Bitrate Efficient 3DoF+ 360 Video View Synthesis for Immersive VR Video Streaming

JongBeom Jeong
Gachon University
(Dept. of Computer Engineering)
Seongnam, South Korea
uof9494@gc.gachon.ac.kr

Dongmin Jang
Gachon University
(Dept. of Computer Engineering)
Seongnam, South Korea
dogzz9445@gc.gachon.ac.kr

Jang-Woo Son
Gachon University
(Dept. of Computer Engineering)
Seongnam, South Korea
sjw6757@gc.gachon.ac.kr

Abstract—Recently, as the demand for Virtual Reality (VR) is increasing, it is not hard to experience immersive contents with VR. However, a tremendous amount of calculation and bandwidth are required when processing 360 videos. Moreover, to enjoy stereoscopic 360 contents, additional information such as the depth of the video must be provided. Therefore, in this paper, it proposes the efficient method of streaming high quality 360 videos. To reduce the bandwidth when streaming and synthesizing the 3DoF+ 360 video which supports limited movements of the user, the proper down-sampling ratio and quantization parameter are offered with the graph between bitrate and Peak Signal-to-Noise Ratio (PSNR). To encode and decode the 360 video, High-Efficiency Video Coding (HEVC) is used. And view synthesizer produces the video of intermediate view to provide the user with an immersive experience.

Keywords—VR, 3DoF+, HEVC, view synthesis, VSRS, RVS, FTV, WS-PSNR, Multi-view video coding

I. INTRODUCTION

As the VR market is getting bigger, efficient immersive VR technology is needed. To play high-quality VR video through Head-Mounted Display (HMD), the resolution of the video must be 4K, at least. In this case, the amount of data to be processed from HMD increases rapidly. That’s why the Moving Picture Experts Group (MPEG) proposed the technology which processes viewport the user views named Motion-Constrained Tile Set (MCTS) [1] in 2016, and one paper contains MCTS implementation for VR streaming was submitted [2]. Also, to provide the user with high quality 360 videos, region-wise packing [3] was proposed. It encodes Region of Interest (ROI) with high quality, and the other is encoded with low quality.

To support the immersive media, MPEG-I group, established by MPEG, divided the standardization associated with VR into three phases, 3 Degree of Freedom (3DoF), 3DoF+, and 6DoF [4]. In 3DoF+ and 6DoF, multi-view 360 videos are needed, and it consists of texture and depth images to support 3D video [5]. Since those provide 360 videos in response to user’s movement, it is inevitable to synthesize the immediate views using existing views. View Synthesis Reference Software (VSRS) for 360 videos [6], Reference View Synthesizer (RVS) [7] and WS-PSNR for 360 video quality evaluation [8] were proposed to MPEG to create virtual views and evaluate them.

When transmitting 3DoF+ or 6DoF 360 videos, it requires lots of bandwidth since those need both high-resolution texture and depth. To overcome this problem, down-sampling or region-wise packing could be applied. In this paper, it proposes the appropriate down-sampling ratio and quantization parameter of 3DoF+ texture and depth in view synthesis. It introduces a pilot test with Super MultiView Video (SMV) [9]. Finally, it provides the graph of bitrate and PSNR obtained by 3DoF+ test sequences using 360lib with HEVC.

II. RELATED WORK

A. Standards of 360 video in MPEG

During the 116th meeting of MPEG, MPEG-I group was established due to the support of immersive media. They started to standardize the format of immersive, omnidirectional video in 2017 [10]. Fig.1 shows the standardization roadmap of MPEG. They divided the standardization into 3 phases [11][12]. Phase 1a aims to provide 360 video and contents including stitching, projection, and video encoding.

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No.2018-0-00765, Development of Compression and Transmission Technologies for Ultra High Quality Immersive Videos Supporting 6DoF)
To overcome the limitation of 3DoF, the concept of 3DoF+, part of phase 1b in MPEG-I, was proposed. 3DoF+ provides limited movements of yaw, pitch, roll, described in Fig.2. So it has increased the degree of freedom in 3DoF, relatively.

In 3DoF+, VR device has to offer the video of view which the user watches. If the view user wants to see is not included in the original video, 3DoF+ system synthesizes the view that doesn’t exist before. So, Reference Intermediate View Synthesizer (RIVS) is needed. Also, to synthesize virtual views, additional depth information, contains distances between camera and objects, must be supplied. Since it requires lots of data to be sent, optimization for data transmits and compression ought to be proposed. For this reason, enhanced technologies of communication such as 5G mobile technology[14] and mobile data offloading[15] are announced recently. View synthesis assumes video transmission from the server to client, so the video needs to be compressed as we can see at Fig.3. The anchor view is used in view synthesis, and it has to be encoded and decoded.

Subsequently, phase 2 of MPEG-I deals with 6DoF, which means 3DoF+ with translational movements to X, Y, and Z axes. It supports user’s movements including walking, as described in Fig.2.

B. Multi-view video Coding

Multi-view video enables the user to have a three-dimensional and immersive experience. It provides diverse views gained from one scene simultaneously. Especially, three-dimensional multi-view video includes both texture and depth information. It enables users to have multiple views which they want to watch. MPEG defined the three-dimensional video system[16] which is a part of Freeviewpoint TV(FTV), and it contains multi-view video acquisition, encoding, transmission, decoding, and display. To process the multi-view video efficiently, multi-view video coding[17][18] is required.

Multi-view videos have common features since they contain the same scene at the same time. The difference between each view is that they have an indigenous point of view. What it implies is a multi-view video of one viewpoint can be made by referencing another view.

Fig.4 shows the hierarchical B frame multi-view video encoding structure between primary view and extended views. The blue box means a key frame which can be referenced by B frame. The I frame is able to be reconstructed while the P frame references one frame. The B frame references two frames when predicting. Joint Multiview Video Model (JMVM)[19] for reference software model of multi-view video coding was proposed to compress multi-view video while containing compatibility with H.264.

C. View Synthesis

Even though multi-view video provides some views, it cannot offer a view which is out of source views. Since multi-view video coding requires lots of computing power to process and data, the number of views multi-view video can support is limited. Accordingly, view synthesis for multi-view video[20][21] was developed to overcome the limitation of multi-view video coding. When we use view synthesis, we don’t have to send all of the source views because it synthesizes dropped views which we didn’t send. Also, if the video provider didn’t acquire many source views due to the limitation of a resource such as a camera and the amount of data, still we can synthesize the other views the provider didn’t offer.

Fig.5 shows the View Synthesis Pipeline of RVS. If we increase or decrease the number of views, the number of frames can't be increased or decreased. For this, RVS deforms an image of the input view to generate an artificial virtual view. It is sure that we need lots of computing power to do view synthesis. The advantage of RVS is the View Synthesizer(RIVS) is independent of a codec, so it can be applied to any kind of video codec.

Fig.6 shows the Image projection in a pinhole camera coordinate system. Fig.7 shows the Image plane in camera coordinate system.

![Fig 3. Anchors generation structure of 3DoF+](Image)

![Fig 4. Multi-view video encoding view reference structure](Image)

![Fig 5. View Synthesis Pipeline of RVS](Image)

![Fig 6. Image projection in a pinhole camera](Image)

![Fig 7. Image plane in camera coordinate system](Image)
Fig. 5 describes how to synthesize the intermediate views with Reference View Synthesizer (RVS) 1.0.2[22]. It requires texture video, depth map, and a camera parameter. Depth map [23][24] represents the distance between the camera and the object shown in texture video. If the depth map format is 8bit, the range of the depth value is between 0 and 255. The depth map can be gained by a depth camera which uses a depth sensor. Otherwise, it is able to be generated by depth estimation software. MPEG-4 group proposed Depth Estimation Reference Software (DERs) [25][26] to make the depth map from the texture video efficiently.

Generally, the multi-view video is obtained from a pinhole camera. It projects the actual object into a 2D plane image, as shown in Fig. 6. To implement the projection, world coordinate system and camera coordinate system is needed. World coordinate system presents three-dimensional space. The camera is located in the world coordinate system, and it also has a three-dimensional coordinate system. The camera center point means the location of the camera in the world coordinate system. Camera coordinate system has X, Y, Z axis. X-axis means horizontal axis, Y axis means the vertical axis, and Z axis means the optical axis, in other words, principal axis. The optical axis is the direction of the camera ray. The principal point is an intersection point between the principal axis and the image plane. The distance from the camera center to the principal is called focal length, as shown in Fig. 7. Each point of the object in three-dimensional space is projected onto a two-dimensional image plane by the camera. At that time, explained parameters are used.

To make the intermediate view, the conversion of point coordinates from reference views into synthesized view must be done. Each reference views which is used to synthesize the intermediate view has its own camera coordinate system. If we know the camera parameter of reference views and intermediate view, it is possible to generate camera coordinate system of intermediate view using world coordinate system. Once the conversion is complete, the texture mapping from the reference views to intermediate view can be done.

III. PILOT TEST

A. Pilot test with FTV multi-view sequences

![Champagne_tower (1280x960), 80 cameras with stereo distance, 30 fps, 300 frames](image1)

![Pantomime (1280x960), 80 cameras with stereo distance, 30 fps, 300 frames](image2)

Fig 8. FTV test sequences from Nagoya University

To reduce the bitrate when transmitting multi-view video, this paper proposes low complexity multi-view video transmit system including down-sampling and up-sampling. The feasibility of this method was proved by a pilot test with FTV multi-view sequences [27]. Champagne_tower and pantomime sequences, as shown in Fig. 8, were used. Their resolution is 1280x960, acquired from 80 cameras, and the number of frames is 300.

Fig. 9 describes proposed system architecture with FTV multi-view test sequences. Firstly, it selects the anchor view, which means the source view used to synthesize the intermediate view. Test sequences provide the depth map of 37, 39, 41 views, mean anchor view which requires both texture and depth. The combination of view synthesis is represented in Table I. Second, it down-samples the selected anchor views. The down-sampling ratio is 0, 20, 40, 50, and 75(%), as shown in Table II. For down-sampling and up-sampling, Joint Scalable Video Model (JSVM) [28] was used.

Third, it encodes and decodes the down-sampled views. For encoding and decoding, HEVC reference software (HM) version 16.16 [29] was used. VSRS 4.2 [30] was used to synthesize the intermediate view. Fourth, it up-samples the decoded views. Fifth, it synthesizes the intermediate view by referencing up-sampled anchor views. Finally, it measures Peak Signal-to-Noise Ratio (PSNR) between original intermediate views and synthesized views for objective quality evaluation. For PSNR measurement, JSVM was used.

![Fig 10. RD-curve between PSNR and average bitrate with different QPs](image3)
C), the QPs used for texture and depth is
d to both
ystem architecture for
0
rMuseum, and TechnicolorHijack as
igher as we can see
Image 45x96 to 290x258
views, and v9, v10, v11, v12, v13, v14 were called far v
conducted to ClassroomVideo.
Image 58x648 to 276x789
B.

For encoding, the Quantization Parameter (QP) values are 22, 27, 32, and 37. In a pilot test with FTV multi-view sequences, experiment for every combination of down- sampling ratio, QP, and view synthesis was executed. The pilot test result is shown in Fig.10. It describes RD-curve between PSNR and average bitrate with different QPs. The reason why the graph shows the combination 0-0 to 20-40 is it only includes the combinations whose difference values with original view combination(0-0) are under 1.

Even though the average down-sampling ratio of the combination 0-40 (left view-right view) is equal to 20-20, the PSNR value of 20-20 was higher than 0-20. What is more, the average bitrate of 20-20 was smaller than 0-40. Fig.10 implies that uniform down-sampling ratio of left and right view is better. Also, the performance of 20-40 was better than 0-50 because of the equality of down-sampling ratio.

Fig.11 describes the RD-curve between PSNR and average bitrate with different down-sampling ratio combinations. In the case of 20-20, the difference value between QP=27 and QP=22 is 0.171615, which is very low while the difference value of bitrate is 862.6038, which is very high.

B. Pilot test with 3DoF+ test sequences

For the 3DoF+ experiment, MPEG provides ClassroomVideo[31], TechnicolorMuseum, and TechnicolorHijack as test sequences described in Fig.12. The pilot test was conducted to ClassroomVideo. View v0 was defined as a synthesized view, v1, v2, v3, v4, v5, v6 were named near views, and v9, v10, v11, v12, v13, v14 were called far views, as shown in Fig.13.

For the 3DoF+ experiment, MPEG provides ClassroomVideo[31], TechnicolorMuseum, and TechnicolorHijack as test sequences described in Fig.12. The pilot test was conducted to ClassroomVideo. View v0 was defined as a synthesized view, v1, v2, v3, v4, v5, v6 were named near views, and v9, v10, v11, v12, v13, v14 were called far views, as shown in Fig.13.

TABLE III. PSNR FOR SYNTHESIZED VIEWS OF CLASSROOMVIDEO

<table>
<thead>
<tr>
<th>Input Views</th>
<th>WS-PSNR_Y</th>
<th>WS-PSNR_U</th>
<th>WS-PSNR_V</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) nearOrg+farDown</td>
<td>31.8250</td>
<td>48.9032</td>
<td>51.4948</td>
</tr>
<tr>
<td>(2) nearOrg+farTextureDown</td>
<td>31.4879</td>
<td>47.8435</td>
<td>50.6719</td>
</tr>
<tr>
<td>(3) nearDown+farOrg</td>
<td>31.4080</td>
<td>48.5565</td>
<td>51.1620</td>
</tr>
<tr>
<td>(4) nearTextureDown+farOrg</td>
<td>31.4425</td>
<td>43.7384</td>
<td>50.5797</td>
</tr>
<tr>
<td>(5) nearOrg+farOrg</td>
<td>32.7309</td>
<td>49.9141</td>
<td>52.4917</td>
</tr>
<tr>
<td>(6) nearOrg</td>
<td>31.8293</td>
<td>48.9084</td>
<td>51.5008</td>
</tr>
<tr>
<td>(7) farOrg</td>
<td>31.4247</td>
<td>48.5634</td>
<td>51.1593</td>
</tr>
</tbody>
</table>

The distances between the synthesized view and the near views are same, the far views as well. For objective quality evaluation, WS-PSNR tool [32] was used.

Table III shows the PSNR for synthesized views of ClassroomVideo. PSNR value of (6) was higher than (1) even though (6) has fewer views. Adding more views which are down-sampled isn’t good for the quality of the synthesized view. If the input views are closer to the synthesized view, its PSNR value was higher as we can see by comparing (1) and (3). Interestingly, the PSNR value of (1) was higher than (2) despite the depth maps of (2) weren’t down-sampled while (1) didn’t.

IV. EXPERIMENT

Fig.14 describes the proposed system architecture for 3DoF+ multi-view video transmission including anchor view selection, down-sampling ratio combination selection, down-sampling, encoding, decoding, up-sampling, view synthesis, and measuring WS-PSNR. In 3DoF+ Common Test Condition (CTC), the QPs used for texture and depth is shown in Table IV. The difference value between texture and depth QP is 5, which was decided by an experiment[33]. Table V shows the resolution of down-sampling ratio for ClassroomVideo. The down-sampling is applied to both texture and depth. 360ConvertStatic of 360lib 5.1 was used in down-sampling. Table VI shows the anchor-coded views per class or ClassroomVideo. Class A1 uses all views for view synthesis, while class A2 and A3 uses the subset of views.
TABLE IV. QPs USED FOR TEXTURE AND DEPTH

<table>
<thead>
<tr>
<th></th>
<th>R1</th>
<th>R2</th>
<th>40%</th>
<th>50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Texture QP</td>
<td>22</td>
<td>27</td>
<td>32</td>
<td>37</td>
</tr>
<tr>
<td>Depth QP</td>
<td>17</td>
<td>22</td>
<td>27</td>
<td>32</td>
</tr>
</tbody>
</table>

TABLE V. RESOLUTION OF DOWNSAMPLING RATIO

<table>
<thead>
<tr>
<th>Down-sampling Ratio</th>
<th>0%</th>
<th>12.5%</th>
<th>25%</th>
<th>37.5%</th>
<th>50%</th>
</tr>
</thead>
<tbody>
<tr>
<td>ClassroomVideo</td>
<td>4096x2048</td>
<td>3584x1792</td>
<td>3072x1536</td>
<td>2560x1280</td>
<td>2048x1024</td>
</tr>
</tbody>
</table>

TABLE VI. ANCHOR-CODED VIEWS PER CLASS

<table>
<thead>
<tr>
<th>Test class</th>
<th>Sequence Name</th>
<th># of source views</th>
<th># of anchor-coded views</th>
<th>Anchor-coded views</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>ClassroomVideo</td>
<td>15</td>
<td>15</td>
<td>All</td>
</tr>
<tr>
<td>A2</td>
<td>ClassroomVideo</td>
<td>15</td>
<td>9</td>
<td>v0,v7…v14</td>
</tr>
<tr>
<td>A3</td>
<td>ClassroomVideo</td>
<td>15</td>
<td>1</td>
<td>v0</td>
</tr>
</tbody>
</table>

TABLE VII. 3DoF+ VIEW SYNTHESIS FRAME RANGE

<table>
<thead>
<tr>
<th>Test class</th>
<th>Sequence Name</th>
<th>Frames</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>ClassroomVideo</td>
<td>89-120</td>
</tr>
<tr>
<td>A2</td>
<td>ClassroomVideo</td>
<td>89-120</td>
</tr>
<tr>
<td>A3</td>
<td>ClassroomVideo</td>
<td>89-120</td>
</tr>
</tbody>
</table>

Table VIII shows the WS-PSNR_Y values of synthesized intermediate views. Result of the regular output was better than the masked output’s. Also, class A2 and A3, which discarded some source views showed low WS-PSNR. If the anchor views must be downsampled, the ratio 12.5% is reasonable.

Fig.16 shows the RD-curve between WS-PSNR_Y and Bitrate of A1 with 0% downsampling.

With QP=27 and 12.5% of down-sampling ratio, it saved about 87.81% bitrate while losing only 8% WS-PSNR.
V. CONCLUSION

This study proposes a bitrate-reducing method for 3DoF+ video synthesis and transmission. Especially, by down-sampling and up-sampling the texture and depth, the proposed method saved lots of bitrate while losing only a few WS-PSNR value. However, since the condition of the experiment wasn’t diverse enough to deduct the optimal parameter for view synthesis, the experiment using video compression methods such as region-wise packing[34] have to be progressed.

ACKNOWLEDGMENT

This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No.2018-0-00765),Development of Compression and Transmission Technologies for Ultra High-Quality Immersive Videos Supporting 6DoF).

REFERENCES


