Implementing 360 Video Tiled Streaming System

Jangwoo Son
Department of Computer Engineering
Gachon University
Korea
sjw6757@gc.gachon.ac.kr

Dongmin Jang
Department of Computer Engineering
Gachon University
Korea
dogzz9445@gc.gachon.ac.kr

Eun-Seok Ryu
Department of Computer Engineering
Gachon University
Korea
esryu@gachon.ac.kr

ABSTRACT

The computing power and bandwidth of the current VR are limited when compared to the high-quality VR. To overcome these limits, this study proposes a new viewport dependent streaming method that transmits 360-degree videos using the high efficiency video coding (HEVC) and the scalability extension of HEVC (SHVC). The proposed SHVC and HEVC encoders generate the bitstream that can transmit tiles independently. Therefore, the bitstream generated by the proposed encoder can be extracted in units of tiles. In accordance with what is discussed in the standard, the proposed extractor extracts the bitstream of the tiles corresponding to the viewport. SHVC video bitstream extracted by the proposed methods consist of (i) an SHVC base layer (BL) which represents the entire 360-degree area and (ii) an SHVC enhancement layer (EL) for selective streaming with viewport (region of interest (ROI)) tiles. When the proposed HEVC encoder is used, low and high resolution sequences are separately encoded as the BL and EL of SHVC. By streaming the BL(low resolution) and selective EL(high resolution) tiles with ROI instead of streaming whole high quality 360-degree video, the proposed method can reduce the network bandwidth as well as the computational complexity on the decoder side. Experimental results show more than 47% bandwidth reduction.

CCS CONCEPTS

• **Multimedia and Communication → Video compression; Video processing; Adaptive streaming;**

KEYWORDS

Virtual reality, HEVC and SHVC tile, Viewport, Motion-constrained tile set (MCTS), Extraction information set (EIS) SEI

1 INTRODUCTION

The head-mounted display (HMD) is the display device worn on the head and has display optic(s) in front of one or each eye, and the device supports head tracking functionality to provide a 360-degree view. To reduce the motion sickness (nausea) with HMD, high-quality video and very low motion delay are required [1].

To increase the video processing efficiency with a limited specification, viewport-based studies have progressed [2-3]. Thus, Motion-constrained tile set (MCTS) was discussed for viewport-based transmission in standardizaion meeting [4]. The MCTS limits the temporal inter prediction (TIP) of the tile at the same position of the current picture and the reference picture for transmitting tiles independently. This paper proposes the implementation method of MCTS in SHVC and HEVC reference software (SHM and HM). The implementation of this paper is adopted in the MPEG standard [5] and further proposes a method of extracting and decoding some tiles among all the tiles. Fig. 1 shows the conceptual architecture of the proposed system considering server and client. The proposed SHM and HM encoder limit temporal motion information for independent tile extraction and transmission. The encoder also generates an extraction information sets (EIS) supplemental enhancement information (SEI) message, which is information for extraction [6]. The extractor obtains the viewport information from the client and extracts the bitstream of the desired tile based on the EIS SEI information. Finally, the client decodes the extracted bitstream.

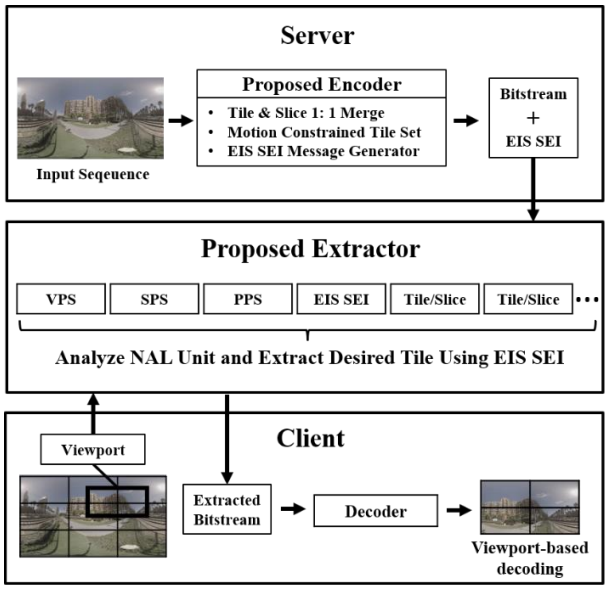


Figure 1: The conceptual architecture of the proposed system.

MMSys'18, June 12–15, 2018, Amsterdam, Netherlands

ACM ISBN 978-1-4503-5192-8/18/06.
https://doi.org/10.1145/3204949.3208119

2 VIEWPORT DEPENDENT 360-DEGREE VIDEO STREAMING

2.1 Challenge: Reference to Un-Decoded Tiles in TIP

The existing SHVC performs TIP within the same layer and inter layer prediction (ILP) between different layers through an upsampling filter [7]. This works well when the decoder decodes all layers into a full picture, but when decoding some tiles, there is a problem with motion estimation and compensation in TIP. Fig. 2 shows an incorrect reference of the problem mentioned above when decoding only the ROI tiles in the EL. Un-decoded tiles are generated when decoding only the ROI tiles. At this time, when the current picture (PicEL t) refers to the previous picture (PicEL t-1), if the motion vector generated by the encoder points to the un-decoded tile, the decoding problem occurs. Therefore, this study proposes to correct motion information on the encoder to resolve issues that arise when decoding only the selected tiles of the entire picture. When the motion vector of TIP points to the position within the same position tile, the PU of the current picture (PicEL t) refers to the PU of the previous picture (PicEL t-1). When calculating the rate-distortion cost (RD cost) to find the optimal PU, both the upsampled BL and the previous picture of the EL are included. The encoder chooses the PU with a more efficient RD cost in either of these options. When referring to a tile at the other position, only the BL is used as a reference picture of the TIP. On the other hand, HEVC encoder does not have ILP. Therefore, HEVC encoder performs intra prediction when the tile temporally refers to the tiles at the other position.

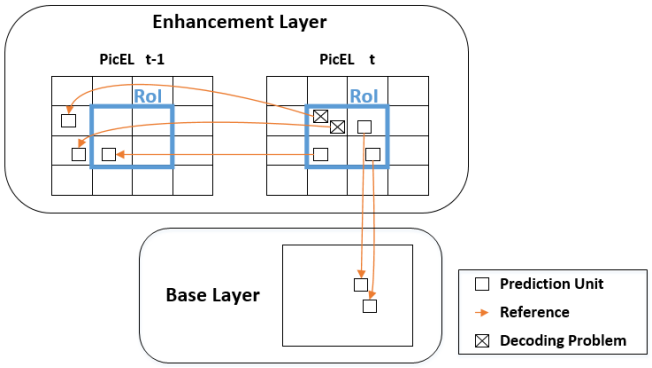


Figure 2: Problem with non-ROI tile references in the SHVC decoder.

2.2 EIS SEI Message Generation for Bitstream Extraction

After generating the bitstream through the encoder using the MCTS, the next step is to extract the bitstream corresponding to the desired tile. The first thing to do is to merge the tile and slice. The reason for merging tile and slice is that tile is not NAL unit. The extractor parses the NAL unit header to distinguish parameter sets, SEI messages and slices. The slice, which is a NAL unit, makes it possible to distinguish tile from tile in the bitstream. The encoder also needs to transmit a replacement parameter set according to the tile being extracted. Table 1 shows the information to be changed in the parameter set. Therefore, MPEG has drafted EIS SEI Message of NAL unit [6]. An EIS SEI Message can contain nearly all extracted cases because it can have about 2,000 EISs. Fig. 3 describes the information that an EIS has. An EIS can contain approximately 2,000 MCTS sets and one parameter set. Slice reordering is optionally used to change the slice address of a tile to render properly [8]. The MCTS set contains a set of tiles to be extracted. Since an EIS has one parameter set, the MCTS sets in an EIS should have the same resolution and the same number of tiles. For example, suppose a 300x300 resolution picture is composed of nine tiles of 100x100. One EIS can generate nine MCTS sets with one tile, and the other EIS can generate six MCTS sets with four tiles in a square shape. The reason why there are six combinations of four tiles in a square is that the left and right sides of the ERP are connected, so two cases are added in four cases [8]. The MPEG draft document [6] contains the syntax of the EIS SEI message, and our research generates the EIS SEI message code based on this syntax.

Table 1: Replacement Information of Parameter Set

Parameter set	Replacement Information
VPS	Level
SPS	Picture width and height
PPS	Tile partition

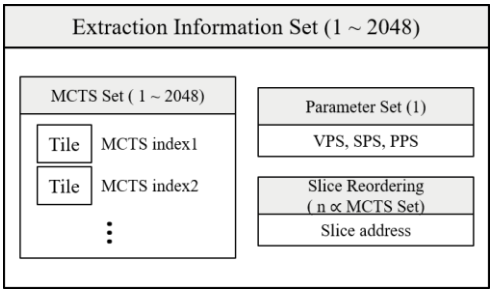


Figure 3: Information(MCTS Set, Parameter Set, Slice Reordering) contained in an EIS.

3 IMPLEMENTATION

Tiles are spatially independent to support parallel processing. However, since the reference pictures have already been decoded, the tiles are not temporally independent. Therefore, Interpolation should be considered when using motion vectors to determine if the referenced PU is within the tile at the same position in TIP. SHM and HM use an eight-tap filter to interpolate luma prediction. When the eight-tap filter is applied horizontally, the three pixels to the left and the four pixels to the right of the current pixel are used. When applied vertically, the top three pixels and the bottom four pixels from the current pixel are used. The left of Fig. 4

describes the interpolation problem of referring to the tile at the same position in TIP. If the PU temporally references to the area of the interpolation problem, the tile cannot be transmitted independently because the PU interpolates using pixels from other tiles. Therefore, the oblique area should be excluded from the TIP reference range.

The SHM and HM use advanced motion vector prediction (AMVP) and Merge to reduce the amount of motion information in the inter prediction. Both modes use spatial and temporal candidate blocks. The block to the bottom right and at the center of the current PU are used as temporal candidates [9]. Basically, when the block to the bottom right of the current PU belongs to a coding tree unit (CTU) beyond the current CTU row, the block is not considered as a temporal candidate [10]. However, there is a problem when the candidate block goes out of the column boundary, not the CTU row. The right of Fig. 4 describes temporal candidate problem at column boundary between Tiles. When the H candidate block is selected, independent tile transmission is not guaranteed because it uses motion information of another tile. We excluded the H block from the candidate under the conditions shown in the right of Fig 4.

The proposed source can be found in HM-16.18 through the #if MCTS_ENC_CHECK directive [16].

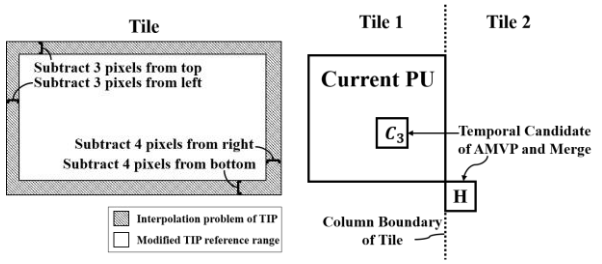


Figure 4: (a) Temporal reference range adjustment considering interpolation, (b) Temporal candidate restriction of AMVP and Merge.

The extractor extracts the bitstream corresponding to the desired tiles from the bitstream generated by the encoder to which the MCTS is applied. First, the executor inputs target EIS id, MCTS set id, and highest temporal id. Thereafter, the extractor starts parsing each NAL units. When parsing the EIS SEI message, the extractor replaces the original parameter set with the replacement parameter set corresponding to the target EIS. To select the desired tiles, the extractor parses the header of the tile / slice NAL units and determines whether the *nuh_temporal_id_plus1* – 1 of the NAL unit is smaller than the target highest temporal id [11]. In the case of a NAL unit generated by the SHM encoder, it is determined whether *nuh_layer_id* is greater than 0 in order to discriminate EL. When the condition of the temporal id is satisfied, the extractor selects tiles corresponding to the MCST index in the target MCTS set. Finally, the extractor modifies *first_slice_segment_in_pic_flag* and *slice_segment_address* of the selected tile / slice header.

4 EXPERIMENTAL RESULTS

This paper uses the test sequences selected by JVET and this test sequences are encoded with general coding options for Random Access (RA) coding structure [12]. This paper set the uniformly 9 tiles configuration considering the field of view (FOV) is 90 ~ 110 degrees [13]. Therefore, the viewport is included in at least one to four tiles. This paper extracts one tile into nine cases and four tiles into six cases.

The bit-rate of the Proposed SHM and HM increased by 8% and 11% on average. The PSNR of the Proposed SHM and HM decreased by 0.04dB and 0.05dB on average. The proposed method increases the bit-rate and decreases PSNR because the motion vector and temporal candidates of AMVP and Merge are limited to allow tiles to be transmitted independently.

Table 2 shows the bit-rate saving ratio when selected tile is transmitted using proposed SHM encoding compare with when transferring the entire picture. Table 3 shows the same measurement method of Table2, but proposed HM encoder was used. When using the proposed SHM encoder, average bit-rate saving is 48% when transmitting 4 tiles and at least 75% up to 97% when transmitting only one tile. For the proposed HM encoder, average 4 tiles bit-rate savings of 47% are achieved. When the server sends only the tiles that correspond to the viewport, a significant bit rate is saved.

Table 2: Comparison Ratio of The Bit-rate to Select and Transmit Tiles Using Proposed SHM Encoding

Name	4 tiles bit-rate saving (average)	1 tile bit-rate saving (min, max)
KiteFlite	49%	77%, 96%
Harbor	46%	63%, 98%
Trolley	50%	80%, 98%
GasLamp	48%	80%, 97%
Average bit-rate saving	48%	75%, 97%

Table 3: Comparison Ratio of The Bit-rate to Select and Transmit Tiles Using Proposed HM Encoding

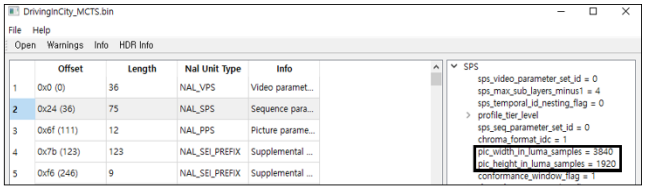
Name	4 tiles bit-rate saving (average)	1 tile bit-rate saving (min, max)
KiteFlite	49%	78%, 96%
Harbor	46%	63%, 98%
Trolley	49%	80%, 98%
GasLamp	47%	79%, 97%
Average bit-rate saving	47%	75%, 97%

5 DEMONSTRATION PLAN

Our demo consists of two 360 video players. Before playback, the proposed encoder generates a single bitstream containing MCTS and EIS SEI Message. Using this single bitstream, the two

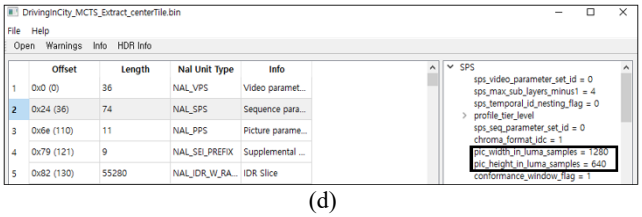
players play the full picture and the extracted picture. Since we only modified HM and SHM encoders, any decoder software can be used. Our implemented player can perform real-time decoding using OpenHEVC of FFmpeg [15]. (a) and (b) of Fig. 5 show the rendering scene and the SPS NAL structure for the entire bitstream [14]. (c) and (d) of Fig. 5 show the rendering scene and the SPS NAL structure for bitstream obtained by extracting one tile from the bitstream of all 9tiles [14].

(a)



(b)

(c)



(d)

Figure 5: (a) Screen capture of player rendering 3840x1920 full picture; (b) SPS NAL structure of 3840x1920 full-screen bitstream; (c) Screen capture of player rendering 1280x640 One extracted tile of 9 tiles; (d) SPS NAL structure of 1280x640 One extracted tile bitstream of 9 tiles bitstream.

6 CONCLUSION

This study proposes an ROI-based SHVC and HEVC tile method. More specifically, the BL encodes a full picture using the SHVC encoder, and the ROI encodes the tiles of the EL referring only to the upsampled BL or a valid tile in TIP. HEVC encoder performs intra prediction when the tile temporally references to the tiles at the other position. The proposed method has a slightly lower encoding efficiency than the original encoding because the motion vector and temporal candidates of AMVP and Merge are limited. However, The proposed method can transmit tiles independently. SHM and HM save average 48% and 47% bit-rate when transmitting only 4 tiles of 9 tiles. Also, SHM and HM save at least 75% up to 97% bit-rate with four sequence averages when only one tile is sent.

7 ACKNOWLEDGMENT

This research was supported by the Korea Electric Power Corporation (Grant number: R17XA05-68).

REFERENCES

[1] M. Champel, T. Stockhammer, T. Fautier, E. Thomas and R. Koenen. 2016. *Quality Requirements for VR*. Technical Report ISO/IEC JTC1/SC29/WG11/ MPEG 116/m39532.

[2] A. Zare, A. Aminlou, M. Hannuksela and M. Gabbouj. 2016. HEVC-compliant Tile-based Streaming of Panoramic Video for Virtual Reality Applications. *In Proceedings of the 24th ACM Multimedia Conference.* Netherlands, 601-605.

[3] Y. Sánchez, R. Skupin and T. Schierl. 2017. Video processing for panoramic streaming using HEVC and its scalable extensions. *Multimedia Tools and Applications*. 5631–5659. DOI: https://doi.org/10.1007/s11042-016-4097-4

[4] Y. K. Wang, Hendry and M. Karczewicz. 2016. *Viewport dependent processing in VR: partial video decoding*. Technical Report ISO/IEC JTC1/SC29/WG11/ MPEG 116/ m38559.

[5] R. Skupin, Y. Sanchez, K. Suhring, T. Schierl, E. S. Ryu and J. Son. 2017. *Temporal MCTS Coding Constraints Implementation*. Technical Report ISO/IEC JTC1/SC29/WG11/MPEG 120/ m41626.

[6] J. Boyce, A. Ramasubramanian, R. Skupin, Gary J. Sullivan, A. Tourapis and Y. K. Wang. 2017. *HEVC Additional Supplemental Enhancement Information (Draft 4)*. Technical Report ISO/IEC JTC1/SC29/WG11/JCTVC-AC1005.

[7] J. Boyce, Y. Ye, J. Chen and A. K. Ramasubramoian. 2015. Overview of SHVC: scalable extensions of the high efficiency video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*. 26, 1(2015), 20-34.

[8] R. Skupin and Y. Sanchez. 2017. MCTS extraction with optional slice reordering. Technical Report ISO/IEC JTC1/SC29/WG11/MPEG 119/m40902.

[9] C. Feldmann, C. Bulla and B. Cellarius. 2013. Efficient stream-reassembling for video conferencing applications using tiles in HEVC. *Proc. of International Conferences on Advances in Multimedia (MMEDIA)*.130-135.

[10] B. Bross, P. Helle, H. Lakshman and K. Ugur. 2014. Inter-Picture Prediction in HEVC. *Integrated Circuits and Systems*. Springer, 113-140.

[11] R. Sjoberg, Y. Chen, A. Fujibayashi, M. M. Hannuksela, J. Samuelsson, T. K. Tan, T. K., Y. K. Wang and S. Wenger. 2012. Overview of HEVC high-level syntax and reference picture management. *IEEE Transactions on Circuits and Systems for Video Technology.* 22, 12(2012), 1858-1870.

[12] J. Boyce, E. Alshina, A. Abbas and Y. Ye. 2017. JVET common test conditions and evaluation procedures for 360° video. Technical Report ISO/IEC JTC1/SC29/WG11/JVET-H1030.

[13] Wikipedia. 2018. Comparison of virtual reality headsets. Retrieved from https://en.wikipedia.org/wiki/Comparison_of_virtual_reality_headsets

[14] Github. 2018. HEVCESBrowser. Retrieved from https://github.com/virinext/hevcesbrowser

[15] Github. 2018. FFmpeg. Retrieved from https://github.com/FFmpeg/FFmpeg.

[16] High Efficiency Video Coding (HEVC) reference software HM. 2017. Retrieved from https://hevc.hhi.fraunhofer.de/