Multimedia Tools and Applications (2021) 80:8309–8329
https://doi.org/10.1007/s11042-020-10050-0

Saliency-driven rate-distortion optimization for 360-degree image coding

Jui-Chiu Chiang [1] · Cheng-Yu Yang [1] · Bhishma Dedhia [2] · Yi-Fan Char [1]

Received: 19 November 2019 / Revised: 17 September 2020 / Accepted: 7 October 2020 /
Published online: 2 November 2020

Abstract
360-degree images allow an immersive experience. They offer multiple views of a scene and the viewpoint can be selected by the user. However, the huge amount of data that is necessary for real-time transmission of 360-degree image and video requires efficient coding techniques, particularly for virtual reality (VR) and augmented reality (AR) applications. The viewer is only interested in a part of the scene so compressing the entire scene with equal quality is inefficient. This study initially constructs a saliency model of the 360-degree image and then a visual attention guided coding scheme is developed using a predicted saliency map. For saliency prediction, two methods of saliency prediction are used and the results are fused, to address the problem of geometry distortion in the ERP (Equirectangular Projection) format. A smoothing-based optimization is then realized in the spherical domain to improve the saliency map. Using the saliency map of the 360-degree image, the distortion of the rate-distortion optimization is modified to ensure a better visual experience. The experimental results show that the viewports of greatest interest are rendered with the highest quality and there is a maximum of 14.33% reduction in the bitrate when the quality measurement is performed in these regions.

Keywords 360-degree image · ERP · Saliency prediction · Rate-distortion optimization

1 Introduction

Advances in hardware and software technology bring more applications of VR and AR to our life, such as computer games, live events, and remote medicine. VR provides users with a new

✉ Jui-Chiu Chiang
rachel@ccu.edu.tw

[1] Department of Electrical Engineering, Advanced Institute of Manufacturing with High-tech Innovations, National Chung Cheng University, Chia-Yi, Taiwan

[2] Department of Electrical Engineering, Indian Institute of Technology, Bombay, India

media experience, offering the freedom to explore 360-degree content and giving users visual experience that is more realistic and immersive. This kind of service is available on popular image/video platforms, such as street exploration through Google map or 360-degree content sharing on social platforms. 360-degree images are usually described in terms of longitudes and latitudes or in terms of 3D coordinates on a spherical surface. They are defined as panoramic images, which include information in 360 degrees in the horizontal direction and 180 degrees in the vertical direction. 360-degree images are termed omnidirectional images and spherical images. For efficient storage and transmission, projections are used for 360-degree images that convert each 3D coordinate to a location in the specific 2D plane. The viewport image is generated on demand when the projected 2D image is converted to the spherical domain, followed by rectilinear projection [45].

Several formats are used to represent 360- degree images using specified projections [45]. Equirectangular projection (ERP) is the most widely used format. This stores the 360-degree image in one 2D image. The horizontal and vertical axis in the ERP image represents the information in the longitudinal and the latitudinal directions on the sphere. The ERP image is represented with high resolution. The test 360-degree video of the Joint Video Exploration Team (JVET) [16] has a maximum resolution 8192×4096, and a maximum frame rate of 60 Hz. Obviously, a 360-degree image or video requires a huge storage and transmission bandwidth. Therefore, many academic and industrial studies have sought to increase the compression efficiency of 360-degree image and video. Many studies concern the development of 360-degree image/video processing [24], coding [15, 19, 21–23, 26, 31, 39–41, 43, 46, 47, 51] and streaming [8, 12, 18, 28, 29].

The Video Coding Experts Group (VCEG, ITU-T Q6/16), worked on the standardization of the 360-degree video coding [35]. The Joint Collaborative Team on Video Coding (JCT-VC) studied the means of signaling for supplemental enhancement information (SEI) when encoding a 360-degree video using High Efficiency Video Coding (HEVC) (https://www.itu.int/rec/T-REC-H.265-201612-S/en). SEI specifies information about the projection format and the region-wise packing. Later on, the JVET, which was jointly established by the ITU Telecommunication (ITU-T) Study Group 16 (SG16) and the Motion Picture Experts Group (MPEG), devised a standard for the future video coding, which is called Versatile Video Coding (VVC) (https://jvet.hhi.fraunhofer.de/). 360-degree video coding is one of the most important technologies, in conjunction with high dynamic range (HDR) video coding. The JVET addresses several problems, including the compression of 360-degree video for different projection methods, coding tool libraries and test methods [45]. These developments in VVC are the reference for MPEG-I standard (https://mpeg.chiariglione.org/standards/mpeg-i).

In terms of the development of immersive media technology, the MPEG (ISO/IEC JTC1/SC29/WG 11) established the standard for immersive technology (Immersive Visual Media): MPEG-I (https://mpeg.chiariglione.org/standards/mpeg-i). This standardizes the virtual environment, the degrees of freedom and the related immersive media formats. Currently, there are ten parts. Part 1, “Technical Report on Immersive Media”, defines the requirements and two phases: phase 1a provides three degrees of freedom (3DoF), which allows the viewer to change the yaw, pitch and roll of the rendered viewport and phase 1b extends the 3DoF coordinate system, called 3DoF+, by enabling a slight translation of the viewer position. The second phase allows a significant change in the viewer position and gives enhanced immersion. This is called 6DoF where the rendered viewport is a combination of point cloud data and 360-degree video. Part 2 “Omnidirectional Media Application Format” (OMAF) and part 3 VVC concern the delivery and coding of 360-degree video. For OMAF, it

provides basic services for monocular and stereoscopic 360-degree video. The input video is processed by projection and optional region-wise packing before encoding. For VVC, it provides a significantly better compression capability than former standards, such as HEVC.

There are several ways to capture a scene with 360 degrees. Capturing and stitching multi-view images is one solution. Another solution is to use fisheye cameras at both the front and the rear sides. This type of product has been widely deployed in the market. The 360-degree image/video can be displayed on computer monitors, smartphones, tablets and head-mounted display (HMD). The viewing experience on different kinds of devices is discussed in [11]. Participants reported that the HMD offers the most immersive experience at the expense of greater cognitive burden, motion sickness and physical discomfort. Therefore, an understanding of the exploration behavior when viewing 360-degree images is crucial for the development of many related technologies, including compression, delivery and free-view rendering, and to ensure the highest quality of experience (QoE) for the viewer [2].

Visual exploration of a 360-degree image is significantly different from that for conventional images. A much greater degree of freedom of viewpoint is offered by a 360-degree image. When viewing a 360-degree image, the human visual system focuses particularly on visually attractive elements and ignores the less important viewports. So, predicting the visually attractive elements is important. The amount of data that is required for 360-degree image and video is quite huge, so it is inefficient to allocate the same resource for each part of the 360-degree image and video. If the viewport image that is selected by the viewer can be predicted, more bits can be assigned to the predicted region and fewer bits to the remaining parts during encoding. Using a saliency map increases the efficiency with which viewport-on-demand is realized. For streaming applications, several versions of the 360-degree video are encoded and stored in a server. A saliency map and the head movement collected from the HMD can be used to predict the viewing direction of the viewer in the next instant. This allows a seamless viewing experience during the change in the viewport.

Predicting areas of visual attention involves determining the significance with respect to the surrounding environment. There are studies of saliency prediction for the conventional 2D images [9, 14, 17, 30, 44, 50]. Itti et al. [17] proposed a method to predict saliency using a bottom-up mechanism, whereby the information about color, intensity and orientation is integrated. Zhang et al. [50] presented a saliency predictor called Boolean Map Saliency (BMS), which determines the significance of each pixel by comparing it to its neighboring pixels. The rapid development of deep learning techniques allows the prediction of a saliency model that uses deep learning [9, 30]. Although the ERP image is represented as a 2D image, these saliency predictors do not perform well for it. The ERP image suffers from the geometric distortion, which is propositional to the latitude. Thus, this issue should be addressed during the development of saliency prediction for the ERP image.

This study proposes two techniques for 360-degree images. The first predicts the saliency map of the ERP image. Then a coding technique for the ERP image uses visual attention as a guide. To predict the saliency, the proposed model uses existing saliency predictors for a 2D image. Pre-processing and post-processing are necessary to address the geometrical distortion of the ERP image. In particular, smoothing-based optimization is realized in the spherical domain. During encoding of the 360-degree image, a saliency map for the 360-degree image is used to modify the distortion definition for the rate-distortion optimization (RDO) process, to provide a better visual experience.

The remainder of this paper is organized as follows: Section II gives an overview of related works on 360-degree image and video. The proposed technique for saliency prediction for an

ERP image and the proposed saliency-based coding for an ERP image are respectively detailed in Section III and Section IV. Section VI details the experimental results and Section VII gives conclusions.

2 Related work on 360-degree images/video

2.1 Projections and viewport generation

There are many projection methods for converting 3D spherical information into a 2D plane [45]. ERP is the most widely used projection method and Youtube supports the ERP format. To describe the ERP conversion, a three-dimensional coordinate system is defined, as shown in Fig. 1(a), where the X axis, Z axis and Y axis respectively points towards the front, the right and the top of the sphere. Any 3D point P(X, Y, Z) on the spherical surface is expressed as $X = \cos(\theta)\cos(\varphi)$, $Y = \sin(\theta)$, and $Z = -\cos(\theta)\sin(\varphi)$, where θ and φ are respectively the latitude and the longitude of the point P. The rectangular plane that is formed by φ and θ is the projection result, where φ is in the range $(-\pi, \pi)$ and θ is in the range $(-\pi/2, \pi/2)$. This projection method is simple and has an obvious artifact towards the pole. As illustrated in Fig. 1(b) (https://blog.google/products/google-vr/bringing-pixels-front-and-center-vr-video/), the projection density at the poles and the equator is uneven, so the geometrical distortion increases with latitude. This creates problems for saliency prediction and compression for ERP images.

In addition to ERP, cube map projection (CMP) which uses six square faces to present the surface of the sphere is also a common projection format. Each face represents a 2D image for a particular viewport with a field of view (FOV) of 90°. Each face in CMP is rendered using rectilinear projection. As shown in Fig. 1 (c) and (d), during rectilinear projection, the viewing angle is along the Z axis and the 2D image is formed by projecting the surface of the sphere

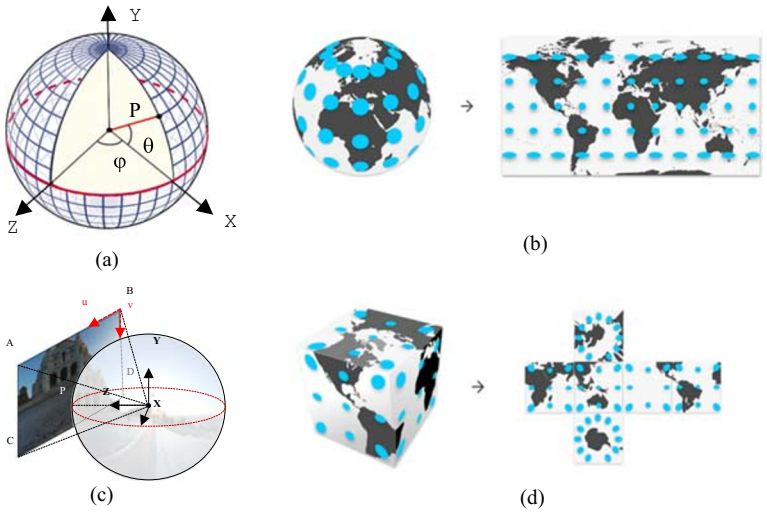

Fig. 1 Projection of a 3D spherical surface to a 2D plane: **a** The 3D coordinate, **b** ERP image (https://blog.google/products/google-vr/bringing-pixels-front-and-center-vr-video/), **c** rectilinear projection [45], and **d** CMP images (https://blog.google/products/google-vr/bringing-pixels-front-and-center-vr-video/)

onto the 2D plane. The pixel value in the 2D image comes from the point on the surface, which is the intersection of the spherical surface and the line that connects the pixel on the 2D plane and the origin of the sphere. If the viewport is not along with the Z axis, the sphere must be rotated so that viewport aligns. The projection is then performed. To allow free-view navigation, when the ERP or the CMP image is projected back to the sphere, a viewport is rendered by rectilinear projection whereby the angles of rotation relative to each axis are specified by the viewer. The JVET has established a 360Lib software package for 360-degree video coding and processing (https://jvet.hhi.fraunhofer.de/svn/svn_360Lib/trunk). The conversion of the projection format and generation of the viewport are realized with 360Lib. In addition to ERP and CMP, the JVET also supports many projections. For more information, please refer to [45].

2.2 Saliency prediction for 360-degree images

There are several techniques for saliency prediction for a 360-degree image [1, 7, 10, 20, 25, 36, 52]. Although ERP is the most commonly used format, it suffers from the geometrical distortion, particularly in regions at high latitudes, so it does not allow an entirely accurate saliency map when a saliency predictor for a traditional 2D image is used. To address this problem, the polar region can be represented using another format, such as a cube map face, which is generated by CMP. The saliency predictor for a conventional 2D image can be used to derive the areas of visual attention for these cube faces and this is then projected to the ERP image. Lebreton et al. [20] proposed GVBS360 and BMS360, which are extended saliency models that use Graph-based Visual Saliency (GVBS) [14] and Boolean Map Saliency (BMS) [50], which are designed for the 2D image. Multi-plane projection is used in [52] to simulate the viewing behavior of human eyes in the HMD. Bottom-up and the top-down feature extractions are performed on each plane. Chao et al. [7] used fine-tuned SalGAN [30] for two image sources, including the original ERP and the cube faces images in several orientations. A fusion process is then used to generate the final saliency map in ERP format. ERP images centered on two different longitudes along the equator and cube map faces generated by rotating the cube center through several angles were used in [36] to generate a saliency map. Ling et al. [25] split the ERP image into patches and extracted sparse features. An integrated saliency map was produced considering the visual acuity and latitude. Abreu et al. [1] determined the saliency using data for eye fixation from subjective experiments. A fusion saliency map was constructed by integrating the saliency maps of the ERP image for various translations.

2.3 360-degree image/video coding

The polar area in an ERP image is stretched so that there are many redundant pixels. Efficient coding must assign fewer bits to regions at high latitudes [19, 21, 24, 31, 41, 46, 47]. Yu et al. [47] divided the ERP image into multiple tiles and adjusted the sampling density by resizing. The sampling rate was determined by rate-distortion optimization (RDO). In [46], a tile-based regional downsampling technique is proposed for inter-frame coding. Three tiles that represent the top, the middle and the bottom parts of the ERP image are rearranged by resizing the top and the bottom tiles. Li et al. [21] used the tile representation, but described the polar region as cambered surfaces and flattened them as circles. The two circular images and tilts are assembled as one 2D image for encoding. A nested polygonal chain mapping was proposed

in [19] to improve the coding efficiency for the polar region. The tile format was also used and tiles are resized according to their locations. The tile nearest to the pole is resampled with a larger factor and placed in the middle of the re-packed rectangular region, surrounded by other resampled tiles from lower latitudes. The rule is enacted for all the tiles and finally tiles with various resampling rates are arranged as one 2D image.

The quality of a region can also be adjusted by assigning an adaptive QP (quantization parameter) [15, 31, 39, 41, 51]. Racapé et al. [31] and Tang et al. [41] expressed the QP as a function of the latitude. Another study [15] computed the QP based on the weight in WS-PSNR (weighted-to-spherically-uniform PSNR) [38]. Other coding-optimization-based techniques have been used in [22, 40]. A spherical domain RDO is realized in [22] and a weighted distortion is used, which depends on the latitude of the pixel in the spherical domain. Luz et al. [26] determined the QP by accessing both the saliency and spatial activity. Several studies focus on the motion model in the sphere domain [23, 43]. Li et al. [23] proposed a spherical motion model, which derives the motion of the block in the 2D plane by projecting to the sphere. A rotational motion model is presented in [43], whereby the motion is described as a rotation on the sphere along geodesics. Viewport adaptive encoding is proposed in [18]. Several viewport dependent projection schemes were studied and multiresolution versions of the ERP and the CMP format were proposed.

2.4 Spherical objective quality metrics

To ensure compatibility with the current video coding standard, a 360-degree image must be projected as a 2D image and compressed. The 2D decoded image is then projected back to a sphere. A free-view image is then generated by rectilinear projection. Since each pixel in the 2D image is not equally important, the specified technique is needed to evaluate the coding performance. The JVET supports three quality assessment metrics, including PSNR, WS-PSNR, S-PSNR-NN [48], and CPP-PSNR [49]. The architecture of the 360-degree video

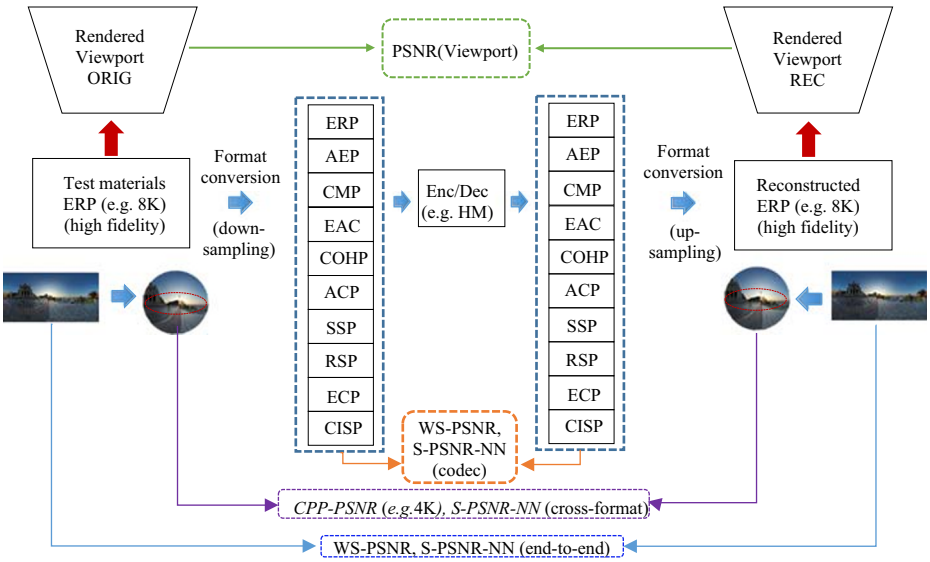

Fig. 2 The testing procedure for 360-degree video [45]

evaluation system specified by the JVET is shown in Fig. 2 [45]. The original ERP image is assumed to be 8 K and it is down-converted to 4 K and then converted to other formats, followed by encoding and decoding. The receiving end performs the calculation for WS-PSNR and S-PSNR-NN on the decoded image and the uncompressed image. Besides, the PSNR computations can also be end-to-end realized.

3 Proposed saliency prediction model for 360-degree images

Omnidirectional images present a scene in a wider range than a conventional image. However, not all of the areas of omnidirectional images received intensive attention. The image feature and the position on the sphere domain can be used to predict accurate saliency maps for 360-degree images.

3.1 Architecture of the proposed model

The proposed saliency prediction model projects the spherical surface into ERP images and multi-view (MV) images. A process then predicts the saliency for each type of image and an initial saliency map is generated by fusing the saliency maps from the ERP image and the MV image. Figure 3 shows the overall architecture of the proposed saliency prediction model, which has four main steps:

(a) ERP-based saliency prediction
(b) MV-based saliency prediction
(c) The use of an equator bias
(d) Optimization in the spherical domain

In the following, each step is introduced with details.

3.2 ERP-based saliency prediction

In the ERP image, the borders on the two sides are connected in the scene. However, if saliency prediction uses a conventional predictor, a visual target is difficult to be recognized if

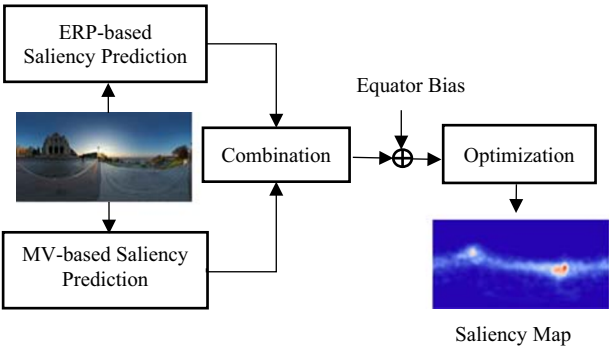

Fig. 3 The architecture of the proposed model

it is on the border. To preserve the saliency in the global context, more than one ERP image is generated by translating the original ERP image. The works in [36, 52], respectively use 2 and 4 ERP images.

Similarly to [36, 52], this study uses 8 orientations along the equator with a longitudinal sampling rate of 45 degrees. As mentioned previously, the region near the equator in the ERP image has good geometry, but other regions suffer from the geometrical distortion, which increases with latitude. Therefore, conventional saliency predictors only yield accurate attention models for the middle portion. The proposed method predicts the saliency for the middle portion and the edges separately for each ERP image. The edge portion is the region that corresponds to the top and bottom faces of CMP and the middle portion denotes the remaining region of the ERP, as illustrated in Fig. 4. Saliency maps for the middle portion and the two edge portions are generated directly by SAM-ResNet [9], which achieves good performance for the conventional image that satisfies the MIT300 benchmark [5]. Since the saliency maps for the cube faces at different orientations with fixed latitude can be viewed as one map at various angles of rotation, the saliency needs not to be predicted for each orientation. Only the saliency maps for the top and bottom faces of the original orientation are predicted. For the middle portion, a saliency map is generated for each orientation and these maps are fused into one map by taking the maximum value. The saliency of the middle portion and the edge portions is assembled after the top and bottom faces of the CMP are projected onto the ERP format. However, before integration, the saliency map for the edge portion is scaled appropriately so that the maximum values in the middle portion and at edges are equal. These procedures are illustrated in Fig. 5.

3.3 Multiview-based saliency prediction

In addition to the ERP-based saliency map, the saliency map is derived using Multiview (MV) images around the sphere. These MV images are rendered by changing the viewport. One viewport corresponds to one image. Using the software 360Lib that is developed by the JVET, arbitrary viewport can be generated by assigning the rotational angles of the three axes. Each viewport image is a 2D image so conventional saliency predictors can be used. Different from CMP images, which have six faces for one orientation, the number of MV images is not fixed

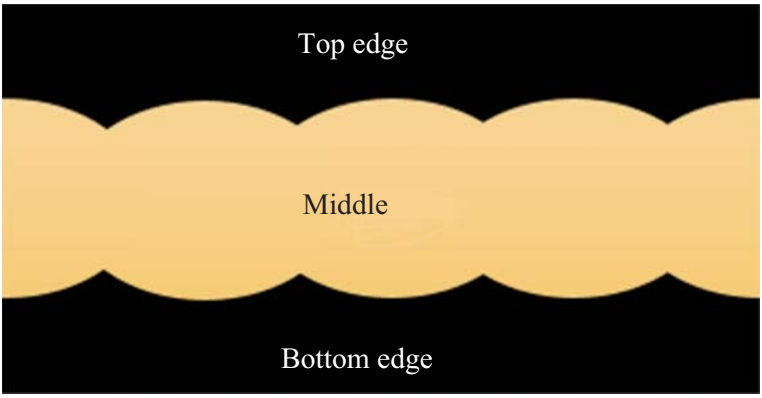

Fig. 4 An ERP image is split into two edges and one middle portion

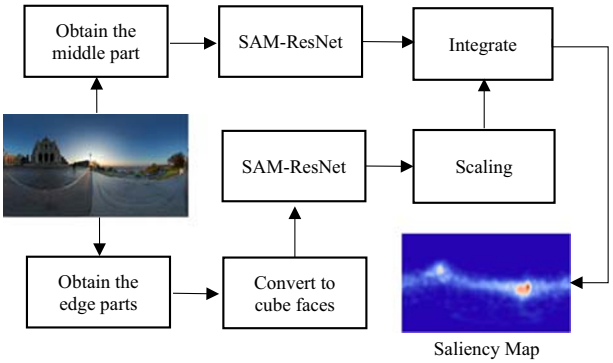

Fig. 5 Procedures for the ERP-based saliency prediction

and more flexibility is allowed. This study uses 62 viewports. One viewport is rendered every 30 degrees along the equator. The procedure is repeated for the circle of latitude 30°, 60°, −30° and − 60°. One viewport is rendered for the north pole and one for the south pole. The FOV for each viewport is 90°. Then the MV images representing the sphere are then produced. Figure 6 illustrates the procedure to generate the viewport along the equator.

BMS [50] is used to predict the saliency of the MV images. Since these MV images overlap, an MV-based saliency map in the ERP format is obtained by combining all the MV saliency maps. A weight is assigned to each pixel in the MV saliency map, as shown in (1)

$$w^{MV}(i,j) = \left(1 + \frac{d^2(i,j)}{r^2}\right)^{-3/2}, \tag{1}$$

where $2r$ is the width of the MV image, and d is the distance between the pixel (i,j) and the center pixel in the MV image. The idea is that the pixels in the center of the MV image have the highest weight, while the boundary pixels have the least weights. After projecting the MV saliency maps back to the ERP format, the obtained saliency is calculated as:

$$S^{ERP2}(i,j) = \frac{\sum_{l=1}^{k} w^{MV}(i_l,j_l) \times S_l^{MV}(i_l,j_l)}{\sum_{l=1}^{k} w^{MV}(i_l,j_l)}, \tag{2}$$

where S_l^{MV} is the saliency map of the *lth* MV image, (i_l,j_l) is the pixel location in the *lth* MV image where its corresponding pixel location is (i,j) in the ERP image and k is the number of MV images involved for current pixel (i,j).

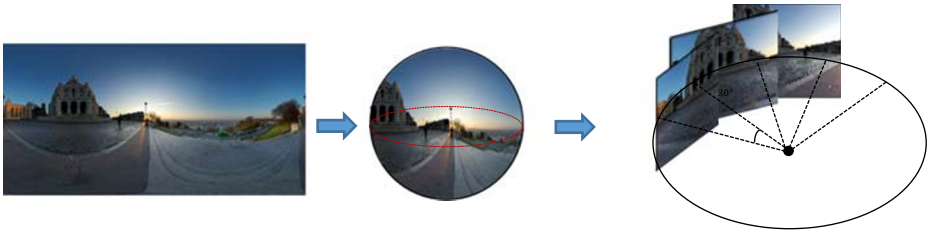

Fig. 6 Multi-view projection for the 360-degree image

When the ERP-based saliency, denoted as S^{ERP1} and the MV-based saliency, denoted as S^{ERP2} are obtained, they are combined into an initial saliency map, denoted as S_{ini}. Before they are combined by averaging, the maps are scaled to ensure the maximum values for each are the same.

3.4 The use of an equator bias

As mentioned in [34] that the regions near the equator are statistically attractive regions for VR navigation, an equator bias is used to predict saliency for 360-degree images. A dataset [32] is used to extract a global latitude-wise subjective attention map. This latitude driven characterization is used to refine the saliency map generated in the previous processes. The equator-bias guided saliency at latitude i is calculated as:

$$S_{EB}(i) = \frac{1}{m \times n} \sum_{t=1}^{n} \sum_{j=1}^{m} S_t(i,j), \tag{3}$$

where $S_t(i,j)$ denotes the subjective saliency value for image t at location (i,j), and n and m respectively denote the image numbers and the width of the image. A weighted average of the equator bias map, denoted as S_{EB},and the initial saliency map, S_{ini}, which is generated from the previous steps, are fused as:

$$S_E = w \times S_{ini} + (1-w) \times S_{EB}, \tag{4}$$

where w is empirically selected as 0.7, which accounts for the contribution of the scene-dependent characteristics and the equator bias.

3.5 Optimization in the spherical domain

The last step involves smoothing the saliency map to remove noise while maintaining the edge. An optimization approach is used [27]. The objective cost function is:

$$J(S_F) = \sum_p \left((S_F^p - S_T^p)^2 + \lambda \sum_{q \in N(p)} (S_F^p - S_F^q)^2 \right), \tag{5}$$

where S_F is the smoothed saliency map, p and q respectively denote specified pixels on S_F. $N(p)$ is the set of four nearest neighbors for the pixel p and S_T is a manipulated version of S_E through a masking operation. This means that the value of some pixels of S_E is retained on S_T, while the remainder is set to 0. The mask is generated by a uniform sampling of the spherical surface using a spiral-based method [6]. In S_T, only the pixel that corresponds to a uniformly sampled point on the sphere is preserved. Because neighboring pixels in the ERP format do not have fixed distance in the spherical domain and not all the pixels in the ERP domain are equally important. Similarly to the metric of S-PSNR which computes the PSNR for selected pixels that are uniformly distributed on a sphere surface, the uniformly sampled pixels on the sphere are projected back to the ERP image to form the mask. These pixels become seeds and the image is smoothed. The number of points that is sampled on the spherical surface is directly proportional to the size of the ERP image.

4 Proposed saliency-driven 360-degree image coding

Since the ERP format is widely used, it is used as the input for the proposed scheme. As mentioned in the previous section, the geometrical distortion in the ERP image is greater at

higher latitudes. To address this problem and to ensure efficient encoding, some works [21, 46, 47] divide the ERP image to several tiles and reduce the amount of resource to the region at high latitude by squeezing the width of the tile or by assigning a larger QP. However, the coding efficiency decreases with a number of tiles. Besides, when the width of the tiles is squeezed, the prediction across the tiles becomes less efficient so that the coding efficiency is reduced. The adaptive-QP-based method uses different QPs across tiles so the playback of a free view is unsatisfactory if the selected viewport covers several tiles that are reconstructed with different quality.

This study proposes a saliency-driven coding technique for a 360-degree image. The saliency map and a weight map that is used to calculate WS-PSNR are combined into a final weight map. The distortion term in the RDO is modified using this final weight map. This ensures that the regions with a high weight are encoded with smaller QPs and high-quality viewports are rendered after reconstruction. The computation of the weight for WS-PSNR is detailed in the next section.

4.1 Weighted-to-spherically-uniform PSNR (WS-PSNR)

In addition to the saliency map, the weight used for the quality metric WS-PSNR is also used to derive the final weight map for the proposed coding technique. A WS-PSNR considers the position on the spherical surface to compute the PSNR. A stretching ratio is defined that represents the area of a point (x,y) on the projection plane over the area of the corresponding longitude and latitude location (θ, φ) on the spherical surface. The stretching ratio for a point (x,y) in the continuous domain for the ERP format is:

$$SR(x,y) = cos(y), \tag{6}$$

where the range of x, and y is $(-\pi$ to $\pi)$ and $(-\pi/2$ to $\pi/2)$ respectively. Since the ERP image is in a digital format, the SR expression in the continuous domain must be discretized. The SR of the pixel (i, j) in the ERP image is calculated as:

$$SR(i,j) = \frac{\iint SR\Big(x(i,j), y\Big(i,j\Big)\Big)dxdy}{\iint dxdy} \tag{7}$$

where $(x(i,j), y(i,j))$ denotes the sampling location on the continuous x-y plane for a discrete point (i,j). The weight is simplified and expressed as the SR for the center pixel. For the ERP, the weight is calculated as:

$$w_s(i,j) = cos\frac{\left(j + 0.5 - \frac{H}{2}\right)\pi}{H} \tag{8}$$

where H is the height of the ERP image.

4.2 Modified distortion for RDO

For image/video coding, rate-distortion optimization [37] is used to determine the best coding mode, in order to ensure a compromise between cost and performance. The RDO is generally expressed as:

$$J = D + \lambda R, \tag{9}$$

where R is the bitrate that is required for the current block and D is the distortion, which is the sum of the squared difference between the original block and the reconstructed block. The Lagrange Multiplier λ controls the balance of R and D and it is modeled as a function of the QP.

A benefit of expressing a 360-degree image using the ERP image is that the ERP is a rectangular image that can be encoded by state-of-the-art coding standards. Although it is feasible to do this, the performance is not optimal. The ERP image is a data format and is not designed to be displayed directly for VR applications. Therefore, the distortion term for RDO must be modified using the specified characteristics of the ERP image.

This study proposes a saliency-driven RDO. The distortion is weighted in terms of the importance of the pixel and expressed as:

$$D_{CTU} = \frac{W \times H}{\sum_{i=0}^{W-1}\sum_{j=0}^{H-1} w(i,j)} \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} w(i,j) \times \left(I(i,j) - \widehat{I}(i,j)\right)^2, \tag{10}$$

where W and H are the width and height of the ERP image, $w(i,j)$ is the weight that represents the importance of the pixel (i,j) and $I, \widehat{I}$ denotes, respectively, the original and the reconstructed block. For HEVC, the coding unit is CTU (coding tree unit) and the CTU size is $N \times N$. The weight w is computed by considering the saliency value, denoted as w_c, and the weight in the WS-PSNR metric, denoted as w_s, as

$$w = w_c \times w_s, \tag{11}$$

The w_c and w_s for the test image P4 in the dataset [32] are shown in Fig. 7. It shows that a high weight appears in the region around the equator for both $w_c(i,j)$ and $w_s(i,j)$.

The distortion term is modified during RDO to reduce distortion for the CTU that is more important. In the normalization term in (10), the denominator sums the weight in the whole image. The QP for each CTU is computed by considering the relative importance within the ERP image. If the weight is uniformly distributed throughout the image, the distortion and RDO are not changed. However, if some regions are more important, they become more distorted if the QP is not changed. Using the new balance between the new distortion and the rate, for a CTU with a higher weight, the QP that is determined by the new RDO is smaller. For a CTU that is less important, which is usually near the polar area, the distortion is reduced and the rate becomes dominant. Therefore, a larger QP is assigned. To ensure a QP adjustment, an adaptive QP is used in the proposed scheme, whereby all QPs within the range of initial QP $\pm \Delta$QP are examined and the one that gives the best R-D performance is selected. In [22], a spherical domain RDO and a weighted distortion is used, as shown in (12),

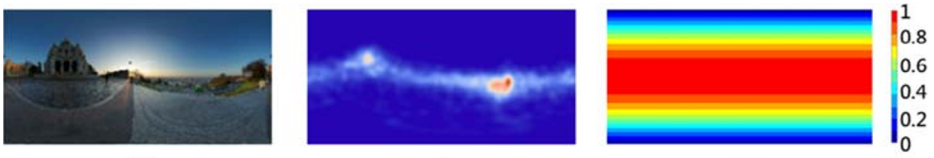

Fig. 7 Illustration of w_c and w_s. (**a**) ERP image, (**b**) w_c for (**a**), and (**c**) w_s

$$J = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1} w_s(i,j)\left(x(i,j)-\widehat{x}(i,j)\right)^2 + \lambda R \tag{12}$$

Then, (12) is rewritten as (13) by considering a block-based operation where w_a is a block-based weight.

$$J = D + \frac{\lambda}{w_a} R \tag{13}$$

In this way, distortion remains and the Lagrange multiplier is changed. This study is different from [22] in three respects: 1) the modified distortion in (10) is defined differently from that in [51], where no normalization term was used; 2) there is no need to modify the Lagrange multiplier for this study but it is modified in [22] by considering the weight in the distortion term, and 3) the QP is automatically determined during the RDO process for this study but it is pre-computed based on the weight in WS-PSNR in [22] and is independent with the input.

5 Experimental results

The performance of the proposed saliency model of the 360-degree image is firstly presented. The saliency model is then used for the proposed coding scheme and the coding performance will be assessed using several objective metrics.

5.1 Saliency prediction

Two datasets [33, 42] are used to evaluate the performance. The dataset in [42] is the test set for the Grand Challenge Salient 360! ICME2017 while the dataset in [33] is the training set for the Grand Challenge Salient 360! ICME2018. These datasets include the data for the original 360-degree images and the head movement and head-eye movement collected from the subjective experiments. The head saliency and head-eye saliency are then served as the ground truth. This study considers the head-eye movement. Four common objective metrics for saliency community are used: Kullback-Leibler Divergence (KLD), Pearson's Correlation Coefficient (CC), Normalized Saliency Scanpath (NSS) and AUC-Judd [4]. The toolbox [13] is used to compute these scores.

The proposed scheme is first evaluated using a database [42] that contains 25 images. Table 1 shows the results. The results of several works are also compared in this table. It is seen that all the model achieves a similar AUC score. The proposed technique outperforms the

Table 1 The head-eye movement prediction using dataset [42]

Method	KLD↓	CC↑	NSS↑	AUC↑
GVBS360 [20]	0.698	0.527	0.851	0.714
[52]	0.481	0.532	0.918	0.734
[7]	0.431	0.659	0.971	0.746
[36]	0.42	0.61	0.81	0.72
[25]	0.477	0.550	0.936	0.736
[10]	0.469	0.570	1.027	0.731
Proposed	0.442	0.566	1.031	0.723

Table 2 The head-eye movement prediction using dataset [33]

Method	KLD↓	CC↑	NSS↑	AUC↑
[7]	0.739	0.642	1.585	0.820
[10]	0.769	0.618	1.616	0.768
Proposed	0.737	0.616	1.615	0.770

other schemes in terms of the NSS score and it is comparable in terms of the KLD and CC metrics. Unlike a previous study by the authors [10], which uses ERP and CMP images as the input for the saliency predictor, this study replaces CMP-based saliency prediction with MV-based saliency prediction and the overall quality is improved. The performance using the dataset [33] is detailed in Table 2. Few studies report the score for the training set for the Grand Challenge Salient 360! ICME2018, so only the results of two studies, [7, 10] are compared. Table 2 shows that the proposed scheme has a smaller KLD score and a higher NSS score, compared to [7].

5.2 Saliency-based coding for 360-degree image

To verify the performance of the proposed coding technique, 12 images from the dataset of omnidirectional images in [33] are used. These images are divided into three groups. Each group has 4 images. The first group has high performance, the second group has medium performance and the third group has low performance in terms of KLD when the predicted saliency is compared to the ground truth one. The mean score for each group is listed in Table 3. There is a significant difference between these results and the average results are shown in Table 3. These three groups are used to verify the coding performance for saliency maps of a different quality that is predicted using the proposed technique.

The proposed coding scheme is implemented in HM16.17. The coding scheme is all intra and the QP is 22, 27, 32 and 37. When the ERP images are decoded, two groups of viewports are rendered using the tools in 360Lib (https://jvet.hhi.fraunhofer.de/svn/svn_360Lib/trunk). The first group renders six viewports along the equator every 60 degrees with a FOV of 75° in both the horizontal and vertical directions. The second group renders viewports that are centered on specified locations, as determined by the ground truth saliency map. Blocks of size 64×64 are determined that present locations with high visual attention. The visual attention for a block is calculated by summing the saliency value inside the block. The centers of the top-3 blocks then serve as the specified viewport locations, and the saliency-based viewport is rendered using rectilinear projection. Only top-3 viewports are used because some 360-degree images do not have many attractive targets, as illustrated in Fig. 8. It is seen that the ground truth saliency has limited regions with high saliency.

Table 3 Saliency prediction scores for three groups of images

	High P5, P26, P.27, P73	Medium P25, P32, P83, P93	Low P4, P7, P48, P57
KLD	0.460	0.598	0.951
CC	0.707	0.541	0.607
NSS	1.615	1.447	1.808
AUC	0.786	0.743	0.779

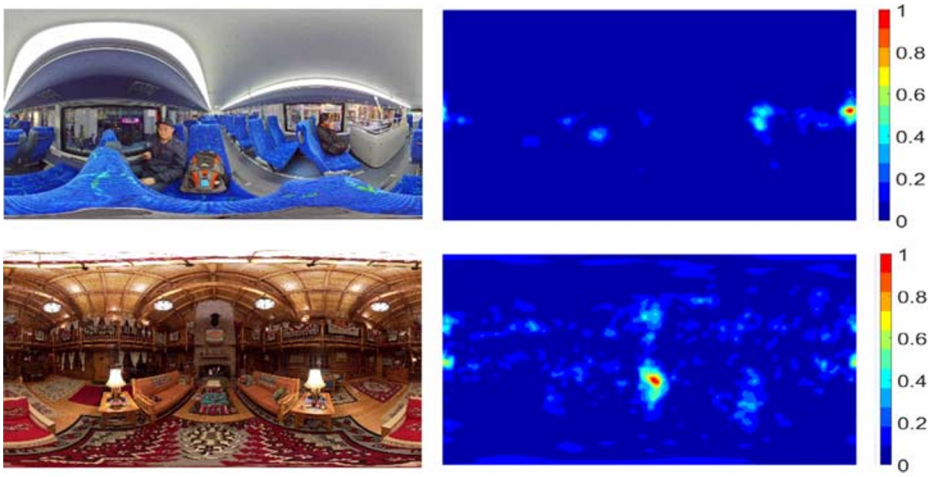

Fig. 8 Two 360-degree images and their saliency maps (ground truth). The top is P57 and the bottom is P25 [33]

Figure 9 shows the rendered viewports for Image P26 [33]. In the first row, the original image and the saliency for the ground truth and the proposed model are shown. When the optimization on the sphere domain is performed, the predicted saliency is more accurate. For the streetlight, the wrong saliency value is corrected. Figure 9 also illustrates three kinds of rendered viewports for Image P26: the equator-based, top-3 saliency-based viewports and the viewports at the pole. The top-3 saliency-based viewports are the viewports that attract the most attention. The viewports at the pole show the sky and the ground, which are seldom required for free-view navigation.

The BD-rate [3] for the proposed coding scheme is defined with respect to the HEVC anchor. Two kinds of PSNR are considered: the PSNR for 6 viewports on the equators, denoted as EQ-PSNR and the PSNR for the top-3 saliency-based viewports, denoted as SM-

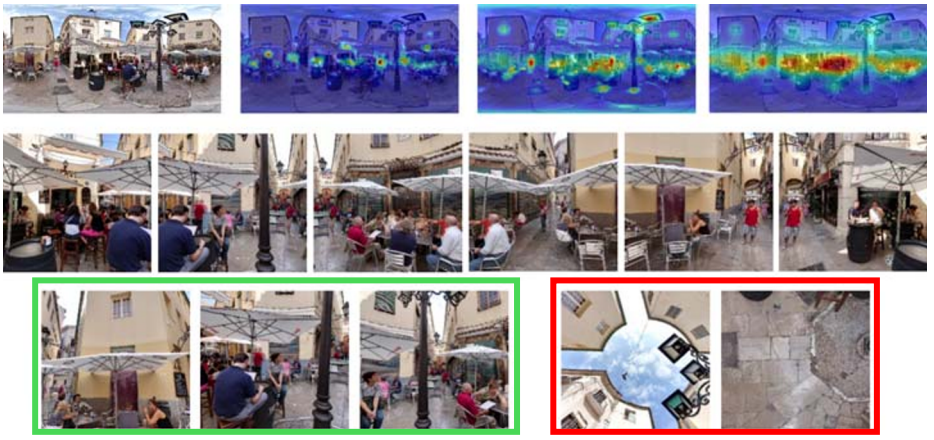

Fig. 9 The rendered viewports for Image P26 [33]. The first row shows the 360-degree image and the saliency overlay on the 360-degree image. From the left to right, they are the ground truth and the predicted saliency without and with the optimization in the sphere domain. The second row shows the equator-based viewports. The third-row shows the top-3 saliency-based viewports inside the green box and the viewports at the pole inside the red box

PSNR. The metrics over the whole ERP image, including WS-PSNR and S-PSNR-NN are also reported. WS-PSNR uses the weighted mean square error to compute the PSNR while S-PSNR-NN measures the quality of a set of uniformly sampled positions on the sphere.

In (11), two types of weight are used. To demonstrate the superiority over the results for one weight, an experiment is conducted. Three images P5, P25 and P7 [33] with a single weight are evaluated. The weight is either w_s or w_c, which is the predicted saliency using the proposed method. P5, P25 and P7 respectively represent the group with high, medium and low KLD performance. Their performance determines whether there is a difference in images with different KLD scores. Table 4 shows the results. The bitrate is reduced for EQ-PSNR and SM-PSNR if both weights are used so the strategy that is defined by (11) increases the quality of the viewport that is highly salient.

The performance of the proposed coding technique is compared with results for [22, 26]. In [26], saliency is used to derive the QP for each CTU while RDO in the spherical domain was adopted in [22]. The saliency used for [26] is the predicted saliency using the proposed work, in order to verify how to use the saliency for 360-degree image coding. In [22], the distortion is modified by considering the weight in WS-PSNR. The Lagrange multiplier is then changed and such a change is equivalent to a QP adjustment. Studies [22, 26] modify the QP in individual ways to achieve improved coding performance for the 360-degree images. For this study, the QP is adjusted automatically and is determined by the RDO when the definition of distortion is modified using the saliency map and the weight in WS-PSNR.

Table 5 summarizes the performance for the 12 images with respect to the HEVC anchor. These results show that the proposed technique achieves a significant bitrate reduction, especially for equator-based and the top-3 saliency-based viewports. The average bitrate reduction for SM-PSNR and EQ-PSNR is 8.73% and 9.76%, respectively, using the proposed scheme. In particular, the maximum bitrate reduction for SM-PSNR is 14.33%. Compared to [26] which is also a saliency-driven coding scheme, Table 5 shows that the modified RDO determines the right QP. For WS-PSNR and S-PSNR-NN metrics, the respective bitrate increment is only 0.95% and 0.99% for the proposed technique and 1.66% and 1.74% for [26]. The bitrate increase for the proposed scheme is smaller than the bitrate reduction in SM-PSNR and EQ-PSNR. Compared to [22], which is also an RDO-based coding scheme, the proposed scheme has a greater bitrate reduction for EQ-PSNR and SM-PSNR. In terms of WS-PSNR an S-PSNR-NN, [22] has a greater bitrate reduction. From the performance for three image groups, Table 5 shows that the quality of the saliency map has an obvious effect on the coding efficiency. Generally, the proposed saliency model is sufficiently accurate in identifying the visual target, so the proposed coding scheme performs well for most of the test images.

The proposed scheme has a bitrate reduction of 11.25% for Image P25. To further analyze the performance, the QP map for the proposed scheme, the anchor, [22, 26]

Table 4 BD-rate (%) for different weight maps (image dataset [33])

	EQ-PSNR			SM-PSNR			WS-PSNR			S-PSNR-NN		
	w_c	w_s	$w_s * w_c$	w_c	w_s	$w_s * w_c$	w_c	w_s	$w_s * w_c$	w_c	w_s	$w_s * w_c$
P5	−8.16	−7.57	−9.75	−7.85	−7.29	−9.28	−1.33	−3.10	−0.22	−1.31	−3.04	−0.18
P25	−7.54	−10.66	−13.94	−7.57	−8.49	−11.25	−0.53	−3.66	−0.65	−0.39	−3.54	−0.60
P7	−5.94	−8.78	−8.52	−5.68	−6.81	−8.33	−0.98	−2.43	−0.16	−0.91	−2.37	−0.10
Ave.	−7.21	−9.00	−10.74	−7.03	−7.53	−9.62	−0.95	−3.06	−0.34	−0.87	−2.98	−0.29

Table 5 BD-rate (%) for different schemes (dataset [33])

Group		EQ-PSNR			SM-PSNR			WS-PSNR			S-PSNR-NN		
		[22]	[26]	proposed	[22]	[26]	proposed	[22]	[26]	proposed	[22]	[26]	proposed
high	P5	−5.30	−4.92	−9.75	−4.08	−4.66	−9.28	−1.40	1.16	−0.22	−1.44	1.16	−0.18
	P26	−5.53	−7.65	−14.64	−6.68	−7.16	−13.97	−0.92	2.13	1.70	−0.86	2.22	1.84
	P27	−4.97	−3.41	−7.46	−5.07	−6.05	**−14.33**	−1.07	2.40	2.84	−0.98	2.68	2.86
	P73	−4.30	−4.38	−11.37	−4.26	−4.90	−11.63	−0.81	2.36	1.48	−0.77	2.52	1.67
mid.	P25	−6.34	−4.84	−13.94	−6.60	−4.46	−11.25	−1.01	0.79	−0.65	−0.89	0.82	−0.60
	P32	−5.14	−4.26	−11.57	−3.36	0.98	−2.86	−1.19	1.34	−0.18	−1.25	1.31	−0.15
	P83	−7.06	−4.26	−4.80	−7.58	−5.28	−6.48	−2.80	1.22	0.78	−3.07	1.33	0.80
	P93	−7.62	−4.21	−11.95	−7.48	−3.61	−11.24	−1.16	2.29	0.59	−1.19	2.30	0.42
low	P4	−6.58	−9.87	−14.16	−5.13	−2.41	−3.67	−1.52	0.87	1.87	−1.62	1.00	2.04
	P7	−4.13	−3.12	−8.52	−4.42	−2.59	−8.33	−0.93	1.04	−0.16	−0.95	1.05	−0.10
	P48	−1.62	2.04	−0.09	−1.88	1.15	−2.08	−0.61	3.05	2.41	−0.60	3.19	2.50
	P57	−7.67	−2.82	−8.83	−7.28	−4.13	−9.69	−2.36	1.33	0.97	−2.36	1.31	0.78
Ave.		−5.52	−4.31	−9.76	−5.32	−3.59	−8.73	−1.32	1.66	0.95	−1.33	1.74	0.99

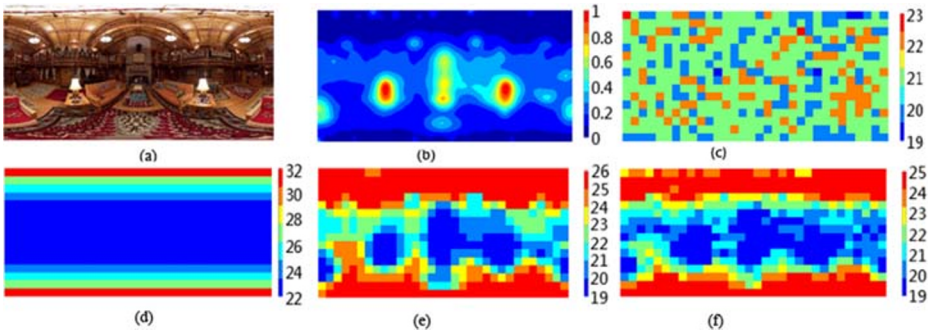

Fig. 10 The QP distribution for Image P25, for an initial QP of 22: (**a**) the 360 degree image, (**b**) predicted saliency map of (**a**), (**c**) the QP map for the anchor, (**d**), (**e**) and (**f**) are the QP maps for [22, 26] and the proposed method

are shown in Fig. 10. The HEVC anchor scheme uses the adaptive QP strategy to give better coding efficiency. The proposed method also uses the adaptive QP, but [22, 26] assign QP to each CTU according to some pre-determined calculations. The ΔQP used is 3.

For [26], although its QP map is also correlated with the saliency value, not many CTUs are encoded with smaller QP. Instead, more CTUs are encoded with the maximum QP. There is no RDO optimization involved in [26] and the coding performance is degraded if the compromise between the distortion and rate during the RDO process is not considered. For [22], the QP is only related to the weight for WS-PSNR and is independent of the image content. The R-D curve for Image P25 is shown in Fig. 11. It is

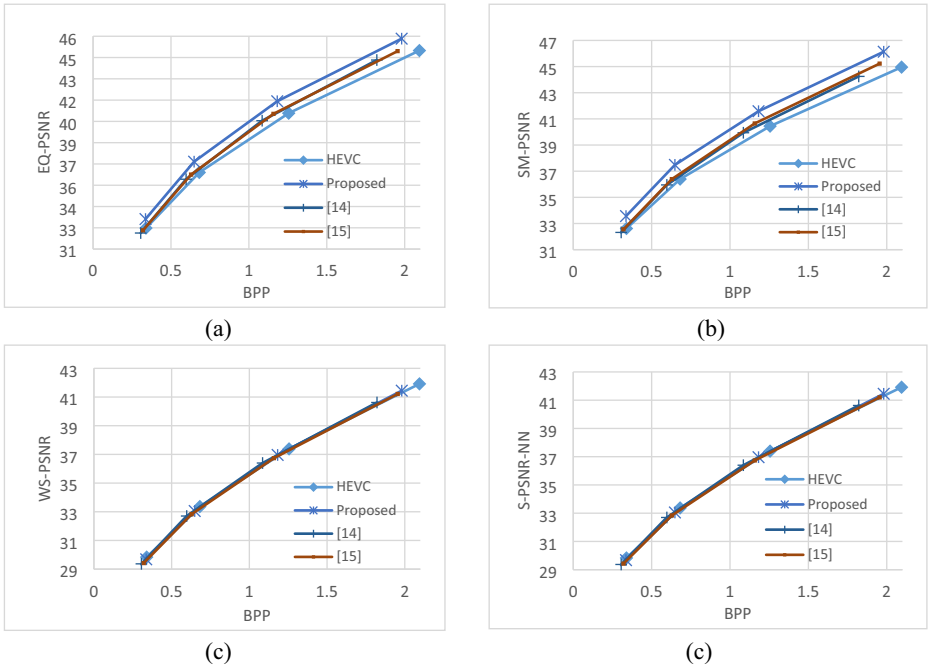

Fig. 11 R-D performance for Image P25, **a** EQ-PSNR, **b** SM-PSNR, **c** WS-PSNR, and **d** S-PSNR-NN

seen that the proposed scheme has better performance in terms of EQ-PSNR and SM-PSNR.

6 Conclusion

This study proposes a saliency-based coding scheme using two techniques: a saliency prediction and a saliency-driven RDO. For the saliency prediction, the saliency map of a 360-degree image is predicted using saliency predictors for a conventional 2D image. ERP-based and MV-based saliency predictions are realized. An optimization in the sphere domain is employed to improve the saliency. The experimental results show that the proposed technique accurately predicts the saliency and particularly it has good performance in terms of the NSS score. For the three other metrics, the proposed technique gives results that are comparable to the best experimental results. For the saliency-driven RDO, a saliency map is a reference and the distortion term is modified during the ROD process to give a better visual experience in the region of interest. Compared to the HEVC anchor, the experimental results show that the proposed technique gives a maximum of 14.33% reduction in the overall bitrate when the image quality in the region with high visual attention is considered. For the WS-PSNR and S-PSNR-NN metrics, the performance is comparable to that for the anchor scheme. In particular, the S-PSNR-NN result shows that the strategy of allocating more resources to the regions that attract the most visual attention does not significantly reduce the quality of the whole image. A comparison with results for other studies shows that the proposed scheme gives much better results for the viewports that contain visual targets. These results confirm the effectiveness of the proposed scheme.

References

1. Abreu AD, Ozcinar C, Smolic A (2017) Look around you: Saliency maps for omnidirectional images in VR applications, *Proc. of International Conference on Quality of Multimedia Experience* (QoMEX), pp. 1–6
2. Bauman B, Seeling P (2019) Spherical image QoE approximation for vision augmentation scenarios. Multimed Tools Appl 78(13):18113–18135
3. Bjontegaard G (2001) Calculation of average PSNR differences between RD curves. VCEG Meeting, Austin
4. Bylinskii Z, Judd T, Oliva A, Torralba A, Durand F (2016) What do different evaluation metrics tell us about saliency models? *arXiv: 1604.03605*
5. Bylinskii Z, Judd T, Borji A, Itti L, Durand F, Oliva A, Torralba A MIT saliency benchmark
6. Carlson C How I made wine glasses from sunflowers," http://blog.wolfram.com/2011/07/28/how-i-made-wine-glasses-from-sunflowers/
7. Chao Y, Zhang L, Hamidouche W, Deforges O (2018) SAlGAN360: visual saliency prediction on 360 degree images with generative adversarial networks, *Proc. of IEEE International Conference on Multimedia and Expo* (ICME)
8. Corbillon X, Simon G, Devlic A, Chakareski J (2017) Viewport-adaptive navigable 360-degree video delivery, *Proc. of IEEE International Conference on Communications* (ICC)
9. Cornia M, Baraldi L, Serra G, Cucchiara R (2018) Predicting human eye fixations via an LSTM-based saliency attentive model. IEEE Transaction on Image Processing 27(10):5142–5154
10. Dedhia J-C Chiang, Char Y-F (2019) Saliency prediction for omnidirectional image considering optimization on sphere domain, *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing* (ICASSP)
11. den Broeck MV, Kawsar F, Schöning J (2017) It's all around you: exploring 360° video viewing experiences on mobile devices, *Proc. of the 25th ACM international conference on Multimedia*

12. Graf M, Timmerer C, Mueller C (2017) Towards bandwidth efficient adaptive streaming of omnidirectional video over HTTP, *Proc. of the 8th ACM on Multimedia Systems Conference*
13. Gutiérrez J, David E, Rai Y, Le Callet P (2018) Toolbox and dataset for the development of saliency and scanpath models for omnidirectional/360 still image, Signal Processing: Image Communication
14. Harel J, Koch C, Perona P (2006) Graph-based visual saliency, *Proc. of Neural Information Processing Systems* (NIPS)
15. Hendry, M. Coban, G. V. Der Auwera, and M. Karczewicz (2017) AHG8: Adaptive QP for 360° video ERP projection, JVET-F0049
16. https://www.itu.int/en/ITU-T/studygroups/2017-2020/16/Pages/video/jvet.aspx
17. Itti L, Koch C, Niebur E (1998) A model of saliency based visual attention for rapid scene analysis. IEEE Transactions on Pattern Analysis and Machine 20(10):99.1254–99.1259
18. K. Kammachi Sreedhar, A. Aminlou, Miska M. Hannuksela, and Moncef Gabbouj (2016) Viewport-adaptive encoding and streaming of 360-degree video for VR application, *Proc. of IEEE International Symposium on Multimedia* (ISM)
19. Kammachi-Sreedhar K, Hannuksela MM (2017) Nested polygonal chain mapping of omnidirectional video, *Proc. of IEEE International Conference on Image Processing* (ICIP), pp. 2169–2173
20. Lebreton P, Raake A (2018) GBVS360, BMS360, ProSal: extending existing saliency prediction models from 2D to omnidirectional images, Signal Processing: Image Communication
21. J. Li, Z Wen, S. Li, Y. Zhao, B. Guo, and J. Wen (2016) Novel tile segmentation scheme for omnidirectional video, *Proc. of IEEE International Conference on Image Processing* (ICIP), pp. 370–374
22. Li Y, Xu J, Chen Z (2017) Spherical domain rate-distortion optimization for 360-degree video coding, *Proc. of IEEE International Conference on Multimedia and Expo* (ICME)
23. Li L, Li Z, Ma X, Yang H, Li H (2019) Advanced spherical motion model and local padding for 360-degree video compression. IEEE Trans Image Process. https://doi.org/10.1109/TIP.2018.2885482
24. Li C, Xu M, Zhang S, Le Callet P (2020) State-of-the-art in 360º video/image processing: perception, assessment and compression. IEEE Journal of Selected Topics in Signal Processing 14(1):5–26
25. Ling J, Zhang K, Zhang Y, Yang D, Chen Z (2018) A saliency prediction model on 360 degree images using color dictionary based sparse representation, Signal Processing: Image Communication
26. Luz G, Ascenso J, Brites C, Pereira F (2017) Saliency-driven omnidirectional imaging adaptive coding: modeling and assessment, *Proc. of IEEE International Workshop on Multimedia Signal Processing* (MMSP)
27. Min D, Choi S, Lu J, Ham B, Sohn K, Do MN (Dec. 2014) Fast global image smoothing based on weighted least squares. IEEE Trans Image Process 23(12):5638–5653
28. Nguyen TC, Yun J-H (2018) Predictive tile selection for 360-degree VR video streaming in bandwidth-limited networks, IEEE Communications Letters
29. Ozcinar C, Cabreray J, Smolic A (2018) Omnidirectional video streaming using visual attention-driven dynamic tiling for VR, *Proc. of IEEE Visual Communications and Image Processing* (VCIP)
30. Pan J, Canton C, McGuinness K, O'Connor NE, Torres J, Sayrol E, Giro-i-Nieto X, SalGAN: Visual saliency prediction with generative adversarial networks, arxiv.org/abs/1701.01081
31. Racapé F, Galpin F, Rath G, François E (2017) AHG8: adaptive QP for 360 video coding, JVET-F0038
32. Rai Y, Gutiérrez J, Le Callet P (2017) A dataset of head and eye movements for 360 degree images, *Proc. of the 8th ACM on Multimedia Systems Conference* (MMSys'17), pp. 205–210
33. Rai Y, Le Callet P, Guillotel P (2017) Which saliency weighting for omnidirectional image quality assessment? *Proc. of the IEEE Ninth International Conference on Quality of Multimedia Experience* (QoMEX'17), pp. 1–6
34. Sitzmann V, Serrano A, Pavel A, Agrawala M, Gutierrez D, Masia B, Wetzstein G (2018) Saliency in VR: How do people explore virtual environments? *IEEE Transactions on Visualization and Computer Graphics*, Vol. 24, No. 4
35. Skupin R, Wang Y-K, Hannuksela MM, Boyce J, Wien M (2017) Standardization status of 360 degree video coding and delivery, *Proc. of IEEE Visual Communications and Image Processing* (VCIP)
36. Startsev M, Dorr M (2018) 360-aware saliency estimation with conventional image saliency predictors, Signal Processing: Image Communication
37. Sullivan GJ, Wiegand T (Nov. 1998) Rate-distortion optimization for video compression. *IEEE Signal Process Mag* 15(6):74–90
38. Sun, A. Lu, and L. Yu. "Weighted-to-spherically-uniform quality evaluation for omnidirectional video," IEEE Signal Processing Letters Vol. 24, No. 9, pp. 1408–1412, 2017.
39. Sun Y, Yu L (2017) AHG8: Stretching ratio based adaptive quantization for 360 video, JVET-F0072
40. Sun Y, Yu L (2017) Coding optimization based on weighted-to-spherically-uniform quality metric for 360 video, *Proc. of the IEEE International Conference on Visual Communications and Image Processing* (VCIP)

41. Tang M, Zhang Y, Wen J, Yang S (2017) Optimized video coding for omnidirectional videos, *Proc. of IEEE International Conference on Multimedia and Expo* (ICME)
42. University of Nantes (2017) Salient360!: visual attention modeling for 360 images grand challenge, *Proc. of IEEE International Conference on Multimedia and Expo* (ICME)
43. Visgwanath B, Nanjundaswamy T, Rose K (2017) Rotational motion model for temporal prediction in 360 video coding, *Proc. of IEEE International Workshop on Multimedia Signal Processing* (MMSP)
44. Wang Y, Peng G, Zhou M (2018) Saliency detection by hierarchically integrating compactness, contrast and boundary connectivity. Multimed Tools Appl 77(10):11883–11901
45. Y. Ye and J. Boyce (2018) Algorithm descriptions of projection format conversion and video quality metrics in 360Lib version 7, JVET-K1004
46. Youvalari G, Aminlou A, Hannuksela MM (2016) Analysis of regional down-sampling methods for coding of omnidirectional video, *Proc. of Picture Coding Symposium* (PCS)
47. Yu M, Lakshman H, Girod B (2015) Content adaptive representation of omnidirectional videos for cinematic virtual reality, *Proc. of International Workshop on Immersive Media Experience* ACM, pp. 1–6
48. Yu M, Lakshman H, Girod B (2015) A framework to evaluate omnidirectional video coding schemes, *Proc. of IEEE International Symposium on Mixed and Augmented Reality*, pp. 31–36
49. Zakharchenko V, Alshina E, Singh A, Dsouza A (2016) AhG8: Suggested testing procedure for 360-degree video, JVET-D0027
50. Zhang J, Sclaroff S (2013) Saliency detection: a boolean map approach, *Proc. of IEEE International Conference on Computer Vision,* pp. 153–160
51. Zhang M, Zhang J, Liu Z, An C (2019) An efficient coding algorithm for 360-degree video based on improved adaptive QP compensation and early CU partition termination. Multimed Tools Appl 78(1):1081–1101
52. Zhu Y, Zhai G, Min X (2018) The prediction of head and eye movement for 360 degree images, Signal Processing*:* Image Communication