Compressed Domain Processing for Stereoscopic Tile Based Panorama Streaming using MV-HEVC

Robert Skupin, Yago Sanchéz, Thomas Schierl
Video Coding and Analytics Department
Multimedia Communications Group
Fraunhofer Heinrich-Hertz-Institute
Berlin, Germany
robert.skupin@hhi.fraunhofer.de

Abstract—Efficient transport schemes for panoramic video delivery were proposed and demonstrated within the past decade. With the recent advances of head mounted displays however, consumers may soon have immersive and sufficiently convenient end devices at reach which could lead to an increasing demand for stereoscopic panoramic video experiences. This paper extends recent compressed domain techniques for efficient panoramic streaming to the domain of stereoscopic panorama video. The proposed technique is capable of reducing peak streaming bitrate, e.g. during changes of the region of interest of a user, which is crucial for allowing a truly immersive and low latency video experience.

Index Terms—Stereoscopic, Panorama, Streaming, MV-HEVC

I. INTRODUCTION

Efficient transport schemes for panoramic video delivery were proposed and demonstrated within the past decade. There are prototypes and deployed systems already showing feasibility of panoramic streaming [1]. With the recent advances of head mounted consumer displays such as the Oculus Rift [2], Samsung Gear VR [3] or Google Cardboard [4] however, consumers may soon have immersive and sufficiently convenient end devices at reach which could lead to an increasing demand of stereoscopic panoramic video experiences. Computer generated content for such applications is long available and the challenges of acquisition of real-world stereoscopic video content [5] have also been addressed and devices have been demonstrated for some time now. However, transport of panoramic video and stereoscopic panoramas also remains an active area of research, especially in context of the recently finalized H.265/High Efficiency Video Coding (HEVC) [6] and its Multiview extension (MV-HEVC) [7].

In the classic single view streaming scenario, only part of the panoramic video content, i.e. the users Region of Interest (RoI), is transmitted to the end device. A straight forward solution is to individually encode the users RoI video. However, such a solution hardly allows large scale deployment and therefore tile-based panoramic streaming was first proposed by Mavlanakar et al. [8]. The main idea is to divide the video panorama horizontally (and even vertically) into smaller video tiles that are encoded independently. Only the tiles that cover the user RoI are sent to a specific user. The granularity of the tiling determines the balance of the coding efficiency versus the picture sample overhead outside the RoI to be transmitted to the user. On the downside, such an approach requires sufficient resources at the end device to decode a potentially high amount of different video tiles [9], leaving behind single hardware decoder devices such as the mobile platforms used in [3] and [4].

A lightweight compressed domain operation that allows for stitching of multiple videos using encoder constraints was proposed in [10]. It enables usage of a single hardware decoder on the end device to efficiently decode multiple videos under battery constraints. This technique can also be applied to tile-based panorama streaming to reduce end device requirements.

An extension of the technique shown in [10] was demonstrated in [11] to reduce the bitrate requirements in tile-based panoramic streaming. Especially during RoI change events, tile-based panoramic streaming using a single bitstream as in [10] leads to bitrate peaks as inter-picture prediction is broken for video tiles that change their position in the RoI. In [11] specially crafted pictures, so called Generated Reference Pictures (GRPs) are inserted into the bitstream that allow for inter-prediction in those events, thereby dramatically reducing bitrate requirements. This paper extends the techniques introduced in [10] and [11] to the domain of stereoscopic panoramic video streaming using MV-HEVC as the latest available codec standard for such content.

Figure 1 shows an exemplary setup in which both views of the panorama video are tiled horizontally and a subset of tiles depending on the user RoI is transmitted to the end user. These tiles are stitched together to form a MV-HEVC compliant bitstream. The proposed technique is capable of reducing the peak streaming bitrate of a user, e.g. during RoI changes, which is crucial for allowing a truly immersive and low latency video experience.
video experience, e.g. with head mounted displays.

The remainder of this paper is organized as follows. Section 2 gives an overview of the system considered for tile based panoramic streaming. In section 3, the proposed technique is explained. Section 4 and section 5 describe the experiments and the results, respectively. Finally, a short conclusion is presented in section 6.

II. STREAMING SYSTEM OVERVIEW

Figure 2 provides an overview of the components of the considered system, i.e. an encoder array located at server side, a single video decoder instance at the client side and an Interactive Bitstream Stitching (IBS) device in between. After synchronous tiling of the two views of the panoramic video, individual 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 as blue rectangle determines which pairs of tiles are required by a user, e.g. the dashed tiles 0 to 3 in Figure 2. The corresponding bitstreams are subsequently 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.

The physical location of the IBS in the system can be either on server side, client side or within network given the IBS is allowed for media access and modification at this point. As the processing steps are of comparatively low complexity, usage of cloud infrastructure should scale well and even low-end client devices should be able to handle the operation alongside decoding.

III. COMPRESSED DOMAIN MULTIVIEW VIDEO PROCESSING

This section provides details about the processing carried out by IBS device that allows to combine the individual MV-HEVC bitstreams into a single RoI bitstream and handle RoI change events efficiently. First an overview of compressed domain video stitching is given in III.A, followed in III.B by a description of the processing for RoI change events in stereoscopic video applying an extension of the GRPs introduced in [11] for single-layer coded panorama videos.

A. COMPRESSED DOMAIN STEREOPTIC VIDEO STITCHING

This initial processing step as first described in this form in [10] creates a single MV-HEVC compliant joint bitstream out of the multiple individually encoded stereoscopic bitstreams, each depicting a pair of corresponding left and right view tiles and all together covering the user RoI. For this purpose, slice segment data of the individual stereoscopic tile bitstreams is copied to tiles of the joint bitstream within each view, i.e. independent and dependent coded layer.

Adjustments to high-level syntax are also required but are comparatively lightweight to carry out. At first, parameter sets need to be rewritten, mainly to reflect the spatial picture dimensions, level and tile setup of the RoI bitstream. Furthermore, adjustments on slice level are necessary, e.g. slice addresses in the merged slice headers need to reflect the tile positions 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 sets.

The compressed domain video stitching procedure further requires the encoders to be constrained in order to avoid prediction mismatches between the multiple individual encoders and the single decoder of the joint bitstream. Below is a short summary of the constraints for HEVC coded bitstreams as detailed in [10] which apply to MV-HEVC analogously.

1) Motion Vector (MV) constraints: MVs should not point to samples outside the picture borders or sub-pel sample positions, for which the encoder-side invoked sub-pel interpolation filter kernel overlaps with the picture borders.

2) Prediction Units: The rightmost prediction units within a picture shall not use the MV prediction candidate that corresponds to a temporal motion vector prediction (TMVP) candidate or the spatial MV candidate at the position of a non-existent TMVP candidate.

3) In-loop filters: Slice segment and tile borders (if present) shall not be crossed by in-loop filters such as the deblocking and SAO filter.

B. INTERACTIVE STITCHING FOR STEREOPTIC PANORAMA VIDEO

Whenever a RoI change event occurs the transmitted subset of tiles changes. Figure 3 shows an illustration of such an event for a single view panorama, where at one time instant \( t_0 \) the tiles 0, 1, 2, 3, and 4 are required for the RoI (blue rectangles) and at time instant \( t_1 \) tiles 3, 4, 5, 6, and 7 are required. Tiles 5, 6, and 7 require random access, i.e. intra-coded pictures, in order to be decoded at the end device. However, the picture content of the dashed red area (tile 3 and 4) is present at decoder side from before the RoI change event and could be used for temporal inter-prediction at \( t_1 \), i.e. sending of intra-coded tiles 3 and 4 at \( t_1 \) that contribute to bitrate peaks could be avoided. However, due to the spatial displacement of this area within the RoI caused by the stitching process, which is of course unknown to the encoder, a decoder cannot utilize the picture content for inter-prediction. Figure 3 illustrates this situation by showing the encoder perspective on a single tile of...
Figure 3: RoI change without GRP.

Figure 5: RoI change with GRP.

Figure 4: Overview of GOPs in RoI change interval.

a single view on the top part of the figure before (t₀) and after (t₁) processing by the IBS device. Temporal prediction between pictures is illustrated by the arrow. On the bottom part of Figure 3, the decoder perspective on the processed bitstream is shown where due to the spatial displacement of tiles within the picture plane, the inter-prediction does use the correct block for reference. This situation results in a prediction mismatch between encoder and decoder and, hence, visual artifacts.

The technique of GRPs as introduced in [11] for single-layer coded panoramas intents to allow a decoder to utilize the displaced reference pictures content for inter-prediction regardless of spatial displacement and hence allows to avoid the overhead of sending intra-coded pictures in such situations. A GRP is a completely inter-predicted picture that is not output by the decoder but replaces an original reference pictures from time instant t₀ used at time instant t₁. Picture content of the original reference picture content is copied and the same displacement that the original tiles undergo in the RoI change is applied to the reference picture content. A decoder is prevented from outputting the GRP by setting the no-output flag in the slice header.

Figure 5 illustrates the described situation after insertion of a GRP by the IBS from encoder and decoder perspectives analogously. The inter-prediction process at encoder and decoder side now use blocks with the same picture content as reference and hence no prediction mismatch or visual artifacts occur.

As GRPs can be encoded very efficiently, e.g. by using large block sizes and invariant motion vectors, the resulting peak bitrate reduction from omission of intra-coded random access for tiles 3 and 4 is significant. The Picture Order Count (POC) at the relevant position within the bitstream has to be adjusted without interfering with the usual POC based reference picture signaling in order to make place for the one or multiple additional pictures.

Usage of TMVP has to be restricted on the encoder side to not use pictures that may be reference-wise replaced with GRPs as reference for TMVP during a RoI change event. Otherwise, visual artifacts may result from erroneous inter-prediction using the invariant GRP motion vectors on decoder side instead of the genuine MV candidate used on encoder side.

As the work in this paper applies MV-HEVC, the mentioned constraints from [11] respectively apply to all individual views or layers to enable usage of Multiview GRPs (MGRPs). Since the spatial displacement of tiles during a RoI change event necessarily happens synchronously in all views, the coding dependencies between views, i.e. inter-layer prediction dependencies, do not require further precaution at the IBS. Hence, also the use of inter-layer TMVP has not to be restricted for MGRPs usage as reference and referencing picture for inter-layer TMVP keep their relative position.

In this work, we define certain pictures within the GOP structure, i.e. pictures of the lowest temporal level within a GOP structure, as so called switching points. These pictures are not used for TMVP within the same layer and can therefore be replaced by MGRPs without breaking inter-prediction.

IV. EXPERIMENTS

The benefit of the proposed MGRPs for stereoscopic panorama streaming will be evaluated through experiments using MV-HEVC coded stereoscopic panoramic video content acquired through a proof-of-concept camera as presented in [5]. The camera consists of three mirror segments and six cameras accounting for a total viewing angle of around 70°. A continuous test sequence with a resolution of 5760×1664 pixels at 30 fps and 1000 frames was created from three individual scenes. The effect of MGRPs in context of sparse versus dense horizontal tiling grids is evaluated through dividing the panorama video into tiles of 640×1664 pixels and 320×1664 pixels respectively. Each pair of left and right view tiles were jointly encoded using a modified version of the MV-HEVC reference software HTM version 14.1 [12] that
incorporates the restrictions described in Section III, using a GOP size of 8 pictures at QP 22 and at QP 32.

A streaming client resolution or RoI resolution of 1920×1080 pixels was simulated which can be covered by three tiles in case of the sparse tiling grid or six tiles for the dense variant. Two RoI change speeds are investigated in order to evaluate the impact of the MGRP technique for slow and fast RoI movement across the spatial plane of the panorama video. First, a complete change of the RoI at the speed of 1200 fps per second represents the fast movement scenario, which corresponds to a RoI change interval of 1.6 seconds or 6 GOPs. A second movement pattern was tested simulating a speed of only 600 pixels per second which accounts for change of around half of the RoI within the investigated RoI change interval of 3.2 seconds. Figure 4 illustrates the distribution of RoI change events over the GOP structures within the RoI change interval. It can be seen for example that for a tile size of 320×1664 at fast movement, the tile set within the RoI changes with every GOP, i.e. every 8 pictures or 0.27 seconds, while for 640×1664 at slow movement only two RoI change events will occur during the RoI change interval which corresponds to a distance of 32 pictures or roughly 1 second.

Multiple test cases were tested by varying the start of the RoI change interval, position of the initial RoI, movement speed, tiling granularity and QP. Every test case includes a single RoI change interval of 1.6 seconds. RoI change has been simulated at six different temporal positions 152 frames apart which corresponds to roughly 5 seconds. Also all possible tile starting positions of the RoI within the panorama plane were simulated applying movement towards the right panorama picture border. Results for all cases are subsequently averaged over the various RoI change intervals and starting positions. RoI changes were either carried out using the presented MGRP technique for spatially displaced tiles or using the conventional approach with random access via intra-coded pictures in each tile of the RoI at each RoI change event.

### V. Results

Four different Bitrates were considered in this work as follows:

- $B_w$ is the bitrate of the complete RoI sequence.
- $B_s$ is the bitrate within the RoI change interval of 1.6 seconds.
- $B_a$ is the bitrate of GOPs within the RoI change interval in which a RoI change event occurs according to Figure 4.
- $B_n$ is the bitrate of the remaining GOPs within the RoI change interval without RoI change events.

![Figure 6: IDR bitrate overhead in $B_w$ and $B_s$ for QP 22.](image)

A more granular measurement of the bitrate peak characteristics can be provided by analyzing $B_s$ and $B_a$. For instance, using MPEG-DASH [13] for panoramic streaming, 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. Alternatively, an MPEG-DASH client could download subsegments by performing an HTTP request with byte-ranges, if segments contain multiple GOPs, to achieve the same minimized end-to-end latency. In such a case, the bitrate peaks of interest are the ones that correspond to the downloaded chunks of data, i.e., the response to each of the requests which corresponds to a GOP. Note that the longer the interval of the requested data the bigger the pixel overhead that has to be downloaded to account for possible RoI movements. In the work one GOP as the smallest segments corresponds to 0.27 seconds of video. Bitrate $B_s$ can be considered as the significant bitrate peaks is this application and reduction of $B_s$ is therefore crucial.

Table 1 provides the four bitrates $B_w$, $B_s$, $B_a$, and $B_n$ in Mbps at QP 22 and both investigated tiling variants and both RoI movement speeds using either the conventional RoI change with IDRs or the proposed MGRP. 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 6 visualizes the overhead of the IDR based RoI change in percent with respect to the proposed MGRP solution regarding $B_s$ on the left and $B_w$ on the right.

Analogously, bitrate results for QP32 are reported in Table 2. Overall, the effect of MGRP at such low bitrate is even more pronounced in the relevant bitrates $B_s$ and $B_a$ while $B_n$ for example does not exhibit significant impact. As seen in Figure 7, the relative overhead of IDR based RoI changing is even more significant at low quality.

A denser tile grid, i.e. comparatively small tiles, results in lower benefits from the MGRP technique as it implicates more frequent RoI change events that introduce intra-coded pictures.
without MGRP. The overhead in \( B_s \) increases with the movement speed as more RoI change events occur while the overhead in \( B_i \) itself, i.e. during RoI change events, is mostly independent of movement speed.

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 3. 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 (\( B_i \)), as is reported in [11] for single view HEVC coded panoramic video. While overhead in \( B_s \) is to some extent weighted by the distribution of RoI change events, i.e. RoI movement speed, as illustrated in Figure 4, 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. 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 the QP setting.

VI. CONCLUSION

This paper extends the technique of GRPs for tile-based panorama streaming to the domain of stereoscopic panoramic video using MV-HEVC. Results based on respective content are reported and the effect of MGRP is analyzed in context of a layered coding scenario. Results show that the MGRP technique allows for significant peak bitrate reduction which can be a key element in future panoramic video streaming services with particularly strict latency requirements as introduced by head mounted displays.

![Figure 7: IDR bitrate overhead in \( B_i \) and \( B_s \) for QP 32.](image)

<table>
<thead>
<tr>
<th>RoI change with</th>
<th>Movement speed</th>
<th>Tile width</th>
<th>( B_{\text{i}} )</th>
<th>( B_{\text{s}} )</th>
<th>( B_{\text{n}} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDR</td>
<td>Fast</td>
<td>320</td>
<td>.484</td>
<td>1.12</td>
<td>1.12 ( 0^1 )</td>
</tr>
<tr>
<td>MGRP</td>
<td></td>
<td>640</td>
<td>.456</td>
<td>.574</td>
<td>.574 ( 0^1 )</td>
</tr>
<tr>
<td>IDR</td>
<td>Slow</td>
<td>320</td>
<td>.414</td>
<td>1.02</td>
<td>.434</td>
</tr>
<tr>
<td>MGRP</td>
<td></td>
<td>640</td>
<td>.404</td>
<td>1.20</td>
<td>.609</td>
</tr>
</tbody>
</table>

\( ^1 \)The RoI change interval in this particular case contains only GOPs in which RoI change events occur.

### Table 3: IDR overhead percentage per layer.

<table>
<thead>
<tr>
<th>QP</th>
<th>Movement speed</th>
<th>Tile width</th>
<th>Overhead ( B_i )</th>
<th>Overhead ( B_s )</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>Fast</td>
<td>320</td>
<td>.604</td>
<td>.744</td>
</tr>
<tr>
<td></td>
<td></td>
<td>640</td>
<td>.274</td>
<td>.320</td>
</tr>
<tr>
<td></td>
<td>Slow</td>
<td>320</td>
<td>.374</td>
<td>.443</td>
</tr>
<tr>
<td></td>
<td></td>
<td>640</td>
<td>.206</td>
<td>.237</td>
</tr>
<tr>
<td>32</td>
<td>Fast</td>
<td>320</td>
<td>.105</td>
<td>.742</td>
</tr>
<tr>
<td></td>
<td></td>
<td>640</td>
<td>.435</td>
<td>.286</td>
</tr>
<tr>
<td></td>
<td>Slow</td>
<td>320</td>
<td>.574</td>
<td>.416</td>
</tr>
<tr>
<td></td>
<td></td>
<td>640</td>
<td>.320</td>
<td>.203</td>
</tr>
</tbody>
</table>

REFERENCES