IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 10, 2018, accepted September 20, 2018, date of publication October 1, 2018, date of current version October 25, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2873367

Scalable Omnidirectional Video Coding for Real-Time Virtual Reality Applications

DEYANG LIU[1,2], PING AN[3], RAN MA[3], WENFA ZHAN[1], AND LIEFU AI[1,2]

[1]School of Computer and Information, Anqing Normal University, Anqing 246000, China
[2]The University Key Laboratory of Intelligent Perception and Computing of Anhui Province, Anqing Normal University, Anqing 246000, China
[3]Key Laboratory of Advanced Display and System Applications, Shanghai University, Shanghai 200072, China

Corresponding author: Deyang Liu (liudeyang@163.com)

This work was supported in part by the National Natural Science Foundation of China, under Grant 61801006, Grant 61571285, Grant U1301257, in part by the Key Project on Anhui Provincial Natural Science Study by Colleges and Universities under Grant KJ2018A0361, in part by the Natural Science Foundation of Anhui Province under Grant 1608085MF144, in part by the Foundation of University Research and Innovation Platform Team for Intelligent Perception and Computing of Anhui Province, and in part by the Open Fund of the Key Laboratory of Advanced Display and System Applications, Shanghai University.

ABSTRACT Virtual reality (VR) can provide users an immersive and realistic visual experience, which leads to the widely use of VR in many fields. However, transmission of the ultrahigh resolution omnidirectional video requires huge bandwidth, which brings great challenges for real-time VR application. In this paper, we propose a scalable omnidirectional video coding method to improve the coding efficiency with the help of the viewer's point of view (POV) and provide three-layer scalability as well. Based on the equirectangular projection (ERP), a down-sampling procedure of ERP video with corresponding super-resolution method is proposed to save bandwidth and provide spatial resolution scalability. With the super-resolution version of the reconstructed down-sampled video as the inter-view reference, the viewer's POV within sphere is mapped and encoded in high quality, while the non POV areas are compressed in low quality to further improve the coding efficiency and provide quality scalability. The correlation of the ERP and cube map projection is utilized in the POV mapping procedure. The proposed scalable coding method is achieved based on the multiview extension of high efficiency video coding, where only a few modifications are operated in the encoder side. Experiments results demonstrate that the proposed method can save approximately 75% average bit rate with no significant decrease in quality of the viewer's POV region compared with HEVC standard.

INDEX TERMS Virtual reality (VR), omnidirectional video coding, scalable coding, point of view, MV-HEVC.

I. INTRODUCTION

Omnidirectional video, also known as the panoramic video, can provide users extraordinary viewing experience by simulating the 3D scene of the real world. By wearing the head-mounted displays (HMDs), a corresponding portion of the omnidirectional videos can be played back according to the head movement of the users [1].There are many HMDs, such as Oculus Rift, HTC Vive and Sony Play station VR. With these products, the users can achieve an immersive experience, which leads to the widely use of virtual reality in many fields [2], [3].

Various types of equipments can be used to derive the omnidirectional video contents, such as fish-eye cameras, multiple wide angle cameras or multiple high definition cameras. The captured videos in different directions are then stitched together into a sphere to form the omnidirectional video. In order to provide an immersive and realistic visual experience, adequate resolution is required, and, consequently, effective coding methods become of paramount importance for such particular type of contents.

Since the omnidirectional video is commonly represented by a spherical surface, we cannot directly use the existing coding standards, such as H.265/HEVC, H.264/AVC or MPEG *et al*, or the existing compression scheme for 2D/3D video [4]–[6] to compress the omnidirectional video. A common method is to project the spherical shape into a 2D plane and then use the existed video encoders to compress the omnidirectional contents. In this context, several omnidirectional video coding methods have been proposed, which can be mainly categorized into three different groups all based

on the existing image/video coding standards. The first kind of coding method, referred to as projection based coding method, considers to project the sphere videos into a 2D space by using various projection methods, then utilizes the generic video coding standards for the subsequent compression. However, different projections may introduce different artifacts [7], such as redundant samples, shape distortion and discontinuous boundary, which reduces the coding efficiency. The second kind of compression method, called as optimization based compression method, tries to improve the coding efficiency of the omnidirectional video by exploring the features of the projected video contents and optimizing the coding standards according to these features. This kind of method can mitigate the problem of low coding efficiency caused by the projection deformation. However, since the optimization based compression method has to encode the entire omnidirectional video, it is still difficult to satisfy a real-time VR application. The third kind of compression method, referred to as the regions-of-interest (RoI) based coding method, aims to improve the coding efficiency by transmitting corresponding portion of the omnidirectional video contents in high quality based on the current user's RoI, while the others in low quality [21]. The RoI based method can save lots of bandwidth. Nevertheless, it is hard to extract the RoI areas and such method may induce a bad immersive experience when the users move quickly in the real-time virtual reality applications.

Inspired by the RoI based coding method, in this paper, we propose a scalable omnidirectional video coding method with the help of the viewer's POV for real-time VR applications. The proposed coding method is based on the equirectangular projection. The main idea of the proposed method is to encode the viewer's POV in high quality and the non POV areas in low quality. Different from the RoI based coding method [21], in the proposed method, a down-sampling version of the ERP video (the obtained video after ERP) is firstly encoded as a basic layer in low quality. The region of viewer's POV is then encoded in a high quality in the enhancement layer by using the super-resolution version of the reconstructed down-sampled video as the inter-view reference. The sphere video is divided into six surfaces, and each surface is referred to as one viewer's POV in our method. The change of viewer's POV is detected by the VR device, and then fed back to the encoder. The proposed scalable coding method has a three-layer structure. The basic layer can be seen as the first layer. The super-resolution version of the reconstructed down-sampled video is referred to as the second layer, while the enhancement layer is the third layer. Spatial resolution scalability is provided from first to second layer, while the quality scalability of the viewer's POV is available from the second to third layer. The main contributions of this paper are summarized as follows: 1) a down-sampling procedure is adopted based on the content property of the ERP video to further reduce bandwidth of basic layer; 2) a simple and fast super-resolution method is utilized to provide the spatial resolution scalability; 3) the correlation of the ERP and CMP is analyzed to map the viewer's POV into a specific-shape region within the ERP video, which will be encoded in high quality to provide the quality scalability; 4) the proposed scalable coding method is based on MV-HEVC, and only a few modifications are operated in the encoder side, which make the structure of the proposed coding method simple.

The rest of the paper is organized as follows. Section II presents a review of the related work. The ERP and CMP are briefly introduced in Section III. The proposed scalable omnidirectional video coding method for real-time virtual reality applications is described in section IV. Section V discusses the simulation results and the last section is devoted to conclusions.

II. RELATED WORK FOR OMNIDIRECTIONAL VIDEO CODING

As mentioned above, the omnidirectional video coding method can be divided into three categories: projection based coding method, optimization based coding method and the RoI based coding method. The projection based coding method aims to utilize the generic video coding standards, such as the H.265/HEVC, H.264/AVC or MPEG *et al*, to compress the omnidirectional video by mapping the sphere video into a 2D space with various projection methods. The ERP [8] and CMP [9] are two common projection methods. The ERP tries to map the sphere to a 2D rectangle by stretching the pixels in the latitudinal direction to construct a rectangle. For the north and south poles, the stretching is extremity severe, which increases the bandwidth consumption. In contrast, CMP tires to construct a rectangle to represent the sphere by mapping the sphere into six faces of a cube, then rearrange these faces into a rectangular image. The CMP can mitigate the scaling-caused geometry distortions. Therefore, the CMP can achieve a higher coding efficiency than the ERP. Other than such two projection methods, in [10], a rhombic dodecahedron map projection scheme is proposed, where the sphere is firstly divided into twelve rhombs, then the divided rhombs is rearranged into a rectangle. In order to alleviate the stretch related distortions, a content adaptive representation of omnidirectional video is proposed in [11], where the sphere videos are divided vertically into tiles. The acquired tiles are then resized based on the latitudes or user preferences. In [12], [13], an octahedron projection method is put forward to map the sphere into octahedron faces to maximally achieve content continuity between each face. In [14], an icosahedron projection is put forward, which can also achieve a compact format after rearranging. In order to decrease the over-sampling areas, a novel octagonal mapping scheme is proposed in [15], where the sphere is firstly mapped into an octagon, then the octagon is reshaped and rearranged into a rectangle before encoding. The main idea of the above mentioned methods is to map the spherical information into a 2D plane with different projection methods before encoding. Even though the acquired rectangular video can be encoded by the generic video coding standards, the coding efficiency

is rather low because of the artifacts introduced by the projection.

In order to further improve the coding efficiency of the projection based coding method, the optimization based coding method is proposed. Such kind of method tries to improve the coding efficiency by exploring the features of the projected video and optimizing the coding standards according to such features. In [16], a new motion model is proposed based on the spherical coordinates transform to compensate the deformation in panoramic videos. To solve the problem that the ERP introduces redundancies in high latitude areas, an omnidirectional video coding method using latitude adaptive down-sampling with pixel rearrangement is proposed in [17]. Likewise, a tile-based segmentation and projection scheme are also proposed in [18] to reduce the redundancies in high latitude areas introduced by ERP. To save the bitrate of the omnidirectional videos after compression, two adaptive encoding techniques are put forward in [1], where the first one is a content adaptive temporal resolution adaptation scheme based on the CMP, and the second one is a quantization and rate-distortion optimization scheme for ERP. In [19], two coding methods including intra-frame and inter-frame coding methods are proposed to improve the compression performance of the pseudo-cylindrical panoramic content. The optimization based coding method can achieve a better coding efficiency. However, it is still infeasible to encode a full panoramicin a limited time or real-time with the optimization based coding method.

Since the users can only view a portion of a full panoramic image at a given moment, RoI based coding method is put forward to transmit a specific RoI area in high quality to further improve the coding efficiency. In [20], two RoI based coding methods, called as tiles based streaming and monolithic based streaming, are proposed. The tiles based streaming firstly splits the video into multiple tiles and transmits a subset of tiles of interest to the viewer subsequently. The monolithic based streaming explores to only transmit the macroblocks that falls within the RoI and their corresponding depended macroblocks to further reduce the bandwidth consumption. In [21], a scalable full-panorama video coding method is proposed, where the RoI is mapped and encoded in high quality and the others are encoded in low quality. A pyramid format of equirectangular layout is proposed in [22], where a pyramid layout is utilized for each of the 30 viewpoints. Instead of encoding these viewpoints, such method stores them on the server and only views in a specific angle are transmitted to users when the client makes a request to the server. The RoI based coding method tries to derive a high coding efficiency combining the feature that users only view a portion of a full panoramic image at a given moment. Nevertheless, there are still some shortcomings. Firstly, it is difficult to extract the RoI areas, since the explicit user input varies from user to user. Secondly, most of the RoI based coding methods are latency-sensitive because of some preprocessing procedure or lots of transform work. Thirdly, the quality scalability is considered in the RoI based coding method. However, the spatial resolution scalability is always ignored, which is also benefit to improve the coding efficiency.

To this end, we propose a scalable omnidirectional video coding method with the help of the viewer's POV for real-time VR applications. A down-sampling procedure and corresponding super-resolution method are adopted to provide spatial resolution scalability. In order to avoid extracting the RoI, we try to derive the viewer's POV by analyzing the correlation of the ERP and CMP. The derived viewer's POV is then encoded in high quality to provide the quality scalability. In the proposed coding method, only a few modifications are operated based on the MV-HEVC standard to make the structure sample. The details will be described in the following sections.

III. EQUIRECTANGULAR PROJECTION AND CUBE MAP PROJECTION

Since the proposed coding method is based on the ERP and the mapping of the viewer's POV to the projected video is based on the correlation of the ERP and CMP, we will briefly introduce the ERP and CMP in this section.

A. EQUIRECTANGULAR PROJECTION

The equirectangular projection is the most common projection method, which aims to project the parallels of spherical shape into rows of a 2D shape. The mapping from spherical surface to rectangular plane is shown in Fig.1. The spherical coordinates (θ, φ) correspond to the horizontal and vertical coordinates (x, y). As shown in Fig.1, the coordinate θ varies from $-\pi$ to π, and the coordinate φ varies from $-\pi/2$ to $\pi/2$, which means that the ERP video presents a 2:1 ratio of width to eight. In order to fit in rectangle, the parallels at φ have to be stretched with a ration of $1/cos(\varphi)$, which results in the over-sampling problem, especially in the areas near the pole.

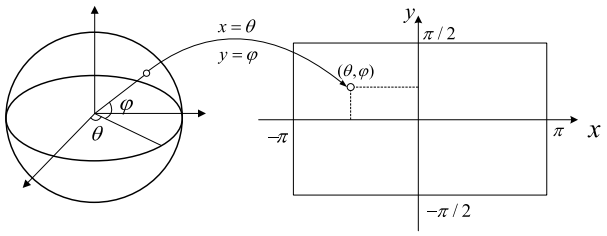

FIGURE 1. Mapping from spherical surface to rectangular plane.

B. CUBE MAP PROJECTION

Different from the ERP, the CMP explores to map the pixels on the sphere into six surfaces of a cube firstly. The obtained six surfaces are then unfolded and rearranged to a rectangular plane. The mapping and unfolded process is shown in Fig.2. Many layout formats are proposed to arrange the unfolded surfaces, such as the 4×3 layout and 3×2 layout shown in Fig.3. Different layout format may lead to different coding efficiency [23]. Compared to the ERP, the CMP has no

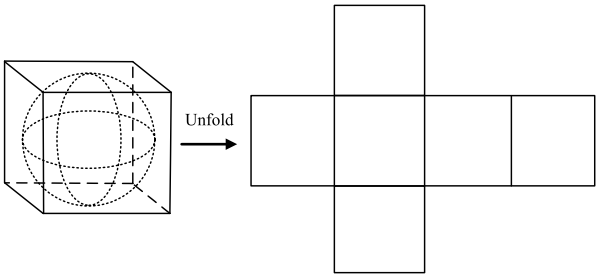

FIGURE 2. The mapping and unfolded process of cube map projection.

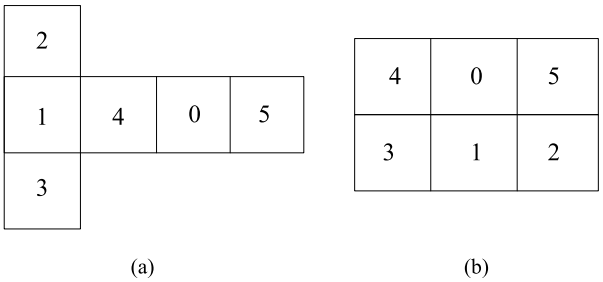

FIGURE 3. Layout format of the unfolded surfaces: (a) 4 × 3 layout; (b) 3 × 2 layout.

geometry distortion within the surfaces, which brings a better Motion Estimation (ME) and Motion Compensation (MC) efficiency within the generic video coding standards. However, the cube map projection still has the over-sampling problem within the edge of each surface. The over-sampling rate is up to 190% compared to the original sphere [18].

IV. PROPOSED SCALABLE CODING METHOD

The proposed scalable coding method aims to improve the coding efficiency of the omnidirectional video, at the same time, provide a three-layer scalability. Since the ERP is the most common projection method of the sphere video, the proposed method is based on the ERP. In our method, the obtained ERP video is firstly down-sampled according to the feature of ERP to form the basic layer (the first layer). The basic layer is then encoded in low quality by using the generic intra and inter prediction based on the MV-HEVC standard. A full ERP video can be reconstructed at the second layer with a simple super-resolution method. In order to provide a good VR experience in the real-time VR applications, the user's POV is mapped and cut to form the enhancement layer (the third layer) based on the user's movement information. With the super-resolution version of the reconstructed down-sampled videos as the inter-view reference, the enhancement layer is encoded in high quality. The user's movement information, such as the visual center coordinate of user's POV, is recorded by the VR device, and fed back to the encoder side subsequently. It is worth mentioning that the coding process of the enhancement layer is delayed by at least one frame than that of the basic layer, since the super-resolution ERP video frame can only be constructed after at least one basic layer video frame decoded. However, since the high frame rate of the VR video, the delay is too short and could be neglected.The detailed processes of the proposed scalable omnidirectional video encoding and decoding system are presented in Fig. 4. The details of each block in the proposed scalable coding system will be explained in the following subsections.

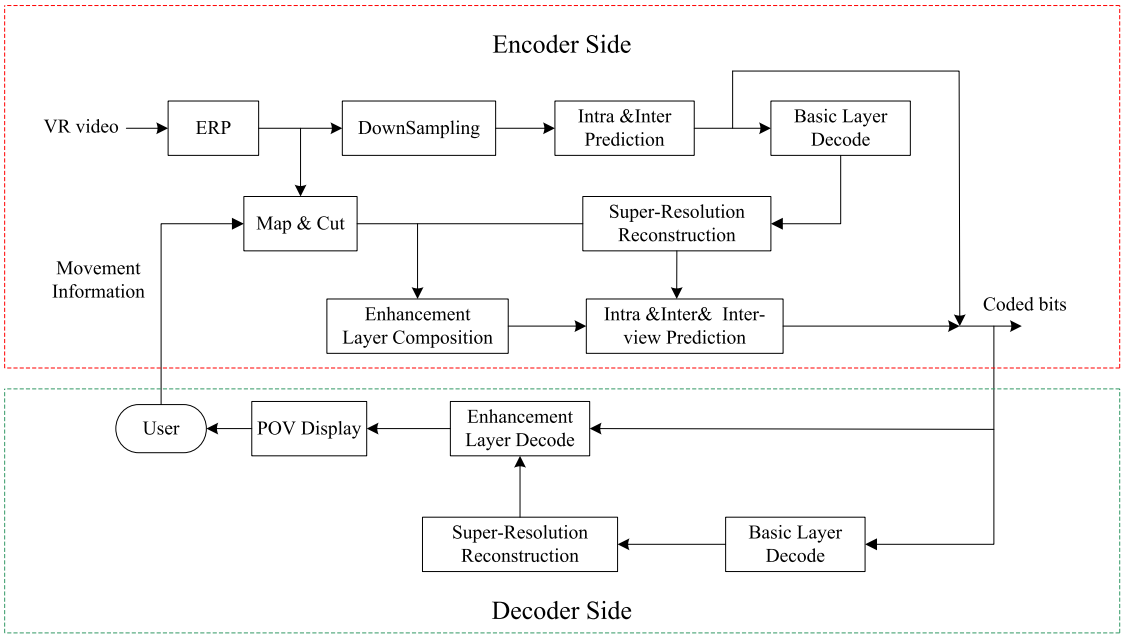

FIGURE 4. The proposed scalable omnidirectional video encoding and decoding system.

A. ENCODER SIDE

1) EPR

The ERP is used to map the sphere video into a 2D plane video. The details of ERP are presented in Section III.A.

2) DOWNSAMPLING

In the ERP, the row pixels in the obtained rectangular image are mapped from the corresponding latitude circle in sphere. The projection results in the high sampling density near the pole, which reduces the coding efficiency of the basic layer. Therefore, we propose a down-sampling method of the ERP video before encoding according to the feature of ERP to improve the coding performance of basic layer. Suppose $f(x, y)$ represents the original continuous EPR video frame with resolution $M \times N$, where (x, y) are the coordinates of the pixels in ERP video frame. A down-sampling image $f(x_s, y_s)$ can be obtained with a sampling factor s. Instead of using a same sampling factor for the entire ERP video frame, we utilize two different sampling factors in the sampling procedure, since the high sampling density exists near the pole. Based on the ERP procedure shown in Fig.1, the down-sampling can be rewritten as

$$f(x_s, y_s) = \begin{cases} f(x \cdot s, y \cdot s) & if\,(|\varphi| < \pi/4) \\ f(x \cdot 2s, y \cdot 2s) & Otherwisw \end{cases} \quad (1)$$

After down-sampling procedure, three low-resolution parts of ERP video frame are acquired, shown in Fig.5(a). We then rearrange the three parts and stitch them together to form a low-resolution rectangular ERP image, shown in Fig.5 (b). The derived low-resolution rectangular ERP video is then encoded as the basic layer.

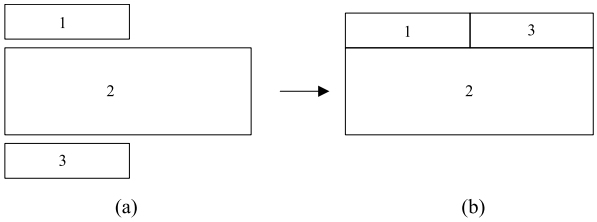

FIGURE 5. The rearrangement and stitching of the down-sampled ERP video frame to form a rectangular image. (a) Three low-resolution parts. (b) The rearranged version.

3) INTRA & INTER PREDICTION

The basic layer is encoded by using the hybrid intra and inter prediction method, which is utilized in the conventional HEVC standard. The obtained bitstream forms the basic layer bitstream.

4) BASIC LAYER DECODE

The encoded basic layer video is decoded for a later super-resolution reconstruction.

5) SUPER-RESOLUTION RECONSTRUCTION

A full ERP video frame is reconstructed by using a Gaussian pyramid based up-sampling method. The low-resolution ERP video frame is firstly upsized to twice resolution in horizontal and vertical directions with the new rows and columns filled with zeros. Subsequently, a convolution procedure is performed with the Gaussian kernel GK, where

$$GK = \frac{1}{16}\begin{bmatrix} 1 & 4 & 6 & 4 & 1 \\ 4 & 16 & 24 & 16 & 4 \\ 6 & 24 & 36 & 24 & 6 \\ 4 & 16 & 24 & 16 & 4 \\ 1 & 4 & 6 & 4 & 1 \end{bmatrix} \quad (2)$$

Regarding to the part 1 and part 3 of ERP video frame, shown in Fig. 5, one more super-resolution procedure is performed by using the same Gaussian pyramid based up-sampling method to make the resolution of the reconstructed ERP video frame be the same as the original ERP image. With a reverse rearrangement used in *DownSampling*, a super-resolution reconstructed ERP video frame is derived. The reconstructed ERP video frame is then be utilized as the inter-view reference to encode the enhancement layer.

6) MAP & CUT

In order to transmit the user's POV in high quality, we have to map the user's POV in the sphere to ERP video frame. In the real-time VR applications, most VR applications can provide a 90° rectangle-view to users, which means that only one six of the sphere is in the user's vision [24]. Therefore, in this paper, we assume that origin sphere video can offer six POVs to the users. The mapping problem of user's POV can be described as how to project the six POVs to the ERP video frame. In our method, we try to find the mapping relationship by utilize the correlation of the ERP and CMP. The reason is that the CMP also uses six cube faces to represent the sphere video, where the six surfaces are consistent with the six POVs. Mapping six POVs in sphere to ERP video frame based on the correlation of the ERP and CMP is shown in Fig.6.

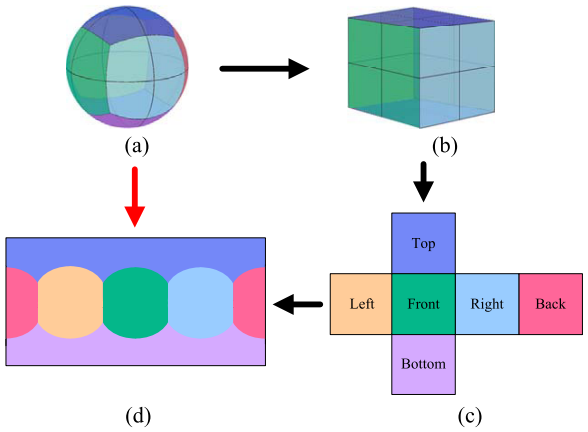

FIGURE 6. Mapping procedure based on the correlation of the ERP and CMP. (a) Shpere. (b) CMP. (d) Projection in ERP. (c) Unfolded cube.

Suppose that the user is on the coordinate origin (the center) of the sphere and faces forward. The conversion of spherical coordinates and 3D Cartesian coordinates can be

given by

$$\begin{cases} x = R cos\theta sin\varphi \\ y = R sin\theta sin\varphi \\ z = R cos\varphi \end{cases} \tag{3}$$

where θ is longitude and φ is latitude, R is the distance between user and sphere. Therefore, the relation between (θ, φ) and the 2D ERP plane coordinates (u, v) can be derived by

$$\begin{cases} u = \lambda \cdot \theta \\ v = \lambda \cdot \varphi \end{cases} \tag{4}$$

where λ is a constant.

As a result, if a point with coordinates (θ', φ') is in the user's POV, then the corresponding point (u', v') in the ERP video frame can be derived by Eq.(4). In order to decide which points in the sphere are within the user's POV, we make full use of the correspondence between CMP and ERP, shown in Fig.6 (c) and Fig.6 (d). Suppose that the horizontal and vertical visual angles of user's POV are all set as 90°, the points in the sphere matching the following constraints are concluded within the user's POV, where the constraints are defined by

$$\begin{cases} \theta' \leq \pi/4 \\ \varphi' \leq cos(\theta') \end{cases} \tag{5}$$

Based on the Eq.(4) and Eq.(5), the user's POV in sphere can be mapped into the ERP video frame. Once the user changes the view direction, the spherical coordinate is changed according to the movement information provided by the VR devices before mapping. After mapping the user's POV in sphere to the ERP video frame, the obtained corresponding areas in ERP video frame are then cut for the subsequent enhancement layer construction.

7) ENHANCEMENT LAYER COMPOSITION

The projected user's POV in the ERP video frame substitutes the corresponding area in the super-resolution reconstructed ERP video frame to form the enhancement layer video frame. The derived enhancement layer video frame is composed of mapped user's POV areas from original sphere and the non POV areas from the super-resolution reconstructed ERP video frame.

8) INTRA & INTER & INTER-VIEW PREDICTION

Other than the intra and inter prediction, the constructed enhancement layer video is encoded by using the inter-view prediction with the super-resolution reconstructed ERP video as the reference. With the inter-view reference, the encoder can avoid some mismatch caused by the stretch deformation using inter-view prediction. Since the non POV areas in basic layer and the enhancement layer are the same, therefore, few bits are needed for such areas. Moreover, by transmitting the non POV areas to the encoder, blank areas will not appear on the user's perspective, which can avoid details missing phenomenon when users move too fast. The acquired bitstream forms the enhancement layer bitstream.

B. DECODER SIDE

1) BASIC LAYER DECODE

The encoded basic layer video is decoded for a later super-resolution reconstruction.

2) SUPER-RESOLUTION RECONSTRUCTION

The process is the same as that in the encoder side. A full ERP video is reconstructed by using a Gaussian pyramid based up-sampling method.

3) ENHANCEMENT LAYER DECODE

The enhancement layer is decoded by using the reconstructed super-resolution ERP video as inter reference.

4) POV DISPLAY

The decoded enhancement layer is then remapped to the sphere for VR display.

C. SCALABILITY

The proposed method can provide a three-layer scalability. The first layer is the down-sampled ERP video according to the feature of ERP. Note that, variable low-resolution rectangular ERP video can be derived by changing the sampling factor to fit different VR devices. The second layer is the super-resolution reconstructed ERP video by using a Gaussian pyramid based up-sampling method. The second layer is then used to construct the enhancement layer. The third layer is the mapped user's POV. With the second layer as the reference, the encoder can avoid some mismatch caused by the stretch deformation and provide a high quality POV to users. The scalability from the first layer to the second layer is a resolution/spatial scalability and the scalability from the second layer to the third layer can be seen as the quality/PSNR scalability of user's POV.

V. EXPERIMENTAL RESULTS

A. SIMULATION SETUP

In order to validate the efficiency of the proposed coding method, six sequences (Dianying, Fengjing_1, Hangpai_1, Tiyu_1, Xinwen_1 and Yanchanghui_1) provided by the IEEE 1857 [25] are used in the test set. The resolution of all the test sequences is 4096×2048, and the frame rate all equals to 30 fps.

The HEVC based multiview extension coding standard, MV-HEVC reference software 14.0 [26], is modified for the proposed scalable coding method. The coding configurations are set as basic encoder configuration, which is defined in [27]. The quantization parameters (QPs) are set as 40 and 22 for basic layer and enhancement layer, respectively. Two views are used, where one is utilized to encode the basic layer and the other is utilized to encode the enhancement

TABLE 1. Results of scalable method compared to HM-13.0 (QP = 22, random access).

Seq.	Projection direction	Bit stream size (KB)			Reduced (%)	Y-PSNR(dB)		Loss (dB)
		Scalable method		HEVC		Scalable method	HEVC	
		Basic layer	Enhancement layer					
Dianying	Top	170	50	4899	95.51	67.86	67.76	-0.10
	Bottom		960		76.93	49.77	49.80	0.03
	Front		2077		54.13	45.81	45.88	0.07
	Back		1680		62.24	44.00	44.01	0.01
	Left		624		83.79	53.01	53.15	0.14
	Right		927		77.61	46.28	46.33	0.05
Fenjing_1	Top	874	4094	21057	76.41	44.00	43.89	-0.11
	Bottom		7068		62.28	41.98	41.83	-0.15
	Front		4048		76.63	42.63	42.40	-0.23
	Back		3594		78.78	42.41	42.27	-0.14
	Left		5647		69.03	41.75	41.60	-0.15
	Right		3661		78.46	42.23	42.05	-0.18
Hangpai_1	Top	699	350	36653	97.14	54.10	54.18	0.08
	Bottom		17317		50.85	41.13	41.11	-0.02
	Front		5796		82.28	42.92	42.94	0.02
	Back		6174		81.25	42.48	42.47	-0.01
	Left		5639		82.71	43.24	43.26	0.02
	Right		6185		81.22	42.19	42.18	-0.01
Tiyu_1	Top	217	781	7286	86.30	49.29	49.25	-0.04
	Bottom		3763		45.37	43.54	43.45	-0.09
	Front		2414		63.89	43.05	42.84	-0.21
	Back		1126		81.57	44.71	44.56	-0.15
	Left		1816		72.10	44.12	43.98	-0.14
	Right		1078		82.23	44.72	44.55	-0.17
Xinwen_1	Top	819	1235	13831	85.15	48.27	48.14	-0.13
	Bottom		3776		66.78	45.04	45.00	-0.04
	Front		3130		71.45	43.74	43.60	-0.14
	Back		3813		66.51	43.15	43.15	0.00
	Left		3593		68.10	43.87	43.83	-0.04
	Right		2264		77.71	43.86	43.76	-0.10
Yanchanghui_1	Top	325	409	6975	89.48	66.85	67.62	0.77
	Bottom		969		81.45	61.63	61.78	0.15
	Front		2575		58.42	48.84	48.98	0.14
	Back		760		84.44	55.54	55.90	0.36
	Left		1682		71.23	58.27	58.37	0.10
	Right		1432		74.81	49.56	49.69	0.13
Average					74.84			-0.01

layer. Six kinds of vision mapping directions (called front, back, left, right, bottom and top) corresponding to the six surfaces of the cube are considered to simulate different situations. The proposed scalable coding method (referred to as Scalable method) is compared with the original HEVC reference software ver. 13.0 (referred to as HEVC). The QP used in HEVC is set as 22 and the coding configurations were set as the ''random access''. 30 frames of each test sequence are encoded by HEVC and the proposed method. In this paper, we use the well-known Bjontegaard delta bitrate (BDBR) metric [28] to evaluate the performances of the proposed method in terms of bitrate reduction and the decoded picture quality. Since the proposed scalable method is focused on the user's POV, we only consider the Y-PSNR of the decoded

POV areas. The sampling factor in the *DownSampling* procedure is set to 2.

B. EXPERIMENTAL RESULTS

The main goal of the proposed scalable coding method is to improve the coding efficiency of the omnidirectional video with the help of the viewer's POV. This subsection, we will verify the effectiveness of the proposed method from three aspects.

1) RD PERFORMANCE

The RD performance comparison of the proposed coding method with HEVC in terms of the bit stream size and the Y-PSNR of user's POV is shown in Table 1. From Table 1, we can see that the proposed scalable coding method is superior to the HEVC. Table 1 gives the bit stream size of basic layer and enhancement layer with six kinds of vision mapping. About 45%-97% bit stream is saved by the proposed method with a little average Y-PSNR gain within POV. According to Table 1, we find that the top and bottom areas always need less bit stream size and have a better Y-PSNR than the other mapping directions. This means that the regions near the polar within the sphere require less bit stream size to be encoded. The main reason is that the pixels near the polar are stretched in the ERP and over-sampling exists in such areas. Moreover, the bit stream size is related to the texture information of the mapped surfaces. Richer texture information always requires more bit stream size to be transmitted. For example, the texture information in the front and back surfaces is rich in sequences Dianying, Fengjing_1 and Xinwen_1, where more bit stream sizes are needed in the transmitting procedure than the other directions. This is mainly because achieving an accurate prediction in such areas by using the intra, inter or the inter-view prediction method is difficult. This is also consistent with the Y-PSNR value of such areas, where the Y-PSNR values in front and back surfaces are lower than that in other surfaces. Compared to the enhancement layer, the basic layer always adds least overhead to the final bit stream, as seen from Table 1. The reason mainly lies in two aspects. One is that the basic layer is a down-sampled version of the ERP video, which can save a lot of bits to be encoded. The other is that the basic layer is encoded in low quality.

2) COMPUTATIONAL COMPLEXITY

A low complexity coding method can give users a better interactive experience in the real-time VR applications. In the proposed method, the computational complexity mainly contains three parts. The first part is the coding of basic layer. The most time-consuming part to encode the basic layer is the intra and inter prediction. Since the basic layer is a down-sampled version of the original ERP video with a low coding quality, the computational complexity is much lower than the HEVC. The second part is the super-resolution reconstruction. Since only a simple convolution procedure is performed with a Gaussian kernel, the computational complexity of the super-resolution reconstruction adds little overhead to the total computational complexity. The third part is the coding of enhancement layer. With the super-resolution reconstructed ERP video as the inter reference, the prediction process comprises the intra prediction, inter prediction and the inter-view prediction. Since most areas of the enhancement layer and the super-resolution reconstructed ERP are the same except the substituted POV regions. The encoder tends to

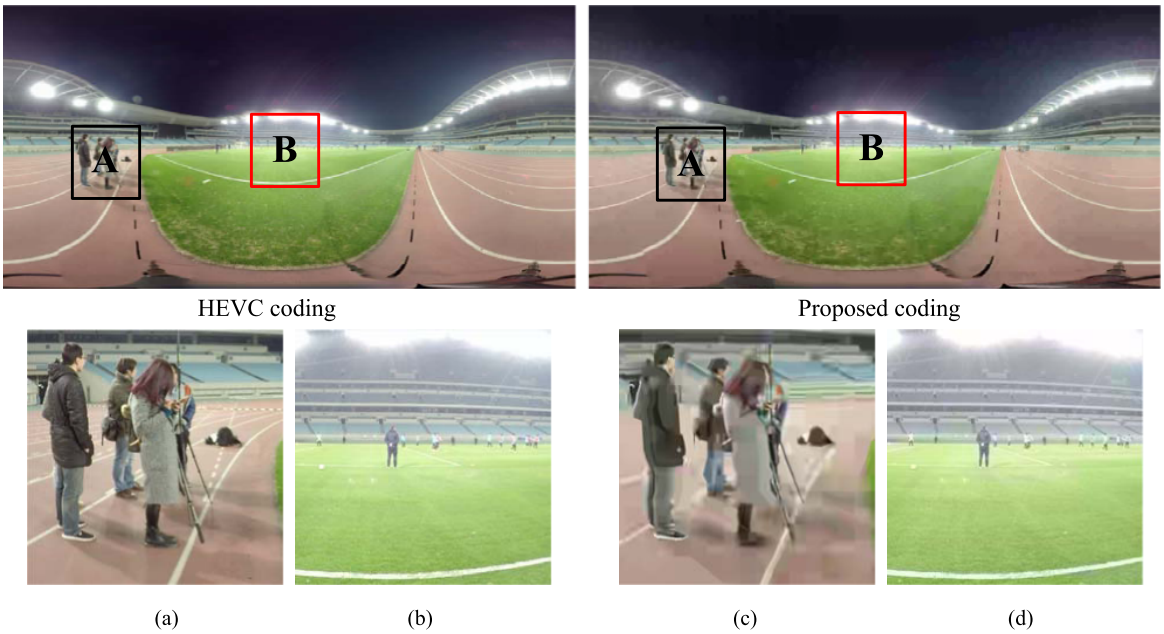

FIGURE 7. Subjective comparison between the decoded output of the HEVC and that of the scalable coding method. (a) Non POV A by HEVC. (b) POV B by HEVC. (c) Non POV A by scalable method. (d) POV B by scalable method.

choose the skip mode for such same areas, since the residual errors are small enough. This can dramatically reduce the computational complexity of coding the enhancement layer.

3) VISUAL QUALITY OF RECONSTRUCTED POVs

In the proposed scalable coding method, we aim to encode the omnidirectional video by transmitting the viewer's POV in high quality and the other areas in low quality with the help of the viewer's movement information. In order to verify the effectiveness of the proposed method, we compare the visual quality of the decoded ERP video frame by using the two methods. We randomly select one part within the reconstructed viewer's POVs and non POVs and compare their visual quality. The visual quality comparison is shown in Fig.7. From Fig.7, we observe that the quality of both parts within the POVs is high, no matter the coding method is the proposed scalable method or HEVC. The reason is that a same QP is used in the enhancement layer of proposed method and HEVC. The same QP ensures a same visual quality for the viewer's POVs. For the non POVs encoded with our method, the visual quality is lower than that encoded with HEVC. This is mainly because the non POVs in our method are reconstructed by using a super-resolution procedure of the decoded basic layer, where the basic layer is compressed with a larger QP. Since the non POVs are not in the viewer's perspective, the low quality non POVs can not affect the immersive experience.

VI. CONCLUSION

In this paper, we propose a scalable omnidirectional video coding to improve the coding efficiency for real-time virtual reality applications. The proposed method encodes the viewer's POV in high quality and the non POV areas in low quality based on the ERP. In order to further reduce the bit-rates, the ERP video is down-sampled and encoded as the basic layer. The region of viewer's POV is mapped to construct the enhancement layer, which is encoded in a high quality by using the super-resolution version of the reconstructed down-sampled video as the inter-view reference. The sphere video is divided into six surfaces, and each surface is referred to one viewer's POV. The change of viewer's POV is detected by the VR device, and then fed back to the encoder. The proposed scalable coding method has a three-layer structure. Spatial resolution scalability is provided from first to second layer, while the quality scalability of the viewer's POV is available from the second to third layer. Experimental results demonstrate that the proposed method can save about 45%-97% bit stream with a little average Y-PSNR gain in POV compared with HEVC. Moreover, the proposed method can achieve a similar visual quality of viewer's POV with HEVC standard in a rather low computational complexity.

REFERENCES

[1] M. H. Tang, Y. Zhang, J. T. Wen, and S. Q. Yang, "Optimized video coding for omnidirectional videos," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2017, pp. 799–804.

[2] B. Kwon *et al.*, "Implementation of a virtual training simulator based on 360° multi-view human action recognition," *IEEE Access*, vol. 5, pp. 12496–12511, 2017.

[3] K. R. Anderson, M. L. Woodbury, K. Phillips, and L. V. Gauthier, "Virtual reality video games to promote movement recovery in stroke rehabilitation: A guide for clinicians," *Arch. Phys. Med. Rehabil.*, vol. 96, no. 5, pp. 973–976, 2015.

[4] Z. Zhang, T. Jing, J. Han, Y. Xu, and X. Li, "Flow-process foreground region of interest detection method for video codecs," *IEEE Access*, vol. 5, pp. 16263–16276, 2017.

[5] S. Zhu, S. Zhang, and C. Ran, "An improved inter-frame prediction algorithm for video coding based on fractal and H.264," *IEEE Access*, vol. 5, pp. 18715–18724, 2017.

[6] J. Qiao, M. Liu, S. Li, Z. He, and Z. Yang, "Highly efficient quality assessment of 3D-synthesized views based on compression technology," *IEEE Access*, vol. 6, pp. 42309–42318, 2018, doi: 10.1109/ACCESS.2018.2859439.

[7] Z. Chen, Y. Li, and Y. Zhang, "Recent advances in omnidirectional video coding for virtual reality: Projection and evaluation," *Signal Process.*, vol. 146, pp. 66–78, May 2018.

[8] J. P. Snyder, "Flattening the Earth: Two thousand years of map projections," *ISIS*, vol. 85, no. 3, pp. 488–489, Sep. 1994.

[9] K.-T. Ng, S.-C. Chan, and H.-Y. Shum, "Data compression and transmission aspects of panoramic videos," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 1, pp. 82–95, Jan. 2005.

[10] C.-W. Fu, L. Wan, T.-T. Wong, and C.-S. Leung, "The rhombic dodecahedron map: An efficient scheme for encoding panoramic video," *IEEE Trans. Multimedia*, vol. 11, no. 4, pp. 634–644, Jun. 2009.

[11] M. Yu, H. Lakshman, and B. Girod, "Content adaptive representations of omnidirectional videos for cinematic virtual reality," in *Proc. 3rd Int. Workshop Immersive Media Exper.*, 2015, pp. 1–6.

[12] H.-C. Lin, C.-Y. Li, J.-L. Lin, S.-K. Chang, and C.-C. Ju, *AHG8: An Efficient Compact Layout for Octahedron Format*, document JVET-D0142, Joint Video Exploration Team of ITU-T SG16 WP3and ISO/IEC JTC1/SC29/WG11, Chengdu, China, 2016.

[13] H.-C. Lin *et al.*, *AHG8: An Improvement on the Compact OHP Layout*, document JVET-E0056, Joint Video Exploration Team of ITU-TSG16 WP3 and ISO/IEC JTC1/SC29/WG11, Geneva, Switzerland, 2017.

[14] S. N. Akula *et al.*, *AHG8: Efficient Frame Packing for Icosahedral Projection*, document JVET-E0029, Joint Video Exploration Team of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, Geneva, Switzerland, 2017.

[15] C. Wu, H. Zhao, and X. Shang, "Octagonal mapping scheme for panoramic video encoding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2402–2406, Sep. 2018, doi: 10.1109/TCSVT.2018.2814074.

[16] Y. Wang, L. Li, D. Liu, F. Wu, and W. Gao, "A new motion model for panoramic video coding," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 1407–1411.

[17] S.-H. Lee, S.-T. Kim, E. Yip, B.-D. Choi, J. Song, and S.-J. Ko, "Omnidirectional video coding using latitude adaptive down-sampling and pixel rearrangement," *Electron. Lett.*, vol. 53, no. 10, pp. 655–657, 2017.

[18] J. Li, Z. Wen, S. Li, Y. Zhao, B. Guo, and J. Wen, "Novel tile segmentation scheme for omnidirectional video," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 370–374.

[19] R. G. Youvalari, A. Aminlou, M. M. Hannuksela, and M. Gabbouj, "Efficient coding of 360-degree pseudo-cylindrical panoramic video for virtual reality applications," in *Proc. IEEE Int. Symp. Multimedia*, Dec. 2016, pp. 525–528.

[20] Y. S. de la Fuente, R. Skupin, and T. Schierl, "Compressed domain video processing for tile based panoramic streaming using SHVC," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2015, pp. 2244–2248.

[21] G. He, J. Hu, H. Jiang, and Y. Li, "Scalable video coding based on user's view for real-time virtual reality applications," *IEEE Commun. Lett.*, vol. 22, no. 1, pp. 25–28, Jan. 2018.

[22] *Next-Generation Video Encoding Techniques for 360 Video and VR*. Accessed: Jan. 21, 2017. [Online]. Available: https://code.facebook.com/posts/1126354007399553/next-generation-video-encoding-techniquesfor-360-video-and-vr

[23] M. Zhou, *AHG8: A Study on Compression Efficiency of Cube Projection*, document JVET-D0022, Joint Video Exploration Team of ITU-T SG16 WP3 and ISO/IEC JTC1/SC29/WG11, Chengdu, China, 2016.

[24] A. T. Nasrabadi, A. Mahzari, J. D. Beshay, and R. Prakash, "Adaptive 360-degree video streaming using layered video coding," in *Proc. IEEE Virtual Reality*, Mar. 2017, pp. 347–348.

[25] *Mapping Core Experiment Description*, IEEE Standard 1857.9-04-N0008, IEEE 1857.9 Working Group, Guizhou, China, Jun. 2016.
[26] *HEVC Multiview Extension Reference Software Ver. 14.0 (MV-HEVC14.0)*. Accessed: Oct. 18, 2014. [Online]. Available: https://hevc.hhi.fraunhofer.de
[27] D. Rusanovskyy, K. Müller, and A. Vetro, *Common Test Conditions of 3DV Core Experiments*, document JCT3 V-C1100, 3rd Meeting ITU-T/ISO/IEC Joint Collaborative Team 3D Video Coding (JCT-3 V), Jan. 2013.
[28] G. Bjøntegaard, *Calculation of Average PSNR Differences Between RD-Curves*, document ITU-T VCEG-M33, 2001.

DEYANG LIU received the B.S. degree from Anqing Normal University, Anqing, China, in 2011, and the M.S. and Ph.D. degrees in signal and information processing from Shanghai University, Shanghai, China, in 2014 and 2017, respectively. He is currently a Lecturer with the School of Computer and Information, Anqing Normal University. His research interests include 3-D video processing and video coding.

PING AN received the B.S. and M.S. degrees from the Hefei University of Technology, Hefei, China, in 1990 and 1993, respectively, and the Ph.D. degree in communication and information systems from Shanghai University, Shanghai, China, in 2002. She is currently a Professor with the School of Communication and Information Engineering, Shanghai University. Her research interests include stereoscopic and 3-D vision analysis, image and video processing, coding and application.

RAN MA received the M.S. and Ph.D. degrees from the School of Communication and Information Engineering, Shanghai University, Shanghai, China, in 2000 and 2008, respectively. In 2000, she joined Shanghai University, where she became an Associate Professor in 2009. Her current research interests include image and video compression.

WENFA ZHAN received the Ph.D. degree from the Hefei University of Technology, Anhui, in 2009. He is currently a Professor with the Department of Educational Technology, Anqing Normal University, Anhui. He has published over 50 papers and holds 10 Chinese patents. His research interests include test data compression and ATPG algorithms.

LIEFU AI received the Ph.D. degree from the Huazhong University of Science and Technology, Wuhan, China, in 2015. In 2015, he joined Anqing Normal University, where he became an Associate Professor in 2018. His research interests include content-based high dimensional indexing and retrieval for large scale image.