up in a situation, where the currently accessed high resolution tiles do not fully cover the current ROI. In that case, the multi-resolution approach is useful to conceal the missing tiles. For example, tiles may be missing at the borders of the ROI in case of a fast panning command by the end user. In that case, lower resolution tiles can be used to approximate those border pixels at a much lower transmission cost. However, the tasks of reconstructing the ROI from multi-resolution tiles is also more complex and might lead to visual quality drops at the border between high resolution tiles and concealed pixel regions. Figure 2 depicts the use of multiple resolution layers.

With tiled streaming, a client device may need as many decoders as it needs tiles to cover a requested ROI. This limitation poses a problem for the forms of multi-resolution tiling as discussed previously. For example, if at least four tiles are required to reconstruct any given ROI, then depending on the tile size, the number of simultaneous tiles needed for a given ROI can be as high as 9, 16 or even 25. To solve this problem, one can resort to changing the tiling scheme in such a way that the client device never needs more than one tile in order to prepare the view to be displayed. This approach uses *overlapping* tiles, i.e. tiles with a certain overlap with each other. By using overlapping tiles, the total number of tiles that together cover the requested ROI can be reduced. 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 less tiles being required to reconstruct a given ROI. The downside of this overlap is that it results in more storage on the server side, which has to store redundant pixels. Another side effect of having overlapping tiles is that tile sizes are no longer equal across a given tiling scheme. That is, for a given overlapping factor, a corner or edge tile has an overlap in respectively one or two planar directions. For example, a tile containing the top-left of a given video stream overlaps with the tile to its immediate right and with the tile directly below it. A tile in the middle of a tiling scheme, however, overlaps in four planar

directions and therefore has a larger tile size. Another approach to tiled streaming is to use coding tools, such as HEVC tiles, to enable a combination of tiles to be decoded by a single decoder as described in [10].

4. MPEG-DASH SRD: DESIGN PRINCIPLES, DEFINITIONS AND EXAMPLES

4.1 Overview and Initial Limitations of MPEG-DASH

DASH is the MPEG standard for describing dynamic adaptive streaming content [1]. In this standard, the Media Presentation Description (MPD) is an XML document which describes the content available in an adaptive streaming session. It enables client-side adaptation strategies because the content is made available in different characteristics (quality, codec, etc.) by simple HTTP servers. The DASH client decides which content to stream depending, for example, on user preferences and client constraints. The MPD content is structured along the media timeline, i.e. the MPD is divided into Period elements which correspond to periods of time during which a given set of media is available. In each Period, and for a given media, AdaptationSet elements group the alternate versions of this media in terms of codec, resolutions or bitrate. Each alternate version is described by a Representation element, which provides the decomposition of the media over time into media segments. Each media segment is described by a URL, either explicitly or through a template.

In addition, the DASH standard allows associating non-timed related information to MPD elements, such as the role of a media asset (e.g. main video or alternate video, subtitle representation, or audio description). The MPD uses so-called Descriptor elements to associate such information. Prior to the definition of SRD, there was no descriptor to associate spatial information with media assets. The existing Viewpoint descriptor was too restricted and mainly oriented toward 3D. It was not possible to describe that two videos were representing spatially related parts of a same scene. The SRD feature solves this problem.

4.2 Standardization of SRD

The standardization effort for SRD in MPEG DASH started in April 2013 at the 104th MPEG meeting. Two individual contributions addressing immersive interaction with video content were submitted in DASH. Consequently, MPEG experts agreed on starting a Core Experiment called SRD within MPEG DASH. The aim of a Core Experiment in MPEG is to steer the group effort by setting the end result goal as well as the framework in order to carry out the work over several meeting cycles. In particular for SRD, MPEG experts first collected a list of use cases covering a wide range of immersive experience use cases based on video user interaction. In an effort to enable as much use cases as possible, participants of the SRD Core Experiment reached consensus on a list of requirements that the solution should meet. In this process, decisions were made to keep the scope of the solution within reasonable boundaries. For instance, experts agreed on the fact the future SRD solution should primarily enable use cases where video positions are expressed in a 2-dimensional space. More complex scenarios would have significantly hindered the chance of a rapid success of a first version SRD. Furthermore, experts favoured an extensible design of the solution such it did not preclude future evolutions of SRD where such more advanced use cases could be addressed.

4.3 Design Principles for MPEG-DASH SRD

The Spatial Relationship Description (SRD) part of the DASH standard obeys the same design principles as DASH. It provides additional information to further help DASH clients in determining which media to choose. Essentially, SRD describes how the content is spatially organized at the origin when created. This information can be leveraged by DASH clients to provide user interactions. In DASH, the MPD provides information about the media assets consumed by DASH clients, but the client behaviour and the adaptation logic is out of scope of the standard and left open to stimulate innovation. The SRD feature was elaborated with the same philosophy. It describes how media assets spatially relate to each other and it does not presume anything on how a player shall use this information. For instance, given an MPD describing two spatially related videos, a player may decide to display both videos while another player can decide to play them one at a time. In other words, MPEG DASH SRD does not define a composition or presentation layer. The exact composition of the media on the screen is left to the application layer, such as an HTML 5 presentation engine. Furthermore, the information in SRD describes how the media assets are spatially related from a content creator perspective. This may even deviate from the way the content was actually captured. For instance, if an MPD contains SRD information describing a grid of videos, there may be either a single camera shooting the entire scene, which is then split and encoded as multiple individual videos; or there may be multiple cameras each shooting a different part of the scene. In the course of the standardization effort, experts submitted numerous use cases with a different nature and complexity, which advocated for a flexible design of SRD. Possible solutions were envisaged such as an explicit grid positioning (e.g. placing media in an *NxN* grid), or a one-to-one positioning with directions (e.g. a media is at the left, or at the north of another media). In the end, the final decision was to position the media in a 2D coordinate system. Hence, it is possible to position any media in a coordinate system providing the common x, y, width and height attributes, respectively the top-left corner coordinate on the x-axis, on the y-axis, the width and the height of the media asset in the reference space. In an effort to cover most of the discussed use cases in standardization, it should be emphasized that this coordinate system does not correspond to the rendering coordinate system but represents an arbitrary coordinate system in which the positioning of the media is provided. The arbitrary nature of the coordinate system enables on the one hand simple use cases where SRD information contains grid cell indices. On the other hand, the relationship between media elements is therefore not direct, i.e. it is not explicitly indicated whether a media constitutes a cell of a given grid. Instead, all the media positioned in the coordinate system need to be examined in order to conclude that, all together, they form a grid. This design choice makes it also possible to represent complex relationships, such as overlapping videos. For instance, SRD allows describing that two cameras capture the same scene: one using a wide angle and the other one capturing only a narrow part of the scene. It should be noted that SRD has been designed to enable describing spatial relationships between any type of media. It is not restricted to describe relationships between video streams. In particular, one could position audio streams relatively to each other, or to a video stream. It should also be noted that the SRD feature does not make any assumption on how the processing of spatially related videos is implemented. In line with the DASH standard, SRD is codec-agnostic. It is therefore possible to describe a tiling composed of videos using different codecs if necessary. Additionally, it does not assume, for instance, that two decoders are necessary when two videos are

selected by the DASH player. This is particularly relevant for tiling when the HEVC standard is used. With additional tools such as compressed-domain stitching [10], specifically encoded independent HEVC videos are combined using the HEVC tiling tool to form a single stream, decodable by a single decoder.

4.4 SRD Syntax

The DASH SRD feature leverages the generic descriptor elements, namely the Essential Property and Supplemental Property elements. Essentially, these descriptors are composed of a scheme URI (`@schemeIdUri` attribute) and a value (`@value` attribute). For a given descriptor in the MPD, its scheme URI normatively defines the syntax and the semantics of its value attribute. In case of SRD, the scheme URI to be used is `urn:mpeg:dash:srd:2014` and the `@value` attribute must then follow the specific syntax as specified in Annex H of MPEG-DASH. In particular, the `@value` attribute must be a comma-separated list of containing always the `source_id`, `object_x`, `obect_y`, `object_width`, and `object_height` parameters while the remaining `total_width` and `total_height` parameters are conditionally mandatory and the `spatial_set_id` parameter is optional. The semantics of each of these parameters is given in Table 1 below.

Table 1 - Syntax and semantics of SRD information.

EssentialProperty@value or SupplementalProperty@value parameter	**Description**
`source_id`	non-negative integer in decimal representation providing an identifier for the source of the content and implicitly defining a coordinate system
`object_x`	non-negative integer in decimal representation expressing the horizontal position of the top-left corner of the associated media assets in the coordinate system
`object_y`	non-negative integer in decimal representation expressing the vertical position of the top-left corner of the associated media assets in the coordinate system
`object_width`	non-negative integer in decimal representation expressing the width of the associated media assets in the coordinate system
`object_height`	non-negative integer in decimal representation expressing the height of the associated media assets in the coordinate system
`total_width`	optional non-negative integer in decimal representation expressing the width of the extent of all media assets in the coordinate system
`total_height`	optional non-negative integer in decimal representation expressing the height of the extent of all media assets in the coordinate system
`spatial_set_id`	optional non-negative integer in decimal representation providing an identifier for a group of media assets.

The MPD author can signal SRD descriptors for each of the different content items in the same MPD. Identical values of `source_id` in different SRD descriptors indicate that these SRD descriptors belong to the same reference space. Conversely, SRD descriptors with different `source_id` have no spatial relationship with each other. The MPD author may also choose to use the same `total_width` and `total_height` parameters for all SRD descriptors. In some of the SRD descriptors these values are optional, provided that there is at least one SRD descriptor in the MPD, for a given `source_id`, that provides these values. Lastly, certain use cases require the efficient selection of tiles within the same resolution layer, e.g. when panning, or from another resolution layer, e.g. when zooming in or out. In this case, the MPD author can help the DASH clients even further by explicitly signalling the grouping of multiple AdaptationSets into the same resolution layer. This grouping is achieved by setting identical values for the `spatial_set_id` of AdaptationSets belonging to the same resolution layer.

Another matter of concern for content providers is to issue a single MPD for both legacy and SRD-aware DASH clients. This is why the standard allows both the EssentialProperty and SupplementalProperty elements to be used for SRD information. The DASH standard allows the MPD author to explicitly signal two cases via these descriptors. With an EssentialProperty element, the MPD author expresses that the successful processing of the descriptor is essential in order to properly process the content of the parent element. It is then expected that DASH clients discard the parent element of an EssentialProperty when they do not recognise its scheme URI. In contrast, with a SupplementalProperty element, the MPD author expresses that the successful processing of the descriptor is not essential. Consequently, DASH clients that do not recognise the scheme URI of a SupplementalProperty element can safely consume the content of its parent element as signalled by the MPD author in the MPD. Applied in the context of SRD, the DASH implementation guidelines recommend the following usage. If the MPD author does not want legacy DASH clients to present video content augmented with SRD descriptors, these SRD descriptors should be using the EssentialProperty descriptor. This way, legacy DASH clients will discard this video content. On the other hand, if a given video content augmented with an SRD descriptor can properly be presented individually then the MPD author should use the SupplementalProperty descriptor.

4.5 Examples

The MPD in Code 1, in which some elements and attributes have been omitted for brevity, uses SRD and offers two videos: a large one, with a resolution of 3840x2160; and a smaller one, with a resolution of 1920x1080, as indicated by the Representation attributes. Since these videos are in different AdaptationSet elements, the MPD indicates that no seamless switching is guaranteed between them, meaning that the DASH client should not use them for adaptation logic. Without any additional information, a DASH player would not know which video to select for display, if both can be downloaded and decoded. The DASH Role elements can further assist the player by indicating that the large resolution is the main video while the smaller resolution video is a supplementary video. However, only the SupplementalProperty Descriptors with the SRD schemeIdUri attribute provide spatial information. In this particular example, the small video captures the centre of the large video, i.e. the rectangle (1920, 1080, 3840, 2160) in the coordinate system while the large video occupies the rectangle (0, 0, 5760, 3240).

```
<Period>
<AdaptationSet>
<SupplementalProperty schemeIdUri="urn:mpeg:dash:s
rd:2014" value="0,0,0,5760,3240,5760,3240"/>
<Role schemeIdUri="urn:mpeg:dash:role:2011" value=
"main"/>
<Representation id="1" width="3840" height="2160"
…>
<BaseURL>full.mp4</BaseURL>
</Representation>
</AdaptationSet>
<AdaptationSet>
<SupplementalProperty schemeIdUri="urn:mpeg:dash:s
rd:2014" value="0,1920,1080,1920,1080,
5760,3240"/>
<Role schemeIdUri="urn:mpeg:dash:role:2011" value=
"supplementary"/>
<Representation id="2" width="1920" height="1080"
…>
<BaseURL>part.mp4</BaseURL>
</Representation>
</AdaptationSet>
</Period>
```

Code 1: Example of an MPEG-DASH MPD with SRD information indicating that one version of a video is a spatial part of the other version of the video.

As mentioned earlier, the SRD coordinate system has arbitrary units such that SRD information is decoupled from the actual rendered sizes of the videos. This way, the same SRD information in a given AdaptationSet element can apply to all the Representations present in this element even if they do not have the same resolution (but the same aspect ratio). In the above example, the MPD author chose SRD values to match the width and height of the actual videos. But another valid approach is to normalize these values by dividing them with the smallest resolution such that the video with the smallest resolution has a width and height of one unit.

```
<Period>
<AdaptationSet>
<SupplementalProperty schemeIdUri="urn:mpeg:dash:s
rd:2014" value="0,0,0,3840,2160,7680,2160"/>
<Representation id="1" width="3840" height="2160"
…>
<BaseURL>UHD-left.mp4</BaseURL>
</Representation>
<Representation id="2" width="1920" height="1080"
…>
<BaseURL>HD-left.mp4</BaseURL>
</Representation>
</AdaptationSet>
<AdaptationSet>
<SupplementalProperty schemeIdUri="urn:mpeg:dash:s
rd:2014" value="0,3840,2160,3840,2160,7680,2160"/>
<Representation id="3" width="3840" height="2160"
…>
<BaseURL>UHD-right.mp4</BaseURL>
</Representation>
<Representation id="4" width="1920" height="1080"
…>
<BaseURL>HD-right.mp4</BaseURL>
</Representation>
</AdaptationSet>
</Period>
```

Code 2: Example of an MPEG-DASH MPD with SRD information describing a horizontal panorama made of 2 videos, each as 2 alternate resolutions. Units in SRD are chosen to match pixels of the highest resolution videos.

Note that SRD only allows integer values, which requires careful consideration when normalizing values from actual videos sizes. This is shown in the MPD examples in Code 2 and Code 3. Both MPD describe a set of four videos: one UHD video and its HD equivalent representing a part of a scene (the left), and another UHD video and its HD equivalent representing the other part of the scene (the right), i.e. the video represents a 2x1 contiguous horizontal panorama composed of two videos positioned next to each other.

```
<Period>
<AdaptationSet>
<SupplementalProperty schemeIdUri="urn:mpeg:dash:s
rd:2014" value="0,0,0,2,2,4,2"/>
<Representation id="1" width="3840" height="2160"
…>
<BaseURL>UHD-left.mp4</BaseURL>
</Representation>
<Representation id="2" width="1920" height="1080"
…>
<BaseURL>HD-left.mp4</BaseURL>
</Representation>
</AdaptationSet>
<AdaptationSet>
<SupplementalProperty schemeIdUri="urn:mpeg:dash:s
rd:2014" value="0,2,0,2,2"/>
<Representation id="3" width="3840" height="2160"
…>
<BaseURL>UHD-right.mp4</BaseURL>
</Representation>
<Representation id="4" width="1920" height="1080"
…>
<BaseURL>HD-right.mp4</BaseURL>
</Representation>
</AdaptationSet>
</Period>
```

Code 3: Example of an MPEG-DASH MPD with SRD information describing a horizontal panorama made of 2 videos, each as 2 alternate resolutions. Units in SRD are chosen such that the value 1 corresponds to the size of the smallest videos dimensions.

5. SERVICE AND DEPLOYMENT EXAMPLES

5.1 High quality zoom-in

An use case for tiled streaming is the possibility to allow high quality zoom-in, as illustrated on Figure 3. In this use case, we assume a video sequence encoded as two independent switchable streams: the first version of the video is a standard HD resolution stream that is used as video preview; the second version of the video is an ultra-high resolution stream encoded in very high quality. This second version is a tiled video. Starting from the first version of the video, the user is browsing the preview that can be augmented with the tiling information to indicate that spatial access is possible. When clicking on a tile, or when selecting a set of tiles, the DASH client automatically switches to the high quality/high resolution stream, streaming only the selected tiles. Through user interactions, the DASH client can dynamically switch back to the full-frame video (the preview). The so-tiled video plus the tile description in the MPD provide a better user experience than with predefined regions of interest. Moreover, this provides better image quality for the selected region than with a simple upscaling of the preview video. Finally, the tiled video with the SRD annotations in the MPD provides a new adaptation dimension next to bandwidth, resolution, quality.

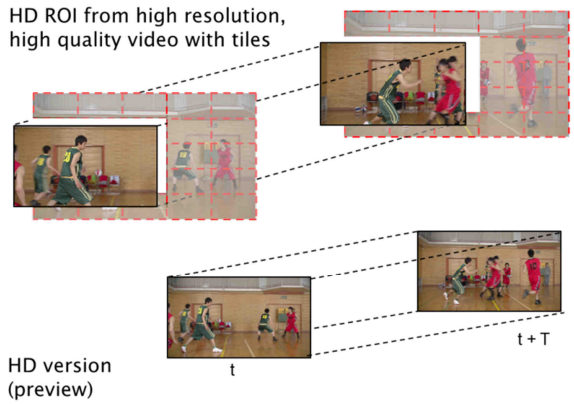

Figure 3: Example of tiled stream for high-quality zoom-in.

That is, the possibility at a given bandwidth to choose between full-frame video in low quality and a spatial area in higher quality. In particular, it is possible to display multiple tiles and mix the qualities, as discussed in [11].

5.2 4K streaming to regular mobile devices

With the interactive Camera-based Coaching and Training (iCaCoT) experiment, it was tested how actual users make use of the concept of tiled streaming in a live and real-world environment. The experiment took place in a testbed situated at the ski resort of Schladming, Austria. High resolution (i.e. HD or higher) cameras were We statically mounted at a suitable location around the Schladming venue, in particular around a ski slope for training or a fun park for winter sport enthusiasts. The video recorded by these cameras was spatially segmented in real-time using tiled streaming ingest components. The experiment started on March 25th and ended on March 28th 2014. The goal of the experiment was twofold:

- Providing a new tool for training and coaching;
- Offering a novel way for Schladming visitors to record themselves coming down the mountain and sharing their experiences with friends and family.

For this purpose, the SRD technology, implemented in a prototype with tiled streaming technology [12] has been enhanced with a number of additional components and features in the following areas: support for live streaming, trick-play functionalities, visual annotation drawing. For the coaching scenario where the precision is essential, the users acknowledged the benefit of a higher resolution content (4K instead of HD).

As explained in section 4.4, it is critical to distinguish different resolution layers when it comes to very high resolutions. This greatly helps the application in its decision logic for selecting the appropriate video(s). Therefore the ingest components produced two layers, each composed of several videos. Translated in SRD terminology, the experiment defined two SRD reference spaces whose dimensions were 1920 by 1080 for the first one and 3840 by 2160 for the second one. Two different spatial_set_id were assigned to each of them. Figure 4 depicts the SRD information structure in a diagram. This way of defining the SRD descriptors provides two essential benefits to the application.

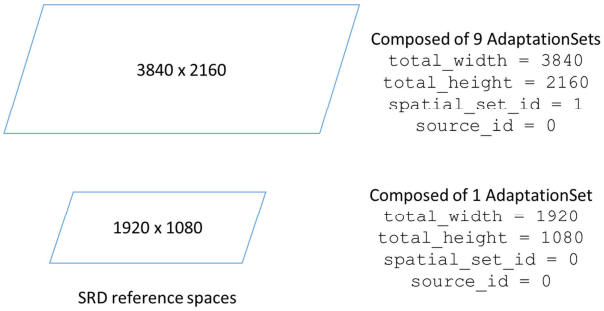

Figure 4: 4K content offered with DASH SRD in 10 tiles.

1. Efficient parsing of the MPD. The AdaptationSets having the same spatial_set_id value belong to the same reference space or resolution layer in our case;
2. Speed improvement of the decision process when the user interacts with the content as described.

In practice, the application can break down the selection process into two steps: 1) which reference space? and 2) which Representation(s)? Without 1), the application would have to perform 2) on the whole set of available AdaptationSets. For 4K and 8K resolutions, where up to 100 AdaptationSets may be described in the MPD, differentiation of reference spaces through the spatial_set_id parameter is essential for guaranteeing low-latency interaction and responsiveness.

5.3 Watching at home a live sport event augmented with UHD tiled streaming

Figure 5: SRD use case for the 2014 Commonwealth Games.

During the 2014 Commonwealth Games, BBC R&D and TNO ran a trial in the UK based on SRD. The trial involved more than 50 participants in UK households who watched this sporting event in ultra-high definition on their smart phone or tablet [12]. These viewers were able to navigate and zoom within the video, hence having full control over what they wanted to watch, see Figure 5. This new way of watching TV used the tiled streaming technology, bringing ultra-high definition video (4K or greater) to mobile devices like smart phones and tablets. After the trial, TNO made the iOS application iXperience available in the App Store [13], so users can freely navigate through high-definition videos. BBC R&D also used tiled streaming as part of their own "Venue Explorer" demonstration [14] that showed interactive coverage of the Commonwealth Games athletics. During this trial, a dedicated ingest node processed the 4K original content to generate 17 tiles in total, constituting as many AdaptationSets.

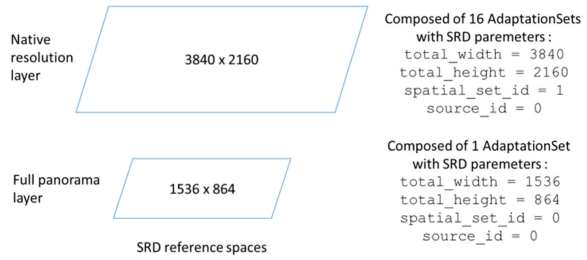

Figure 6: 4K content offered with DASH SRD in 17 tiles.

Figure 6 presents the corresponding SRD description using MPEG DASH terminology. In addition to the video, an audio feed accompanied the full panorama video to provide both the ambient sound of the event and the commentary. Questionnaire results showed that participants were very much inclined to use the app again and to recommend it to others. In addition, the interactive functionalities of the app (zoom in/out and navigate through the video area) were considered far more likable than traditional playback functionalities providing a real immersive experience in the stadium. It is indeed very common that traditional broadcast feeds only show a fraction of what is happening in the stadium whereas the tiled streaming feed was capturing the full stadium.

5.4 Transmission and rendering of panoramic video on a video wall

The previous examples focussed on videos with conventional aspect ratios. For a fully immersive experience, wide field-of-view panoramic videos with horizontal or vertical wide angles are a necessity. Such videos come with irregular aspect ratios typically require a dedicated panoramic video player. MPEG-DASH SRD can be an appropriate candidate to enable a standardized video player for panoramic videos. For rendering of the full panoramic video on e.g. a video wall, the synchronization performance of the video client consuming the multiple video tiles is essential. That is, a general and significant problem of multi-session streaming methods is the synchronization of multiple tiles. This is because each tile segment is separately downloaded and therefore needs a synchronization point for the rendering module. Fortunately, we can have several assumptions to manage video frames. First, we ensure that all segments sizes are equal, i.e. the numbers of video frames in each segment is equal. Second, we ensure that the first video frame of a segment is an intra-coded frame to sever any dependent links to other frames. Now we have enough synchronization points in every segment. A video player should wait for the interrupts from all video download buffers, until all segments at the same time instance are fetched in the buffers. Then it delivers the downloaded segments to the decoding and rendering modules simultaneously. The rendering buffers store the decoded pictures and they are rendered onto the video wall, as depicted in Figure 7.

6. CONCLUSIONS AND OUTLOOK

The MPEG-DASH standard has been enhanced with a feature that enables spatial random access in streaming video. This feature, referred to as Spatial Representation Description, or SRD, enables the streaming of spatial sub-parts of a video to display devices, in combination with adaptive multi-rate streaming that is intrinsically supported by MPEG DASH. Specific design principles have been taken into account to associate spatial information with media assets, i.e. to describe that multiple videos represent spatially related parts of a same scene. The SRD syntax takes into account MPD authoring requirements and hybrid

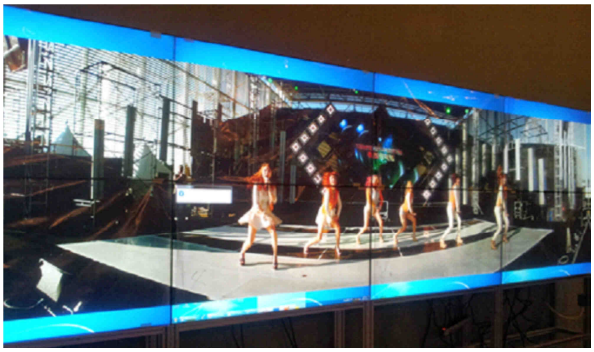

Figure 7: Experiment for streaming and rendering 8 HD video tiles to a video wall using the SRD feature.

deployments with both legacy and SRD-aware DASH clients. The feature enables a variety of advanced use case beyond conventional MPEG DASH, such as high quality zoom-in, interactive streaming of UHD video to mobile devices, and transmission of wide-angle panoramic video to large tiled displays. To promote industry take-up by key players, the authors now focus on providing SRD test vectors and developing reference implementations of SRD generation and validation tools.

7. ACKNOWLEDGMENTS

Many thanks to Arjen Veenhuizen, Ray van Brandenburg and David G. Mico at TNO, to Frédéric Maze at Canon, to Jean Le Feuvre at Télécom ParisTech, as well as to the many colleagues in MPEG DASH for the collaboration on the matter and their contributions to a hopefully successful and widely deployed standard.

8. REFERENCES

[1] ISO/IEC 23009-1:2014/Amd 2:2015, "Spatial relationship description, generalized URL parameters and other extensions".

[2] Mavlankar, A., Agrawal, P., Pang, D., Halawa, S., Cheung, N.-M., and Girod, B. An interactive region-of-interest video streaming system for online lecture viewing. In *Packet Video Workshop (PV)*, 2010 18th International, IEEE (2010), 64–71. DOI=http://dx.doi.org/10.1109/PV.2010.5706821.

[3] Mavlankar, A., and Girod, B. Spatial-random-access-enabled video coding for interactive virtual pan/tilt/zoom functionality. *Circuits and Systems for Video Technology, IEEE Transactions on*, 21, 5 (2011), 577–588. DOI= http://dx.doi.org/10.1109/TCSVT.2011.2129170.

[4] Pang, D., Halawa, S., Cheung, N.-M., and Girod, B. Classx mobile: region-of-interest video streaming to mobile devices with multi-touch interaction. In *Proceedings of the 19th ACM international conference on Multimedia, ACM* (2011), 787–788. DOI= http://dx.doi.org/10.1145/2072298.2072457.

[5] Pang, D., Halawa, S., Cheung, N.-M., and Girod, B. Mobile interactive region-of-interest video streaming with crowd-driven prefetching. In *Proceedings of the international ACM workshop on Interactive multimedia on mobile and portable devices, ACM* (2011). DOI= http://dx.doi.org/10.1145/2072561.2072564.

[6] Quang Minh Khiem, N., Ravindra, G., Carlier, A., and Ooi, W. T. Supporting zoomable video streams with dynamic region-of-interest cropping. In *Proceedings of the first*

annual ACM SIGMM conference on Multimedia systems, ACM (2010), 259–270. DOI= http://dx.doi.org/10.1145/1730836.1730868.

[7] Quang Minh Khiem, N., Ravindra, G., and Ooi, W. T. Adaptive encoding of zoomable video streams based on user access pattern. *Signal Processing: Image Communication* 27, 4 (2012), 360–377. DOI= http://dx.doi.org/10.1145/1943552.1943581.

[8] Quax, P., Issaris, P., Vanmontfort, W., and Lamotte, W. Evaluation of distribution of panoramic video sequences in the explorative television project. In *Proceedings of the 22nd international workshop on Network and Operating System Support for Digital Audio and Video, ACM* (2012), 45–50. DOI= http://dx.doi.org/10.1145/2229087.2229100.

[9] Niamut, O., Prins, M., van Brandenburg, R., and Havekes, A. Spatial tiling and streaming in an immersive media delivery network. *Adjunct Proceedings of EuroITV* (2011).

[10] Sanchez, Y.; Globisch, R.; Schierl, T.; Wiegand, T., Low complexity cloud-video-mixing using HEVC, in *Consumer Communications and Networking Conference (CCNC), 2014 IEEE 11th* , vol., no., pp.213-218, 10-13 Jan. 2014. DOI= http://dx.doi.org/10.1109/CCNC.2014.6866573.

[11] Wang, H., Nguyen, V.-T., Ooi, W. T., and Chan, M. C. Mixing tile resolutions in tiled video: A perceptual quality assessment. In *Proceedings of Network and Operating System Support on Digital Audio and Video Workshop, ACM* (2014), 25. DOI= http://dx.doi.org/10.1145/2578260.2578267.

[12] Niamut, O.A.; Thomas, G.A.; Thomas, E.; van Brandenburg, R.; D'Acunto, L.; Gregory-Clarke, R.: 'Live event experiences - Interactive UHDTV on mobile devices', *IET Conference Proceedings*, 2014, IET Digital Library. DOI=http://dx.doi.org/10.1049/ib.2014.0033.

[13] "TNO iXperience", https://itunes.apple.com/us/app/tno-ixperience/id895475926?ls=1&mt=8 Last visited: March 11, 2016.

[14] "BBC R&D Venue Explorer", http://www.bbc.co.uk/rd/projects/venue-explorer Last visited: March 11, 2016.