COMPRESSED DOMAIN VIDEO PROCESSING FOR TILE BASED PANORAMIC STREAMING USING HEVC

Y. Sánchez, R.Skupin, T. Schierl

Fraunhofer HHI, Einsteinufer 37, 10587 Berlin

ABSTRACT

Panoramic streaming is a particular way of video streaming where an arbitrary Region-of-Interest (RoI) of high-spatial resolution videos is transmitted. It allows users to navigate interactively around the video and thus select anytime the portion of it they are interested in. The most basic approach consists of each user interacting with the system indicating the desired RoI. Then an encoder associated with each user encodes the desired RoI. However, such a system does not scale well. Instead, we consider tile based panoramic streaming, where users receive a set of tiles that match their RoI, and propose a low-complexity compressed domain video processing technique for tile based panoramic streaming using H.265/HEVC that generates a single video bitstream out of the selected tiles so that single hardware decoders can be used to decode the RoI video stream.

Index Terms— Panoramic Streaming, HEVC, Tiles, Processing, Stitching

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 around the video by choosing at any time the RoI they are interested in. Figure 1 shows an example, where the spatial plane of the panorama video is shown with two RoIs. The left region is the content displayed to the user at time t_0 and the right region is displayed at time t_1 after user interaction, i.e. a RoI switching event.

There are prototypes and deployed systems already showing panoramic streaming's feasibility [1][2]. We envision that interactive panoramic streaming will be a popular application in a few years from now. In fact, techniques already exist that allow capturing 360° video in real time [3], by stitching multiple HD views from multiple cameras. Furthermore, the recent success of virtual reality headsets such as Oculus Rift [4] is changing the way multimedia is experienced, emphasizing user interaction.

Having a dedicated encoder per user that encodes a specific RoI and transmits it to user based on the user's interest does not scale well. An alternative is to use tile based panoramic streaming, introduced by Mavlankar et al. in [5]. The main idea is to divide the panorama picture horizontally and vertically into smaller regions that are encoded independently. Then the regions $r \subset P$ (where P is the set of all regions in the panorama picture) that fulfill that $r \cap \text{RoI} \neq \emptyset$ are transmitted to the user. In other words, the regions that contain the content belonging to the RoI are transmitted to the user. Although some other approaches have been pursued in the past, using e.g. some hierarchical prediction of a thumbnail view [6] using H.264/SVC, tile based panoramic streaming using H.265/HEVC has the benefit that standard video decoders can be used.

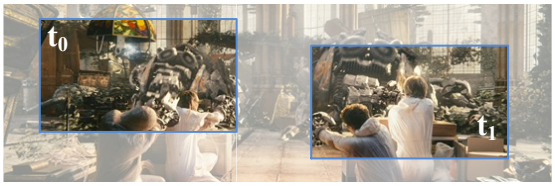

Figure 1: RoI before (t_0) and after (t_1) user interaction.

[6] or [7] present an optimization of the dimension of the tiles in which the panorama video is split. Note that the borders of these regions/tiles do not necessarily coincide with the RoI borders and therefore some extraneous data (not RoI) might be transmitted. Spatially large tiles increase compression efficiency but also lead to the transmission of additional data that is not part of the RoI.

These previous works assume instantiation of a decoder per received spatial region. However, in many cases, devices only use a single hardware-accelerated decoder to achieve real-time decoding capabilities or to save battery power. Therefore, in this paper, we use the technique in [8] to generate a single video stream out of the received regions encoded with H.265/HEVC [9]. We will present a solution that reduces the transmission bitrate at RoI switching events, which is critical for panoramic streaming services due to their strict low-latency requirements: RoI switching must happen seamlessly and almost instantaneously.

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 their evaluation, respectively. Finally, the conclusion is presented in section 6.

2. STREAMING SYSTEM OVERVIEW

An overview of the components involved in the considered system is depicted in Figure 2. The system consists of a set of encoders, an Interactive Bistream Stitching (IBS) device and the client. First, the video is split into multiple tiles indexed at the bottom. Then, each tile is encoded by a different encoder, generating independent H.265/HEVC bitstreams. Once the RoI of each user is determined as indicated through the blue rectangle, a set of bistreams corresponding to the tiles that match the RoI (dashed region) is transmitted to the user and stitched at the IBS device generating a single bitstream.

The IBS device is depicted between the encoders and the decoder. It can be placed either at the receiver side before the decoder or at a proxy in the network, depending on the approach used for transmission. In any case, the process has low complexity and therefore would scale well even if performed in the network.

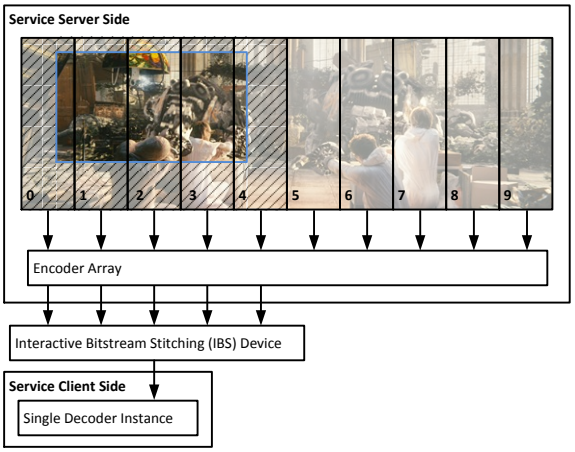

Figure 2: System overview

3. COMPRESSED DOMAIN VIDEO PROCESSING

The goal of the proposed compressed domain video processing technique IBS is to achieve an efficient panoramic streaming service with a single coded stream and a reduced bitrate at RoI switching points. IBS entails stitching the videos of the tiles belonging to the RoI and the insertion of Generated Reference Pictures (GRP) as described below.

3.1. Overview of compressed domain video stitching

A single bitstream is generated by applying the low-complexity compressed domain stitching process described in [7]. Each of the independently encoded received spatial regions is converted into an HEVC tile of a common bitstream. This is done by rewriting the slice segment header of NAL units and parameter sets. Parameter sets are rewritten, for instance, to indicate the level of the resulting bitstream or the dimensions of the HEVC tiles within the stitched bitstream. Additionally, it is a requirement that the original videos are encoded fulfilling a set of constraints so that the compressed domain stitching can be performed without leading to any decoding drift:

- 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.

For more information the reader is referred to [7].

3.2. Interactive Bitstream Stitching

The technique presented hereafter aims to reduce the transmission bitrate during RoI switching events. Figure 3 depicts a RoI switching event, where 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 changes over time. Figure 3 shows an example where the dashed tiles 3 and 4 belong to both of the transmitted bitstream sets at time instants t_0 and t_1 but are located at a different position within in the respective RoI of the time instants.

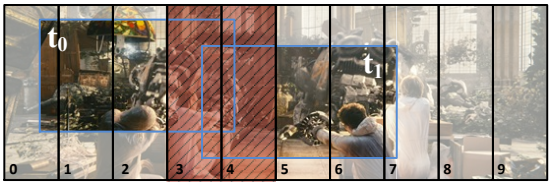

Figure 3: Overlap of RoIs before user interaction (t0) and after (t1)

Although new spatial regions (see non-shaded tiles with index 5 to 7 in Figure 3) require random access at the RoI switching event, the set of tiles that remain displayed albeit displacement (tiles 3 and 4) would benefit from using temporal prediction. However, since their position has changed in the stitched picture, temporal prediction cannot be used. This is illustrated as seen from encoder and decoder perspective in Figure 4, with 2 bitstream sets corresponding to time instants t_0 and t_1, the latter using the picture at t_0 as reference. Spatial regions using random access are 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. 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.

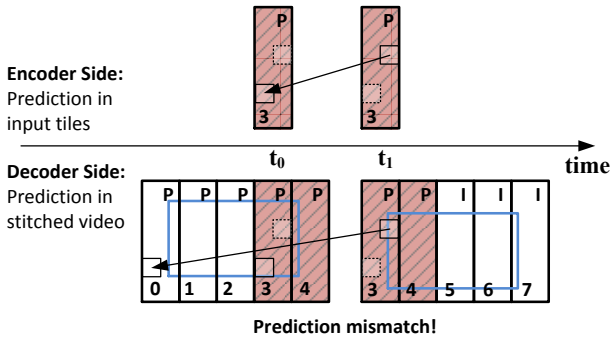

Figure 4: Prediction mismatch after IBS without GRP.

In order to avoid random access for all spatial regions, which would lead to big 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. The content of a regular reference picture is used and copied at the position that matches the new spatial region setup from the RoI switching event onwards.

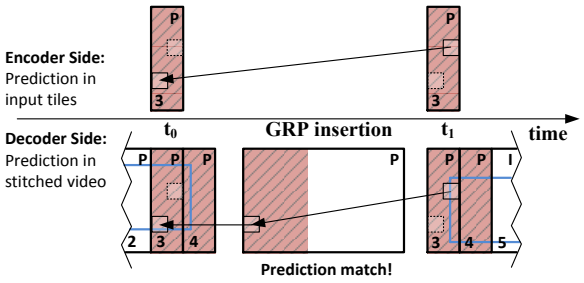

Figure 5: Prediction match after IBS with GRP

Figure 5 shows how the block in tile 3 at *t1* uses the block in the GRP that contains the content the desired block from t_0 (belonging to tile 3 as well). A GRP is encoded without residual data and all Prediction Units (PUs) have a MV associated that compensates the movement performed by users at an RoI switching point, so that the desired content is at the right position in the new spatial region setup. GRPs can be encoded very efficiently using the skip mode, since all PUs have the same MV.

The Picture Order Count (POC) of the pictures has to be modified, as new pictures are added to the bitstream. Further, GRP pictures have to be marked as non-output pictures, so that they are not displayed but only used as reference. In order to avoid any decoding drift, the original bitstreams have to fulfill the following constraint:

- TMVP has to be restricted so that no pictures that may be reference-wise substituted by GRPs are used for TMVP.

MV prediction in H.265/HEVC [10] is performed from neighbor or temporal MV candidates. The latter (TMVP) 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, as they 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 DPB are never selected for TVMP.

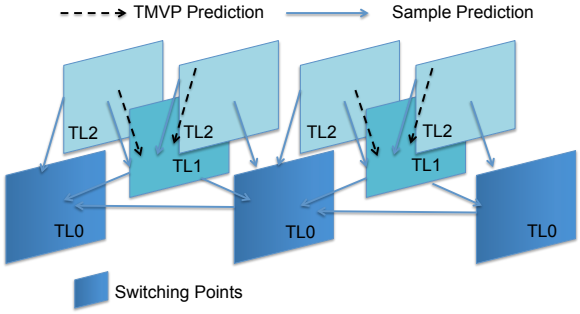

Figure 6: Example of TMVP restriction

Figure 6 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.

4. EXPERIMENTS

Two panorama videos (captured with [3]) with a spatial resolution of 8192x1600 pixel have been used for the experiments. They consist of 1425 frames at 25fps and have been encoded with the GOP structure shown in Figure 6 and two different QPs: 22 and 32. 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: the first one resulting into spatial regions of 512x1600 pixels and the second one into spatial regions of 256x1600 pixels. These sizes have been exemplarily selected. The reader is referred to [6][7] if interested in how to perform optimization of the tile sizes.

Interactivity has been simulated by selecting two screen movement patterns, which consist of a movement towards the right: a high-speed movement of 1600 px/s and a low-

speed movement of 800 px/s. They correspond to a complete RoI change in 1.2s and 2.4s respectively, for the considered client screen of 1080p. In both cases the movement duration, i.e. RoI switching interval, is 1.2s, which means that the fast movement leads to a complete RoI change, while the slow movement leads only to an RoI change of half of its dimension. Figure 7 illustrates for each tile dimension and movement speed the frequency of switching point GOPs, i.e. GOPs where the tile setup changes and random access or GRP are used.

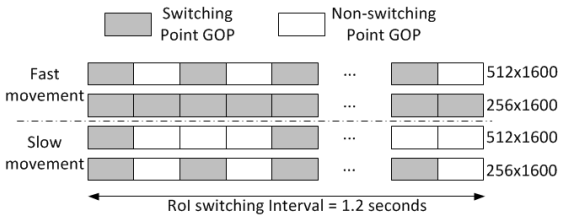

Figure 7: Switching Point GOP distribution

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 section 5.

5. RESULTS

The following figures 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.

Figure 8 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 bigger is the gain of using GRPs, which is reasonable as the number of switching Point GOPs is higher (see Figure 7). 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.

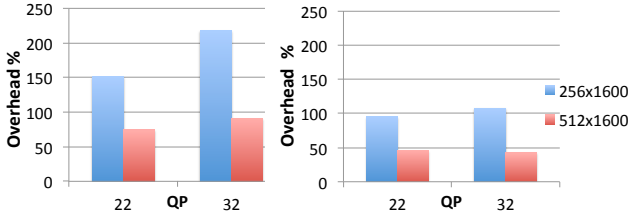

Figure 8: Transmission overhead RA vs. GRP for fast movement (left) and slow movement (right) for 1st video

Similar conclusions can be drawn from the results of the second video with respect to the influence of the QP, tile sizes and speed of the movement. However, as shown in Figure 9, the overhead is slightly smaller: 50-100% for 256x1600 spatial regions and 20-40% for 512x1600 regions. The gains provided by GRP are significant.

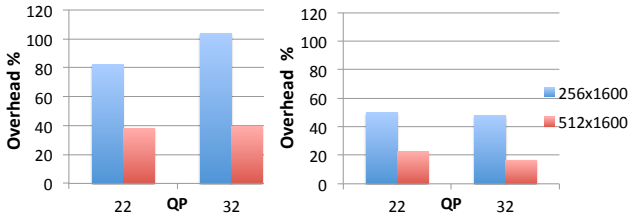

Figure 9: Transmission overhead RA vs. GRP for fast movement (left) and slow movement (right) – 2nd video

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 1 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.

Table 1: 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

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.

6. 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 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.

We showed that transmission bitrate savings between 80%-200% or 40%-100% during the screen 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.

7. REFERENCES

[1] ClassX, Stanford University. http://classx.stanford.edu/.

[2] R. van Brandenburg, O. Niamult, M. Prins, H. Stokking, “Spatial Segmentation For Immersive Media Delivery”, 2011.

[3] Christian Weissig, Oliver Schreer, Peter Eisert, Peter Kauff, “The Ultimate Immersive Experience: Panoramic 3D Video Acquisition”, Advances in Multimedia Modeling, Lecture Notes in Computer Science Volume 7131, 2012, pp 671-681.

[4] http://www.oculus.com/rift/.

[5] A. Mavlankar, P. Agrawal, D. Pang, S. Halwa, N. Cheung, B. Girod “An interactive Region-of-Interest Video Streaming System for Online Lecture Viewing”, Special Session on Advanced Interactive Multimedia Streaming, Proc. of 18th International Packet Video Workshop (PV), Hong Kong, Dec. 2010.

[6] A. Mavlankar, P. Baccichet, D. Varodayan, and B. Girod, "Optimal Slice Size for Streaming Regions of High Resolution Video with Virtual Pan/Tilt/Zoom Functionality," Proc. of 15th European Signal Processing Conference (EUSIPCO), Poznan, Poland, Sept. 2007

[7] N. Quang , G. Ravindra, W.T. Ooi, “Adaptive Encoding of Zoomable Video Streams based on User Access Pattern”, MMSys 2011, San Jose, Feb. 2011.

[8] Y. Sanchéz, R.Globisch, T. Schierl, and T. Wiegand: Low Complexity Cloud-video-Mixing Using HEVC, Proceedings of IEEE Consumer Communications and Networking Conference, Las Vegas, NV, USA, January 2014.

[9] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand: Overview of the High Efficiency Video Coding (HEVC) Standard, IEEE Transactions on Circuits and Systems for Video Technology, December 2012, Best Paper Award.

[10] J. Lin, Y. Chen, Y. Huang, S. Lei, “Motion Vector Coding Techniques for HEVC”.