IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 6, 2019, accepted September 17, 2019, date of publication September 20, 2019,
date of current version October 2, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2942771

Towards 3DoF+ 360 Video Streaming System for Immersive Media

JONG-BEOM JEONG[1], SOONBIN LEE[2], DONGMIN JANG[1], AND EUN-SEOK RYU[1], (Senior Member, IEEE)
[1]Department of Computer Education, Sungkyunkwan University, Seoul 03063, South Korea
[2]Department of Computer Engineering, Gachon University, Seongnam 13120, South Korea

Corresponding author: Eun-Seok Ryu (esryu@skku.edu)

This work was supported in part by the Institute of Information Communications Technology Planning Evaluation (IITP) supported by the Korea Government (MSIT) under Grant 2018-0-00765, the development of compression and transmission technologies for ultra high-quality immersive videos supporting 6DoF, and in part by the Korea Electric Power Corporation under Grant R17XA05-68.

ABSTRACT Immersive video streaming has become very popular. To increase the quality of experience (QoE) with immersive media, user movement adaptive video streaming, three degrees of freedom plus (3DoF+), has emerged and is expected to meet this growing demand. Satisfying the limit of the bandwidth, providing high-quality immersive experience is challenging because 3DoF+ system requires high resolution, multi-view video transmission. This paper proposes a stride based 3DoF+ 360 video streaming system and introduces two main ideas: (i) a multi-view video redundancy removal method using view synthesis, (ii) a multi-view video residual packing method. The proposed multi-view video compression method removes redundancy between videos and packs them into one video, and it exhibits a BD-rate saving of 36.0% in maximum compared to the results of the high-efficiency video coding reference model. In addition, the proposed system requires fewer number of decoders for the clients, and it decreases the burden for immersive video streaming.

INDEX TERMS Virtual reality, 3DoF+, view synthesis, region growing, dilation, pruning, packing.

I. INTRODUCTION

Currently, the virtual reality (VR) market is consistently growing. Therefore, the necessity for efficient immersive VR technology such as streaming has become more important because of the large amount of data to process in VR systems. For instance, the minimum resolution required for VR video that is played through a head-mounted display (HMD) is 4K. This means that the amount of data handled by the HMD increases rapidly, which is very challenging. For this reason, asymmetric core processing [1], [2], data offloading over mmWave [3], [4] for a mobile devices, and view location based asymmetric down-sampling method [5] have proposed. Consequently, the motion-constrained tile set (MCTS) [6] has been reported to process the viewport of the user. In addition, studies have described the MCTS implementation for VR streaming [7], [8].

Consequently, moving picture experts group (MPEG) has established the MPEG-I subgroup to solve the problems of immersive media. To standardize immersive media step by step, MPEG-I has defined three phases for VR : (1) 3DoF, (2) 3DoF+, and (3) 6DoF [9], [10]. In 3DoF, an user sitting in a chair can watch a 360 video without the support of user's movement; it provides limited immersive experience. In 3DoF+, the user is also supposed to sit in a chair. However, it supports the user's head movement, which increases user's QoE. In 6DoF, it provides 360 video considering both user's head and body movement. In 3DoF+ and 6DoF, synthesizing the virtual view in accordance with the user's movement is required; it is called view synthesis [11], [12]. In MPEG, view synthesis reference software (VSRS) [13], reference view synthesizer (RVS) [14], and versatile view synthesizer (VVS) [15] have proposed as the software for view synthesis. Among them, RVS has been adopted as the view synthesis reference software in 3DoF+. If several input views and the metadata for the virtual view are given to RVS, it synthesizes a virtual view texture and depth for each input view using 3D warping. This texture contains the color information of the video, and an additional depth map is required when synthesizing a virtual view. This depth map contains the distance

The associate editor coordinating the review of this manuscript and approving it for publication was Zhaoqing Pan.

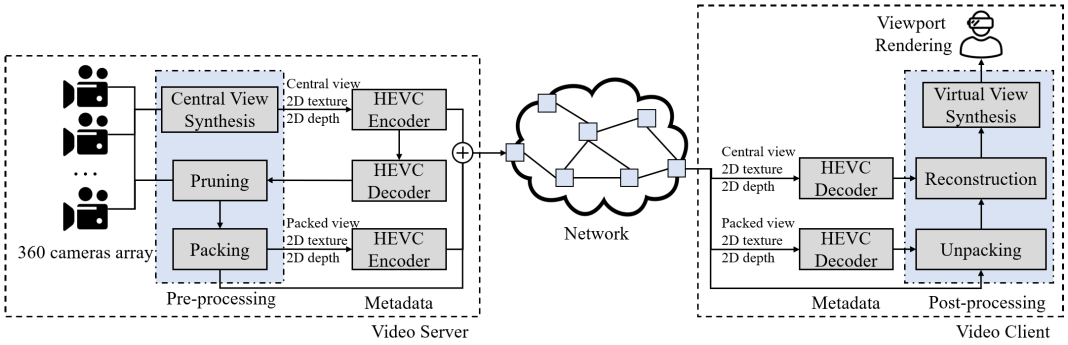


FIGURE 1. Conceptual diagram of the stride based 3DoF+ 360 video streaming system.

between the camera and objects shown on the texture. Then, it blends the synthesized virtual views from each input view to integrate the results.

3DoF+ system requires a large bitrate, which means large files must be transmitted every second, to the user because they contain multiple videos. However, it is challenging to meet the bandwidth required to the user with the high-efficiency video coding (HEVC) codec. Consequently, MPEG-I proposed a call for proposals (CfP) on 3DoF+ [16] to build a system on the existing HEVC codec with compressed texture and depth, and the metadata required will be standardized in MPEG-I part 7.

This paper proposes a stride based 3DoF+ 360 video streaming system for an immersive experience that satisfies the conditions of the 3DoF+ CfP. In the proposed method, an existing redundancy removal algorithm using RVS is adopted to remove the redundancy between source views. Source view means a video acquired from a camera; it is provided to the user. This process is defined as pruning. Because source views for 3DoF+ have similarity between each other, one source view contains pixels that are already included in another one. Consequently, the proposed system chooses a central view among the source views to remove the redundancy efficiently. Pixels from the source views are removed when they are already conveyed by the central view. The position of the central view is the center of the source views; that represents them the most in usual cases. If there is no candidate for the central view among the source views, the proposed system generates the central view. To prevent an artifact when reconstructing the redundancy, region growing [17] is used, and it increases the number of pixels sent, which improves the quality of the source views while requiring more bitrate. A decision making process for the pixel value for the removed pixel area is also considered to reduce the bitrate.

After pruning, the bitrate required to transmit the source views decrease. However, the number of decoders is still a burden for the 3DoF+ system because it has to simultaneously decode multiple source views. Moreover, although the pixels that are included in the central view are removed, the empty areas where the pixels are removed requires additional bitrate even though they are unnecessary. To solve these problems, the proposed system conducts packing. The function of packing is to remove the empty areas and integrate the remaining areas, which are informative, into a packed view. The number of pixels of a packed view does not exceed the maximum pixel rate of the HEVC encoder. The pixel rate is the number of pixels in a video to be encoded. If the number of pixels for a packed view exceeds the limit, the proposed system splits it into multiple packed views to satisfy this limit. Accordingly, metadata for packing is needed to reconstruct the source views, and this will be explained in Section III. As a result, a lower bitrate and number of decoders are needed. Fig. 1 shows the conceptual diagram of the stride based 3DoF+ 360 video streaming system.

II. RELATED WORK

This section introduces related works of 3DoF+ systems. To compress the multi-view video, there have been some approaches such as multi-view video coding (MVC) [18], multi-view HEVC (MV-HEVC) [19], and 3D-HEVC [20]. However, at the discussion on 3DoF+ system, the listed compression methods have not been adopted. Because they are extensions of existing standards, they may cause the fragmentation of 3DoF+ system. Instead, HEVC has been adopted as a video encoder for 3DoF+, and some approaches with pre-processing and post-processing for the videos has been allowed. Approaches to 3DoF+ systems have been recently published in MPEG, which are not public yet, because the CfP for the 3DoF+ system was published in January 2019. In March 2019, some institutes proposed their responses on 3DoF+ CfP. Philips introduced hierarchical pruning based system [21], and Technicolor & Intel proposed depth map refinement included system [22]. Nokia reported point cloud based system [23], and poznan university of technology (PUT) and the electronics & telecommunications research institute (ETRI) submitted multi-layer based system [24]. Among them, this section introduces point cloud based system in Section II-A and multi-layer based system in Section II-B.

A. POINT CLOUD BASED 3DOF+ STREAMING SYSTEM

In March 2019, Nokia proposed their response to the CfP for the 3DoF+ system; it is based on scene simplification

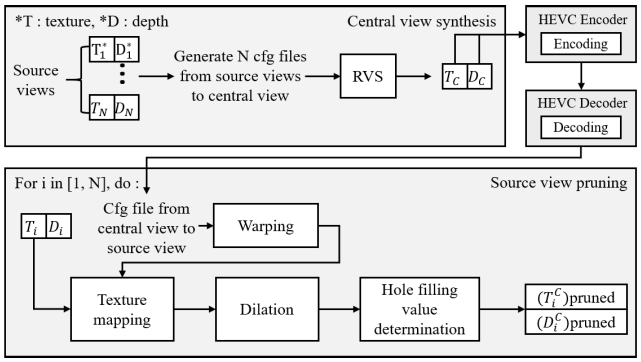


FIGURE 2. Block diagram of the multi-view location based pruning.

using point cloud. Once the input views are given to the view optimizer, it generates a point cloud for one intraperiod. Then, they are divided into shards, which represent smaller-sized images like tiles, using a greedy algorithm iteratively to reduce the pixel rate. After that, shards and the following metadata are received by a mosaic generator. A set of shards is defined as a mosaic. To obtain gain from compression, the mosaic that represents color is dilated. On the contrary, the depth mosaic is not dilated. Instead, to reduce some noise, a spatial median filter with a 3×3 size is applied. Because packing shards into mosaics is a NP-hard problem, the proposed method used heuristics to pack shards while minimizing the size of the mosaics. The packed views are encoded and decoded with HEVC. The following metadata is then serialized into a binary format, and compressed. The decoded views are given to a view synthesizer, and it generates a view using the metadata while casting a ray for each pixel and computing the tiles that intersect the ray. Using the geometry intersection points of the ray and colors related to the hit points, the output view is generated. Otherwise, a hole filling algorithm is applied.

B. MULTI-LAYER BASED 3DOF+ STREAMING SYSTEM

PUT and ETRI submitted their response to the CfP in March 2019, and their proposed system exploits a multi-layer structure as scalable HEVC [25]. Input textures are separated into a base layer and a residual layer in the spatial frequency domain. The base layer includes spatially low-pass filtered contents. Residual layer contains high-frequency residual contents. Both layers are encoded and transmitted to the view synthesis module. Because main idea of proposed method is to reduce the redundancy between source views, it contains an unified scene representation (USR); it generates an optimized set of views. If multiple overlapping rectangular views are given, USR gathers them and generates a base view and supplementary views. The base view contains objects that are not occluded by other objects. Supplementary views have residual objects that are not represented by the base view, and it is located at the center of the input views. As a result, the proposed system reduces the number of views to transmit.

III. STRIDE BASED 3DOF+ 360 VIDEO STREAMING SYSTEM

This section explains the stride based 3DoF+ 360 video streaming system. This study's contributions are as follows: (i) multi-view location based pruning method for inter-view redundancy removal by finding the redundancy between a central view and source views and removing the redundancy using view synthesis which will be explained in Section III-A, and (ii) a packing method that integrates the pruned views into one packed view using region growing and region allocation with stride to reduce both the bitrate and the number of decoders; it will be introduced in Section III-B.

A. MULTI-VIEW LOCATION BASED PRUNING

This section explains the multi-view location based pruning method. Fig. 2 shows a block diagram of pruning. The purpose of pruning is to remove the redundancy among the source views. Consequently, it selects a central view among the source views which is located in the center among the source views similar to a multi-layer based streaming system, and it represents most of the pixels from the source views. If there is no central view among the source views, a central view synthesis using RVS and the existing views is conducted.

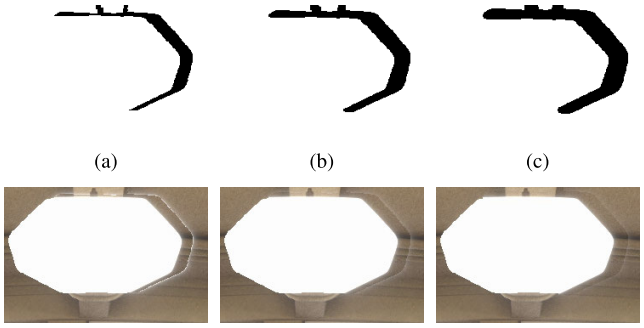
Pruning is composed of three parts. Firstly, image warping based on triangles method is conducted, and this is explained in Section III-A1. Second, region dilation of the warped image is processed; it is described in Section III-A2. Finally, hole filling value determination is processed; this is introduced in Section III-A3.

1) TRIANGLES METHOD BASED WARPING

This section explains the triangles method based warping process. It is based on RVS [14], which was adopted as a reference software for 3DoF+ in 2018. In view synthesis using RVS, the following processes are conducted. First, pixels of an input view are unprojected to the world coordinate system using camera metadata. Camera metadata includes its positions and rotations. Resolution, horizontal angular range and vertical angular range of the input view are also included. To get the world coordinates, ray direction are deduced using spherical coordinates. The world coordinate system used in RVS is the same as that of the MPEG-I omnidirectional media format(OMAF) [26]. Second, it applies an affine transformation [27], which moves the coordinates from the input view to the target view. A single affine transformation is applied. Third, the world coordinates are projected onto the virtual image, which means that the spherical 3D coordinates are projected onto the rectangular 2D coordinates. Fourth, pixels from the input views are warped to the resulting virtual view using the triangles method that was proposed by Universite Libre de Bruxelles [28]. Using the triangles method, the input view is split into triangles, and each pixel is the center of the vertex. When the triangles from the input view are warped to the virtual view, they are distorted if a triangle made of

(a) (b)

FIGURE 3. Image warping in pruning: (a) Central view and (b) warped central view towards a source view.



(a) (b) (c) (d) (e) (f)

FIGURE 4. Cropped first frame of a pruned view mask and reconstructed view texture in position v1 of *ClassroomVideo* with different dilation sizes (a) mask size = 0, (b) mask size = 2, (c) mask size = 4, (d) texture size = 0, (e) texture size = 2, and (f) texture size = 4.

pixels from the input view cannot represent the virtual view properly. If distortion of a warped triangle exceeds the threshold, the triangle is removed. For example, in Fig. 3, a yellow triangle on man's shoulder is warped without distortion. At the same time, red triangle on the left side of man's elbow shows distortion in warping. In other words, the red triangle can be only represented by the virtual view position, not the input view. In pruning, the central view is set to the input view, and source views are set to the virtual views. Distorted triangles are only represented by the source views. Consequently, the distorted triangles are filled with the pixels of the input view because they cannot be represented with the central view. In contrast, areas that do not belong to the distorted triangles are filled with neutral gray because these areas are able to be represented by the central view. After these processes, pruning generates a pruned view as a result. In reconstruction, the proposed system receives the pixels from the central view to the pruned view to reconstruct the holes of the pruned view.

2) BINARY IMAGE DILATION

After pruning, the holes and informative areas are computed. However, when reconstructing the pruned image, there are some artifacts at the edges of the objects. Because of the warping issue, pixels at the edge are stretched, which causes decrease in the reconstructed image quality. Consequently, this section introduces binary image dilation to avoid this quality decrease. Binary image dilation is based on mathematical morphology [29], and it can be used when expanding the informative area, and vice versa. Pruning generates mask and it consists of 8-bit pixels with a black or white color binary value, as shown in Fig. 4. Black represents informative areas that can be represented by a source view and white represents holes that are conveyed by a central view. During dilation, the black pixel areas are expanded to include more pixels of a source view. As a result, distortion on the edge of the objects decreases. As shown in Fig. 4, when the dilation size increases, the distortion on the edges reduce in the reconstructed image. Nevertheless, there are tradeoffs between quality and bitrate in dilation because increasing the dilation size increases the quality of the reconstructed image and bitrate. Therefore, finding the appropriate dilation size depending on the bandwidth is important. in this study, 2 dilation sizes, 0 and 4 were used.

TABLE 1. Bitrates of pruned view position v1 in *ClassroomVideo* with different hole filling value determination.

Sequence name	Pos. of view	Method	QP	Bitrate (Kbps)
ClassroomVideo	v1	**Average value**	22	**3746**
			37	**212**
		Neutral gray	22	3840
			37	235
		Nearest	22	11088
			37	688
		Interpolation	22	21122
			37	750

3) HOLE FILLING VALUE DETERMINATION

In RVS, after removing the invalid areas with the triangles method and processing dilation, the invalid areas, i.e., holes, are filled with neutral gray. However, determination of the appropriate hole filling value after dilation leads to a bitrate gain. This study introduces four hole filling value determination method. First, the holes are filled with neutral gray, as previously conducted in RVS. Second, the hole filling value is determined by the average pixel values of the valid areas that do not belong to the hole. Third, hole filling by referencing the nearest pixels is introduced. Finally, hole filling with interpolation is conducted.

As a test sequence, *ClassroomVideo* was used which has been proposed as a 3DoF+ test sequence by Philips [30]. It consists of 15 views with 4096×2048 resolution, and they were captured from different camera positions simultaneously. In this experiment, the central view was set to position v0, and position v1 was used as a target view. Among the introduced methods, hole filling with the average pixel value shows the best result, as shown in Table 1. Accordingly, the proposed system adopted hole filling with the average pixel value.

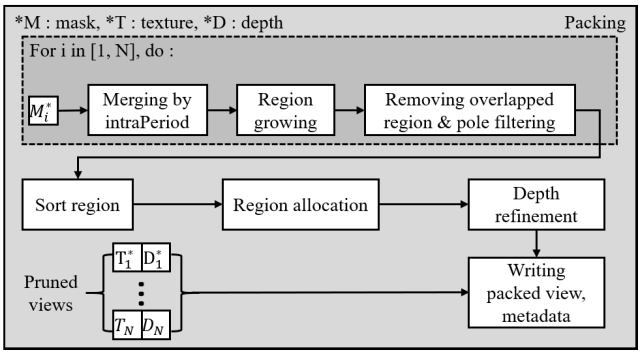


FIGURE 5. Block diagram of the stride based intraperiod level multi-view packing method.

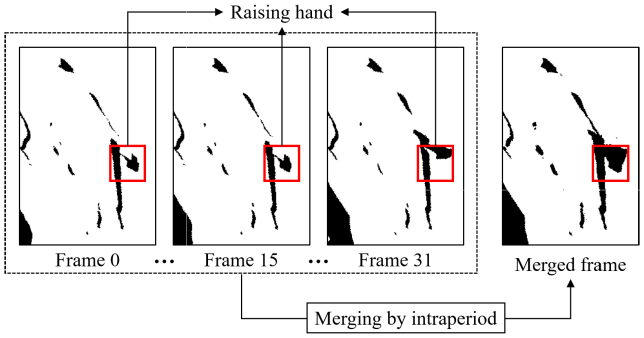


FIGURE 6. Merging by intraperiod in a pruned view mask of *TechnicolorMuseum*.

B. STRIDE BASED INTRAPERIOD LEVEL MULTI-VIEW PACKING

This section introduces the stride based intraperiod level multi-view packing method for 3DoF+ 360 videos. Fig. 5 shows a block diagram of the packing process. After pruning, there are lots of holes that are useless when reconstructing the pruned view. Furthermore, even though pruning removes useless pixels, the number of required decoders is unchanged; it becomes a burden for streaming 3DoF+ video. Consequently, this paper proposes packing; it extracts the informative area and merges them together. As a result, the bitrate and the number of required decoders decrease.

Packing consists of four parts. Firstly, mask merging by intraperiod is processed, which is described in Section III-B1. Second, adaptive block based region growing is conducted; this is explained in Section III-B2. Third, stride based region allocation is processed, and this is introduced in Section III-B3. Lastly, digital filter based depth refinement is operated, which is explained in Section III-B4.

1) MASK MERGING BY INTRAPERIOD

This section introduces mask merging by intraperiod to compute the informative regions. The proposed system uses a mask to gather pixels that are not conveyed by a central view and merge them into a packed view. For 3DoF+ test sequences, the intraperiod was set to 32 frames, which is defined in the common test conditions (CTC) [9] for 3DoF+. In the HEVC encoder, inter prediction [31] is used to compress a video and remove the redundancy between frames. This is possible because the positions of objects in a video changes only slightly frame-by-frame in the intraperiod range. Packing merges the regions that include the informative areas, and it packs them in a packed view while generating the following metadata. However, with the proposed method which is explained in Section III-B3, the position of a region in a packed view can be changed frame-by-frame because the movements of objects in pruned views change the size of the region frame-by-frame. Therefore, the correlation between the frames can be lost; it requires lots of bitrate because of inefficient compression.

To preserve the similarity between frames, the proposed system merges the mask of a pruned view in the range of the intraperiod, as shown in Fig. 6. In this figure, a hand is raising. Without merging, the position of the hand will be changed frame-by-frame. By merging the frames, the position of the hand in the packed view can be fixed, and the correlation between frames is preserved. Furthermore, metadata that consists of pruning and packing information should be transmitted. If the metadata contains information about a region by the intraperiod, the required bitrate for the metadata decreases. Therefore, it will satisfy the bitrate for the bandwidth.

2) ADAPTIVE BLOCK BASED REGION GROWING

After merging masks, informative areas of pruned views are computed. However, if pixels are extracted from the pruned views and packed with pixel-by-pixel, the correlation between frames can be lost. To avoid this problem, the proposed system divides the pruned views into regions using adaptive block based region growing [32], which is a part of image segmentation, and computes regions.

If the region growing is conducted with pixel level, a lot of small regions are obtained from the masks, and this increases the bitrate for the metadata. Consequently, in the proposed system, each mask is divided into blocks. The size of a block was set to 16×16 in this study; it can be adjusted. Then, each block is checked as to whether it contains black pixels, in other words, informative pixels. If the number of informative pixels of a block exceeds the threshold, the block is added to the set of seeds, and it is marked as informative. Otherwise, it is marked as uninformative. The threshold was set to 0 in this study, which means that a block is excluded from a packed view if it does not have informative pixels. After checking all the blocks, region growing is conducted. All the seeds have eight neighbors surrounding each seed. Seeds are expanded into neighbors recursively if they are informative, and they are marked as assigned. If a block is assigned, it cannot be assigned again. Region growing for one seed ends when there are no neighbors that are informative or the size of the region exceeds a maximum width or height. The maximum size of a region should be declared in the region growing; that was set to $256x256$. Otherwise,

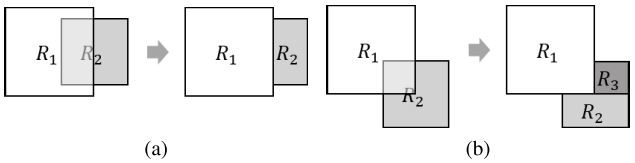


FIGURE 7. Removing redundancy between regions: (a) Removing a side overlapped region and (b) dividing a vertex overlapped region.

the number of pixels from a region may exceed the maximum pixel rate of the HEVC encoder.

Region growing is also used in pole filtering. In pruning, there are occlusions on the poles in the pruned view mask. In test sequence *ClassroomVideo* and *TechnicolorMuseum*, the top and bottom areas are not hidden by any objects. This means that these areas are already included in the central view. Consequently, those areas do not need to be included in the pruned and packed views. However, the pruned view contains poles because of the warping issue. The proposed system uses the triangles method to warp an image. However, because the test sequence uses the ERP format, the pixels of the top and bottom areas are overlapped and distorted when projecting from a spherical to a rectangular view. Therefore, warping is not operational in the pole areas, and these areas have to be excluded by pole filtering. Once the proposed system divides a pruned view into blocks, the blocks at the top and bottom areas are set to seeds, and region growing is conducted. Region growing for pole filtering ends when the height of a region exceeds the maximum pole height or there is no neighbor from the seeds. The maximum height of the poles should be defined before pole filtering; that was set to 256 in the experiments of this study.

After computing all of the regions, removing redundancy between the regions is conducted. Even though the duplicated assignation of a block is forbidden, there are some overlapping areas between regions because each region has a rectangular shape, as shown in Fig. 7. The process consists of three cases, and there is a big region R_1 and small region R_2. First, if R_1 contains the whole of R_2, R_2 is removed. Second, as represented in Fig. 7a, if R_2 has a side that is totally overlapped by R_1, the overlapping region is removed from R_2. Lastly, if two regions have overlapping vertices as shown in Fig. 7b, R_2 is divided horizontally, and the area that is totally included in R_1 is removed. After that, the remaining area on the right of R_1 is assigned as R_3.

3) STRIDE BASED REGION ALLOCATION

This section explains the stride based region allocation method; it takes the regions computed from the region growing and packs them into a packed view. The main idea of this process is to use a stride which is introduced in convolutional neural networks [33]. The proposed system sorts the regions in the descending order by height and confirms whether they can be included in a specified position of a packed view. If the black pixels of a region are not overlapping those of a packed view, the region is included in a packed view. Otherwise,

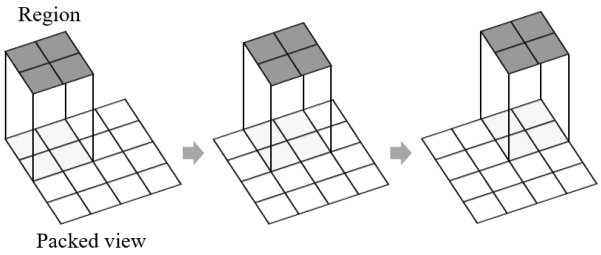


FIGURE 8. Region allocation on a packed view with stride.

TABLE 2. Packed view size with different region allocation method.

Sequence name	Frames	QP	Packed view size	
			No stride	With stride
ClassroomVideo	1-32	22	7680×4672	**7680×2560**
TechnicolorMuseum	1-32	22	7680×3904	**7680×1664**

TABLE 3. Bitrates of packed view depth with depth refinement.

Sequence name	QP	Bitrate (Mbps)	
		No refinement	With refinement
ClassroomVideo	20	11.5	**7.2**
	25	8.5	**5.0**
	28	6.3	**3.5**
	32	4.2	**2.1**
TechnicolorMuseum	14	13.3	**9.8**
	20	10.7	**7.6**
	26	8.0	**5.6**
	29	6.0	**4.2**

TABLE 4. Anchor views per class.

Sequence name	Class	Resolution	No.of source views	No.of anchor views
ClassroomVideo	A1	4096×2048	15	15
	A2			9
TechnicolorMuseum	B1	2048×2048	24	24
	B2			8

the proposed system searches all of the positions in a packed view with stride. In other words, it checks positions with a certain interval, as shown in Fig. 8. Region allocation with stride decreases a packed view height to meet the pixel rate for the encoder, as shown in table 2. For *ClassroomVideo*, it decreases 45.2% of packed view height, and 57.4% for *TechnicolorMuseum*. In this study, stride width and height are set to 16.

4) DIGITAL FILTER BASED DEPTH REFINEMENT

This section introduces the digital filter based depth refinement. 3DoF+ CfP recommends to use the same configuration for texture and depth. However, depth contains sharp edges and large blocks with similar pixel values [34]; it requires a large amount of bitrate. To solve this problem, this paper applies a digital filter to the depth map of a packed view. 3×3 spatial median filter is used to smoothen the edge of the depth map. The result of depth refinement is shown in table 3, with dilation size 0. The proposed method shows bitrate saving approximately 43.2% for *ClassroomVideo* and 28.8% for *TechnicolorMuseum*.

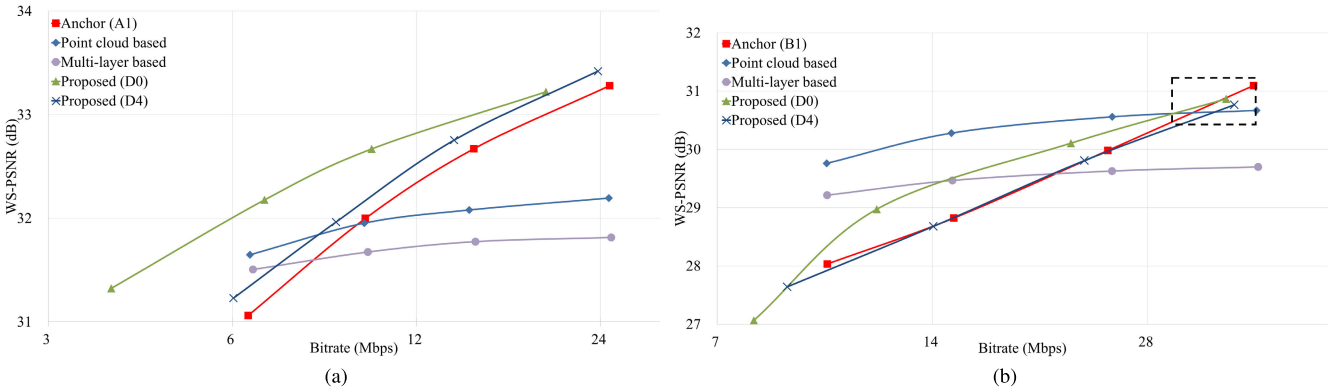


FIGURE 9. RD-curves of source view synthesis: (a) *ClassroomVideo*. (b) *TechnicolorMuseum*.

TABLE 5. Target bitrate points not to be exceeded in the 3DoF+ CfP.

Sequence name	Target bitrates [Mbit/s]					
	R1	R2	R3	R4	R5	R6
ClassroomVideo	6.5	10	15	25	40	65
TechnicolorMuseum	10	15	25	40	65	100

TABLE 6. Frame ranges for the experiment.

Sequence name	Encoding frames	Evaluation frames
ClassroomVideo	1-120	1-32, 89-120
TechnicolorMuseum	1-300	1-32, 269-300

IV. IMPLEMENTATION AND EXPERIMENTAL RESULTS

This section introduces the experimental results of the stride based 3DoF+ video streaming system. Two servers were used for this experiment, one had two Intel Xeon E5-2687w v4 CPUs and 128 GB of memory, and another has 2 Intel Xeon E5-2683 v4 CPUs and same memory capacity with the first one. The used reference tools for the experiment are as follows: RVS version 3.1 was used as a view synthesizer, with OpenCV 3.4.2 [35]. For encoding, HEVC test model (HM) version 16.16 [36] was adopted. If the input video is ERP, HM with 360lib version 5.1-dev [37] was used. For objective quality evaluation, weighted-to-spherically-uniform peak signal-to-noise ratio (WS-PSNR) version 2.0 [38], [39] was used to evaluate the quality of the results. For the test materials, MPEG-I provided 5 test sequences for 3DoF+. However, there were 2 perspective 2D test sequences and 1 test material which captured only 1 direction and it was not omnidirectional. Because the aforementioned 3 sequences are not appropriate for 3DoF+ system, these are not introduced in this paper. Therefore, 2 test sequences, *ClassroomVideo* [30] and *TechnicolorMuseum* [40] were used for the experiment. In the CTC for 3DoF+, the views selected to transmit are called anchor views, and they are defined in Table 4. In this experiment, A1 and B1 were used as anchors; they send all of the source views. The CfP on 3DoF+ recommends to meet the target bitrate, as shown in Table 5. Consequently, the parameters used for the experiment is shown in Table 7. In the experiment, DeltaQP was used to use the different quantization parameter (QP) value for texture and depth.

TABLE 7. Parameters used for the experiment.

Module	Parameter	Value
Pruning	PoleMaskingSize	256 pixels
	DilationSize	0, 4 pixels
Packing	BlockSize	16×16 pixels
	StrideSize	16×16 pixels
	IntraPeriod	32 frames
Encoding	Depth DeltaQP	10

If DeltaQP is set to -10 and 22 for texture QP is used, QP for depth is 12. DeltaQP -10 was used for the experiment because this showed the best result for 3DoF+ anchor which was reported in MPEG-I [41]. In this experiment, dilation sizes 0 and 4 were used. For encoding, all of the frames were used. To encode the views with ERP format including a central view, HM with 360lib was used. If the ERP view is omnidirectional, it was encoded with padded ERP format [42]; it adds padding at both left and right sides. On the contrary, HM without 360lib was used to encode a packed view because its format is not ERP. A subset of the frames was used for objective quality evaluation, as described in the CTC for 3DoF+ and Table 6. In pruning, a central view position must be defined. If there is no candidate, the proposed system synthesizes the central view using RVS. For *TechnicolorMuseum*, the central view at the center of the source views was synthesized. For *ClassroomVideo*, the position v0 was used as a central view that is located at the center.

Fig. 9 shows the rate-distortion curve of the anchors and mentioned methods. In Table 8, the results of the experiments compared to the anchor are shown. For all methods, rates 1 to 4 were tested. The proposed method with dilation size 0 showed a 36.0% BD-rate saving, and 78.75% pixel rate saving. For *TechnicolorMuseum*, the proposed method showed a 18.9% BD-rate saving, and 86.72% pixel rate saving. When the bandwidth was over 15 Mbps, the proposed method showed a better result than the multi-layer based method. As shown in dotted line box in Fig. 9b, the proposed method performed better than the point cloud based method for the high-bitrate. For both *ClassroomVideo* and *TechnicolorMuseum*, the proposed method outperformed the anchors, and showed better results than the related works for

TABLE 8. Results of the experiments compared to the anchor.

Sequence	Test class	Method	BD-rate	No.of decoders required	Encoding time saving	Pixel rate saving
ClassroomVideo	A1	Anchor	0.00%	30	0.00%	0.00%
		Proposed-D0	-36.00%	4	-81.68%	-78.75%
		Proposed-D4	-11.10%	4	-80.48%	-76.88%
		[23]	6.90%	2	-88.66%	-91.61%
		[24]	21.20%	4	-81.96%	-86.61%
TechnicolorMuseum	B1	Anchor	0.00%	48	0.00%	0.00%
		Proposed-D0	-18.90%	4	-85.95%	-86.72%
		Proposed-D4	1.30%	4	-83.45%	-85.16%
		[23]	-44.50%	2	-84.64%	-87.50%
		[24]	-20.00%	8	-56.98%	-66.67%
Average	X1	Anchor	0.00%	39	0.00%	0.00%
		Proposed-D0	**-27.45%**	4	83.81%	82.73%
		Proposed-D4	-4.90%	4	81.96%	81.02%
		[23]	-18.80%	**2**	**86.65%**	**89.55%**
		[24]	0.60%	6	69.47%	76.64%

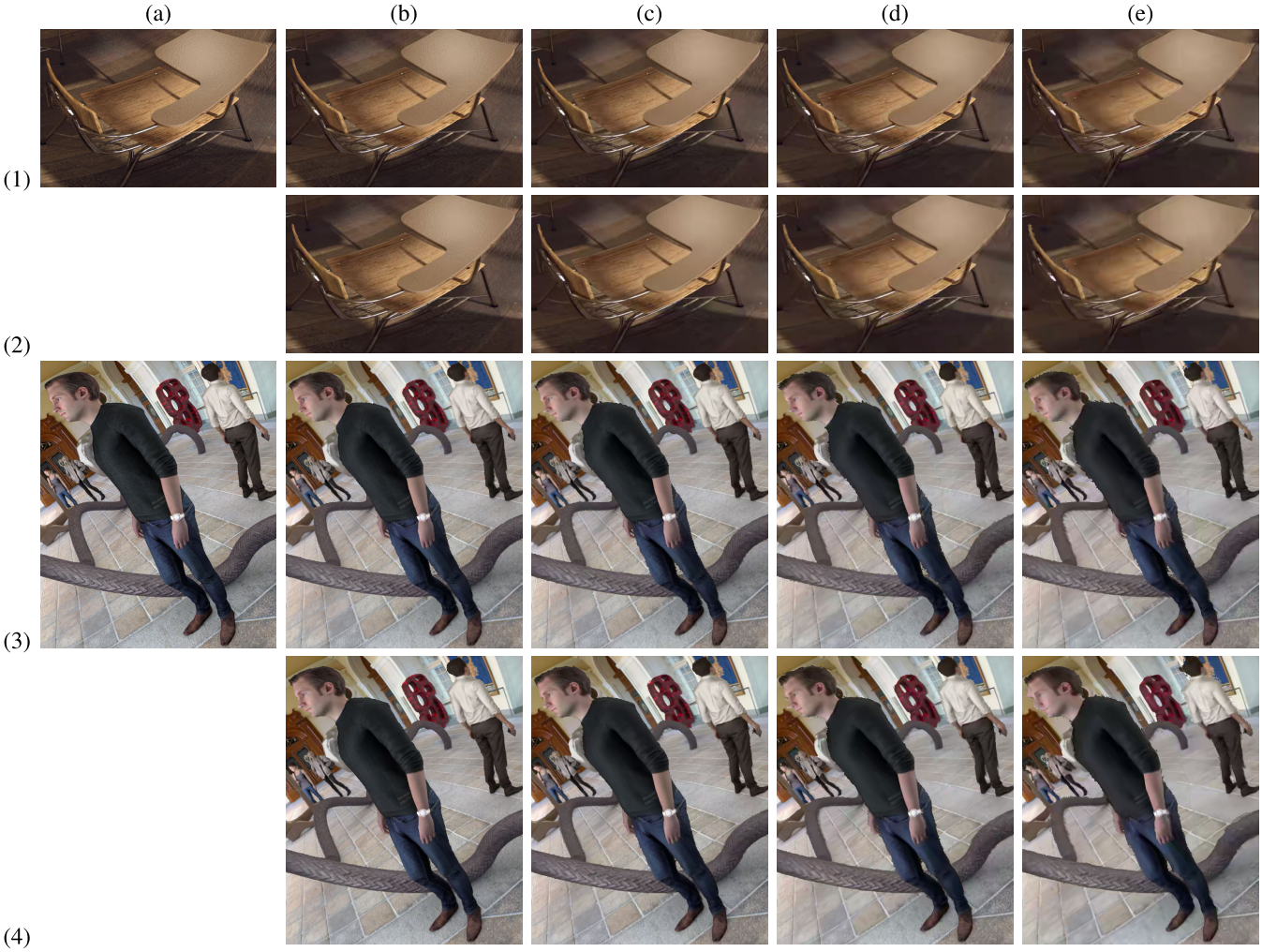


FIGURE 10. Cropped result of source view synthesis: (a) uncompressed source view, (b) synthesized source view of Rate 4, (c) synthesized source view of Rate 3, (d) synthesized source view of Rate 2, and (e) synthesized source view of Rate 1, (1) position v1 of class A1 with DilationSize=0, (2) position v1 of class A1 with DilationSize=4, and (3) position v0 of class B1 with DilationSize=0, (4) position v0 of class B1 with DilationSize=4.

the high-bitrate, which provides the high-quality video. For the low bitrate, the proposed method with small dilation size was promising. In contrast, for the high bitrate, the proposed method with large dilation size was efficient. In average, the proposed method with dilation size 0 showed a 27.45% BD-rate saving, which requires the smallest bandwidth. For the client devices, the proposed method required only 4 decoders while the anchor needed 30 decoders for *ClassroomVideo* and 48 decoders for *TechnicolorMuseum*. Because the proposed system decreased the number of decoders, it can

be easily applied to mobile devices. Although the point cloud based method shows better results in number of decoders required, encoding time, and pixel rate, the proposed method still shows the promising results. At the client side, encoding time and pixel rate are less important than the BD-rate because encoding is conducted at the server side, not the client side. Also, for the proposed method, the required number of decoders can be reduced by concatenating the central view and the packed view, this is possible because the pixel rate of these views generated by the proposed method does not exceed the limit of the HEVC encoder. Therefore, the proposed method, which requires the smallest bandwidth, can be used for 3DoF+ 360 video streaming system.

Fig. 10 shows the cropped result of the experiment. As shown in Fig. 1, the proposed system synthesized the virtual views at the source view positions using reconstructed source views, and compared them with the uncompressed source views. Column (a) shows the uncompressed source views, and (b), (c), (d), (e) correspond to the synthesized views with the proposed method at rate 1, rate 2, rate 3, and rate 4. Row (1), (2) shows the result of *ClassroomVideo*, and row (3), (4) shows *TechnicolorMuseum*. Noise at the object's edge and distortions of the objects in views increased when the target bitrate decreased from rate 4 to rate 1. In low bitrate, result of large dilation size showed less artifacts in the edge of the objects, especially in *TechnicolorMuseum*. The point cloud based method, multi-layer based method, and the proposed method were evaluated by comparing them with the anchor, which was encoded by the HEVC encoder. Any other methods were not applied to the anchor except HEVC.

V. CONCLUSION

This paper proposes a stride based 3DoF+ video streaming system with two main concepts: (i) a multi-view location based pruning method to remove the redundancy among the multi-view video, (ii) an stride based intraperiod level multi-view packing method to acquire a bitrate gain and reduce the number of decoders. In the experiment with HM, 360Lib, RVS, and WS-PSNR, the proposed method shows a BD-rate saving up to 36.0%. In addition, the proposed method requires fewer number of decoders for the clients. Intensive experiments need to be conducted with diverse parameters to find a compromise point between traditional video streaming system and the proposed method. Consequently, color refinement for central view synthesis also needs to be developed to achieve better QoE.

REFERENCES

[1] H.-J. Roh, S. W. Han, and E.-S. Ryu, "Prediction complexity-based HEVC parallel processing for asymmetric multicores," *Multimedia Tools Appl.*, vol. 76, no. 23, pp. 25271–25284, Dec. 2017.

[2] S. Yoo and E.-S. Ryu, "Parallel HEVC decoding with asymmetric mobile multicores," *Multimedia Tools Appl.*, vol. 76, no. 16, pp. 17337–17352, 2017.

[3] D. Van Nguyen, T. T. Le, S. Lee, and E.-S. Ryu, "SHVC tile-based 360-degree video streaming for mobile VR: PC offloading over mmWave," *Sensors*, vol. 18, no. 11, p. 3728, 2018.

[4] T. T. Le, D. Van Nguyen, and E.-S. Ryu, "Computing offloading over mmWave for mobile VR: Make 360 video streaming alive," *IEEE Access*, vol. 6, pp. 66576–66589, 2018.

[5] J. Jeong, D. Jang, J. Son, and E.-S. Ryu, "3DoF+ 360 video location-based asymmetric down-sampling for view synthesis to immersive VR video streaming," *Sensors*, vol. 18, no. 9, p. 3148, 2018.

[6] R. Skupin, Y. Sanchez, K. Sühring, T. Schierl, E.-S. Ryu, and J. Son, *Temporal MCTS Coding Constraints Implementation*, document MPEG 122/m42423, 122th MPEG Meeting of ISO/IEC JTC1/SC29/WG11, 2018.

[7] J. Son, D. Jang, and E.-S. Ryu, "Implementing 360 video tiled streaming system," in *Proc. 9th ACM Multimedia Syst. Conf.*, 2018, pp. 521–524.

[8] J. Son and E.-S. Ryu, "Tile-based 360-degree video streaming for mobile virtual reality in cyber physical system," *Comput. Elect. Eng.*, vol. 72, pp. 361–368, Nov. 2018.

[9] J. Jung, B. Kroon, R. Dore, G. Lafruit, and J. Boyce, *Common Test Conditions on 3DoF+ and Windowed 6DoF*, document MPEG2018/n18089, 124th MPEG Meeting of ISO/IEC JTC1/SC29/WG11, 2018.

[10] X. Wang, L. Chen, S. Zhao, and S. Lei, *From OMAF for 3DoF VR to MPEG-I Media Format for 3DoF+, Windowed 6DoF and 6DoF VR*, document MPEG 119/m44197, 119th MPEG Meeting of ISO/IEC JTC1/SC29/WG11, 2017.

[11] K. Müller, A. Smolic, K. Dix, P. Merkle, P. Kauff, and T. Wiegand, "View synthesis for advanced 3D video systems," *EURASIP J. Image Video Process.*, vol. 2008, no. 1, 2009, Art. no. 438148.

[12] A. Smolic, K. Mueller, P. Merkle, C. Fehn, P. Kauff, P. Eisert, and T. Wiegand, "3D video and free viewpoint video—Technologies, applications and MPEG standards," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2006, pp. 2161–2164.

[13] T. Senoh, N. Tetsutani, and H. Yasuda, *MPEG-I-Visual: View Synthesis Reference Software (VSRSX)*, document MPEG 123/m42911, 123rd MPEG Meeting of ISO/IEC JTC1/SC29/WG11, 2018.

[14] *Reference View Synthesizer (RVS) Manual*, document MPEG 124/n18068, 124th MPEG Meeting of ISO/IEC JTC1/SC29/WG11, 2018.

[15] P. Boissonade, J. Jung, and P. Nikitin, *[MPEG-I Visual] View Synthesis Algorithm for Windowed-6DoF*, document MPEG2018/m44788, 124th MPEG Meeting of ISO/IEC JTC1/SC29/WG11, 2018.

[16] *Call for Proposals on 3DoF+ Visual*, document MPEG/n18145, 125th MPEG Meeting of ISO/IEC JTC1/SC29/WG11, 2019.

[17] R. Adams and L. Bischof, "Seeded region growing," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 16, no. 6, pp. 641–647, Jun. 1994.

[18] P. Merkle, A. Smolic, K. Muller, and T. Wiegand, "Multi-view video plus depth representation and coding," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, vol. 1, Sep./Oct. 2007, pp. I-201–I-204.

[19] M. M. Hannuksela, Y. Yan, X. Huang, and H. Li, "Overview of the multiview high efficiency video coding (MV-HEVC) standard," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2015, pp. 2154–2158.

[20] K. Müller, H. Schwarz, D. Marpe, C. Bartnik, S. Bosse, H. Brust, T. Hinz, H. Lakshman, P. Merkle, F. H. Rhee, "3D high-efficiency video coding for multi-view video and depth data," *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3366–3378, Sep. 2013.

[21] B. Kroon and B. Sonneveldt, *Philips Response to 3DoF+ Visual CfP*, document MPEG2019/m47179, 126th MPEG Meeting of ISO/IEC JTC1/SC29/WG11, 2019.

[22] J. Fleureau, F. Thudor, R. Dore, B. Salahieh, M. Dmytrychenko, and J. Boyce, *Technicolor-Intel Response to 3DoF+ CfP*, document MPEG2019/m47445, 126th MPEG Meeting of ISO/IEC JTC1/SC29/WG11, 2019.

[23] V. K. M. Vadakital *et al.*, *Description of Nokia's Response to CFP for 3DoF+ Visual*, document MPEG2019/m47372, 126th MPEG Meeting of ISO/IEC JTC1/SC29/WG11, 2019.

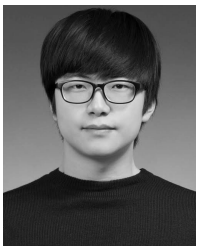
[24] M. Domanski *et al.*, *Technical Description of Proposal for Call for Proposals on 3DoF+ Visual Prepared by Poznan University of Technology (Put) and Electronics and Telecommunications Research Institute (ETRI)*, document MPEG2019/m47407, 126th MPEG Meeting of ISO/IEC JTC1/SC29/WG11, 2019.

[25] J. M. Boyce, Y. Yan, J. Chen, and A. K. Ramasubramonian, "Overview of SHVC: Scalable extensions of the High Efficiency Video Coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 1, pp. 20–34, Jan. 2016.

[26] (2018). *MPEG-I OMAF*. [Online]. Available: https://mpeg.chiariglione.org/standards/mpeg-i/omnidirectional-media-format

[27] C. A. Glasbey and K. V. Mardia, "A review of image-warping methods," *J. Appl. Statist.*, vol. 25, no. 2, pp. 155–171, 1998.

[28] S. Fachada, D. Bonatto, A. Schenkel, and G. Lafruit, "Depth image based view synthesis with multiple reference views for virtual reality," in *Proc. 2018-3DTV-Conf. True Vis.-Capture, Transmiss. Display 3D Video (3DTV-CON)*, Jun. 2018, pp. 1–4.
[29] R. M. Haralick, S. R. Sternberg, and X. Zhuang, "Image analysis using mathematical morphology," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-9, no. 4, pp. 532–550, Jul. 1987.
[30] B. Kroon, *3DoF+ Test Sequence Classroomvideo*, document MPEG2018/m42415, 122nd MPEG Meeting of ISO/IEC JTC1/SC29/WG11, 2018.
[31] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.
[32] M. M. S. J. Preetha, L. P. Suresh, and M. J. Bosco, "Image segmentation using seeded region growing," in *Proc. Int. Conf. Comput., Electron. Electr. Technol. (ICCEET)*, Mar. 2012, pp. 576–583.
[33] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.
[34] K. Müller, P. Merkle, and T. Wiegand, "3-D video representation using depth maps," *Proc. IEEE*, vol. 99, no. 4, pp. 643–656, Apr. 2011.
[35] (2018). *Open Computer Vision*. [Online]. Available: https://github.com/opencv/opencv/releases/tag/3.4.2
[36] (2018). *HM Reference Software 16.16*. [Online]. Available: https://hevc.hhi.fraunhofer.de/svn/svn_HEVCSoftware/tags/HM-16.16
[37] (2018). *360lib 5.1 Software Package*. [Online]. Available: https://jvet.hhi.fraunhofer.de/svn/svn_360Lib/tags/360Lib-5.1
[38] Y. Sun, A. Lu, and L. Yu, "Weighted-to-spherically-uniform quality evaluation for omnidirectional video," *IEEE Signal Process. Lett.*, vol. 24, no. 9, pp. 1408–1412, Sep. 2017.
[39] (2018). *ERP Weighted-to-Spherically-Uniform Peak Signal-to-Noise Ratio (WS-PSNR)*. [Online]. Available: http://mpegx.int-evry.fr/software/MPEG/Explorations/3DoFplus/WS-PSNR
[40] R. Dore, *Technicolor 3DoF+ Test Materials*, document MPEG2018/m42349, 122nd MPEG Meeting of ISO/IEC JTC1/SC29/WG11, 2018.
[41] B. Wang, L. Yu, B. Kroon, and J. Jung, *[MPEG-i Visual] Results on Depth QPS in CTC of 3DoF+ Video*, document MPEG 124/m44688, 124th MPEG Meeting of ISO/IEC JTC1/SC29/WG11, 2018.
[42] J. Boyce and Z. Deng, *EE4: Padded ERP (PERP) Projection Format*, document JVET-G0098, Joint Video Exploration Team of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, 2017.



JONG-BEOM JEONG received the B.S. degree from Gachon University, in August 2018. He is currently pursuing the master's degree with the Department of Computer Education, Sungkyunkwan University (SKKU). His current research topics include video compression standards, view synthesis, and deep learning-based video coding.

SOONBIN LEE is currently pursuing the master's degree with the Department of Computer Engineering, Gachon University. His current research topics include video compression standards, view synthesis, and deep learning-based video coding.

DONGMIN JANG received the B.S. degree from Gachon University, in February 2019. He is currently pursuing the master's degree with the Department of Computer Education, Sungkyunkwan University (SKKU). His current research topics include video compression standards, view synthesis, and deep learning-based video coding.

EUN-SEOK RYU received the B.S., M.S., and Ph.D. degrees in computer science from Korea University, in 1999, 2001, and 2008, respectively. He was also a Principal Engineer with Samsung Electronics, Suwon, South Korea, where he led a multimedia team. He was a Staff Engineer with InterDigital Labs, San Diego, CA, USA, from January 2011 to February 2014, where he researched and contributed to next generation video coding standards, such as HEVC and SHVC. From September 2008 to December 2010, he was a Postdoctoral Research Fellow of the School of Electrical and Computer Engineering, Georgia Centers for Advanced Telecommunications Technology (GCATT), Georgia Institute of Technology, Atlanta, GA, USA. In 2008, he was a Research Professor with the Research Institute for Information and Communication Technology, Korea University, Seoul, South Korea. He is currently an Assistant Professor with the Department of Computer Education, Sungkyunkwan University (SKKU), Seoul, South Korea. Prior to joining SKKU in 2019, he was an Assistant Professor with the Department of Computer Engineering, Gachon University, Seongnam, South Korea, from March 2015 to August 2019. His research interests include areas of multimedia communications, including video source coding and wireless mobile systems.