PEAK BITRATE REDUCTION FOR MULTIPARTY VIDEO CONFERENCING USING SHVC

Y. Sánchez, R. Skupin, C. Hellge, T. Schierl

Fraunhofer HHI, Einsteinufer 37, 10587 Berlin

ABSTRACT

Real-time video applications, such as multi party video conferencing, involve the simultaneous transport of multiple and potentially multi-layered video sources to participating or interested parties. It is desirable to mix these multiple source videos into a single video stream at intermediary nodes in the network, e.g. at Multipoint Control Units (MCU). This has the advantage of reduced application and transport complexity on the client device while allowing single hardware decoder devices to consume the content. This paper proposes a solution, which uses the scalable extension of H.265/HEVC (SHVC), and that generates a single bitstream out of several sources by a low-complexity operation. In addition, the presented technique drastically reduces the peak bitrate at layout change events in comparison to state-of-the-art solutions.

Index Terms— SHVC, video mixing, compositing, MCU, video conferencing.

1. INTRODUCTION

Video services, such as multi-party video conferencing, entail simultaneous presentation of multiple videos. Such an application can be implemented in software by running individual decoders per involved video in a parallel fashion. However, many devices use a single hardware decoder to achieve real-time decoding under battery constraints. Therefore, respective end points require a single coded stream that needs to be generated within the network from the streams of each party.

Such an operation can be accomplished by applying pixel-domain video stitching, where the different video streams are transcoded into a single bitstream. However, it has a high computational complexity which compromises service scalability and might result in quality degradation of the original source videos due to the repeated encoding procedure [1]. The functionality described within the paper is similar in nature to that of RTP mixers [2] or RTCP-terminating MCU [3]. However, in the case of both RTP mixers and MCUs, the original streams are either kept as independent streams, identified by different CRSC fields [2], or the individual video streams are decoded, mixed together in the pixel-domain and re-encoded at intermediate nodes in the network.

In order to address the complexity issues and device constraints mentioned above, a technique is proposed in this paper that is applied in the compressed domain, without requiring decoding of the original video streams. This low-complexity technique allows real-time operation in a scalable and cost-effective manner in contrast to the traditional transcoding approach, which requires more resources at the MCU (or equivalent stream merging device). The proposed solution (cf. Section 3) is based on the scalable extension of H.265/HEVC (SHVC) [4]. As the scalable extension of H.264/AVC (SVC) has previously been proven to be very valuable in video conferencing systems, it is expected that SHVC will also be used in multi-party conversational scenarios in the near future.

The technique proposed in this paper provides a low-complexity solution for video merging that generates a standard-compliant bitstream containing video of all participants of a multi-party conference session. The proposed technique is compared to another similar technique [8] while providing a significant improvement by focusing on reducing the bitrate peaks that occur when changes in the video layout happen, i.e. at a speaker change event. Avoiding bitrate peaks is a crucial aspect in conferencing scenarios, since video data at these bitrate peaks is more susceptible to be affected by packet losses or can lead to an delay increase, which is not acceptable in a scenario with strict delay constraints of around 150 ms [5].

The remainder of the paper is organized as follows. Section 2 gives an overview of related work. In Section 3, an overview of the compressed domain SHVC video compositing technique is given. Section 4 describes the simulations performed and discusses the results and Section 5 presents the conclusions.

2. RELATED WORK

In [6] and [7], the authors propose a technique to perform compressed domain stitching of several videos. The technique is based on a lightweight adjustment of high-level syntax elements. It entails re-writing slice headers and parameter sets, so that each of the original pictures is

mapped to a tile in the stitched bitstream. The reader is referred to [6] for further details.

In addition to the lightweight processing mentioned above, it is required that the bitstreams to be stitched together fulfill a set of encoder constraints so that the compressed domain stitching does not lead to prediction mismatches on decoder side. Below, there is a short summary of the constraints for HEVC coded bitstreams as detailed in [6].

- Motion Vector (MV) Constraints: MVs should not point to samples outside the picture borders or sub-pel sample positions that require samples outside the picture border in the sub-pel interpolation process.

- 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.

- 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 at which the TMVP predictor would be if it existed.

In [8] a technique has been developed that allows for generating a single scalable bitstream out of multiple bitstreams by stitching all input bitstreams in the compressed domain in a similar manner as explained before. Figure 1 illustrates the working principle of such a technique. Three streams are represented, which correspond to the speaker and two additional participants in a video conferencing session. All participant video streams feature a basic quality l0 (base layer). The video stream of the speaker features an additional layer l1 (enhancement layer) at a higher spatial resolution, since, typically, it is shown in a larger picture area compared to the other participants. The compressed domain stitching (CDS) process, operates in such a way that the output bitstream contains two layers: L0 with layer l0 of the speaker and L1 where the layer l1 of the speaker and layers l0 of the other participants are stitched together, e.g. following the process describe in [6].

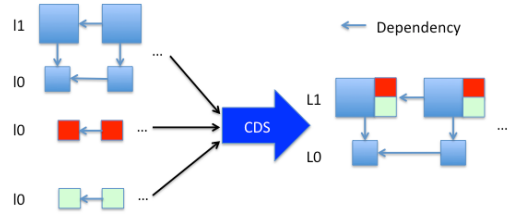

Figure 1: Compressed Domain Stitching of SHVC streams

Although such an approach provides a lightweight processing solution that generates a single bitstream that can be decoded by single hardware decoders, it has the drawback that whenever a layout change occurs (e.g. at a speaker change event) mid-stream decoding, i.e. so-called Random Access (RA), of several (or even all) bitstreams is required. This is illustrated in the following figure.

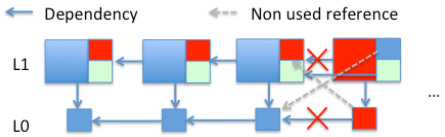

Figure 2: Layout change at speaker change event

As shown in Figure 2, whenever the video layout is changed, prediction to previous pictures cannot be used on encoder-side since the samples of the reference picture are displaced at decoder-side (see dashed arrows). At the speaker change event in the figure (the participant in red provides l0 and l1) the prediction to previous pictures in L0 is broken, even though the previous picture in the figure comprises samples that could have been used as reference (indicated with the dashed line). However, since the content of the participant in red (l0) was previously stitched in L1 in the merged bitstream, it is not possible to use it as reference for L0. Similarly, the layer l0 of the previous speaker (participant in blue) is stitched in L1 from the layout change onwards, forbidding usage of the previous picture as reference. In the shown example only the green participant can still use inter-prediction, while for the blue and red participants, who undergo a positional change, full RA (l0 and l1) is required

3. PROPOSED SOLUTION

As previously discussed, requiring RA in all layers of the bitstreams affected by the layout change, significantly increases the instantaneous bitrate resulting in a peak. This has the negative effect that packet losses or additional delay may arise, which are detrimental for a service such as video conferencing where the end-to-end delay needs to be kept below 150 ms. In order to avoid such peak rates, a solution is proposed that entails extending the technique in [6] and [8], using the concept of Extended Scalability or ROI scalability in SHVC. This is done by inserting Multi-Layer Composition Pictures (MLCP) into the merged bitstream and modifying the Picture Parameter Set (PPS), in addition to high-level syntax adjustments mentioned before. The concept of MLCP is an extension of the work that authors proposed in [9], where a slice in the MLCP contains the enhancement layer of the speaker and additional slices that copy sample values via inter-layer prediction from lower layers that contain the actual video streams of the other participants.

The main idea behind the proposed technique using MLCP is illustrated in Figure 3. Contrary to the technique [8] described previously, the idea is to keep the base layer of each of the participants independent from each other. I.e., each of the base layers is assigned to a different layer from L0 to L(N-1), where N is the number of participants. This

has two main advantages in comparison to the previous approach:

- None of the constraints described earlier is required to be fulfilled in the base layer since the content is not stitched.
- A layout change (speaker change) does not have any impact on these base layers.

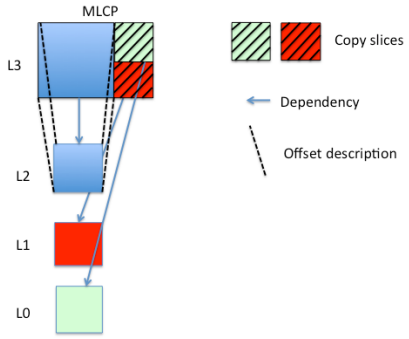

Figure 3: Illustration of MLCP technique

In Figure 3, L3 corresponds to the MLCP. Its first component, i.e. the blue square in L3, corresponds to the enhancement layer slice of the speaker. Since this slice is stitched together with further slices, it has to fulfil the previously described constraints with a potential relaxation of the TMVP usage constraint at the rightmost prediction units. In SHVC, the TMVP predictor may correspond to a motion vector of another picture in the same access unit but belonging to a lower layer or to a previous picture of the same layer. In the latter case, the TMVP constraint for rightmost PUs has to be fulfilled. However, if the TMVP candidate belongs to the base layer the described constraint is no longer required since the base layers are not stitched together.

The dashed lines in Figure 3, which for simplicity are only depicted for the blue participant, represent the scaled reference layer offsets and reference region offsets. In SHVC, two types of offsets are described. Scaled reference offsets determine the area within the referencing picture that is predicted from another layer and reference region offset determine the area that is used for prediction within the referenced picture in a lower layer. This information is included in the Picture Parameter Set (PPS). Therefore, in the proposed method, a PPS is generated for the highest layer that includes the respective offsets that correspond to the composition layout. I.e. as many scaled reference layer offsets and potentially referenced region offsets are included as participants to describe the regions in the output picture that correspond to each of the participants. Since a single PPS can only describe one layout, whenever a layout change happens a new PPS is required. It is important to remark that it is not necessary that a PPS is generated each time the layout/speaker changes. A set of suitable PPSs could be generated and transmitted at the beginning of the services so that they are available at decoder-side when required.

The other components of an MLCP are the copy slices (see dashed boxes in Figure 3). A copy slice performs a direct copy of the samples values of the reference picture. It consists of Coded Tree Units (CTUs) with a constant zero motion vector that in combination to the specified scaled reference layer offsets and referenced region offsets point to the samples of the intended region of the reference pictures of the non-speaker participants, i.e. layers L0 and L1 in the figure. Using region reference offsets allows flexible compositing of the layout since parts of the original content can virtually be “cropped out”, i.e. the composition of all participants can fit a determined resolution that does not match the resolution of a regular stitching procedure, where the complete pictures are stitched together. These copy slices can be very efficiently encoded, by using the largest possible Coding Unit (CU) sizes, encoding the first CU with zero motion vectors and no residual data, and the rest of the CUs in the slice as skip.

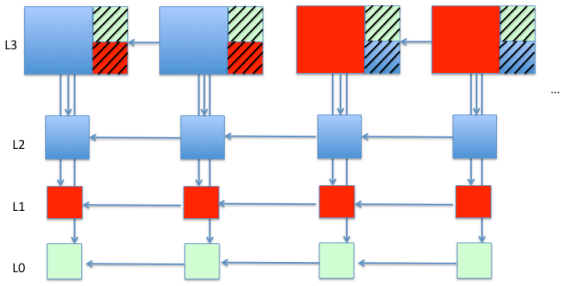

Figure 4: Speaker change event with MLCP

Figure 4 depicts a speaker change event when the proposed technique using MLCPs are applied. It can be seen that, in contrast to the previous solution, the use of temporal prediction in lower independent layers (L0, L1, L2 in the figure) is not affected through layout changes. RA is only required at the highest dependent layer (L3), into which the enhancement layer data of the speaker is written, as references for temporal prediction are not available at decoder side in this case.

In summary, the proposed technique allows for a more flexible composition in comparison to basic compressed domain stitching techniques, since the data that is displayed from the original content can be flexibly selected by assigning an appropriate reference region offset. Besides, the encoding efficiency of the original SHVC streams (especially in the base layer) is higher when the proposed technique is used, since base layers of the original streams do not have to fulfil any constraints. Finally, and more important, the bitrate peaks at speaker change events are reduced due to the relaxation on RA requirements when MLCPs are used.

4. SIMULATIONS AND RESULTS

Three sequences have been selected from the Common Test Conditions [10] used in SHVC standardization in order to analyse the performance of the proposed technique. The "Low-Delay P – Main" configuration has been used to represented the targeted application scenario for the sequences of Class E (namely Johnny, Kristen and Sara and Four People). The test sequences have been downsampled to a resolution of 320x160 samples in the base layer and 640x320 samples in the enhancement layer so that the stitched resolution (using a layout as shown in Figure 3) has a width of 960 samples, which accommodates the targeted single decoder end devices such as smartphones.

To evaluate the proposed MLCP technique, first a series of coding performance evaluations regarding the mentioned encoder constraints was carried out using SHM-8.0 [11]. A Full Constrained (FC) configuration with the mentioned constraints fulfilled within all layers and an Enhancement-Layer Constrained (EC) configuration, where the constraints only apply to the enhancement layer with the described TMVP relaxation (cf. Section 3) are compared to a Non-Constrained (NC) anchor configuration, in which encoder decisions were not constrained. While the solution in [8] would require FC encodings and the proposed MLCP technique could be realized with EC encodings, the anchor NC encodings would require multiple decoders on the end device. The loss in terms of BD-rate of the base layer (L0) and the whole bitstream (L0+L1) for FC and EC configurations compared to NC are shown in Table 1.

Table 1: BD-Rate Overhead of EC and FC wrt. NC

Sequence		BD-rate losses			
Name	Frame rate	EC		FC	
		L0	L0+L1	L0	L0+L1
Johnny	60 fps	0%	7.62%	3.35%	7.71%
Kristen & Sara	60 fps	0%	5.85%	4.63%	6.06%
Four People	60 fps	0%	2.61%	1.39%	2.70%

As can be seen in the table, the overhead of the whole bitstream over unconstraint encodings, irrespective of whether we are considering EC or FC, varies between 2.61% to 7.71% BD-rate depending on the sequence. Although EC provides a better compression efficiency than FC, the differences are below 0.25% and can be considered negligible. However, the base layer for EC is more efficiently encoded, since it is not constrained. More concretely, FC has from 1.39% to 4.6% overhead compared to EC. This means that the merged stream will have a lower bitrate if EC is used, which is only possible with the proposed MLCP technique. However, the magnitude of the difference is rather negligible and hardly affects the resulting quality of a videoconference application.

Much more important is the behaviour of the peak bitrates with MLCPs at speaker change events in comparison to the approach in [8]. The test streams have been encoded in FC and EC configurations with four different QPs (22, 27, 32, and 37). For each stream, a variant with RA at picture n*16 with n=1…25 has been encoded to provide for a speaker change at the time of the RA picture. In case of the basic stitching approach (solution in [8]), RA is required in both layers with FC configuration. In case of the proposed approach based on MLCP, RA is only required at the enhancement layers using the EC configuration. The streams have been merged using both discussed methods and a speaker change has been carried out at each of the 25 different positions aforementioned (picture n*16-th).

Table 2: Bitrate reduction of MLCP over the solution in [8] at speaker change events

	QP=22	QP=27	QP=32	QP=37
Switching-picture	40.17 %	50.17 %	51.45 %	44.23 %
66.6 ms period from Switching-picture	34.17 %	36.03 %	36.16 %	38.24 %

Table 2 shows the peak bitrate reduction achieved by the proposed MLCP solution over the method presented in [8]. The first row of Table 2 shows the bitrate reduction for the actual Switching-picture (picture at which the speaker change occurs). It can be seen that there is a peak bitrate reduction of 40%-50% on average for the instantaneous bitrate at the Switching-picture. Since it could be argued that some kind of rate shaping could be applied during transmission and there is 150 ms end-to-end delay that can be accepted, we have computed the peak bitrate savings for a time interval of 66.6ms (i.e. 4 pictures at 60fps) beginning at the Switching-picture, which would correspond to around half of the 150 ms delay constraint. The results are reported in the second row for the 66.6 ms interval. Still the bitrate saving are around 34%-38%, which are substantial savings under the given constraints. We argue that computing the peak bitrate for longer intervals is not meaningful since some time of the 150 ms is needed for transmission (propagation delay) and processing.

5. CONCLUSION

This paper proposes a low-complexity operation, which generates a single bitstream from several bitstreams using the scalable extension of H.265/HEVC (SHVC). Thereby, the proposed technique allows usage of single decoder devices, e.g. smartphones. The technique has been studied for a video conversational scenario, showing that it significantly reduces the peak bitrate at layout changes, i.e. speaker change events and outperforms state-of-the-art solutions. More concretely, peak bitrate savings of above 30% can be achieved, which is very valuable in a scenario where an end-to-end delay of 150 ms is required and bitrate variations can hardly be compensated by a buffer.

6. REFERENCES

[1] K.-Y. Yoo; K.-d. Seo, "Syntax-based mixing method of H.263 coded video bitstreams," Consumer Electronics, 2005.

ICCE. 2005 Digest of Technical Papers. International Conference on , vol., no., pp.403,404, 8- 12 Jan. 2005

[2] H. Schulzrinne, S. Casner, R. Frederick, V. Jacobson, “RTP: A Transport Protocol for Real-Time Applications”, RFC 3550.

[3] M. Westerlund, S. Wenger, “RTP Topologies”, RFC 5117.

[4] ITU-T, Recommendation H.265 (04/15), Series H: Audiovisual and Multimedia Systems, Infrastructure of audiovisual services – Coding of Moving Video, High Efficiency Video Coding.

[5] ITU-T Rec. G.114, "SERIES G: TRANSMISSION SYSTEMS AND MEDIA, DIGITAL SYSTEMS AND NETWORKS International telephone connections and circuits – General Recommendations on the transmission quality for an entire international telephone connection - One way transmission time", May 2003.

[6] Y. Sanchez, 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, Jan 2014.

[7] C. Feldmann, C. Bulla, B. Cellarius, "Efficient bitstream-Reassembling for Video Conferencing Applications using Tiles in HEVC." MMEDIA 2013, The Fifth International Conference on Advances in Multimedia. 2013.

[8] Eleftheriadis, Alexandros, et al. "System and method for videoconferencing using scalable video coding and compositing scalable video conferencing servers." U.S. Patent No. 8,436,889. 7 May 2013.

[9] R. Skupin, Y. Sanchez, and T. Schierl, “Compressed Domain Video Compositing with HEVC”, Picture Coding Symposium (PCS 2015), Cairns, Australia, May 31 - June 3, 2015.

[10] Bossen, Frank. "Common test conditions and software reference configurations." m28412, Joint Collaborative Team on Video Coding of ISO/IEC and ITU-T, JCTVC-L1100, 12th Meeting: Geneva, CH, 14-23 January 2013

[11]https://hevc.hhi.fraunhofer.de/svn/svn_SHVCSoftware/tags/SHM-8.0