Implementing Motion-Constrained Tile and Viewport Extraction for VR Streaming

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

[1] 360-degree video streaming for virtual reality (VR) is emerging. However, 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 360 video tiled 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. 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

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

ACM ISBN 978-1-4503-5772-2/18/06..$15.00
https://doi.org/10.1145/3210445.3210455

1 INTRODUCTION

Recently, many kinds of head-mounted display (HMD) devices and 360-degree cameras are introduced on the market. According to the emerging technologies on virtual reality, new 360-degree video coding and file format standards are actively studying [15].

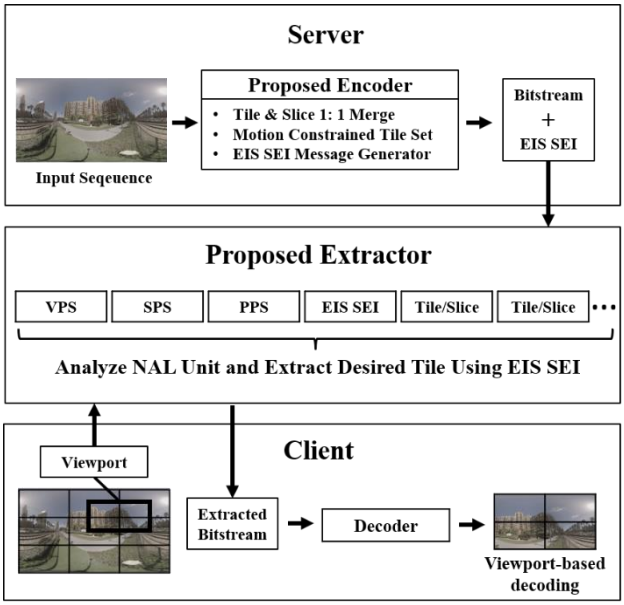


Figure 1: The conceptual architecture of the proposed system.

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. The resolution of high quality VR is 12K [1]. Its resolution is 36 times larger than Full high-definition (HD). The high quality VR requires high computing power, but the specification of VR that is currently on the market is not fully supported. Therefore, it needs to work efficiently in these limited conditions. Thus, Motion-constrained tile set (MCTS)

was discussed for viewport-based transmission in standard meetings [2]. 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. We propose the implementation method of MCTS in SHVC and HEVC. The proposed implementation is adopted in the MPEG standard 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 SHVC reference software (SHM) and HEVC reference software (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 [3]. 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.

2 RELATED WORK

2.1 Prior Work for 360 Video Streaming System

The goal of merciless video processing (MVP) project of Gachon University is to provide the highest quality video that can be reached in a limited mobile VR performance. Fig. 2 shows the conceptual architecture of MVP project. Through the proposed implementation, the entire picture is transmitted in a low quality and ROI is transmitted in a high quality. The entire picture is communicated with the PC using mmWave and decoded using digital signal processor (DSP) [4]. ROI is decoded by parallel processing using mobile asymmetric big/little cores [5]. The size of the tile is optimally allocated to big/little core to improve decoding efficiency. The proposed implementation is part of the MVP project and is applicable to both mobile VR and tethered VR.

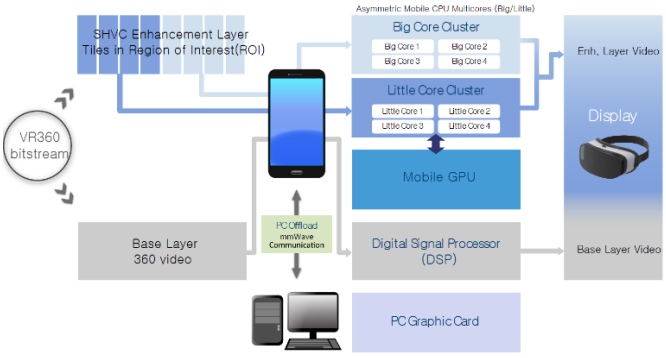


Figure 2: The conceptual architecture of Merciless video processing (MVP) project

2.2 Viewport Dependent Studies on 360 Video

The area that the user is looking at in the 360 picture is part of the entire picture. The bitrate is significantly reduced when the server only transmits the viewport area. A simple way to do this is to split the picture and encode them separately, but it requires a step of segmenting the picture and a lot of storage space. Conversely, in order to transmit only the area corresponding to the viewport in one bitstream, motion vector (MV) should be considered. Because a decoding error occurs by referring to the area that is not transmitted. Y. Sanchez *et al.* [6] proposed Generated Reference Picture (GRP) to prevent prediction mismatch. All prediction unit (PU) of GRP have a MV associated that compensates the movement caused by sending only selected tiles. This study fixed MV reference errors through GRP, but there is still overhead for GRP generation. The MPEG standard discussed the MCTS, a way to limit MV of the current picture [2]. MCTS limits the temporal motion information in the encoder so that the encoding efficiency is slightly lower. However, a single bitstream using MCTS can decode only the desired tiles of the full picture without additional picture generation. A. Zare *et al.* [7] modified the HEVC encoder as a concept of MCTS. This study was conducted with three tiling methods and resulted in 3% to 6% penalties in compression. However, when only the tile corresponding to the field of view (FOV) is transmitted, the streaming bit-rate saving from 30% to 40% was achieved. Compared with this study, our study modified the SHVC reference software(SHM) and HEVC reference software (HM), and details the implementation of the proposed MCTS in SHVC and HEVC encoder according to the standard. We also describe in detail the process of extracting a single bitstream composed of MCTS using the EIS.

3 360 VIDEO TILED STREAMING

3.1 Challenge: Reference to Un-Decoded Tiles in TIP using Existing SHVC

The existing SHVC performs TIP within the same layer and inter layer prediction (ILP) between different layers through an up-sampling filter [8]. 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. 3 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, we propose 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 up-sampled 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. Since the HEVC encoder is a single layer, it does not

perform ILP. Therefore, HEVC encoder performs intra prediction when the tile temporally refers to the tiles at the other position.

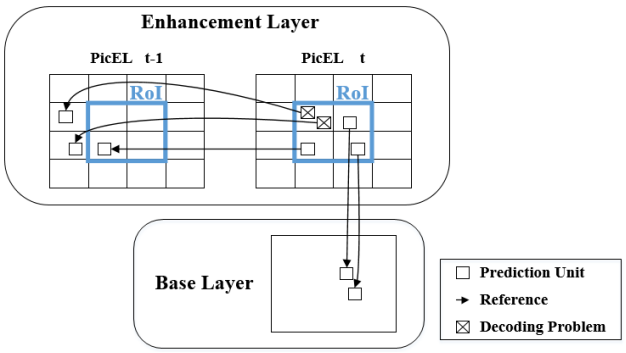


Figure 3: Problem with non-ROI tile references in the SHVC decoder.

3.2 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 extractor parses the NAL unit header to distinguish parameter sets, SEI messages and slices. Since the tile is not a NAL unit, we merge the slice and the tile in the encoder configuration options. 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. MPEG has drafted EIS SEI Message of NAL unit to replace the parameter set [3]. An EIS SEI Message can contain nearly all extracted cases because it can have about 2,000 EISs. Fig. 4 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 [9]. 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 300×300 resolution picture is composed of nine tiles of 100×100. 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 [9]. Our research implements a method of generating EIS SEI messages based on MPEG draft documents [3].

Table 1: Replacement Information of Parameter Set

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

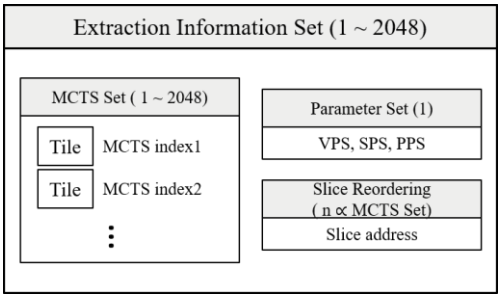


Figure 4: Information (MCTS Set, Parameter Set, Slice Reordering) contained in an EIS.

4 IMPLEMENTATION

4.1 Encoder: Modifying TIP Information for MCTS

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. 5 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.

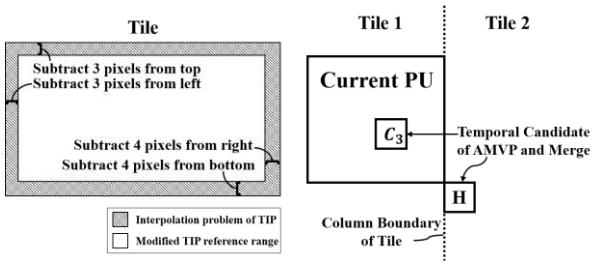


Figure 5: (a) Temporal reference range adjustment considering interpolation, (b) Temporal candidate restriction of AMVP and Merge.

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 [10]. 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 [11]. However, there is a problem when the candidate block goes out of the column boundary, not the CTU row. The right of Fig. 5 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 5.

The proposed source can be found in HM-16.18 through the *#if MCTS_ENC_CHECK* directive [14].

4.2 Encoder: EIS SEI Message Generation

The SHM and HM encoders perform three steps to generate SEI NAL units: variable management class generation, setting of values, and output to file. Fig. 6 shows the class and functions that performed the above three steps for EIS SEI message generation. *SEIMCTSExtractionInfoSets* class declares and manages variables according to the MPEG EIS SEI draft document [3]. *initSEIMCTSExtractionInfoSets()* function sets the values of all variables related to the SEI EIS message. Section 3.2 explained that the EIS generates replacement parameters according to the information in Table 1. This function creates raw byte sequence payload (RBSP) of the replacement parameter set. Depending on the size of the picture to be extracted, *general_level_idc* of the VPS NAL unit and *pic_width_in_luma_samples*, *pic_height_in_luma_samples* of the SPS NAL unit are modified. Also, *num_tile_columns_minus1*, *num_tile_rows_minus1*, *uniform_spacing_flag*, *column_width_minus1* and *row_height_minus1* of the PPS NAL unit are modified according to the extracted tiles. Finally, *xWriteSEIMCTSExtractionInfoSets()* function performs entropy coding on the values and generates an EIS SEI message NAL unit.

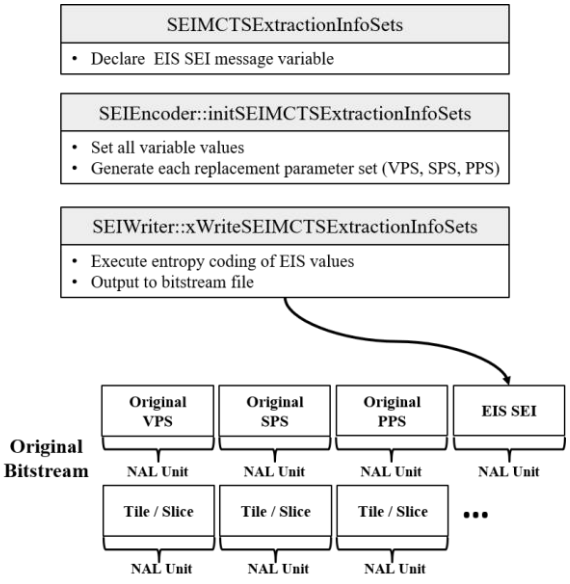


Figure 6: Key class and functions for EIS SEI message generation.

4.3 Extractor: Extracting Desired Tiles

The extractor extracts the bitstream corresponding to the desired tiles from the bitstream generated by the encoder to which the MCTS is applied. Fig. 7 illustrates the functional flow chart of the extractor. The input values are target EIS id, MCTS set id, and highest temporal id. Depending on the input value, the extractor starts parsing each NAL units. When parsing the original PPS, the extractor gets the number of original tiles to parse NAL units per picture. Next, the extractor parses the EIS SEI message generated by the encoder. Using the EIS SEI message, the extractor replaces the original parameter set with the replacement parameter set corresponding to the target EIS id. 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 [12]. *nuh_temporal_id_plus1* enables an adaptive response to the bandwidth. 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. *nuh_layer_id* of EL has a value greater than 0. When the condition of the temporal id is satisfied, the extractor selects tiles corresponding to the target MCTS set id. Finally, the extractor modifies *first_slice_segment_in_pic_flag* and *slice_segment_address* of the selected tile / slice header. Set *first_slice_segment_in_pic_flag* to 1 if the selected tile / slice is first on a picture, otherwise 0. If *slice_reordering_flag* of the EIS SEI message is 1, set *slice_segment_address* of tile / slice header to *output_slice_segment_address* of EIS SEI message. Otherwise, *slice_segment_address* is set based on the raster scan addressing scheme. When all NAL units are parsed, the extractor generates an extracted bitstream and ends.

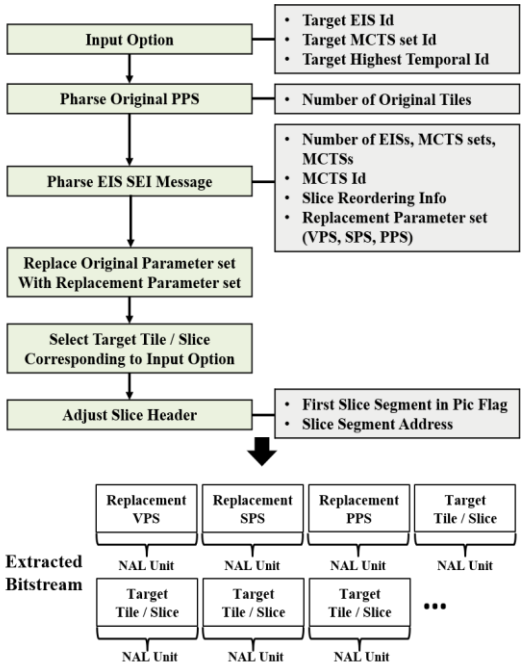


Figure 7: Functional flow chart of extractor.

5 EXPERIMENTAL RESULTS

We uses the test sequences selected by JVET as shown in Table 2 and this test sequences are encoded with general coding options for Random Access (RA) coding structure as shown in Table 3 [13-15]. We set the uniformly 9 tiles configuration considering the field of view (FOV) is 90~110 degrees [16]. Therefore, the viewport is included in at least one to four tiles. We extract one tile into nine cases and four tiles into six cases. Fig. 8 shows six cases when extracting four tiles and a picture divided into nine tiles. Case 5 and 6 are added because both sides of the ERP are attached.

Table 2: Information of Test Sequences

Name	Resolution	Frame length	Frame rate
KiteFlite	8192×4096	300	30 fps
Harbor	8192×4096	300	30 fps
Trolley	8192×4096	300	30 fps
GasLamp	8192×4096	300	30 fps

Table 3: Coding Options

Coding option	SHM Parameter	HM parameter
Version	12.3	16.16
CTU size	64×64	
Coding structure	RA	
QP	-	22
Base Layer QP	22	-
Enhancement Layer QP	22	-
Tile	Uniformly 3x3 = 9 tiles	
Slice mode	3 : Tile in a Slice	
Slice Argument	1	

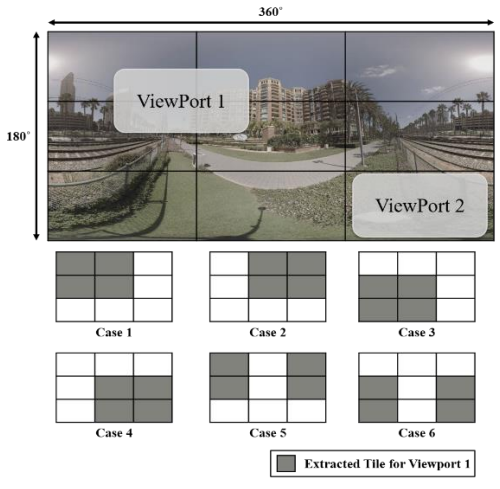


Figure 8: Tile partition and extraction considering viewport.

Table 4 and 5 shows the bit-rate and peak signal-to-noise ratio (PSNR) difference between original and proposed entire picture encoding before extraction. 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. Although PSNR was very slightly lost, since tiles can be transmitted independently, bit-rate can be saved when only some tiles are transmitted.

Table 4: Bit-rate Differences between Original and Proposed Entire Picture Encoding

Name	Proposed SHM	Proposed HM
KiteFlite	6%	8%
Harbor	5%	8%
Trolley	10%	13%
GasLamp	13%	17%
Average bit-rate increase	8%	11%

Table 5: PSNR Differences between Original and Proposed Entire Picture Encoding

Name	Proposed SHM	Proposed HM
KiteFlite	-0.04dB	-0.05dB
Harbor	-0.03dB	-0.02dB
Trolley	-0.06dB	-0.07dB
GasLamp	-0.06dB	-0.06dB
Average PSNR decrease	-0.04dB	-0.05dB

We compared the bit-rate when transmitting the bit-stream of the extracted tiles by applying the proposed method and the original bit-stream without applying the MCTS. Table 6 shows the bit-rate saving ratio when selected tile is transmitted using proposed SHM encoding compare with when transferring the entire original bit-stream. Table 7 shows the same measurement method of Table 6, 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% and 1 tiles bit-rate savings of at least 75% up to 97% are achieved. When the server sends only the tiles that correspond to the viewport, a significant bit rate is saved. Fig. 9 shows a scene in which the whole and extracted bitstream are decoded / rendered using mmWave communication in mobile VR. We experimented with uniformly dividing nine tiles. In the future, it is necessary to allocate the tile size based on the viewport by reducing the size of the tile of both poles. We also plan to experiment with other projections besides ERP for various experiment comparisons.

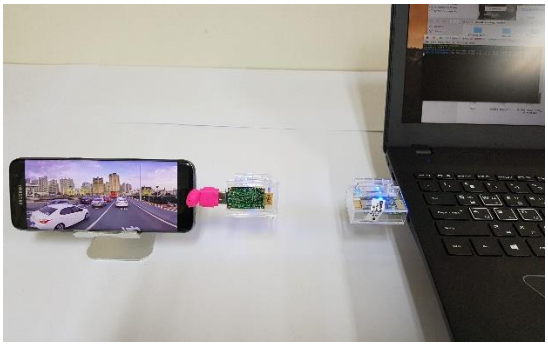
Table 6: Comparison Ratio of The Bit-rate to Select and Transmit Tiles Using Proposed SHM Encoding

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 7: 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%

(a)



(b)

Figure 9: Implemented mobile VR 360 video player using mmWave communication; (a) 3840x1920 full picture consisting of 9 tiles; (b) 1280x640 one extracted tile of 9 tiles.

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 up-sampled 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 HM encoding efficiency average 11% bitrate and 0.05dB PSNR 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 ROI 4 tiles among 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 MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2018-2017-0-01630).

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] 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.

[3] 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.

[4] T. T. Le, D. N. Van and E. S. Ryu. 2018. Real-time 360-Degree Video Streaming over Millimeter Wave Communication. *International Conference on Information Networking 2018 (ICOIN2018)*. Malaysia, 5-17.

[5] Y. Ryu and E. S. Ryu. 2017. Video on Mobile CPU: UHD Video Parallel Decoding for Asymmetric Multicores. *Proceedings of the 8th ACM International Conference on Multimedia System (MMSys 2017)*. ACM, Taiwan 229-232.

[6] 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

[7] 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.

[8] 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.

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

[10] 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

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

[12] 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.

[13] HEVC Scalability Extension (SHVC) reference software SHM. 2016. Retrieved from https://hevc.hhi.fraunhofer.de/shvc.

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

[15] 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.

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