Computers and Electrical Engineering 72 (2018) 361–368

Contents lists available at ScienceDirect

Computers and Electrical Engineering

journal homepage: www.elsevier.com/locate/compeleceng

Tile-based 360-degree video streaming for mobile virtual reality in cyber physical system☆

Jangwoo Son, Eun-Seok Ryu*

Department of Computer Engineering, Gachon University, 1342 Seongnamdaero, Sujeong-gu, Seongnam, Gyeonggi 13120, Republic of Korea

ARTICLE INFO

Article history:
Received 18 September 2017
Revised 30 June 2018
Accepted 1 October 2018

Keywords:
VR
Tile
region of interest (ROI)
motion-constrained tile set (MCTS)
SHVC

ABSTRACT

Today, the demand for and interest in virtual reality (VR) is increasing, since we can now easily experience VR in many applications. However, the computational ability of mobile VR is limited compared to that of tethered VR. Since VR represents a 360-degree area, providing high quality only for the area viewed by the user saves considerable bandwidth. Therefore, we propose a new tile-based streaming method that transforms 360-degree videos into mobile VR using high efficiency video coding (HEVC) and the scalability extension of HEVC (SHVC). While the SHVC base layer (BL) represents the entire picture, the enhancement layer (EL) can transmit only the desired tiles by applying the proposed method. By transmitting the BL and EL using region of interest (ROI) tiles, the proposed method helps reduce not only the computational complexity on the decoder side but also the network bandwidth.

1. Introduction

In recent years, many companies have launched head-mounted displays (HMDs), and new standards are being created for 360-degree video streaming.

HMDs are display devices worn on the head and have a display optic in front of one or each eye. These devices support head tracking to provide a 360-degree view and therefore require high-quality and high-performance hardware. The recommended specification for Oculus Rift, a type of tethered virtual reality (VR) system, is the Intel i5-4590, Nvidia GeForce GTX 970 processor with 8 GB RAM. In contrast to a tethered VR system, which is a PC-based HMD, a mobile VR system such as Samsung's Gear VR headset has limited video processing capabilities. Table 1 illustrates the differences between mobile VR and tethered VR systems.

Table 1 shows that, while a mobile VR system is more convenient, its performance is poor compared to that of a tethered VR system. To increase the video processing efficiency with a limited specification, we propose a method to solve the problems of bitrate and computational complexity through region of interest (ROI)-based SHVC tile processing. We propose a solution to the motion compensation problem that occurs when the enhancement layer (EL) sends selected ROI tiles and the base layer (BL) sends a full picture using an SHVC encoder. Furthermore, we propose a method of sending selected ROI tiles in a single layer using the HEVC encoder.

☆ Reviews processed and recommended for publication to the Editor–in–Chief by Guest Editor Dr. Jia–Shing Sheu.

* Corresponding author.
E-mail address: esryu@gachon.ac.kr (E.-S. Ryu).

https://doi.org/10.1016/j.compeleceng.2018.10.002
0045-7906/

Table 1
Differences between mobile VR and tethered VR systems.

	Mobile VR	Tethered VR
Pros	Wireless	Computing power
	Portability	Various content
Cons	Less-capable tracking	Limited portability
	Performance ∝smartphone	Expensive

The composition of the paper is as follows. Section 2 gives a brief description of 360 video standards and ROI-related research. Section 3 describes the architecture of the proposed methods. Section 4 describes the implementation process, and Section 5 shows the performance of each technology.

2. Related work

2.1. 360 video standards of MPEG, JCT-VC, and JVET

The Moving Picture Experts Group (MPEG), the Joint Collaborative Team on Video Coding (JCT-VC), and the Joint Video Exploration Team (JVET) have discussed various 360-degree video streaming standards for VR. JVET defines Common Test Conditions (CTC) and evaluation procedures for 360 video [1]. Since VR requires high quality resolution, the test sequence is composed of 4 K and 8 K. In addition, MPEG-I (MPEG Immersive media) introduced the three-step goals for 360 video [2]. In the first phase, the aim of MPEG-I was to complete a 3 Degree of Freedom (3DoF) standard by 2017. In the second phase, their goal is to activate VR commercial services and to support 3DoF+by 2020. The objective of the last phase is to support 6DoF by 2022. This allows the user's movements to be reflected in VR. In addition, MPEG DASH-VR standardized the dynamic adaptive streaming over http (DASH) syntax for VR. They configured five used cases for compatibility and efficient streaming, one of which is viewport-based DASH streaming for VR content [3]. In addition to DASH, the viewport users observe is one of the key points according to VR standards for reducing bandwidth. To this end, the standardization groups have discussed the possibility of motion-constrained tile sets (MCTS) [4].

2.2. Single encoding based on MCTS

Unlike the conventional encoder, the MCTS-applied encoder does not temporally refer to tiles that have different positions on the current picture and the reference picture. Thus, the tiles can be separated in one bitstream, although the bitrate increases slightly. A. Zare et al. explains a method of saving bitrate when sending only the field of view (FOV) area using the MCTS-applied encoder [5]. In their study, the MCTS method is applied and the encoding efficiency is reduced by from 3% to 6%. However, the study reduces the bitrate by from 30% to 40% when transmitting tiles corresponding to FOV. Compared with their study, our study embodies the installation process for applying MCTS to the HEVC reference software (HM) and SHVC reference software (SHM), and describes implementation issues.

2.3. Tile based panoramic streaming using SHVC

Y. Sanchez et al. proposed a technique to minimize picture transition delay and bitrate according to the change of ROI, which is a point seen by the user when using SHVC [6]. Their technique involved dividing BL and EL into multiple tiles, and only the tiles corresponding to ROI are streamed. However, if streaming only the corresponding tiles, a prediction mismatch occurs when decoding. Fig. 1 depicts the prediction mismatch and its solution. At the encoder, the second tile of the t1 picture refers to the second tile of the t0 picture. Considering the ROI, the t0 picture transmits from the second to the fourth tiles, and the t1 picture transmits from the first to the third tiles. The decoder encounters prediction mismatch with reference to the same second tile using the encoder's motion vector. This study creates a Generated Reference Picture (GRP) between the reference and the current picture in order to correct the motion vector. The GRP holds motion vector information that compensates for the prediction mismatch that occurs when decoding only some of the tiles. The Multi Layer GRP (MLGRP) is similar to GRP, but utilizes the characteristics of SHVC to obtain motion vector information through the lower layer. This study solves the problem of motion vectors, but there is an overhead of generating GRP. In contrast, we solve the motion vector problem in the encoder and perform a single encoding.

2.4. Viewport independent studies on 360-degree video

The viewport independent methods transmit whole 360-degree video such as equirectangular projection (ERP) and cube-map projection (CMP). These methods reduce bitrates and computational complexity by down-sampling and/or increasing the number of quantization parameters (QPs) of lesser important regions. Fig. 2 shows the efficient preprocessing studies using the ERP and CMP regions. The adaptive-QP ERP is encoded into different qualities for each region in consideration of the user's gaze. The region-wise packing also considers the user's gaze. The ERP down-samples the top and bottom regions,

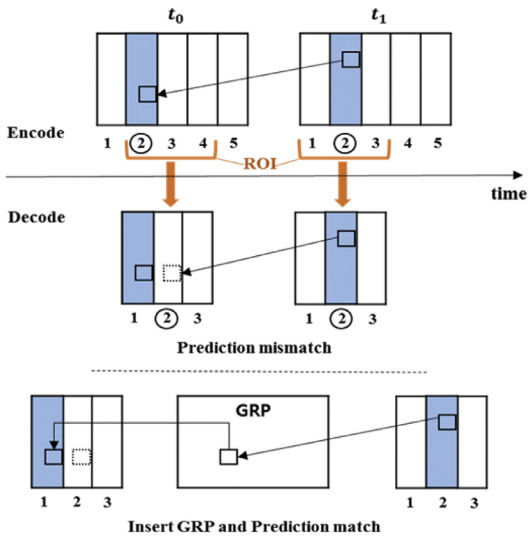

Fig. 1. Prediction mismatch and GRP concept for solving mismatch.

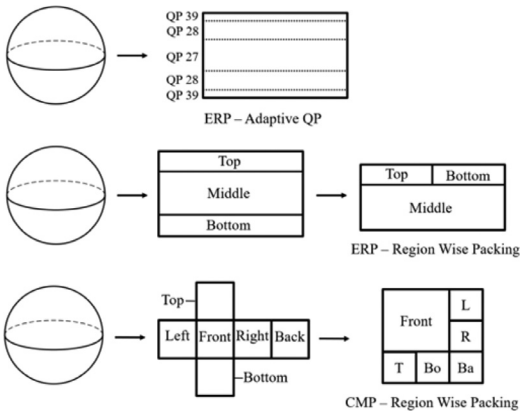

Fig. 2. Viewport independent studies of ERP and CMP.

and the CMP down-samples all regions except for the front region [7]. K. K. Sreedhar et al. presents a comparison of the results for a viewport independent projection [8]. In contrast, our method is viewport-dependent since only the area viewed by the user is transferred to the original ERP.

3. MVP: SHVC tile-based 360-degree video streaming for mobile VR in cyber physical system

3.1. Proposed system architecture of MVP

Since 2015, extensive studies have been carried out on Merciless Video Processing (MVP) projects related to 360-degree video streaming [9,10]. Furthermore, two standardizations (No. MSS.S-Y16-001,002) of ROIs were submitted to Multi-Screen Service Forum Specification in August 2016.

This section describes the architecture of the methods used in the encoder of MVP. First, we explain how to solve problems related to un-decoded tiles in two steps. Through the proposed method, the BL transmits the full picture and the EL transmits only ROI tiles without error. Second, comparing the SHVC and HEVC encoder, we explain how to apply MCTS in the HEVC encoder.

3.2. Challenge: reference to un-decoded tiles in TIP in existing SHVC

When the tile is applied to an existing SHVC encoder, the intra prediction uses only the pixels of the current tile, but the inter prediction refers to all the regions of the reference picture [11]. In the existing encoder, since all the regions of the reference picture are decoded, the tiles temporally refer to the other position tiles as well as the current position tiles in the

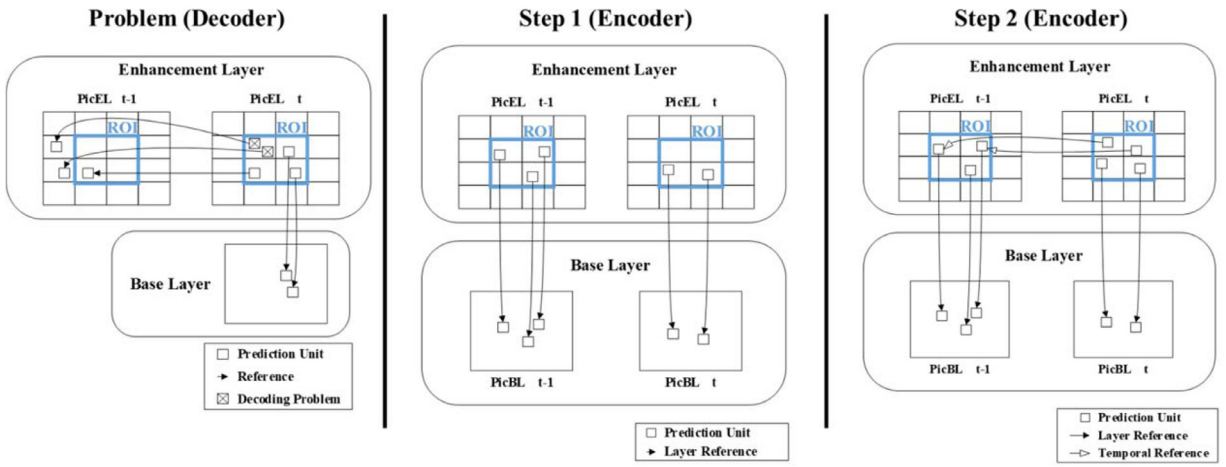

Fig. 3. Problem with non-ROI tile references in the SHVC decoder. In Step 1, the EL refers only to the picture up-sampled by the BL, and in Step 2, the current picture refers to the prediction unit when TIP points to the tile at the same position in the EL.

reference picture. Therefore, when the decoder decodes the selected tiles in the picture, a problem occurs with the motion estimation and compensation in the Temporal Inter Prediction (TIP). The left side of Fig. 3 depicts a temporally reference problem that occurs when the decoder decodes only the center four tiles using a bitstream created by an existing encoder. The decoding problem occurs because the Prediction Unit (PU) of the current picture (PicEL t) temporally refers to the non-decoded tile of the reference picture (PicEL t-1) using the motion vector determined by the existing encoder. Therefore, our study solves the problem of temporally referring to tiles that are not decoded through Step 1 and Step 2. Through the proposed method, the encoder implements MCTS.

3.2.1. Proposed step 1: tile encoding in EL using up-sampled BL

Step 1 solves the limitations of the above problem using an up-sampled BL. As shown in the center of Fig. 3, the encoder considers only the pictures of the up-sampled BL as a reference picture, and does not consider the EL. Therefore, the ROI tiles refer to the up-sampled BL in the same picture. As the BL is encoded in the entire picture, the ROI tiles selected in the EL can refer to all areas of the reference picture. This eliminates reference errors in decoding the ROI tiles in the EL. However, since EL does not use TIP, the bitrate increases significantly.

3.2.2. Proposed step 2: available tile encoding in EL using up-sampled BL and decoded tile of EL

Step 1 solves the problem of referring to the outer region of the ROI. However, as Step 1 uses only an up-sampled BL as a reference list, the PU where TIP is possible still uses the Inter Layer Prediction (ILP). Therefore, as shown on the right side of Fig. 3, when the motion vector of the temporal reference refers to the same position tile of the current picture (PicEL t) and the reference picture (PicEL t-1), the PU of the current picture (PicEL t) refers to the PU of the reference picture (PicEL t-1) considering the interpolation method, Advanced motion vector prediction (AMVP), and MERGE mode. Consequently, Step 2 demonstrates an optimized encoding result compared to Step 1.

3.3. Available tile encoding using the HEVC encoder

This section proposes the MCTS architecture of HEVC to enable tile selective decoding. Fig. 4 shows the proposed architecture of the HEVC encoder. We modify the motion estimation process in the existing encoders. As described in the previous section, the SHVC encoder performs the up-sampled BL as a reference picture when the tile temporally references to the tiles at the other position. On the other hand, the HEVC encoder does not have ILP. Therefore, the proposed HM is modified to use the intra prediction when referring to different position tiles between the current and reference pictures.

4. Implementation

4.1. Modifying range of motion vectors for MCTS

The interpolation is applied to improve the precision of the prediction and the compression performance. The encoder uses the top, bottom, right and left pixels of the current pixel as interpolation, while the nearest pixel in the tile boundary uses other tile pixels outside the boundary for interpolation. For MCTS implementations, temporal references should ensure that motion vectors are not interpolated using pixels from tiles in other positions. SHVC and HEVC use an 8-tap filter on the luma for interpolation. The 8-tap filter interpolates using 7 surrounding pixels, including the current pixel. When

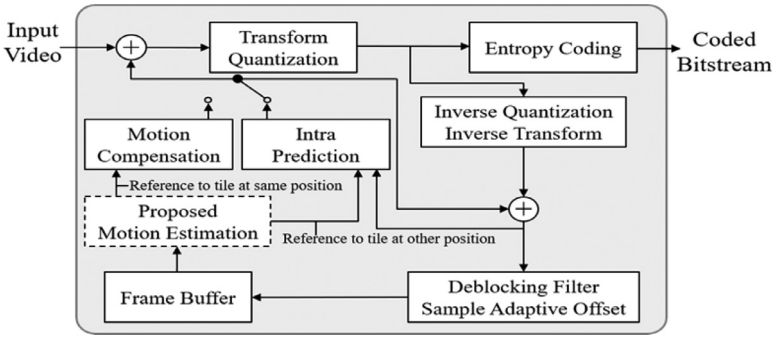

Fig. 4. Proposed HEVC encoder architecture.

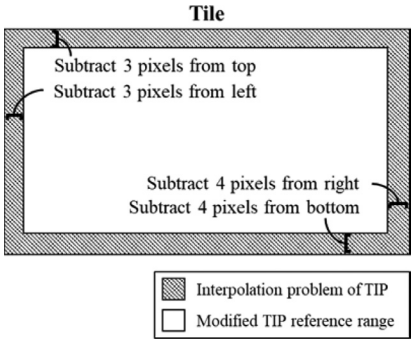

Fig. 5. Interpolation problem of referring to a tile at the same position in TIP.

interpolating horizontally, 3 left and 4 right pixels based on the current pixel are used for interpolation. When interpolating vertically, 3 top and 4 bottom pixels based on the current pixel are used. Fig. 5 describes the modification of the reference range to avoid pixels of different tiles being used for interpolation. Our study excludes the oblique area from the reference range for tile independence when the PU temporally references tiles in the same position on the current and reference pictures.

When implementing MCTS in SHM and HM, the position of the current PU should be considered. The x and y pixel values at the top and left of the current PU can be obtained using the getCUPelX() and getCUPelY() functions, and the x and y pixel values at the bottom and right can be obtained by adding the values obtained through the getWidth() and getHeight() functions in the HM and SHM. However, if the current PU is not in the $2N \times 2N$ mode, its position should be changed. Because the four functions discussed above return a value based on the $2N \times 2N$ mode, the position and size of the PU can be obtained by considering the position of the PU in eight partition modes.

4.2. Temporal candidate of AMVP and MERGE at the column boundary between tiles

AMVP and MERGE increase the encoding efficiency by using the motion information of neighbor candidate blocks. Candidates in both modes include temporal blocks as well as spatial blocks. As mentioned in the previous section, temporal candidates should be considered for MCTS implementation. As shown in Fig. 6, the center (C3) and bottom right (H) blocks on the current PU are used as temporal candidates for AMVP and MERGE [12]. The bottom right block is automatically excluded from the temporal candidate when it crosses the row criteria of the current Coding Tree Unit (CTU). However, if the bottom right block crosses the current CTU column criteria, it is not automatically excluded. Fig. 6 describes the problem using the H block as a temporal candidate at the column boundary between the tiles. In Fig. 6, the H block is a block belonging to a tile at another position, so the tile is not independent. Our study excluded the H block at the column between the tiles.

The modified HM and SHM first determines whether the current CTU is located on the right side of the tile. The CTU position to the right side of the current tile is obtained using the getRightEdgePosInCtus() function, and the current CTU position is obtained using the getFrameWidth InCtus() function and the getCtuRsAddr() function. Next, it is determined whether the current PU in the CTU is located on the right side of the CTU. The position of the current PU is obtained by using the deriveRightBottomIdx(), while the getNumPartInCtuWidth() function and the getNumPartIn CtuHeight() function are used to determine whether or not the position of the current PU is located on the right side of the CTU. If both conditions are met, the H block is excluded from the candidate.

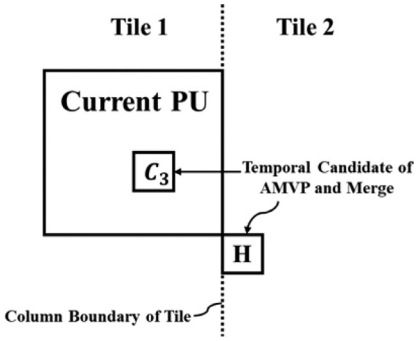

Fig. 6. Temporal candidate problem at column boundary between tiles.

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 3 × 3 = 9 tiles	
Slice mode	Disable all slice options	
WPP mode	Disable all wpp options	
SAO	On	
AMP	On	

Table 4
Bitrate increase ratio compared to original encoding.

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

The implemented MCTS codes with updated HM [13] were presented in the JCT-VC standard meeting in October 2017. This paper extends the efforts to introduce SHM and conducts various coding performance tests.

5. Experimental results

The test sequence in Table 2 is selected by JVET and used in our experiments. Our experiments in Table 3 use the Random Access (RA) coding structure as the coding option, and the sequence is divided into 9 tiles of uniformly 3 × 3 [14,15].

Tables 4 and 5 show the bitrate increase rate and peak signal-to-noise ratio (PSNR) decrease compared to the existing encoding. The bitrate of the proposed SHM increased by 8% and the PSNR decreased by 0.04 dB on average for 4 sequences. For the proposed HM encoder, the bit rate increased by 11% and the PSNR decreased by 0.05 dB. Compared with the existing method, the proposed method restricts temporal reference information, consequently reducing the bitrate and PSNR efficiency. However, the proposed method is able to extract selected tiles by ensuring the independence of tiles.

Tables 6 and 7 show the results of comparing the transmission of some tiles using the proposed encoder and the transmission of all 9 tiles using the existing encoder. When the proposed SHM encoder independently transmits tiles correspond-

Table 5
PSNR decrease compared to original encoding.

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

Table 6
Comparison ratio of the bitrate to select and transmit tiles using proposed SHM encoding.

Name	4 tiles bitrate saving	1 tile bitrate saving
KiteFlite	49%	88%
Harbor	46%	88%
Trolley	50%	87%
GasLamp	48%	87%
Average bitrate saving	48%	87%

Table 7
Comparison ratio of the bitrate to select and transmit tiles using proposed HM encoding.

Name	4 tiles bitrate saving	1 tile bitrate saving
KiteFlite	49%	87%
Harbor	46%	87%
Trolley	49%	87%
GasLamp	47%	86%
Average bitrate saving	47%	87%

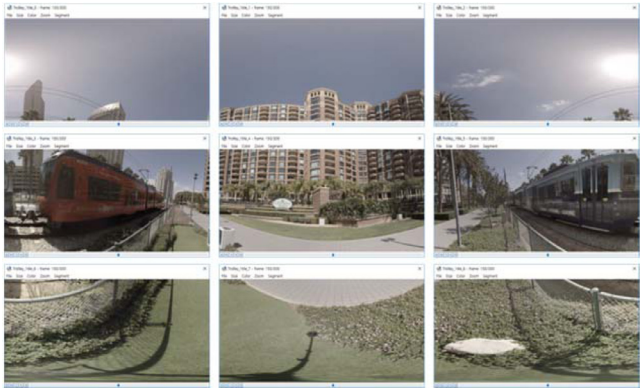

Fig. 7. Independent docoding of extracted tile.

ing to the ROI, average bit rate savings of 48% and 87% are achieved for 4 tiles and 1 tile, respectively. For the proposed HM encoder, average bit rate savings of 47% and 87% are achieved. Using the proposed encoding, the bitrate is significantly reduced when the server only transmits some tiles of the entire picture. Fig. 7 shows that the extracted bitstream is decoded without problems.

6. Conclusion

When decoding only some of the tiles in a bitstream generated by the existing encoder, the decoding problem occurs because the decoder refers to a non-decoded tile. In order to transfer tiles independently, the proposed SHM uses ILP and the proposed HM uses intra prediction when temporally referencing different position tiles of the current and reference pictures. In addition, when referring to the same position tile of the current and reference pictures, the proposed encoder reduces the reference range by considering interpolation and limits the temporal candidates of AMVP and MERGE at the tile boundary. The proposed method restricts temporal reference information, so bitrate and PSNR efficiency are slightly lower. However, the proposed method is able to extract selected tiles by ensuring independence of the tiles. When applying the

proposed method to the SHM encoder and transmitting 4 tiles and 1 tile, the bit rate saves 48% and 87%, respectively. In the case of the proposed HM, the bit rate is reduced by 47% and 87% when transmitting 4 tiles and 1 tile, respectively.

Acknowledgment

This paper was supported by the Gyeonggi-do Regional Research Center (GRRC) program of Gyeonggi province (GRRC-Gachon2017(B01)).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.compeleceng.2018.10.002.

References

[1] Boyce J, Alshina E, Abbas A. JVET common test conditions and evaluation procedures for 360°video. In: 5th JVET meeting of ISO/IEC JTC 1/SC 29/WG 11, no. JVET-F1030; 2017. p. 1–7.
[2] Koenen R, "*Working Draft 0.1 of TR: technical report on immersive media*", 117th MPEG meeting of ISO/IEC JTC1/SC29/ WG11, MPEG2017/ W16718, 2–4.
[3] Gu Y, Higgs P, Zhang E, Gao Y, "*Multiple angle VR streaming*", 117th MPEG meeting of ISO/IEC JTC1/SC29/ WG11, MPEG2017/ M39994, 1–2.
[4] Hannuksela M, Vadakital K.M, Grüneberg K, Sanchezon Y, "*Extractor design for HEVC files*", 114th MPEG meeting of ISO/IEC JTC1/SC29/WG11, MPEG2016/M38147, 1–10.
[5] Zare A, Aminlou A, Hannuksela M, Gabbouj M. HEVC-compliant tile-based streaming of panoramic video for virtual reality applications. In: MM '16 Proceedings of the 2016 ACM on Multimedia Conference; Oct. 2016. p. 601–5.
[6] Sánchez Y, Skupin R, Schierl T. Compressed domain video processing for tile based panoramic streaming using SHVC. In: ImmersiveME '15 Proceedings of the 3rd International Workshop on Immersive Media Experiences; oct. 2015. p. 13–18.
[7] Oh S, Hwang S, "OMAF: Generalized signaling of region-wise packing for omnidirectional video", 118th MPEG meeting of ISO/IEC JTC1/SC29/ WG11, MPEG2017/ m40423, 1–4.
[8] Sreedhar KK, Alireza A, Hannuksela M, Gabbouj M. Viewport-adaptive encoding and streaming of 360-degree video for virtual reality applications. In: Multimedia (ISM), 2016 IEEE International Symposium on. IEEE; 2016. p. 583–6.
[9] Roh HJ, Han SW, Ryu ES. Prediction complexity-based HEVC parallel processing for asymmetric multicores. Multimedia Tools and Applications 2017;76(23):25271–84.
[10] "Overall Research Goal: Merciless Video Processing (MVP): Video decoding speed-up for mobile VR by using Tiled-SHVC as well as asymmetric mobile CPU multicores.". Available: http://mcsl.gachon.ac.kr/?page_id=1620.
[11] Boyce J, Ye Y, Chen J, Ramasubramoian AK. Overview of SHVC: scalable extensions of the high efficiency video coding standard. IEEE Trans Circuits Syst Video Technol Jul. 2015;26(1):20–34.
[12] Feldmann C, Bulla C, Cellarius B. Efficient stream-reassembling for video conferencing applications using tiles in HEVC. In: Proc. of International Conferences on Advances in Multimedia (MMEDIA); Jan. 2013. p. 130–5.
[13] Skupin R, Sanchez Y, Sühring K, Schierl T, Ryu E-S, Son J. Temporal MCTS Coding Constraints Implementation. In: 29th JCT-VC Meeting: Macao, of ISO/IEC JTC 1/SC 29/WG 11, JCTVC-AC0038; Oct. 2017. p. 19–25.
[14] HEVC Scalability Extension (SHVC) reference software SHM., Fraunhofer Institute for Telecommunications,Heinrich Hertz Institute, https://hevc.hhi.fraunhofer.de/shvc.
[15] High Efficiency Video Coding (HEVC) reference software HM., Fraunhofer Institute for Telecommunications,Heinrich Hertz Institute, https://hevc.hhi.fraunhofer.de/.

Jangwoo Son is a Masters student at the Department of Computer Engineering at Gachon University, Seongnam, Republic of Korea. His research focuses on the Tile-based real-time virtual reality (VR) video streaming system for a head-mounted display (HMD). He is a Student Member of the IEEE, IEEE Computer Society.

Eun-Seok Ryu is an Assistant Professor at the Department of Computer Engineering in Gachon University, Republic of Korea. He was a Principal Engineer at Samsung Electronics, Republic of Korea, and a Staff Engineer at InterDigital Labs, California, USA, where he researched on next generation video coding standards such as HEVC and SHVC. IEEE Senior Member.