Multimed Tools Appl (2017) 76:5631–5659
DOI 10.1007/s11042-016-4097-4

Video processing for panoramic streaming using HEVC and its scalable extensions

Y. Sánchez de la Fuente[1,2] · **R. Skupin**[1] · **T. Schierl**[1]

Received: 4 January 2016 / Revised: 27 September 2016 / Accepted: 27 October 2016 /
Published online: 1 December 2016

Abstract Panoramic streaming is a particular way of video streaming where an arbitrary Region-of-Interest (RoI) is transmitted from a high-spatial resolution video, i.e. a video covering a very "wide-angle" (much larger than the human field-of-view – e.g. 360°). Some transport schemes for panoramic video delivery have been proposed and demonstrated within the past decade, which allow users to navigate interactively within the high-resolution videos. With the recent advances of head mounted displays, consumers may soon have immersive and sufficiently convenient end devices at reach, which could lead to an increasing demand for panoramic video experiences. The solution proposed within this paper is built upon tile-based panoramic streaming, where users receive a set of tiles that match their RoI, and consists in a low-complexity compressed domain video processing technique for using H.265/HEVC and its scalable extensions (H.265/SHVC and H.265/MV-HEVC). The proposed technique generates a single video bitstream out of the selected tiles so that a single hardware decoder can be used. It overcomes the scalability issue of previous solutions not using tiles and the battery consumption issue inherent of tile-based panorama streaming, where multiple parallel software decoders are used. In addition, the described technique is capable of reducing peak streaming bitrate during changes of the RoI, which is crucial for allowing a truly immersive and low latency video experience. Besides, it makes it possible to use Open GOP structures without incurring any playback interruption at switching events, which provides a better compression efficiency compared to closed GOP structures.

Keywords Panoramic streaming · HEVC · SHVC · MV-HEVC · Tiles · Stitching

✉ Y. Sánchez de la Fuente
yago.sanchez@hhi.fraunhofer.de

[1] Multimedia Communications Group, Video Coding and Analytics Department, Fraunhofer HHI, Einsteinufer 37, 10587 Berlin, Germany

[2] Image Communication Group, Department of Telecommunication Systems, Technische Universität Berlin, 10587 Berlin, Germany

1 Introduction

Panoramic streaming is a specific case of video streaming, in which an arbitrary Region-of-Interest (RoI) of a high-spatial resolution video is transmitted. Users navigate within a wide-angle video (e.g. 180° or 360°) by choosing at any time the RoI they are interested in. Figure 1 shows an example of a high-resolution video, where the spatial plane of the panorama video is shown with two RoIs marked as blue rectangles. The two rectangles show two different RoIs displayed at two different time instants: namely at time t_0 (left rectangle) and at time t_1 (right rectangle) after user interaction, i.e. after a RoI switch event.

There are prototypes and deployed systems already showing panoramic streaming's feasibility [2, 27]. Taking into account that there exist techniques and commercial products that allow capturing 360° video in real time [28], by stitching multiple HD views from multiple cameras, and the recent market availability of plenty of head mounted consumer displays such as the Oculus Rift [6], Samsung Gear VR [7] or Google Cardboard [8], we envision that interactive panoramic streaming will be a popular application in a few years from now. In fact, consumers may soon have immersive and sufficiently convenient end devices and content at reach, which could lead to an increasing demand of panoramic video experiences.

Panorama (or very wide-angle) videos require a very high-spatial resolution in order to allow for a truly immersive experience. Although, there is no specification yet about the formats required for Virtual Reality (VR) and panoramic streaming, there is some discussion ongoing about the proper formats for such services in order to have a good QoE. In (http://dashif.org/wp-content/uploads/2015/08/4a-Harmonic-5G-Video.pdf) potential formats for VR are described. More concretely the formats described have a spatial resolution of 2560x1440 pixels at 60 or 120 fps or even 3840x2160 pixels at 120 fps for the section shown at the Head Mounted Displays (HMD), i.e. the RoI. Although current HMDs do not yet support ultra-high resolution formats, the trend points towards such a direction. Taking into account that the field-of-view of HMDs cover around 110°, a more than three times larger panorama video can be expected in the horizontal resolution. This accounts to an 8Kx2K or even higher-resolution video for the full panorama. Despite the better coding efficiency of new video codecs such as H.265/HEVC, there is a need of optimization in the way the panorama video is coded and transmitted in order to reduce the very high throughput that would be required if the whole panorama was transmitted. In fact, transmitting the whole panorama at full resolution would waste resources since only a smaller part of it, i.e. the RoI, is used at the receiver side.

The most basic approach to tackle the aforementioned problem consists of a fully client-server coupled system. In such a system, each user indicates to the panorama streaming system the desired RoI at any time and an encoder associated with each user encodes only the desired RoI. Then, the RoI is transmitted back to the user. Since having a dedicated encoder per user does not scale well, tile-based panoramic streaming (introduced by Mavlankar et al. in [12])

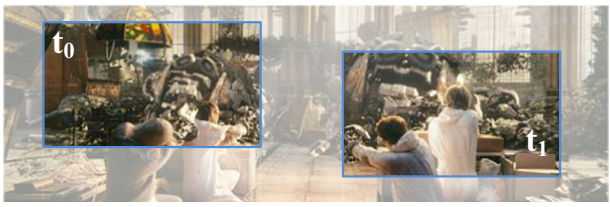

Fig. 1 RoI before (t_0) and after (t_1) user interaction

has drawn the attention of the research community and standardization bodies (see [10]). The main idea is to divide the panorama picture horizontally and vertically into smaller regions that are encoded independently. Then the regions that contain the content belonging to the RoI are transmitted to the user. Therefore, the number of streams encoded only depends on the granularity, with which the content has been tiled, and not on the number of users and trajectories chosen by them when making use of the interactivity.

However, while tiled streaming solves the scalability issue described before, it assumes instantiation of a decoder per received spatial region, typically using software decoders that have an impact on power consumption or real-time decoding capability. Power consumption is a very critical issue for head mounted displays based on mobile devices such as the case of Samsung Gear VR or Google Cardboard, since without power supply the battery life would be too short to be able to provide a satisfactory service. Additionally, UHD software decoding can be challenging in software, especially looking at the resolution requirements that are discussed in the field for VR. Therefore, even for HMDs with power supply such as Oculus Rift, hardware decoding is an important component of the end-to-end chain, since GPU decoding can be used. Note, for instance, that for PlayStation VR, where power consumption is not an issue, video decoding is done in hardware. Typically, devices only use a single hardware-accelerated decoder to achieve real-time decoding capabilities and save battery life. Therefore, in order to provide a good technical solution, it is a requirement that a single bitstream is provided to the end-device so that hardware decoding can be performed. An alternative to the solution presented within this paper is definitely to use multiple parallel decoders, which is currently not widely supported in hardware and seems not to be realistic for software decoding taking into account the upcoming trends in terms of resolutions and their impact for real-time decoding.

The goal of this paper is to provide a technique that makes use of tiled streaming for scalability, but allows using a single decoder for efficiency considerations, i.e. real-time decoding of high resolution videos and power efficiency. In addition, several aspects are investigated in order to provide an efficient way of encoding the panorama video.

The technique, described within this paper, overcomes the multiple decoding issue described above. It operates in the compressed domain and generates a single H.265/HEVC [26] bitstream (or H.265/SHVC [25] or H.265/MV-HEVC [25]) that can be fed into a single hardware decoder. The novel method processes the video bitstreams of the different tiles in the compressed domain, by performing a simple manipulation of some information in the header. This simple manipulation is key component for the system scalability. The proposed solution consist of merging coded video bitstreams into a common output bitstream by re-writing some high-level syntax of the bitstreams and inserting some potentially pre-encoded, inter-predicted pictures. This process is of a very low-complexity, which allows doing it either in the network with no scalability issue or at the clients prior to the decoding process. It simply requires that an application is available before the decoding to perform the described process. In any case, the presence of an application is required to take care of the streaming of the video and to serve as an interface to the display. An example of such applications is a DASH player written for a browser. With this respect, the described technique can be implemented at the file parsing level and has been standardized in [14].

In addition, several aspects that allow for a more efficient encoding and transmission of the data are discussed within this paper. One of the aspects proposed within the paper focuses on avoiding or reducing peak bitrates at RoI switch events, which are detrimental for low-latency

services such as panorama streaming. This is achieved by reducing the Random Access Points (RAPs) of the bitstreams. Reducing the number of RAPs has the clear benefit of reducing the overall bitrate of the bitstreams and the peak rates. However, there is always a trade-off between the coding efficiency and the availability to random access a stream when seeking or tuning-in when the services are broadcasted. This paper describes how the proposed technique enables usage of open GOP structures, which provide a higher coding efficiency than closed GOP structures, while maintaining seamless playback of the content during RoI switch events.

The remainder of this paper is organized as follows. Section 2 gives an overview of related work in the field. Section 3 describes the system considered for tile-based panoramic streaming. In section 4, the general concept of compressed domain bitstream processing is explained. Section 5 to Section 7 explain the techniques in detail and report on the experiments and results. Finally, the conclusion is shown in section 8.

2 Related work

As aforementioned, research has been carried out during past years on panorama streaming. In [4, 5], authors describe a real-time system for panoramic video that allows for zooming and panning. They present the whole transmission chain from panorama video generation to the interactive view presentation. For the (potentially zoomed) video, a pin-hole camera model is used to project the per-user camera view to a cylindrical panorama video.

Since it allows for not sending the whole panoramic video at high-resolution and it solves the scalability issue as mentioned above, tile-based streaming has drawn the attention of many researchers. When tile-based streaming is used, two factors have to be taken into account. On the one side, splitting the high-resolution content into different tiles leads to a reduction of the prediction efficiency, since a smaller part of the video can be used as reference for each of the tiles. On the other side, it allows for a better fit of the RoI. Clearly, the smaller the tiles the smaller is the amount of unnecessary pixels (pixel overhead) that are transmitted but also the lower is the compression efficiency of the transmitted pixels. On the contrary, spatially large tiles increase compression efficiency but also lead to the transmission of higher amount of additional data that is not part of the RoI, i.e. higher pixel overhead. The pixel overhead

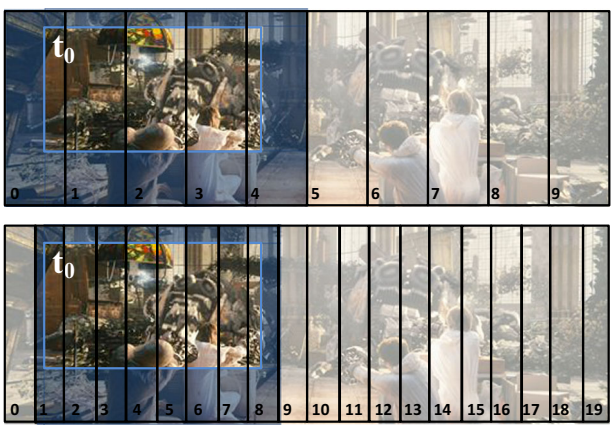

Fig. 2 Finer (*bottom*) and coarser (*top*) tiling and impact on pixel overhead

issue is illustrated in Fig. 2. Note that borders of the regions/tiles do not necessarily coincide with the RoI borders and therefore some extraneous data (not RoI) might be transmitted, depicted in blue in the figure. For a deeper insight on how to dimension tiles in an optimum way, the reader is referred to [13, 15], where authors present an optimization of the dimension of the tiles in which the panorama video is split.

When multiple decoders are used in parallel and several tiles are presented simultaneously together with a thumbnail of the panorama video at lower resolution, it might become an implementation drawback to synchronize the different videos. With this regard, authors in [9] propose to use H.264/MVC to have multiple parts (encoded into layers) of the video. Layers are encoded without coding dependency and are properly synchronized. One of the layers corresponds to the navigation video (video thumbnail). The rest of the layers (with a given view_id) correspond to the different tiles of the high-resolution panorama video offered at different bitrates. Thus, the users can select a set of view_ids that correspond to a given RoI at a given bitrate that matches the available throughput.

Similarly, some approaches that use some hierarchical prediction of a thumbnail view have been pursued in the past, see [13]. The authors use a similar procedure as H.264/SVC. However, in their solution, enhancement layers do not use temporal prediction but only inter-layer prediction. However, since H.264/SVC is not used, only proprietary software decoding can be used.

In [21] a system using multi-resolution panorama streaming is described. The main idea is to encode the panorama video into tiles and at different resolutions. Then the client downloads one or another resolution depending on the zoom factor. Based on the user interactivity, if the desired RoI is not available at the desired resolution, a lower resolution might be still obtained and is upscaled to the proper resolution for display. In [17], authors show the impact of using tiles in an interactive system that allows for panning and zooming in terms of overhead.

An alternative solution to tile-based streaming is described in [15], referred to as monolithic streaming. The idea is to encode videos as done conventionally, without splitting it into tiles. Then all macroblocks required for decoding a RoI are transmitted to the user. In order to do so, a dependency map has to be built for each RoI. This dependency map, determines all macroblocks required for decoding a given RoI. It follows the dependencies of each macroblock that corresponds to a given RoI. Additionally, the encoding process can be constrained in such a way that the dependency maps can be built with a reduced complexity. Note that if the encoder is not constraint, a full search needs to be performed to find all macroblocks of previous frames that a given RoI depends on. However, although the transmission bandwidth savings are similar to the tile-based approach, this approach has two main drawbacks. First, it has a higher complexity than the tile-based approach. But more important is the fact that the subset of macroblocks transmitted cannot be decoded by a standard-conform decoder, since they do not form a rectangular representation and therefore they do not form a valid bitstream.

There has been further work on tile-based streaming (e.g. [16]) that in contrast to [13] consider overlapping tiles. While in [13] all tiles are non-overlapping, in [16] tiles might have some overlapping content. The focus of such works lay on how to optimize the video transmission, for instance, by multicasting some of the tiles while transmitting others via unicast connections. However, we focus only on non-overlapping tiles, since our objective is to generate a single bitstream in the compress domain that contains the RoI of each user, which is not feasible with overlapping tiles.

3 Streaming system overview

The solution described within this paper consists of a low-complexity video processing technique that generates a single bitstream out of multiple independently encoded tiles. Besides, a key contribution of this paper is the way the content is generated so that the content is encoded efficiently. With this respect, the target of the proposed technique is to reduce the peak-bitrate during switching events, and to be able to use open GOP structures. The technique entails stitching the videos of the tiles belonging to the RoI and the insertion of Generated Reference Pictures (GRP), Multiview Generated Reference Pictures (MGRP) or Multi-Layer Generated Reference Pictures (ML-GRP) as described in sections 5, 6 and 7 respectively.

Figure 3 provides an overview of the considered system, which consists of an encoder array located at server side, a single video decoder instance at the client side and an Interactive Bitstream Stitching (IBS) device. The figure shows the case, where stereoscopic content is considered. After synchronous tiling of the two views of the panoramic video, individual H.265/MV-HEVC compliant Multiview video encoders create video bitstreams from every pair of two corresponding left and right view tiles. The user RoI, as illustrated with a blue rectangle, determines which pairs of tiles a user requires, e.g. the dashed tiles 0 to 3 in Fig. 3. The corresponding bitstreams are processed by the IBS device to form a single bitstream that can be decoded on the end device by a single MV-HEVC decoder instance.

Note that the mentioned H.265/MV-HEVC encoder array can be replaced by H.265/HEVC or H.265/SHVC depending on the use case considered, as seen later. Note also that the physical location of the IBS in the system can be either on server side, client side or within

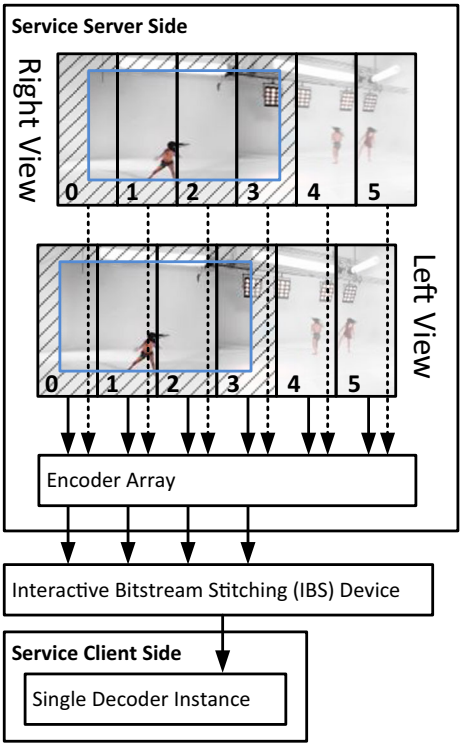

Fig. 3 System overview

the network, if the bitstreams are unencrypted and therefore accessible for modification. As the processing steps are of low complexity, usage of cloud infrastructure should scale well. Even low-end client devices would be able to handle the operation alongside decoding.

4 Compressed domain video processing overview

The proposed compressed domain video processing technique, IBS, is based on the stitching process described in [18] used to generate a single bitstream. Each of the independently encoded spatial regions belonging to the RoI is converted into an HEVC tile of a common bitstream. For this purpose, only adjustments to high-level syntax are required, which are lightweight to carry out.

First, the parameter sets need to be rewritten, mainly to reflect the spatial picture dimensions, level and tile setup of the RoI bitstream. Second, adjustments on slice level are necessary, e.g. slice addresses in the slice headers of the merged bitstream, which identify the HEVC tile that a slice belongs to. Each tile corresponds to the position of the independently encoded tiles within the merged RoI picture plane. Slice delta Quantization Parameters (QPs) might also need adjustment to reflect the common initial QP value as signaled in the rewritten parameter set.

Additionally, in order to be able to stitch the different bitstreams in the compressed domain, it is a requirement that the original videos are encoded according to a set of constraints:

- Motion vectors (MVs) cannot require samples that lie outside picture boundaries for temporal prediction.
- The rightmost Prediction Units (PU) cannot use the MV prediction candidate that corresponds to the Temporal MV Prediction (TMVP) candidate if it exists or that would correspond to the TMVP candidate if it existed.
- In-loop filters across slices and tiles have to be disabled.

The coding efficiency loss of these constraints was reported to be around 1 % [22] when applied to a single picture border and 3 % for all picture borders [18]. For more information the reader is referred to [18].

The technique described in [18] requires some extensions in order to be used for H.265/SHVC and H.265/MV-HEVC. In case of H.265/MV-HEVC it is a straightforward extension where MVs are constraint so that no sample that lies outside picture boundaries are used for inter-layer prediction. For H.265/SHVC this is not an issue since the MVs for inter-layer prediction are constrained to be equal to zero. However, in case of H.265/SHVC with spatial scalability, when inter-layer sample prediction is performed and the described technique is applied, the resampling process could use samples from multiple tiles. Therefore, the second constraint has to be rewritten as follows:

- The rightmost Prediction Units (PU) cannot use the MV prediction candidate that corresponds to the Temporal MV Prediction (TMVP) candidate if it exists or that would correspond to the TMVP candidate if it existed. Furthermore, the rightmost PU cannot use inter-layer prediction to a lower layer that is resampled.

An advantage of using H.265/SHVC is that users can download a larger region at the base than at the enhancement layer. The additional base layer data can be prefetched and presented

when a RoI change event occurs and the higher layer data is not available. Prefetching only the lower layers reduces the downloading throughput while enabling the data to be presented at a slightly lower quality. The described approach with a different number of tiles in an H.265/SHVC stream with two layers is illustrated in Fig. 4.

If a larger area is downloaded at the base layer, it is important that the IBS process adjusts signaling to indicate the area, which the enhancement layer uses as prediction. This is done by including the region reference offsets in the PPS.

The following sections describe the presented techniques in detail: namely GRP [19], ML-GRP [20] or MGRP [23]. The work in [19, 20, 23], has been extended by additional simulation results, i.e. a higher number of sequences have been tested. Besides, the work has been extended by combining GRPs with open GOP structures (see Section 5.1) and by extending the usage of ML-GRP for a full non-tiled base layer with unequal RAPs periods (see Section 7).

5 Generated reference pictures (GRP)

The technique presented hereafter aims to reduce the transmission bitrate during RoI switch events and was first introduced in [19]. The main idea behind it is to insert some pictures into the bitstream that are not output and perform a content displacement of reference pictures at occurrences of RoI switch events. Thus, temporal prediction can still be used for some parts of the video from the RoI switch event onwards.

Figure 5 illustrates a RoI switching event. Time instant t_1 represents the switching point at which the presented RoI changes compared to t_0. It can be seen that the position of the received spatial regions at the receiver screen (RoI view) changes over time (see dashed tiles 3 and 4 within the blue rectangles representing the RoI at different time instants).

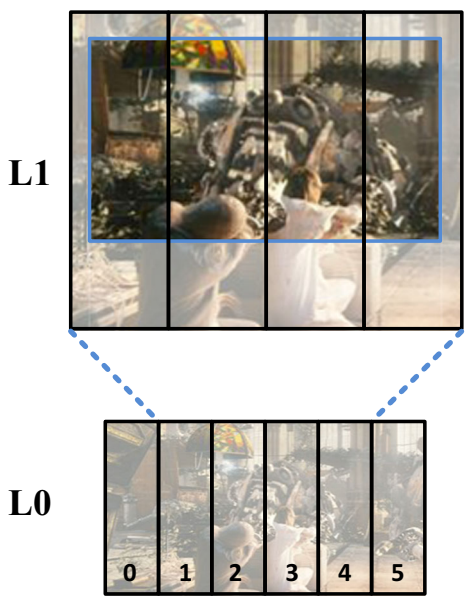

Fig. 4 Tile setup for 2 layers n H.265/SHVC with a larger amount of tiles in L0

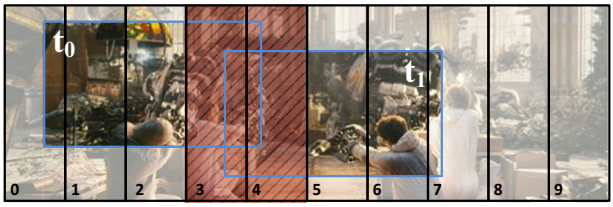

Fig. 5 Overlap of RoIs before user interaction (t0) and after (t1)

Obviously, new spatial regions (see non-shaded tiles with index 5 to 7 in Fig. 5) require random access, since they were not present in the RoI view previously. However, the set of tiles that remains displayed albeit displacement (tiles 3 and 4) would benefit from using temporal prediction. Since their position changes in the stitched picture based on the user movement, temporal prediction cannot be used in a straightforward manner. Figure 6 illustrates the effect of using temporal prediction for the stitched picture at time t_1 for a RoI switch event as depicted in Fig. 5. The figure shows spatial regions using random access depicted with an I (I slices) while spatial regions using temporal prediction are depicted with a P (P slices in this example). The MV for a block at the encoder side is shown at the top of the figure for the tile with index 3. The bottom part of the figure shows that after the stitching process, the block of the tile with index 3 uses a wrong reference at the decoder side, due to the change of the position of tile 3 within the stitched picture.

In order to avoid random access for all spatial regions, which would lead to large transmission bitrate peaks at switching points, we propose to insert Generated Reference Pictures (GRP): one per reference picture at the Decoded Picture Buffer (DPB). A GRP is a picture that performs a displacement of the content of a regular reference picture and substitutes it so that following pictures (from the RoI switching point onwards) can use temporal prediction.

Figure 7 shows how the block in tile 3 at t_1 uses the block in the GRP that has the same content as the block from t_0 (belonging to tile 3 as well) that is used at the encoder side. For more information on how to generate a GRP the reader is referred to [19].

In order to avoid any decoding drift, in addition to the constraints listed in Section 4, the original bitstreams have to fulfill the following constraint:

- Temporal Motion Vector Prediction (TMVP) has to be restricted so that no pictures that may be reference-wise substituted by GRPs are used for TMVP.

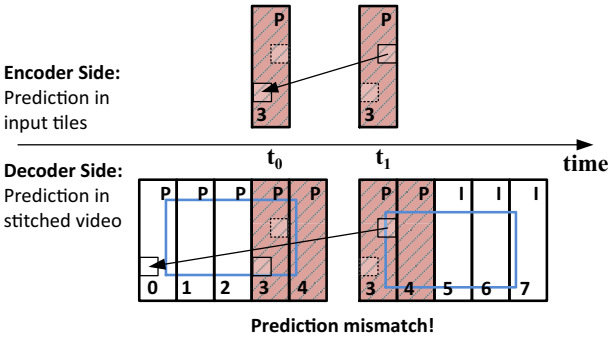

Fig. 6 Prediction mismatch after IBS without GRP

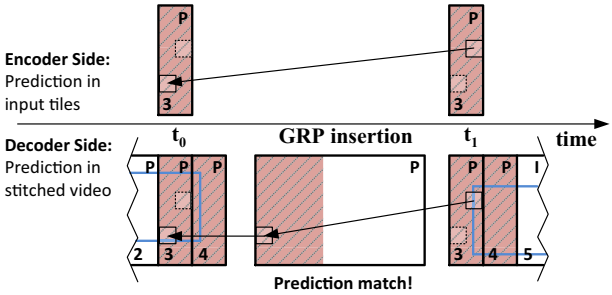

Fig. 7 Prediction match after IBS with GRP

Motion Vector (MV) prediction in H.265/HEVC [11] is performed from neighbor or temporal MV candidates. The latter (TMVP candidate) refers to the right-bottom collocated block in a reference picture. If the reference picture substituted by a GRP were used for TMVP, the TMVP predictors derived after GRP insertion would be wrong. Note that the derived TMVP predictor would belong to the GRP instead of to the substituted reference picture. Therefore, the constraint above must be fulfilled. An effective way to achieve it is to define switching points, at which the reference pictures at the Decoded Picture Buffer (DPB) are never selected for TVMP.

Figure 8 shows a typical hierarchical Group Of Pictures (GOP) structure of 4 pictures and three temporal levels (TL0, TL1 and TL2). The solid arrows in the figure represent the sample prediction and the dashed arrows represent the dependencies for MV prediction when TMVP is used. It can be seen that no TL0 picture is used for TVMP. Therefore, pictures from TL0 can be defined as switching points and can be used for RoI switching and GRP insertion.

In [19], GRPs were used so that only Random Access Points (RAPs) were only inserted for tiles that correspond to areas of the panoramic video that are new when a RoI switch event occurs. A drawback of the proposed solution is that for each potential RAP offered to the clients a new version of the content needs to be encoded and made available at the server. This is not practical in a real scenario. In a practical scenario, it is necessary to set Random Access Points (RAPs) with a given granularity. In the following subsection it is discussed how the solution in [19] can be extended to enhance the coding efficiency of streams using frequent RAPs.

5.1 Open GOP RoI switching with GRP

As aforementioned, only having RAPs at switching events would require encoding a separate stream per potential switching point, which would not be feasible. A solution is to offer a

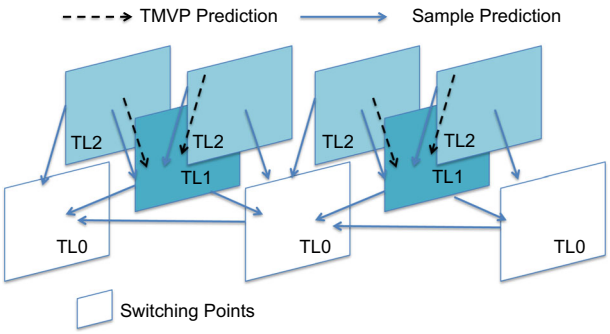

Fig. 8 Example of TMVP restriction

limited number of bitstreams with a given RAP interval, e.g. 5 seconds, and offer at each of alternative bitstreams RAPs at different positions. Thus, when required RAPs for new tiles could be obtained by switching from one of the alternative bitstreams to another where an RAP is available. I.e., RoI switching can be performed at a finer granularity than the RAP interval at the price of having multiple alternatives offered at the server side.

In H.265/HEVC, there are two possibilities to encode RAPs. The first one is to use Instantaneous Decoding Refresh (IDRs) as RAPs, which correspond to closed GOP structures. It means that any picture following a RAP (in decoding order) does not use any picture preceding the RAP (in decoding order) as a reference (see Random Access Decodable Leading – RASL – pictures in the figure). The second one is to use Clean Random Access (CRAs) as RAPs, which correspond to open GOP structures. Open GOP structures mean that pictures, following a RAP in decoding order, but preceding it in presentation order, can use pictures preceding the RAP in decoding order as reference (see prediction arrows and pictures marked as Random Access Skip Leading – RASL - in the figure). The prediction structure of both approaches is illustrated in Fig. 9, for a GOP of size 4.

It is well known that using open GOP structures with CRAs provide a better compression efficiency (around 5 % as reported in [3]). During the Random Access procedure using CRAs, some pictures, namely the RASL pictures in Fig. 9, must be discarded from output, since they require a previous picture (leftmost picture in the figure) that is not available at Random Access procedure. This is not an issue with traditional video services where skipping whole video pictures is only done when tuning-in into a service. However, it is a problem in context of tiled panoramic streaming, where switching from a RoI to another is considered. Such a switch must happen seamlessly without any interruption, i.e. without pictures being discarded or parts thereof corrupted.

If a Random Access procedure were started using CRAs for a newly encompassed tile of the new RoI, some pictures corresponding to that tile would not be able to be presented and would have to be discarded. This would prevent a seamless playout and thus open GOP structures. Therefore, it is necessary to use only closed GOP structures (i.e. IDRs) with the corresponding lower compression efficiency for the newly encompassed tiles. However, for the set of tiles that remains displayed in the RoI albeit displacement, the same issue as described above occurs. In a similar manner as described previously, GRPs can be used so that RASL pictures that use CRAs can reference the correct region when a RoI switch occurs, thus allowing for open GOP structures.

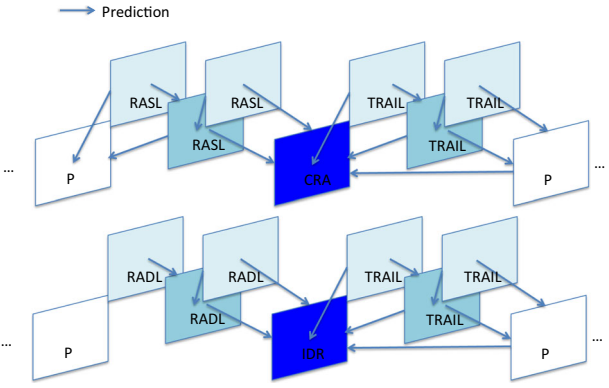

Fig. 9 Closed GOP (*bottom*) and Open GOP (*top*) structures in H.265/HEVC

5.2 Experiments with GRP

In the following, the experiments and results for the proposed techniques using GRPs are described. Four panorama videos (captured with [28]) have been used for the GRP experiments. Two panorama videos have a resolution of 8192x1600 pixels and the other two have a resolution of 6912x1920. All of them consist of 1425 frames at 25fps. All sequences correspond to fixed cameras. The first 2 sequences have low-motion content where several people play different instruments but stay at the same position for the whole sequence. The last 2 sequences have higher-motion with people dancing and moving around the stage, among which the first one (referred to as 3rd video in the results) has the highest motion. The HEVC reference software HM-14 was modified to include the constraints described before. The default random access configuration has been used with a GOP size of 4 and a single RAP at the beginning. The GOP size of 4, which corresponds to a 160 ms reordering delay at the decoding, has been chosen exemplarily. A higher GOP size could have been chosen in order to achieve a higher efficiency but this would increase the reordering delay, as well as the download time of GOPs due to its larger size. Although, it has not been analyzed in this work, a larger GOP size would require downloading a larger amount of non-displayed pixels to compensate for the higher end-to-end latency. In the end the selected GOP size is a compromise about encoding efficiency and additional data that needs to be downloaded (prefetch). Two tiling variants have been used to analyze the impact of a finer or coarser tiling process. The videos have been tiled only vertically, to limit parameter space, into spatial regions of 512 pixels width and spatial regions of 256 pixels width. These sizes have been exemplarily selected. The reader is referred to [13, 15] if interested in how to perform optimization of the tile sizes.

For the GRP experiment, the peak bitrate reduction during RoI switch events has been measured using the following metrics:

- B_w is the bitrate of the whole RoI sequence.
- B_r is the bitrate within the RoI change interval, time during which the user carries out the RoI movement.
- B_s is the bitrate of GOPs within the RoI change interval in which the tile setup changes.
- B_n is the bitrate of the remaining GOPs within the RoI change interval for which the tile setup does not change in comparison to the previous GOP.

B_s and $B_{n,}$ are of a high relevance since they describe the bitrate variability during RoI switching intervals. The motivation of such an analysis is of special interest when considering HTTP streaming techniques such MPEG-DASH [24]. In MPEG-DASH based streaming, media segments are transmitted. The larger the media segments the higher the end-to-end latency. In order to reduce the latency to the minimum the smallest possible segment would be selected and this is a single GOP. GOPs could be mapped to video segments, without requiring an IDR picture at segment start allowing individual transmission and thus reduction of the end-to-end latency of the video transmission. The bitrate peaks of interests are the ones that correspond to the downloaded chunks of data, i.e. the response to each of the requests that correspond to a GOP. In fact, it is crucial to minimize peak bitrates, since the downloaded quality selected by the client would follow the worst-case, which correspond to the biggest segments.

For these two cases, a simple interactivity model has been simulated. It is based on patterns with limited interactivity but sufficient to prove the validity of the proposed techniques. In fact, the proposed technique aims at reducing the RAPs within the session due to RoI interaction. Other movement patterns would lead to different regions requiring RAPs. However, the factor

that plays the most important role is the speed with which the user navigates around the panorama video. Therefore, two client screen movement patterns with different movement speeds have been considered. Both consist of a movement towards the right picture border: a constant movement at high or low speed for a given time interval of 1.2 seconds, referred to as switching interval in the experiments. These patterns were selected to be easily parameterizable. They correspond to a complete RoI change within 1.2s for the fast movement, since the considered client screen is 1080p. The slow movement pattern, on the other hand, results in a change of half of the RoI within the switching interval.

Figure 10 illustrates, for each tile dimension and movement speed, the frequency of switching point GOPs, i.e. GOPs where the tile setup changes and either full random access (IDRs) or GRP are used.

RoI switching has been performed at different times in the bitstream, every 248 frames and the average bitrates have been computed and are presented in subsection 5.3. For this purpose, several streams have been encoded. As aforementioned each of these streams is encoded with a single IDR at the beginning and no more RAPs within the stream. For tiles of width 256, the panorama videos have been encoded starting at frame number $240 + 4\,i + 248\,j$, where $0 \le i \le 8$ and $0 \le j \le 4$, until the end of the sequences. For the wider tiles, the starting frame corresponds to $240 + 8\,i + 248\,j$, where $0 \le i \le 5$ and $0 \le j \le 4$. With this encoding, the switching point GOP distribution in Fig. 10 can be achieved.

5.3 Results for GRPs

Figures 11 and 14 summarize the performance of GRPs as shown in [19]. They show the average transmission overhead during RoI switching intervals of 1.2 s of the full random access (RA) solution compared to usage of GRPs. The test sequences have been encoded with QP 22 and 32 to show the impact of the QP on the results.

Figure 11 shows the average overhead for a fast and slow screen movement on the left and right plot respectively. It shows that the faster the screen movement is or the higher the QP is, the higher is the overhead of RA compared to the usage of GRPs. Additionally, the smaller the spatial regions are, the larger is the gain of using GRPs, which is reasonable as the number of switching point GOPs is higher (see Fig. 11). In general, it can be seen that a transmission bitrate of around a 100–200 % higher for 256x1600 spatial regions or 50–100 % higher for 512x1600 regions can be expected if RA is used instead of GRPs.

Similar conclusions can be drawn from the results of the other three videos shown from Figs. 12, 13 and 14 with respect to the influence of the tile sizes and speed of the movement.

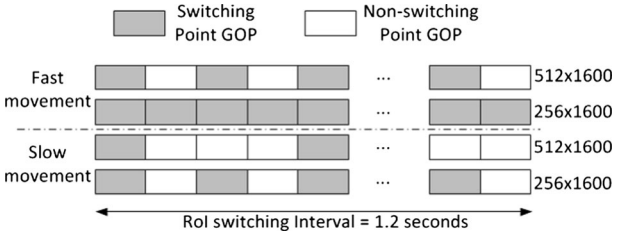

Fig. 10 Switching point GOP distribution

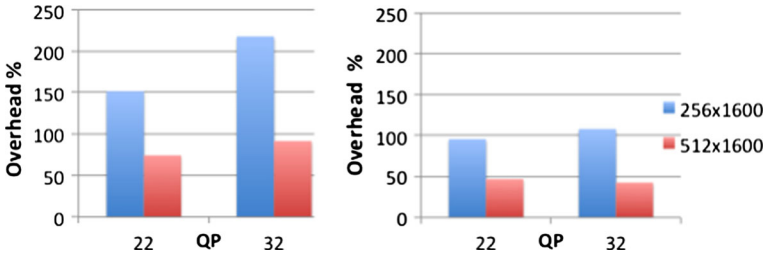

Fig. 11 Transmission overhead RA vs. GRP for fast movement (*left*) and slow movement (*right*) for 1st video

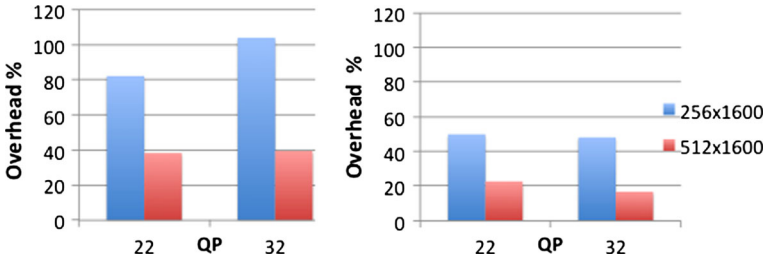

Fig. 12 Transmission overhead RA vs. GRP for fast movement (*left*) and slow movement (*right*) for 2nd video

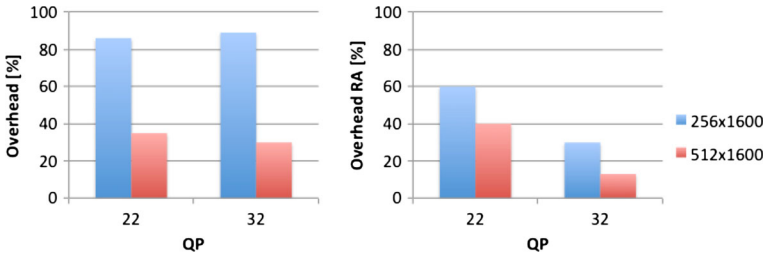

Fig. 13 Transmission overhead RA vs. GRP for fast movement (*left*) and slow movement (*right*) for 3rd video

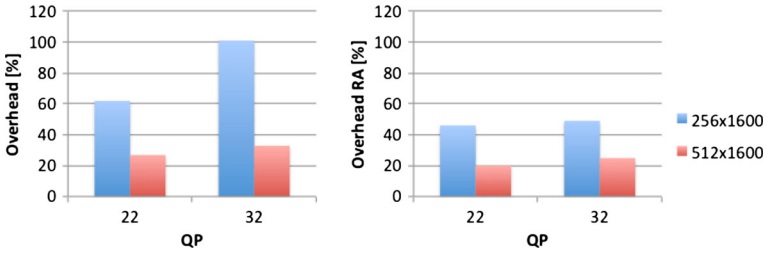

Fig. 14 Transmission overhead RA vs. GRP for fast movement (*left*) and slow movement (*right*) for 4th video

Table 1 Standard deviation for Br overhead of the 4 videos

Movement speed	Tile width	Video 1		Video 2		Video 3		Video 4	
		QP22	QP32	QP22	QP32	QP22	QP32	QP22	QP32
Fast	256	.09	.06	.21	.27	.39	.41	.09	.22
	512	.03	.03	.09	.09	.16	.14	.04	.08
Slow	256	.15	.04	.24	.20	.32	.17	.06	.09
	512	.06	.04	.11	.06	.15	.07	.04	.09

However, as shown in figure, the overhead is slightly smaller: 50–100 % for 256 pixels width spatial regions and 20–60 % for 512 pixels width regions. In any case, the gains provided by GRP are significant.

Table 1 shows the standard deviation of the Br overheads shown in Fig. 11, 12, 13 and 14. In general, it can be seen that the reported values are small, which means that the variability of Br overhead is not so high and the reported values in previous figures are of high reliability.

In order to analyze this issue in more detail, we focus on the absolute values of the transmitted bitrate, taking only the values for the second video and large tiles for brevity. Table 2 shows the average bitrates over the whole sequence B_w, over RoI switching interval B_r, over the switching point GOPs B_s and over the non-switching point GOPs B_n. It can be seen that most of highest bitrate values correspond to the switching point GOPs, for which the gain of GRP is even higher than the values discussed before.

Due to the lack of space the numbers for the other cases are not presented. However, very similar values have been obtained. In the case of slow movement, the transmission bitrate for the switching interval is slightly lower but still very similar values have been obtained for switching point GOPs. Overall, a reduction from around 1.9 up to 7.3 Mbps at the switching point GOPs was achieved for the studied sequences and QPs.

Table 2 Comparison of average bitrates (Mbps) for the 2nd video

	QP =32				QP = 22			
	B_w	B_r	B_s	B_n	B_w	B_r	B_s	B_n
RA	1.3	2.3	3.4	1.2	6.0	10.6	13.8	7.3
GRP	1.2	1.3	1.5	1.0	5.3	5.8	6.5	5.1

6 Multiview generated reference picture (MGRP)

Multiview Generated Reference Pictures (MGRPs) are a straightforward extension of GRPs for stereo video and were presented first in [23]. The only extension that needs to be mentioned in comparison to the GRP is that the Temporal Motion Vector Prediction (TMVP) restriction aforementioned can be relaxed if used for an inter-layer predicted picture. In other words, for the switching points previously defined (TL0 pictures) which are not used for TVMP candidates, this restriction only applies when temporal prediction is considered. For higher layers, a TL0 picture can use the TL0 picture from a lower layer for TMVP. This is due to the fact that replacement of that picture happens after decoding the given whole Access Unit and therefore is not affected by MGRPs. As described in Section 5.1, open GOP switching is facilitated through MGRP likewise.

6.1 Experiments for MGRP

Experiments for MGRP follow the setup of experiments in section 5.2 in spirit. However, a single stereoscopic video sequence has been used with a resolution of 5760 × 1664 pixels. The sequence consists of 1000 frames at 30fps. For this case, the video has been tiled into spatial regions of 640x1664 pixels and spatial regions of 320x1664 pixels and a GOP size of 8 has been chosen exemplarily, which corresponds to a 267 ms reordering delay at the decoding. This parameter selection corresponds to the common test conditions used by JCT-VC during standardization. Correspondingly, these parameters result in a complete RoI change in 1.6 s for the considered fast movement pattern. Therefore, the switching interval in this experiment is slightly larger compared to the experiment for GRP. RoI switching has been performed at different times in the bitstream, every 152 frames and the results are averaged over these instances. Similar to the experiments in Section 5.2, the test sequence has been encoded with QP 22 and 32 to show the impact of the QP on the results. As for the experiments in Section 5.2, several streams have been encoded. For tiles of width 320, the panorama video has been encoded starting at frame number $120 + 8i + 152j$, where $0 \le i \le 8$ and $0 \le j \le 5$. For 640 pixels width tiles, the starting frame corresponds to $120 + 16i + 152j$, where $0 \le i \le 5$ and $0 \le j \le 5$.

6.2 Results for MGRPs

This section summarizes the performance of MGRP as shown in [23]. Table 3 below provides the four bitrates B_w, B_r, B_s and B_n in kbps at QP 22. Results are summarized for both views, i.e. independent layer L0 and dependent layer L1. It can be seen that in particular the critical peak bitrates within the RoI change interval are considerably reduced using the proposed MGRP technique, which is sensible keeping in mind that much of the unnecessary intra-coded data is omitted from the bitstream.

Figure 15 visualizes the overhead (in percent) of the IDR-based RoI change with respect to the proposed MGRP solution, showing B_r on the left and B_s on the right. It can be seen that by not using MGRPs, around 60 % or 20 % more bits needs to be transmitted during the RoI switching interval for the fast movement for the finer and coarser tiles respectively. For the slow movement, a 40 % higher bitrate needs to be transmitted for the coarser tiling approach. It can be seen that the faster the movement the more effective is the MGRP approach in terms of saved bitrate. However, if we focus on the peak bitrate values, it can be seen how the gain obtained by the proposed approach does not depend so much on the moving speed. The

Table 3 Summary of bitrates in Mbps for QP 22

RoI change with	Movement speed	Tile width	B_w	B_r	B_s	B_n
IDR	Fast	320	2.45	4.73	4.73	0[a]
MGRP			2.39	2.87	2.87	0[a]
IDR		640	2.36	3.61	4.56	2.65
MGRP			2.32	2.80	3.10	2.5
IDR	Slow	320	2.52	3.83	4.91	2.75
MGRP			2.47	2.75	2.94	2.55
IDR		640	2.39	3.27	4.68	2.57
MGRP			2.36	2.69	3.17	2.46

[a] The RoI change interval in this particular case contains only GOPs in which RoI change events occur

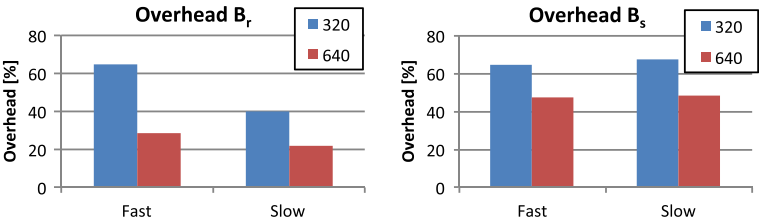

Fig. 15 IDR bitrate overhead in B_r and B_s for QP 22

Table 4 Summary of bitrates in Mbps for QP 32

RoI change with	Movement speed	Tile width	B_w	B_r	B_s	B_n
IDR	Fast	320	.484	1.12	1.12	0[a]
MGRP			.456	.574	.574	0[1]
IDR		640	.414	.726	1.02	.434
MGRP			.404	.523	.609	.437
IDR	Slow	320	.488	.825	1.17	.480
MGRP			.473	.542	.599	.485
IDR		640	.423	.639	1.06	.428
MGRP			.415	.498	.632	.431

[a] The RoI change interval in this particular case contains only GOPs in which RoI change events occur

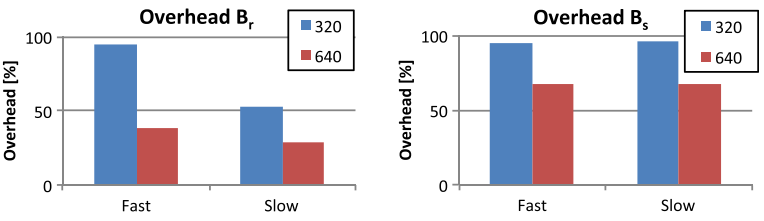

Fig. 16 IDR bitrate overhead in B_r and B_s for QP 32

Table 5 Standard deviation for Br overhead

Movement speed	Tile width	QP22	QP32
Fast	320	.077	.140
	640	.045	.071
Slow	320	.071	.122
	640	.032	.063

overhead of using IDRs instead of MGRPs based on the bitrate of the switching GOPs (B_r) is around 60 and 40 % for finer and coarser tiles respectively, irrespective of the moving speed.

Analogously, bitrate results for QP32 are reported in Table 4. Overall, the effect of MGRP at such low bitrate is even more pronounced in the relevant bitrates B_r and B_s, while B_n for example does not exhibit significant impact. As seen in Fig. 16, the relative overhead of IDR based RoI changing is even more significant at low quality.

Overall it can be seen that a denser tile grid, i.e. small tiles, benefits more from use the MGRP technique as they implicate more frequent RoI change event that introduce intra-coded pictures without MGRP. The overhead in B_r increases with the movement speed as more RoI change events occur while the overhead in B_s itself, i.e. during RoI change events, is mostly independent of the movement speed.

Table 5 shows the standard deviation of the overhead of Br when comparing the solution using IDRs with the proposed MGRP. It can be seen that the values are low, which means that the average Br overhead shown in Fig. 16 is of high reliability.

Of interest for the present stereoscopic use case with MV-HEVC is particularly the distribution of bitrate between the two views or layers as well as the effect of MGRP on each layer. Results of the per-layer IDR overhead relative to MGRP are reported in Table 6. It can be seen that the benefits from MGRP in the independent view, i.e. layer L0, are quality, tile grid and speed dependent over the switching interval (B_r), as for the single layer case. While overhead in B_r is in some respect weighted by the distribution of RoI change events (or speed) as illustrated in Fig. 10, it can be seen from the behavior of B_s, which omits this weighting, that the QP is the main determining factor for MGRP benefits on GOP level.

On the other hand, in the dependent view, i.e. layer L1, again focusing on B_s, the IDR overhead almost only depends on the tiling grid granularity and the benefits are independent of QP.

Table 6 IDR overhead percentage per layer

QP	Movement speed	Tile width	Overhead B_r	Overhead B_s		
			L0	L1	L0	L1
22	Fast	320	60.4	74.4	60.4	74.4
		640	27.4	32.0	44.5	53.0
	Slow	320	37.4	44.3	63.2	75.6
		640	20.6	23.7	45.7	53.0
32	Fast	320	105.0	74.2	105.0	74.2
		640	43.5	28.6	74.1	50.9
	Slow	320	57.4	41.6	105.0	74.5
		640	32.0	20.3	74.9	51.3

7 Multi-layer generated reference picture (ML-GRP)

Although GRP and MGRP techniques described above are sufficient to allow for using open GOP structures and thus reduce the transmitted bitrate in interactive panoramic streaming scenarios, they have two limitations. First, it is necessary to encode and store at the server side two different versions of the content: one using open GOP structures and one using closed GOP structures. Second, the faster the movement, the less effective is the usage of GRPs and MGRPs. This is due to the fact that for fast movements, it is more probable that the RoI after a RoI switching event is a fully new RoI or a RoI with very few content present both before the switch event and after.

H.265/SHVC allows for an enhancement of the described technique that overcomes the aforementioned limitations, as explained below. Additionally, H.265/SHVC provides a good alternative for efficiently coping with fast RoI movements. Note, that it is very inefficient to offer the content encoded with very frequent RAPs. Therefore, RAPs are usually offered at a given granularity, e.g. every 1 second. For interactive panoramic streaming the RAP frequency might be higher (RAP interval lower than 1 second) but still not so frequent that at each picture a new tile can be random accessed. This implies that a user has to compensate the RAP granularity by downloading additional data that is not intended to be shown. This additional data acts as a backup in case respective user interaction shifts it into the RoI. This additional data downloaded only as backup can be offered in a lower resolution when using H.265/SHVC as shown in Fig. 17, where tiles with index 1 to 4 (intended to be shown in the user RoI) are encoded at the highest resolution, i.e. at the higher layer, while tiles with index 0 and 5 are only included in the base layer in case they are required during a RoI switch event.

The extension of GRPs described here and first proposed in [20], allows for using open GOP structures even for content that was not present at the highest resolution but present in the low-resolution layer. The proposed solution consists in inserting so-called Multi Layer Generated Reference Picture (ML-GRP). The main motivation is shown in Fig. 17, where a RoI switch event is depicted. The RoI change represents a move to the right by an amount equal to the width of a tile. As can be seen in the figure, there is a tile that is new at the enhancement layer (see tile with index 5 marked in red), which has its corresponding base layer tile present in previous pictures in the stream.

Multi-Layer Generated Reference Pictures (ML-GRPs) aims at exploiting this fact and using that data available at the base layer so that open GOP switching at RoI switch events can be used at the enhancement layers. ML-GRPs consist of several tiles, which contain the GRP information, i.e. movement information to compensate the RoI change event. I.e., as for GRP, tiles that are presented before the RoI switch event and after need to the shifted. Besides, for newly encompassed tiles an ML-GRP contains a copy slice for those tiles that reference to a lower layer, which leads to inheritance of the sample values from the correct region of the lower layer. Since the reference layer is at a lower resolution, the process is not a simple sample copy but entails a resampling process as defined in H.265/SHVC. It requires indication of the scaled reference layer offsets and referenced region offsets defined in the PPS that indicate, which region of the base layer is to be upsampled. The resulting L1 ML-GRP picture area can then be used as reference by the RASL picture as illustrate in Fig. 18. Depending on the quality of the ML-GRP used as reference by the RASL picture, no noticeable or only minor decoding drift may occur, despite significant coding efficiency gains.

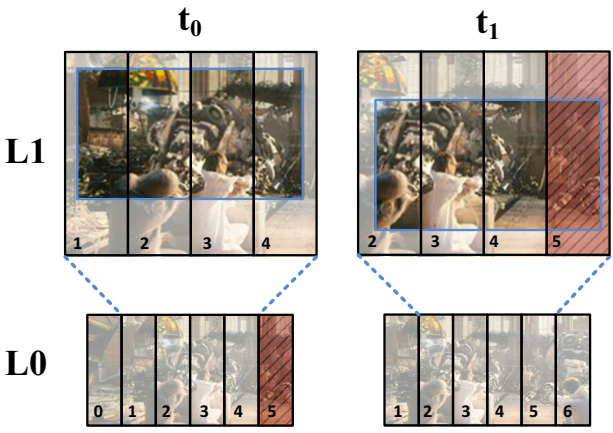

Fig. 17 Tile setup during a RoI change to the right

ML-GRP are inserted into the bitstream only at RoI change events and only for reference by following pictures, i.e. ML-GRP are not output by the decoder, as for the GRPs. For more information the reader is referred to [20].

An issue of the solution provided in [20] is that when the potential movement speed is very high the whole base layer will be needed and tiles at the base layer will add an unnecessary overhead. In this paper, the technique in [20] is extended so that only the enhancement layer is tiled and the whole base layer is always transmitted. In order to achieve a higher coding efficiency, it is combined with unequal RAP periods. H.265/SHVC allows, besides having different types of RAPs at different layers as discussed above, having a different RAP periodicity. This is of especial interest when the base layer at low resolution contains the whole panoramic video. In fact, RAPs periodicity at the base layer can be of a much coarser granularity than at higher layers, since no random access to new regions is required, as the whole panoramic video is always present. Still, the RAPs at the enhancement layer needs to be frequent enough (as frequent as in previous cases) to allow for random access of new tiles at highest resolution.

The ML-GRP is extended to so that a non-tiled whole base layer is used in combination with unequal RAP periods. Thus, open GOP switching is available at the base layer and enhancement layer as well. In comparison to the ML-GRP technique presented in [20], the constraint that the rightmost PU cannot use inter-layer prediction to a lower layer that is resampled mentioned in Section 4 is removed. This leads to a better compression efficiency

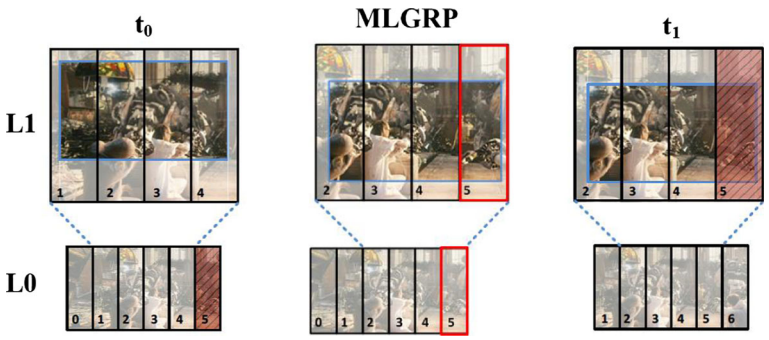

Fig. 18 MLGRP with open GOP in tile-based panorama streaming

than for the ML-GRP encoding described in [20]. The results corresponding to this new feature are presented in Section 7.2.

7.1 Experiments for ML-GRPs

The experiments carried out for ML-GRPs are again similar in fashion to the experiments carried out for GRP. Particular differences in terms of parameters or test sequences are described hereafter. Results for this experiment are reported applying the Bjontegaard metric [1], aka. BD-rate.

In order to analyze the benefits of using ML-GRP, the same four sequences as in Section 5.2 have been used. They have been encoded with a GOP size of 8 and at four different QPs: 22, 27, 32 and 37, which corresponds to the common test conditions using in standardization in JCT-VC. A single vertical tiling variant of 512 pixels width at the enhancement layer has been considered. The vertical resolution is 1600 for the first 2 videos and is 1920 for the last 2 videos. The sequences have been encoded with 2 layers and a layer configuration that corresponds to an upsampling factor of 2. The base layer in section 7.2 consists of tiles of 256x800 and 256x960 pixels respectively. For these experiments, the base layer covers a larger area than the enhancement layer. More concretely, there are 2 more tiles at the base layer than at the enhancement layer, as shown in Fig. 17. If the RoI is not the leftmost or rightmost possible position, i.e. the RoI corresponds to the leftmost or rightmost position, the two additional tiles correspond to the side at which further tiles are available, i.e. right and left respectively. As in the previous experiments, these sizes have been exemplarily selected without optimization. In case of the results for section 7.3, the whole base layer is always present and, therefore, the whole base layer is encoded as a single tile.

A RoI of 1080p has been targeted, which means that the enhancement layer of the bitstream, produced by the IBS, is conformed by 4 tiles. Streams have been produced for RoIs at different positions. This leads to 13 bitstreams with a size of 2048x1600 pixels or 10 streams of 2048x1920 pixels. Bitstreams contain tiles (i, i + 1, i + 2 and i + 3) with i = 0…12 for the first two videos or with i = 0..9 for the last two videos.

7.2 Results for ML-GRPs

Although the quality of the Random Access Skip Leading (RASL) pictures when using the ML-GRPs has not been subjectively measured, our assessment by looking at the resulting videos is that the technique performs very good. In Fig. 19, a small part of the first picture that belongs to a new tile after a RoI change event is shown for comparison. The left picture

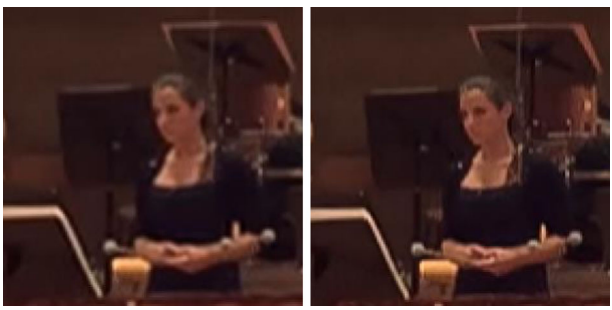

Fig. 19 Example of quality comparison of proposed method

Table 7 BD-rate comparison using the first sequence and GOP size 8 and 512x1600 size tiles

Leftmost tile index for RoI	BD-rate (%) of proposed method
0	3.267
1	3.506
2	3.707
3	3.724
4	3.859
5	3.909
6	3.919
7	4.062
8	4.061
9	4.028
10	3.904
11	3.538
12	2.882

corresponds to the proposed method when the RASL picture uses ML-GRP as reference and the right picture shows the same are of the picture when closed GOP is used. That is, the figure on the left is a part of a picture using the ML-GRP as reference instead of its original reference picture (which is shown on the right).

It can be seen that the quality of the picture when the proposed method is used is slightly lower but still very good. This slightly lower quality is the result of using another reference that entails upsampling of part of the base layer for being used as reference. It can be argued that such a low quality degradation does not affect the streaming experience, especially because it only happens for a few leading pictures when RoI change events occur.

The Random Access Point period used for the results shown below is 24 pictures, which corresponds to around 1 second. As a reference, we use the setup with the base layer and

Table 8 BD-rate comparison using the second sequence and GOP size 8 and 512x1600 size tiles

Leftmost tile index for RoI	BD-rate (%) of proposed method
0	8.867
1	8.846
2	9.051
3	7.569
4	7.857
5	7.792
6	6.773
7	6.042
8	6.712
9	6.625
10	7.582
11	9.591
12	9.421

Table 9 BD-rate comparison using the third sequence and GOP size 8 and 512x1920 size tiles

Leftmost tile index for RoI	BD-rate (%) of proposed method
0	5.561
1	5.633
2	4.362
3	4.293
4	4.606
5	4.123
6	5.665
7	5.995
8	6.280
9	7.188

enhancement layer RAPs using a closed GOP structure, i.e. IDRs. We compare it with the case where the base layer RAPs have been encoded as IDRs, while enhancement layer RAPs have been encoded as CRAs with an open GOP structure. The latter is the one used in combination with the ML-GRPs. The BD-rate [1] of all the bitstreams described above when using closed GOP structures vs. the proposed solution has been measured. The ML-GRPs show to provide an average BD-rate reduction of around 3 to 10 %, in comparison to using closed GOP structures without the proposed ML-GRP insertion. Results in [20] (shown in Tables 7 and 8) have been extended with new sequences, as shown in Tables 9 and 10.

Table 7 shows the BD-rate values of the proposed technique against the solution using closed GOP for the resulting 13 streams described above at different RoIs of the first 8192x1600 video sequence with 512x1600 tiles.

On average a 3.72 % BD rate reduction has been measured for the first sequence. Table 8 shows the BD-rate values of the proposed technique against the solution using closed GOP for the resulting streams at different RoIs of the second 8192x1600 video with 512x1600 tiles. The results shown for the second 8192x1600 stream are much higher: between 6.0 and 9.6 % approximately. For this sequence, a 7.902 % BD rate reduction on average can be seen when using the ML-GRPs.

Table 10 BD-rate comparison using the fourth sequence and GOP size 8 and 512x1920 size tiles

Leftmost tile index for RoI	BD-rate (%) of proposed method
0	8.249
1	9.117
2	9.171
3	7.817
4	7.263
5	7.149
6	6.793
7	7.638
8	8.151
9	7.985

Table 9 shows the BD-rate values of the proposed technique against the solution using closed GOP for the resulting streams at different RoIs of the first 6984x1920 video sequence with 512x1920 tiles. The results shown for this video are between 4.1 and 7.2 % approximately. For this sequence, an average of 5.371 % BD rate reduction can be seen when using the ML-GRPs

Table 10 shows the results for the last video sequence of a resolution of 6984x1920 pixels. The BD-rate values shown in this table correspond to higher values than for the previous sequence corresponding to an average of 7.93 %.

In any case, it can be seen that the reported gains are content dependent. For some sequences, gains up to 9 % can be obtain while for other gains up to 4 %. This depends on the temporal correlation of the pictures within a video sequence. The more temporal prediction can be used for the RASL pictures from previous pictures in presentation order the more effective the technique is. However, the gains shown between 3 % and almost 10 % show that the proposed technique can achieve a higher compression efficiency.

7.3 Results for RoI switching with ML-GRP and unequal RAP period

The results in Tables 11, 12, 13 and 14 show the benefits of using ML-GRPs for open GOP switching with unequal RAP periods. The same four sequences as in Sections 5.2 and 7.2 have been used. The Random Access Point period used for the results shown in Tables 11–14 is 24 pictures for the enhancement layer and 48 or 96 pictures for the enhancement layer. As a reference, we use the setup with the enhancement layer RAPs using a closed GOP structure, i.e. IDRs. We compare it with the case where the enhancement layer RAPs have been encoded as CRAs with an open GOP structure. For both cases the base layer is encoded using an open GOP configuration, i.e. encoded with RAPs of type CRA. As already mentioned, for this case the base layer has not been tiled and therefore consist of a single tile.

The results shown in Tables 11–14 show that the gains of ML-GRP do not almost vary with respect to the periodicity of the RAPs in the base layer. In general, ML-GRP with unequal RAP shows a gain of around 4 % up to 10 % in BD-rate. However, a RAP period of 96 provides between 8.5 % and 20.1 % bitrate savings in comparison to a RAP period of 48 in the base layer.

Table 11 BD-rate comparison using the first sequence and GOP size 8 and 512x1600 size tiles

Leftmost tile index for RoI	BD-rate (%) for RAP period 48	BD-rate (%) for RAP period 96
0	6.48	6.74
1	6.97	7.25
2	7.92	8.13
3	8.67	8.69
4	9.31	9.33
5	9.66	9.73
6	9.68	9.87
7	10.13	10.16
8	10.11	10.23
9	9.78	10.20

Table 12 BD-rate comparison using the second sequence and GOP size 8 and 512x1600 size tiles

Leftmost tile index for RoI	BD-rate (%) for RAP period 48	BD-rate (%) for RAP period 96
0	7.39	7.25
1	7.47	8.01
2	7.67	8.32
3	8.11	8.68
4	8.42	8.00
5	8.59	7.89
6	8.44	7.89
7	8.27	7.89
8	8.22	7.87
9	8.50	8.05

Table 13 BD-rate comparison using the third sequence and GOP size 8 and 512x1920 size tiles

Leftmost tile index for RoI	BD-rate (%) for RAP period 48	BD-rate (%) for RAP period 96
0	4.80	4.95
1	4.18	4.08
2	3.72	3.72
3	3.52	3.61
4	3.32	3.59
5	3.46	3.72
6	3.83	4.11
7	4.39	4.41
8	4.62	4.58
9	4.91	4.98

Table 14 BD-rate comparison using the fourth sequence and GOP size 8 and 512x1920 size tiles

Leftmost tile index for RoI	BD-rate (%) for RAP period 48	BD-rate (%) for RAP period 96
0	5.72	5.73
1	5.83	5.84
2	5.94	5.95
3	5.90	6.03
4	4.32	4.54
5	4.38	4.39
6	4.20	4.37
7	5.10	5.18
8	5.23	5.27
9	5.26	5.30

8 Conclusion

In this paper, we have proposed a technique to perform stitching of multiple H.265/HEVC bitstreams of a tiled panoramic video, so that devices with a single hardware decoder can be used. We have described how this works for streams encoded with H.265/HEVC, as well as for streams encoded with its scalable versions H.265/SHVC and H.265/MV-HEVC. We provide a solution that reduces the transmission bitrate at RoI switching points significantly. It consists of inserting Generated Reference Pictures (GRP) that allows using temporal prediction even at RoI switching points for some spatial regions instead of requiring full random access for the single layer case.

We showed that transmission peak bitrate savings between 80-200 % or 40-100 % during the RoI moving interval can be achieved depending on the video content and movement speed. Such bitrate savings are very beneficial for switching events where a drastic increase in the bitrate can lead to a big delay that can make an interactive streaming system unfeasible.

Since we consider that the success of head mounted displays is imminent and this is going to come along with stereo panorama streaming, tile-based panorama streaming for stereoscopic panorama video has been analyzed. Therefore we have used H.265/MV-HEVC and have extended the GRP technique towards stereo with MGRPs. Results are reported and the effect of MGRP is analyzed in context of a layered codec scenario. Results show that the MGRP technique allows for significant peak bitrate reduction, which can be a key in future panoramic video services with particularly strict latency requirements such as head mounted displays.

Such techniques can be implemented very easily without any concerns about real-time constraints since are of low complexity. They can be implemented in an application before feeding the decoder. For instance, they can be performed at the demultiplexing stage as discussed in the paper. In fact, such a technique can be implemented easily following the ISO base Media File Format [14]. One of the issues of GRPs and ML-GRPs is that they require different versions of the content at the server side, increasing the storage capacity that would be required, for instance, at CDNs.

In order to overcome this issue H.265/SHVC can be used. With this regard, we present Multi-Layer Generated Reference Pictures (ML-GRP). An ML-GRP allows using open GOP encoding structures, instead of requiring usage of closed GOP structures. We show that with the proposed solution the average BD-rate performance can be improved by around 3 to 10 %. Such bitrate savings happen mainly at random access points, which are very beneficial for switching events where a drastic increase in the bitrate can lead to a big delay that can make an interactive streaming system unfeasible.

References

1. Bjøntegaard G (2001) Calculation of average PSNR differences between RD-curves. Proc VCEG-M33 Meeting: 1–4
2. ClassX, Stanford University (2015) http://classx.stanford.edu/. Accessed 18 December 2015
3. Fujibayashi A, Tan TK (NTT DOCOMO) (2011) Random access support for HEVC. JCTVC-D234, 4th JCT-VC Meeting, Daegu
4. Gaddam VR, Langseth R, Ljødal S, Gurdjos P, Charvillat V, Griwodz C, Halvorsen P (2014) Interactive zoom and panning from live panoramic video. Proc Netw Operating Syst Support Digit Audio Video Workshop (p. 19). ACM
5. Gaddam VR, Langseth R, Stensland HK, Gurdjos P, Charvillat V, Griwodz C, Halvorsen P (2014, March) Be your own cameraman: Real-time support for zooming and panning into stored and live panoramic video. Proc 5th ACM Multimed Syst Conf (68–171). ACM
6. http://www.oculus.com/rift/. Accessed 18 December 2015
7. http://www.samsung.com/global/microsite/gearvr/. Accessed 18 December 2015
8. https://www.google.com/get/cardboard/. Accessed 18 December 2015
9. Inoue M, Kimata H, Fukazawa K, Matsuura N (2010) Interactive panoramic video streaming system over restricted bandwidth network. Proc Int Conf Multimed (1191-1194). ACM
10. ISO/IEC 23009-1:2014/Amd 2:(2015) Spatial relationship description, generalized URL parameters and other extensions
11. Lin J, Chen Y, Huang Y, Lei S (2013) Motion vector coding techniques for HEVC. IEEE J Sel Top Sign Proces 7(6):957–968
12. Mavlankar A, Agrawal P, Pang D, Halwa S, Cheung N, Girod B (2010) An interactive region-of-interest video streaming system for online lecture viewing. special session on advanced interactive multimedia streaming. Proc 18th Int Packet Video Workshop (PV), Hong Kong
13. Mavlankar A, Baccichet P, Varodayan D, Girod B (2007) Optimal slice size for streaming regions of high resolution video with virtual Pan/Tilt/Zoom functionality. Proc 15th Europ Sign Process Conf (EUSIPCO), Poznan, Poland
14. MPEG#115, Text of ISO/IEC FDIS 14496-15 4th edition, W16169
15. Quang N, Ravindra G, Ooi WT (2011) Adaptive encoding of zoomable video streams based on user access pattern. MMSys 2011, San Jose
16. Ravindra G, Ooi WT (2012) On tile assignment for region-of-interest video streaming in a wireless LAN. Proc NOSSDAV '12, 22nd Int Workshop Netw Operat Syst Support Digit Audio Video, Toronto, Canada
17. Reddy Gaddam V, Ngo HB, Langseth R, Griwodz C, Johansen D, Halvorsen P (2015) Tiling of panorama video for interactive virtual cameras: overheads and potential bandwidth requirement reduction. Pict Coding Sym (PCS), 2015 (204-209). IEEE
18. Sanchéz Y, Globisch R, Schierl T, Wiegand T (2014) Low complexity cloud-video-mixing using HEVC. Proceedings of IEEE Consumer Communications and Networking Conference, Las Vegas
19. Sanchéz Y, Skupin R, Schierl T (2015) Compressed domain video processing for tile based panoramic streaming using HEVC. Imag Process (ICIP), 2015 22nd IEEE Int Conf. IEEE
20. Sanchéz Y, Skupin R, Schierl T (2015) Compressed domain video processing for tile based panoramic streaming using SHVC. Proc Immersive Med Exper 2015 Workshop - ACM Multimed, Brisbane, Australia
21. Seo D, Kim S, Park H, Ko H (2014) Real-time panoramic video streaming system with overlaid interface concept for social media. Multimedia Systems 20(6):707–719
22. Skupin R, Sanchez Y, Schierl T (2015) Compressed domain video compositing with HEVC. Picture Coding Sym (PCS 2015), Cairns, Australia
23. Skupin R, Sanchez Y, Schierl T (2015) Compressed domain processing for stereoscopic tile based panorama streaming using MV-HEVC. Proc IEEE Int Conf Consumer Electron-Berlin (ICCE-Berlin), Berlin, Germany
24. Stockhammer T (2011) Dynamic adaptive streaming over HTTP–: standards and design principles. Proc Second Ann ACM Conf Multimed Syst. ACM
25. Sullivan GJ, Boyce JM, Chen Y, Ohm J-R, Segall CA, Vetro A (2013) Standardized extensions of high efficiency video coding (HEVC). Select Topics Sign Process, IEEE J 7(6):1001–1016
26. Sullivan GJ, Ohm J-R, Han W-J, Wiegand T (2012) Overview of the High Efficiency Video Coding (HEVC) standard. IEEE Trans Circ Syst Video Technol, December 2012, Best Paper Award
27. van Brandenburg R, Niamult O, Prins M, Stokking H (2011) Spatial segmentation for immersive media delivery, 15th International conference on Intelligence in Next Generation Networks (ICIN), Oct. 2011
28. Weissig C, Schreer O, Eisert P, Kauff P (2012) The ultimate immersive experience: panoramic 3D video acquisition. Adv Multimed Model, Lect NotesComput Sci 7131:671–681

Yago Sánchez de la Fuente received his MSc.-Ing. degree in telecommunications engineering from Tecnun-Universidad de Navarra, Spain, in September 2009. In 2008–2009 he carried out his master-thesis on P2P application layer multicast using SVC at the Fraunhofer HHI. He is currently working as a researcher in the Image Communication Group of Prof. Thomas Wiegand at the Technische Universität Berlin and is a guest researcher at Fraunhofer HHI. His research interests include adaptive streaming services for IPTV and OTT services. One of the main topics he is currently working on is adaptive streaming over mobile networks, in particular LTE networks, focusing on application layer and resource management on mobile networks. In relation to this, he steered the work of Fraunhofer HHI in the ESA project COSAT, focus on satellite backhauling for LTE, e.g. for O3M. He has been also working on compressed domain video processing of HEVC and SHVC. Furthermore, he has been project manager responsible for the FP7 OPTIBAND Project, which he also contributed to technically. Currently, Yago is working on adaptive streaming services for Mobile TV over LTE networks and Virtual Reality 360° video streaming. In addition, Yago is also actively participating in and contributing to standardization committees such as 3GPP, IETF and MPEG. Yago is the co-author of the IETF RTP Payload Format for H.265/HEVC Video.

In 2013, Yago was visiting the End2End Mobile Video Research group of Alcatel Lucent-Bell Labs, USA, where he was doing research on mobile video delivery optimization for low delay HTTP streaming.

Robert Skupin graduated with a Dipl.-Ing. (FH) degree in Electrical Engineering from BRSU in Sankt Augustin, Germany in 2009 and received his MSc degree in Computer Engineering from Technical University of Berlin (TUB), Germany in 2014. Since 2009, he is with the Multimedia Communications Group in the Video Coding and Analytics Department of Dr. Thomas Schierl and Dr. Detlev Marpe at the Fraunhofer Heinrich Hertz Institute (HHI), Berlin, Germany. His research interests lie in the area of video data coding and transport. Currently, his work is focused on techniques towards high quality 360° video services with HEVC and beyond. Robert is an active participant in standardization activities in JCT-VC and MPEG and has as such successfully contributed to HEVC and MPEG-DASH. He has also technically contributed to various scientific and industry-funded research projects such as SVConS and COAST in the EU FP7 program and is steering work in the German BMBF project CODEPAN.

Thomas Schierl received the Diplom-Ingenieur degree (passed with distinction) in Computer Engineering from the Berlin University of Technology (TUB), Germany in December 2003 and the Doktor der Ingenieurwissenschaften (Dr.-Ing.) degree in Electrical Engineering and Computer Science (passed with distinction) from Berlin University of Technology (TUB) in October 2010. Since 2010, Thomas is head of the research group Multimedia Communications in the Image Processing Department at Fraunhofer Heinrich Hertz Institute (HHI), Berlin. Before Thomas was responsible as Senior Researcher for various scientific as well as Industry-funded research projects in the Image Processing department of Dr. Ralf Schäfer and Prof. Dr.-Ing. Thomas Wiegand at HHI. Since 2015, Thomas is heading together with Detlev Marpe the new department Video Coding & Analytics at Fraunhofer HHI. The new department is covering all research groups of the former Image Processing Department with a relation to video coding, communications and analytics. Thomas is the co-author of various IETF RFCs, beside others he is author of the IETF RTP Payload Formats for H.264 SVC as well as for High Efficiency Video Coding (HEVC aka H.265).In the ISO/IEC MPEG group, Thomas is as co-editor responsible for e.g. the MPEG-2 Transport Stream standards on transport of H.264 SVC, MVC and MPEG-HEVC / ITU-T Rec. H.265. Thomas is also a co-editor of the AVC File Format. Typically, he is participating standardization meetings such as JCT, MPEG, IETF, 3GPP or DVB meetings. Thomas and his team – as part of JCT-VC – also contributed to the standardization process of MPEG - HEVC / ITU-T Rec. H.265, mainly in the area of high level parallelism and high level syntax. In 2007, Thomas visited the Image, Video, and Multimedia Systems group of Prof. Bernd Girod at Stanford University, CA, USA for different research activities. Thomas' research interests include system integration of video codecs as well as delivery of real-time media over mobile IP networks such as mobile media content delivery over HTTP. In 2014, Thomas received together with the ISO/IEC JCT1/SC29/WG11 Moving Picture Experts Group (MPEG) the Technology and Engineering Emmy Award by the National Academy of Television Arts & Sciences for the Development of the MPEG-2 Transport Stream.