INTERNATIONAL ORGANISATION FOR STANDARDISATION
ORGANISATION INTERNATIONALE DE NORMALISATION
ISO/IEC JTC1/SC29/WG11
CODING OF MOVING PICTURES AND AUDIO

ISO/IEC JTC1/SC29/WG11 MPEG2013/M29232
April 2013, Incheon

Source **Canon Research Centre France and Telecom ParisTech**
Status **Input Document for 104th MPEG meeting**
Title **Interactive ROI streaming with DASH**
Author Franck Denoual, Hervé Le Floch, Frédéric Mazé, Eric Nassor, Nael Ouedraogo (Canon Research Centre France), Cyril Concolato, Jean Le Feuvre (Telecom ParisTech)

1 Introduction

The resolutions of compressed videos are rapidly increasing: 4k2k cameras are now widely available [5] and even 8k4k demonstrations of live streaming have been realized [6] . With these very high resolution contents, new usages in video streaming are now possible, like interactive zooming features.

The current adaptation means in DASH enable dynamic adaptation of the streaming in terms of bandwidth, spatial resolution, scalability layers… but do not enable to dynamically switch to a user-selected area in the video being streamed. Thus, there is a need for spatial access in video streams as we illustrate via use cases in the next section.

Based on the use cases described below, we would like to highlight the need in DASH to support spatial addressing. We also propose a description based on an extension of the Role element that has a reduced impact on the current MPD syntax.

2 Use case description

A use case of interest that is not possible with current DASH is the possibility for a user to dynamically select and switch to a spatial area, for example to zoom in.

2.1 Dynamic high-quality zoom-in

In this use case, we assume a video sequence encoded as 2 independent switchable streams (Figure 1):

- 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 a ultra-high resolution stream with very high quality.

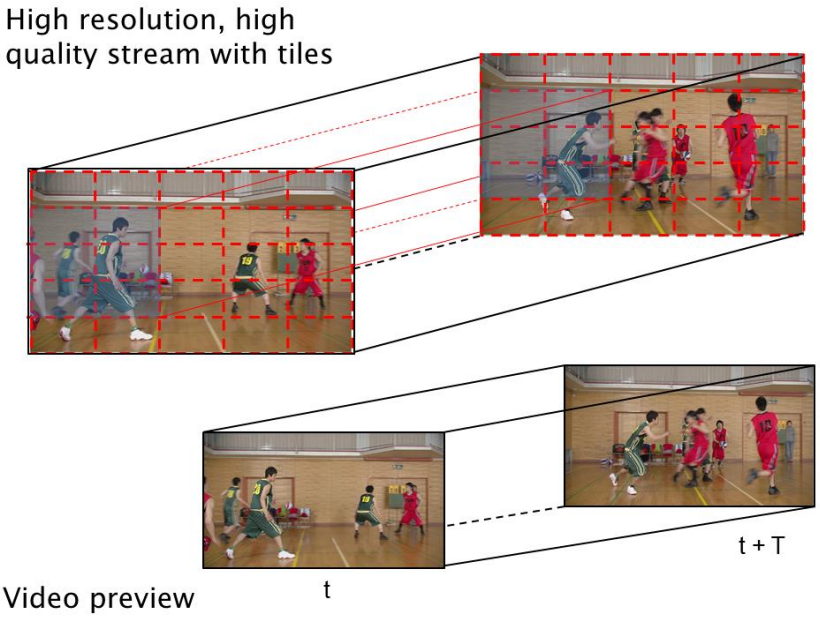

Figure 1: Example of tiled stream for high-quality zoom-in

The user starts streaming with the first version of the video, i.e. the preview. The video preview is augmented with the tiling information to indicate that a 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 graphical zoom. Finally, it provides a new adaptation dimension after bandwidth, resolution, quality: the possibility at a given bandwidth to choose between full-frame video in low quality and a spatial area in higher quality.

2.2 Various applications

2.2.1 Hybrid delivery of panoramic images

The above use case could be applied for personalized experience in panoramic images or panning in super high resolution videos. The Figure 2 below is an example of application from the European project Fascinate [5] . A panoramic video built from multiple cameras could be broadcasted to clients. A MPD could be available to describe spatial access possibilities in this panorama so that each user, via broadband can have a personalized browsing experience by focusing only on his preferred spatial area in the panorama.

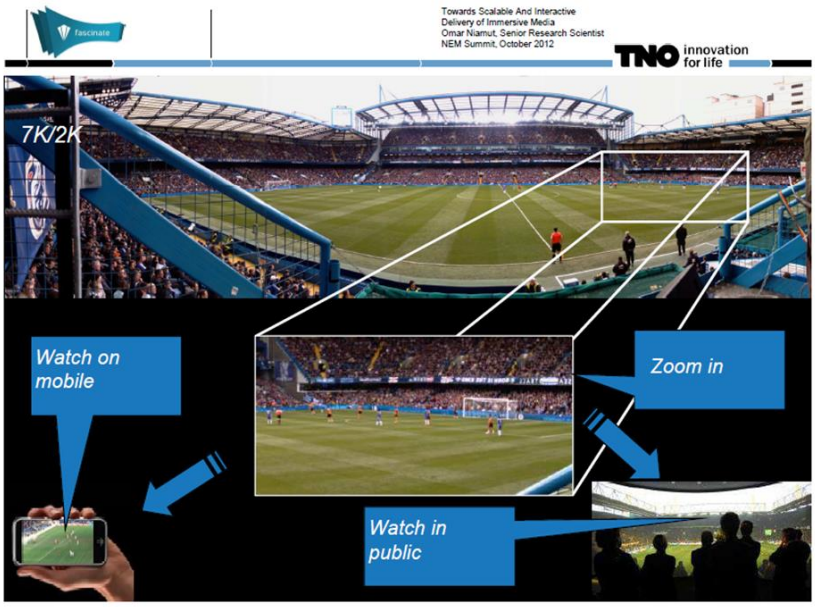

Figure 2: Example of spatial access for personalized browsing experience

2.2.2 Video surveillance system

New video surveillance cameras are able to record a scene with very high resolution providing simultaneously to several users and for recording a global view of the scene and the possibility to zoom on user selected region of interest [9] .

When a motion is detected in a surveillance area, a specific recording of this area begins with high quality and Intra only pictures for random access. If a face detection module is available, the face appearing on the surveillance area can also be tracked and encoded as a specific region of interest. These streams are made available on the surveillance server. The specific area is described as an alternate representation of the full video. The user can remotely check when there is an alarm by switching to the stream representing the specific area but he can also select any part of the global video and seamlessly zoom onto the face of the detected person or on any interesting part of the scene.

2.2.3 E-learning

ClassX player [10] has been developed in Stanford University. It is part of an e-learning project and relies on slices in AVC to provide zooming features on regions of interests. The lecture is recorded as a high spatial resolution H.264/AVC video stream with tiles to provide interactive region of interest (as depicted on Figure 3).

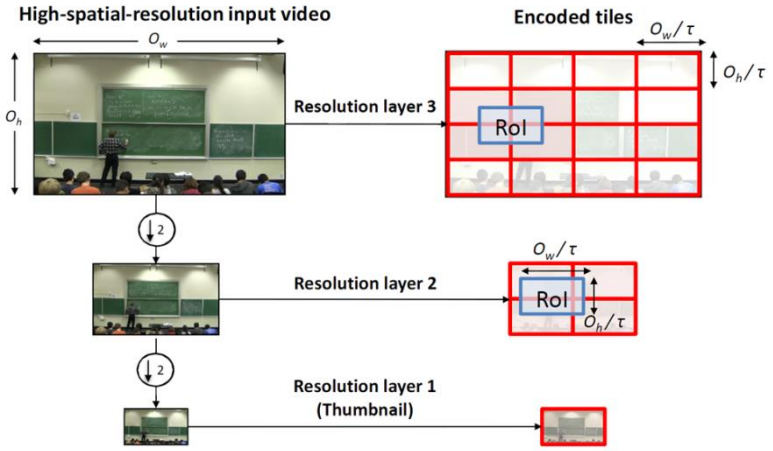

Figure 3: Alternative video representations in ClassX

The ultra-high version of the video is down-sampled by 2 in both dimensions to provide an alternate version of the video, still with tiles and region of interest. The tiles are independently encoded. A third version, down-sampled by a factor of 2 is provided as a thumbnail video. The user starts streaming the thumbnail and can zoom on specific region of interest as illustrated below in Figure 4.

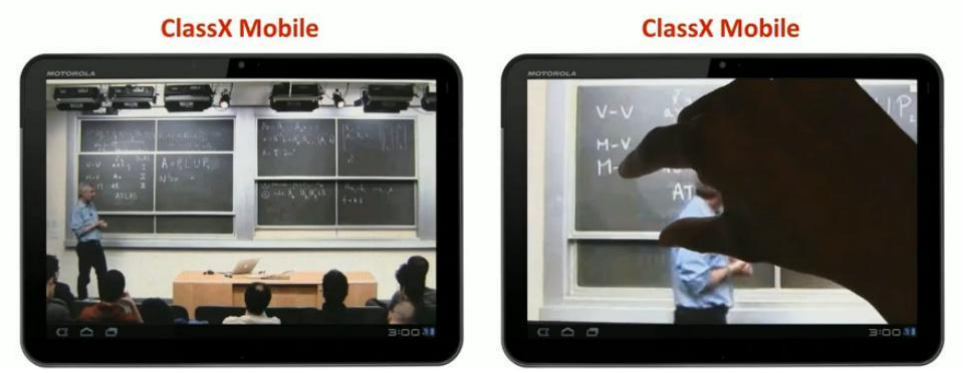

Figure 4: ROI feature from ClassX player

2.2.4 Medical imaging

Medical imaging devices generate high number of images with increasing resolutions. Usage of high compression rates associated with access to region of interest encoded with very high quality has been proposed in order to allow simultaneously easy access to images and correct diagnostic usage [11]

3 Requirements

The new HEVC standard provides tools in the main profile that can be used for spatial access: the slices and the tiles. With these tools it is possible to generate independently decodable region of interest (as described in [8]). While current MPD already provides tools for spatial resolution adaptation, it does not allow describing the addressing of spatial areas, i.e. tiles, in the videos. In order to enable this and to realize the above use cases, the following requirements are considered as important:

- Inform in the MPD that a video (tile) is a spatial part of a video (full-frame video)
- Describe each tile in terms of characteristics (e.g. relative position, bandwidth, tile resolution, grid resolution …).
- A composition engine should be able to use tiling information from the MPD. Tiling description should not restrict application developers to combine tiles.

4 Tile description in DASH/MPD

We propose below to use the *Role* element from the MPD syntax to fulfill the requirements we introduced in the previous section. We extend the *Role* element with a new scheme in order to describe tiled videos. Each tile is described as a *Representation* in its own *AdaptationSet* as shown on Figure 5 below. In the annex, we provide an MPD example with four tiles and full-frame video representations.

```
<AdaptationSet id='AS1' mimeType='video/mp4' codecs='hev1' >
  <Role schemeIdUri=”urn:mpeg:DASH:tiling:2013” id='TS1'  value =“0,0,960,540”/>
  <Representation width='960' height='540' frameRate='30' id='R1' bandwidth='64000'>
    <SegmentList duration='10'>
      <SegmentURL media='seg-tile1.mp4'/>
       …
    </SegmentList>
  </Representation>
</AdaptationSet>
```

Figure 5: Example of tile description in MPD

In the section for the “Specific scheme definitions”, the list of schemes defined in ISO/IEC 23009-1:2012 would be extended with the new scheme below:

Scheme Identifier	Clause in ISO/IEC 23009	Informative description
urn:mpeg:dash:mpd:tiling:2013	TBD	Tiling configuration as defined in a section TBD

Table 1: New scheme identifier for the Role element

The semantics for this new “ tiling Role” are defined in the table below:

Parameter	Description
id (optional)	Tiling *Roles* having the same @id value indicate that the associated *Representations, within a period*, are spatial tiles corresponding to the same video. It is assumed that all Representations, within a Period for which a tiling Role is present but with no ‘id’ attribute, are spatial tiles of the same video. NOTE: An adaptation set containing representations corresponding to the full, non-tiled video could contain a tiling role with the same id attribute as the adaptation set containing the representations corresponding to the associated tiles.
value (optional)	specifies a coma-separated list of 4 integers, in the following order: the x and y position of the top-left corner and the width and height of a tile. The coordinate system used to express the position of a

	tile is a 2D coordinate system, where the origin is arbitrary but shall be the same for all representations sharing the same id for the tiling role. Additionally, the x-axis is assumed to be oriented towards the right of the screen and the y-axis oriented towards the bottom. The width and height may be used by a viewer to render multiple tiles and align them spatially NOTE: an adaptation set containing representations corresponding to the full, non-tiled video could have x and y values set to the same values as the x and y values of the top-left tile and the width and height should be set such that it encompasses all tiles. When @value is not present, this means that the Representation corresponds to the full, non-tiled video whose width and height values can be get from <Representation> element.

Table 2: Semantic for the Role element for a scheme with a value "urn:mpeg:dash:mpd:tiling:2013"

It has to be noticed that no schema modification is required to support this new tile description since we rely on *DescriptorType* through a new *@schemeIdUri* value.

5 Conclusion

We would like DASH experts to consider the support of this new Role for tile description to enable spatial adaptation.

6 References

[1] ISO/IEC 23008-2 High Efficiency Video Coding
[2] ISO/IEC 14496-12 ISO base media file format
[3] ISO/IEC 14496-15 Carriage of NAL unit structured video in the ISO Base Media File Format
[4] ISO/IEC 23009-1 Dynamic Adaptive Streaming on HTTP: Media presentation description and segment formats
[5] "Canon EOS C500" http://cinemaeos.usa.canon.com/products.php?type=Camera-c500
[6] "London's digital Olympics". The Telegraph. http://www.telegraph.co.uk/technology/news/9433163/Londons-digital-Olympics.html
[7] Towards-Scalable-And-Interactive-Delivery-of-Immersive-Media, by Omar Niamut et al., NEM Summit Oct. 2012
[8] SEI Message: Independently decodable regions based on tiles, JCTVC-L0049, 12th JCT-VC meeting, January 2013
[9] "Panomera multifocal system" Dallmeier http://www.dallmeier-electronic.com/en/cctv-ip-video-surveillance/cameras/models/panomera.html

[10] A. Mavlankar, B. Girod, “Spatial-Random-Access-Enabled Video Coding for Interactive Virtual Pan/Tilt/Zoom Functionality,” IEEE Transactions on Circuits and Systems for Video Technology. vol. 21, no. 5, pp. 577-588, May 2011

[11] Gokturk, S. B., Tomasi, C., Girod, B., & Beaulieu, C. (2001). Medical image compression based on region of interest, with application to colon CT images. In *Engineering in Medicine and Biology Society, 2001. Proceedings of the 23rd Annual International Conference of the IEEE* (Vol. 3, pp. 2453-2456). IEEE.

[12] *“An iterative region of interest video streaming system for online lecture viewing*” by Mavlankar et al., in Packet Video 2010

7 Annex : Example of MPD for a video with 4 tiles

```
<?xml version=”1.0”?>
<MPD
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xmlns="urn:mpeg:DASH:schema:MPD:2011"
  xsi:schemaLocation="urn:mpeg:DASH:schema:MPD:2011 DASH-MPD.xsd"
  profiles="urn:mpeg:dash:profile:full:2011"
  minBufferTime="PT5.0S"
  type="static"
  mediaPresentationDuration="PT3256S">

 <BaseURL>http://www.example.com/server/Segments/</BaseURL>

 <Period start="PT0.00S" duration="PT3256S">
  <Subset contains=”AS1”/>
  <Subset contains=”AS2 AS3 AS4 AS5”/>
  <SegmentList duration='10'>
    <Initialization sourceURL='seg-init.mp4'/>
  </SegmentList>

<!—Description of full-frame video -->
  <AdaptationSet id=“AS1” mimeType=’video/mp4’ codecs=’hev1’ >
    <Role schemeIdUri=”urn:mpeg:DASH:role:2011” value=”main”/>
    <Representation id=’R0’ width=’1920’ height=’1080’ frameRate=’30’ bandwidth=’256000’>
      <SegmentList duration='10'>
      <SegmentURL media='seg-full-1.mp4'/>
       …
    </SegmentList>
    </Representation>
  </AdaptationSet>

<!—Description of first tile -->
<AdaptationSet id=“AS2” mimeType='video/mp4' codecs='hev1' >
  <Role schemeIdUri=”urn:mpeg:DASH:role:2011” value=”alternate”/>
  <Role schemeIdUri=”urn:mpeg:DASH:tiling:2013” id=‘TS1’  value =“ 0,0,960,540”/>
  <Representation width='960' height='540' frameRate='30' bandwidth='64000'>
    <SegmentList duration='10'>
      <SegmentURL media='seg-tile1-1.mp4'/>
       …
    </SegmentList>
  </Representation>
</AdaptationSet>

<!—Description of second tile -->
<AdaptationSet id=“AS3” mimeType='video/mp4' codecs='hev1' >
  <Role schemeIdUri=”urn:mpeg:DASH:role:2011” value=”alternate”/>
  <Role schemeIdUri=”urn:mpeg:DASH:tiling:2013” id=‘TS1’  value =“ 960,0,960,540”/>
  <Representation width='960' height='540' frameRate='30' bandwidth='64000'>
    <SegmentList duration='10'>
      <SegmentURL media='seg-tile2-1.mp4'/>
       …
    </SegmentList>
  </Representation>
</AdaptationSet>

<!—Description of third tile -->
<AdaptationSet id=“AS4” mimeType='video/mp4' codecs='hev1' >
  <Role schemeIdUri=”urn:mpeg:DASH:role:2011” value=”alternate”/>
  <Role schemeIdUri=”urn:mpeg:DASH:tiling:2013” id=‘TS1’  value =“0,540,960,540”/>
```

```
  <Representation width='960' height='540' frameRate='30' bandwidth='64000'>
    <SegmentList duration='10'>
      <SegmentURL media='seg-tile3-1.mp4'/>
       …
    </SegmentList>
  </Representation>
</AdaptationSet>

<!—Description of fourth tile -->
<AdaptationSet id="AS5" mimeType='video/mp4' codecs='hev1' >
  <Role schemeIdUri="urn:mpeg:DASH:role:2011" value="alternate"/>
  <Role schemeIdUri="urn:mpeg:DASH:tiling:2013" id='TS1'  value ="960,540,960,540"/>
  <Representation width='960' height='540' frameRate='30' bandwidth='64000'>
    <SegmentList duration='10'>
      <SegmentURL media='seg-tile4-1.mp4'/>
       …
    </SegmentList>
  </Representation>
</AdaptationSet>
</Period>
</MPD>
```

Figure 6: Example of MPD with 4 tiles and full-frame video representations

The above MPD example describes a HD video (Representation 'R0') that is composed of 4 spatial tiles of 960x540. Each tile has its own AdaptationSet and associated Representation plus a "tiling" Role. We have only one set of 4 composable tiles identified by "TS1". We defined 2 Subsets to indicate the alternative between the full-frame video in the 1st Subset and any tile or composition of tiles from the 2nd Subset.