Saliency-driven Omnidirectional Imaging Adaptive Coding: Modeling and Assessment

Guilherme Luz, João Ascenso, Catarina Brites and Fernando Pereira
Instituto Superior Técnico, Universidade de Lisboa - Instituto de Telecomunicações
Lisboa, Portugal
guilherme.luz@tecnico.ulisboa.pt, joao.ascenso@lx.it.pt, catarina.brites@lx.it.pt, fp@lx.it.pt.

Abstract— **Omnidirectional imaging, also known as 360^o and spherical imaging, records all 360^o of a scene from a specific spatial position, thus offering the user the capability to enjoy three rotational degrees of freedom (3-DoF). To offer a good quality of experience, omnidirectional imaging requires very high bitrates as high spatial resolution are a must and, ideally, also high frame rates. Due to the lack of video coding solutions specifically designed for omnidirectional imaging, this type of content is typically coded with the available image and video coding standards, such as JPEG, H.264/AVC and HEVC, after applying a 2D rectangular projection. In this context, this paper proposes an omnidirectional imaging coding solution allowing to reach improved coding performance by using an adaptive coding solution where the most visually salient image/video regions are coded with higher quality in a process appropriately controlled by the quantization parameter. To determine the saliency of the various omnidirectional imaging regions, a machine-learning based saliency detection model is proposed. The proposed coding solution achieves compression gains as measured by a novel objective quality metric also driven by saliency. This novel objective quality metric is validated by formal subjective testing where very high correlations with the subjective tests scores are achieved.**

Keywords— omnidirectional video; saliency detection; HEVC; adaptive coding; objective quality metric; subjective evaluation.

I. INTRODUCTION

In recent years, omnidirectional imaging popularity has been dramatically increasing, motivated by the rising processing capacity of computers and mobile devices and the emergence of high-density displays. Currently, even a common user has easy access to an omnidirectional camera and can produce omnidirectional images and video to broadcast it through the Internet to be visualized in Head-Mounted Displays (HMDs) such as Google Cardboard and Oculus Rift. The recent hype around Virtual Reality (VR), which basically uses omnidirectional images and videos, has also significantly contributed for the raising relevance of this type of visual data. After the failure of 3D TV, meaning stereo TV, omnidirectional imaging offers the viewers more freedom and control over the viewing experience; notably, it offers the possibility to freely exploit the 3D space around a specific spatial position, although limited to the three head rotations, this means 3-DoF. The remaining three DoF, this means the translations, are not yet supported but the gains in immersiveness regarding 3D TV are spectacular. Considering the large scene area covered by omnidirectional imaging, it needs a very high resolution and consequently demands for high bitrates. Due to the sudden growing popularity of omnidirectional imaging, there are no standard coding solutions specifically designed for omnidirectional/spherical video. For this reason, the currently available standard image and video codecs, e.g. JPEG, H.264/Advanced Video Coding (AVC) [1] and High Efficiency Video Coding (HEVC) [2], have been used for omnidirectional imaging coding, naturally after the spherical content is converted into 2D flat rectangular content by using some projection. As a consequence, the encoder input data is a projection-distorted representation of the omnidirectional content, which does not exploit specific omnidirectional imaging characteristics. While the equirectangular projection (ERP) is the most used projection, it is well known that different projections impact the coding performance in different ways and this led to the design and assessment of other projections. For example, ERP stretches the content in the horizontal direction near the poles, thus enlarging the number of pixels to code in regions less important in terms of user impact. The Lambert cylindrical equal area projection has also been studied, where besides the ERP stretching, the content is vertically shrunk to compensate the increasing area [3]. Other projections use geometric solid surfaces where the spherical image/video is virtually placed inside a geometric solid that is then stretched to the solid surfaces and finally, the several surfaces are organized in a rectangular layout. The most studied projection of this type is the cube map projection, but other solid projections have been designed such as the Pyramid [4], Octahedron [5] and Rhombic Dodecahedron projections [6]. In [7], Youvalary *et al.* propose a coding scheme where the Pseudo-Cylindrical projection is used with some specific proposed coding tools to handle the projection characteristics. In [8], Yu *et al.* propose a content adaptive tiling scheme where the video is divided into several vertical tiles according to the latitude. In [9], Li *et al.* use ERP based tiling schemes to reduce the resolution of polar regions and compensate the ERP stretching by reducing the resulting representation area. Regarding coding performance assessment for omnidirectional content, Yu *et al.* propose a framework to study different projections using different metrics, notably taking into account the importance of the various image regions depending on their viewing probabilities [3]. In [10], Upenik *et al.* developed a testbed for subjective testing of omnidirectional content where several metrics adopted by the Joint Video Experts Team (JVET) [11] (the standardization group currently addressing omnidirectional video coding) are evaluated in terms of correlation with the subjective tests results.

This paper proposes a saliency-driven omnidirectional imaging adaptive coding solution where a saliency detection

978-1-5090-3649-3/17/$31.00 ©2017 IEEE

model drives the quantization process. The idea is to invest more quality on the regions where the viewer tends to fixate his/her attention and vice-versa [12], thus reducing the overall bitrate for a target perceptual quality. The coding performance of the proposed solution is assessed by a novel objective quality metric for omnidirectional imaging, which is validated by formal subjective testing. The remainder of this paper is organized as follows: Sections II and III propose a saliency detection model developed for omnidirectional imaging and the adaptive quantization process, respectively. Section IV presents the rate-distortion (RD) performance using an appropriate assessment methodology and meaningful test conditions and metrics, notably a novel objective quality metric which is validated with formal subjective testing. Finally, Section V presents the conclusions and suggestions for future work.

II. Proposing A Saliency Detection Model For Omnidirectional Images

Considering the large viewing area offered to the viewers in omnidirectional imaging, it is natural that not all parts of it deserve the same attention by the users, e.g. depending on the specific content, position on the sphere, etc. This justifies the adoption of a saliency detection model for omnidirectional images, which shall be later exploited to allocate the rate/quality within the image/video by controlling the coding quantization step, and possibly to reach rate gains for the same perceptual quality.

A. *Architecture and Concepts*

Fig. 1 shows the architecture of the proposed Saliency Detection Model (SDM) for omnidirectional images. The input for this architecture is the omnidirectional image dataset provided in the context of the *Salient360! Visual Attention Modelling for 360° Images Grand Challenge* organized at ICME'2017 [13]. This dataset includes a set of 20 omnidirectional images and their ground truth saliency maps, which were experimentally obtained using human fixation maps based on the measured head trajectory. The architecture in Fig. 1 considers three branches:

1. The **ground truth branch** on the left represents the saliency ground truth for each omnidirectional image, in this case expressed by means of the so-called *Viewport integrated Head Direction Map* (VHDM) which has been experimentally collected for the Salient360! Grand Challenge.
2. The **proposed SDM branch** appears at the center and is an image-specific algorithm, which determines the saliency of a specific omnidirectional image expressed by the output *Latitude biased Viewport integrated Saliency Map* (LVSM).
3. Finally, on the right, there is an **alternative SDM branch**, which is not image-specific and only considers the latitude impact, and expresses its output with the so-called *Viewport based Latitude Importance Map* (VLIM).

The algorithms for the several steps in the proposed SDM are described in the following, after presenting the alternative latitude driven SDM, which will also integrate the proposed SDM in a final fusion module.

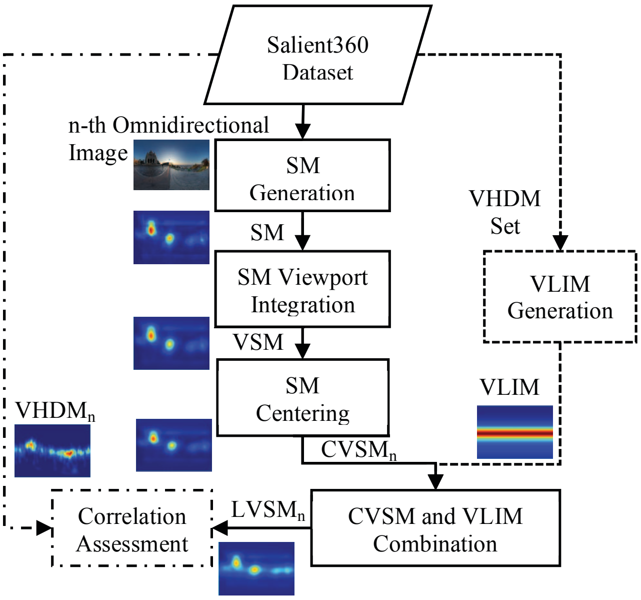


Fig. 1. Architecture of the proposed Saliency Detection Model.

B. *Latitude driven SDM: VLIM Generation*

The set of VHDMs (one per image) in the Salient360! Grand Challenge dataset allows extracting some statistics to globally characterize the importance of each latitude in terms of user attention, independently of the specific omnidirectional image. The VLIM map is computed as the average fixation intensity for each latitude (considering all longitudes) for the full set of the available VHDMs, thus providing a latitude-driven characterization of the saliency within an omnidirectional image.

C. *Proposed SDM: LVSM Generation*

1) *Saliency Map Generation*

The MLNET solution proposed by Cornia *et al.* [14] is one of the top scored SDM for 2D images at the MIT Benchmark [15]. This model combines features extracted at low, medium and high levels of a Convolutional Neural Network (CNN). First, the features are extracted to be used on a network that builds saliency-specific features, where feature weighting functions are learned to generate saliency-specific feature maps, producing a temporary saliency map. Finally, a learned prior is considered to build the final saliency map (SM). The learning consists on a training process of the SDM based on a large set of images and corresponding ground truth saliency maps that adjusts the extracted features and the learned priors, forcing the network to minimize a square error loss function. For the purposes of the omnidirectional imaging saliency detection, the MLNET pre-trained model is used as the first step of the proposed SDM. However, as the MLNET was trained for conventional 2D images, and it was impossible to retrain it due to the lack of a good number of omnidirectional images with saliency ground truth, the saliency detection results from this first step have to be further processed in order the overall SDM is adapted and highly performing for omnidirectional images.

2) *Saliency Map Viewport Integration*

Although the viewers usually look at the center of the viewport, this means the area projected in the 2D display, the eye and head directions are not always the same as the users move the eyes around, even when the head is fixed. Thus, the SM

Viewport Integration module targets to consider this effect: as the eyes move within the viewport, the saliency of a point is impacted by the saliency of its neighbors.

First, each pixel position in the ERP saliency map is converted to spherical coordinates with longitude and latitude. Then, this position is assumed as the head direction or viewport center and the full viewport area is computed with its associated saliency scores. After, the weighted average of all viewport saliency scores is computed with weights defined by a foveation function, in this case a 2D Gaussian distribution with a standard deviation $\sigma = 4^{\mathrm{o}}$ regarding the center of the viewport; this viewport weighted average is then adopted as the saliency viewport integration score for the pixel under consideration. Finally, this process is repeated for every ERP pixel position, thus obtaining the full Viewport integrated Saliency Map (VSM).

3) Saliency Map Centering

The SM Centering module intends to take into account that the viewers do not move the head completely to fixate areas close to the poles, instead the viewers complete the fixation by moving the eyes. For example, when the viewer looks at the north pole, typically the head is turned to a lower longitude relative to the top pole while the eye direction is at the upper part of the viewport. In practice, the detected salient regions should become closer to the Equator to consider this effect. Thus, the designed SDM includes this processing module that moves the previous VSM rows to a latitude closer to the Equator, thus creating a shrinking effect. To avoid creating empty areas near the poles, some appropriate form of padding is used to obtain the Centered Viewport based Saliency Map (CVSM).

4) CVSM and VLIM Combination

As the regions near the Equator are statistically speaking the most viewed regions in an omnidirectional image, it is essential to also introduce this bias in the proposed SDM which still does not consider it. Thus, the objective of this combination module is to introduce a latitude bias in the saliency detection process. As already explained, the VLIM expresses precisely the statistical latitude bias (experimentally computed), and thus it is proposed here to combine the computed CVSM with the experimentally obtained VLIM, more precisely by a weighted averaging the two maps according to:

$$LVSM = w \times CVSM + (1 - w) \times VLIM. \qquad (1)$$

A value $w = 0.8$ has experimentally found as a good trade-off. This combination also plays the role of a safety measure when the CVSM is less accurate, as it increases the importance of the regions near the Equator, an effect with strong experimental validation. To simplify the notation, the LVSM will be simply represented by S in the following.

D. SDM Performance Assessment

The proposed SDM and thus the LVSM may be assessed by comparing it with VHDM as the computed saliency map should replicate the experimentally obtained saliency map. Moreover, the latitude driven VLIM may be also compared with the ground truth VHDM. In principle, the LVSM should achieve better performance than the VLIM, since the VLIM is not computed for any specific image in particular.

To compare the proposed LVSM and the VLIM with VHDM, the Root Mean Square Error (RMSE) metric is used to assess the difference between each pair of maps. Table I presents the average RMSE considering the entire set of 20 Salient360! Grand Challenge images. As expected, the results in Table I show that the average RMSE for VLIM is higher than for LVSM, thus hinting to a good saliency detection performance for the proposed SDM.

TABLE I. AVERAGE RMSE FOR VLIM, VSM, CVSM AND LVSM VERSUS VHDM.

	VLIM	VSM	CVSM	LVSM
Average RMSE	0.3298	0.1638	0.1538	**0.1348**

III. SALIENCY-BASED ADAPTIVE QUANTIZATION MODEL

The key parameter to control the rate and the associated quality in HEVC coding is the quantization parameter (QP) as it defines how much distortion is inserted in the coding process. Naturally, using a QP which is adaptive to the image/video characteristics may allow achieving better coding performance at the price of some additional encoder complexity. This section proposes an adaptive quantization model for the HEVC coding standard that is able to consider the image characteristics to better allocate the rate (through the QP variation) and perceptual quality within an omnidirectional image/video to finally reach better compression efficiency.

As the HEVC standard allows defining the QP at the Coding Unit (CU) level, it is necessary to design an appropriate model to select the QP depending on the saliency map, e.g. depending on the average saliency for each CU and, naturally, to change the reference software encoder to compute the QP based on this model. In [16], a QP selection model based on saliency maps is proposed for the H.264/AVC standard where the QP may vary at the macroblock level. The adaptive QP selection model here proposed for HEVC coding is inspired on [16] and also extended to consider the spatial activity. The model proposed applies to the Largest Coding Unit (LCU) as the saliency map does not usually have drastic intensity variations within each LCU and, thus, the involved complexity is not unnecessarily increased. As the QP selection model targets to reduce the QP in the more salient areas and vice-versa, it is important that there are no significant QP increases in areas where the viewers are very sensitive to quantization noise, such as uniform and slowly varying regions. For this reason, the proposed adaptive QP selection model considers not only the saliency, but also the spatial activity of each LCU. In summary, the QP for the i-th LCU is given by:

$$QP_i = \mathrm{round}\left(\frac{QP_{slice}}{\sqrt{w_i}}\right) \qquad (2)$$

where QP_{slice} is the reference/default QP defined for the current slice (which defines the target quality). The weight w_i depends on both the saliency and spatial activity as it is defined by:

$$w_i = \begin{cases} a + \dfrac{b}{1 + \exp\left(-c \times (S(X_i)/n - \bar{S})/\bar{S}\right.} & , if \;\; l \leq 10 \\ a + \dfrac{b}{1 + \exp\left(-c \times (S(X_i) - \bar{S})/\bar{S}\right.} & , if \;\; l > 10 \end{cases} \qquad (3)$$

where $a = 0.7$, $b = 0.6$ and $c = 4$ (the same values as in [16]), $\bar{S}$ is the average saliency for the whole set of LCU within the current frame, and $S(X_i)$ is the average saliency for the i-th LCU. The normalized spatial activity n is computed as:

$$n = \frac{h \times l + t}{l + h \times t} \tag{4}$$

where h is a scaling factor associated to the Uniform Reconstruction Quantization (URQ) QP adaptation range, l corresponds to the spatial activity in a luma CB, and t refers to the mean spatial activity for all 2N×2N CUs. The variable h is computed as:

$$h = 2^{r/6} \tag{5}$$

where $r = 6$ is the default value in the HEVC reference software, regardless of the YCbCr color channel, and l is given by:

$$l = 1 + \min(\sigma_{Y,k}^2), \text{ where } k = 1, \ldots, 4 \tag{6}$$

where $\sigma_{Y,k}^2$ denotes the spatial activity of the pixel values in sub-block k (of size N×N) in a luminance CB (2N×2N). The spatial activity assumes a great importance as it allows to identify regions which may be more sensitive to larger quantization noise; these regions should not have too increased QP values even if the saliency is low. For this reason, for low spatial activity l, the normalized spatial activity is considered to compute w_i and reduce the QP in these regions.

IV. Performance Assessment

This section intends to assess the tools proposed in the previous sections. First, the coding performance will be assessed using objective quality metrics, notably a novel objective quality metric for omnidirectional imaging. Next, the correlation with the human quality perception of the proposed metric is assessed by formal subjective tests.

A. Objective Quality Evaluation

1) Test Conditions

The JVET Group, which is currently addressing omnidirectional video coding, has defined a set of assessment methodologies to conduct omnidirectional video coding experiments [11]. These methodologies are implemented in the 360Lib reference software, which is an extension of the HEVC HM16.15 reference software. The JVET common test conditions and software reference configuration document [17] define the basic QP values to be used in this paper, notably 22, 27, 32 and 37. The test material is composed by the same 20 omnidirectional images from the Salient360! Grand Challenge dataset [13] with a spatial resolution of 8192×4048 pixels.

2) Objective Quality Metrics

The currently most used objective metrics are those adopted by the JVET Group.

a) JVET Omnidirectional Objective Quality Metrics

Most of the objective quality metrics adopted by JVET aim to consider the distortion due to the projections from the spherical domain. The most relevant metrics are:

- **WS-PSNR** which compares the input (after the projection) and output of the codec, while considering the different importance of the various regions according to the used projection distortion relative to the spherical domain.
- Viewport based PSNR (**V-PSNR**) which compares the rendered viewports using the original material and the corresponding decoded image/video. In this case, two viewports are studied: viewport 0 at longitude $\theta = 0$° and latitude $\varphi = 0$° (at the Equator) and viewport 1 at $\theta = 0$° and $\varphi = 90$° (at north pole).

b) Proposing a Saliency biased PSNR Quality Metric

Although the previous metrics are specifically designed for omnidirectional video, the fact that different regions of the omnidirectional image may have a different viewer impact is not taken into account. In this context, a new metric is proposed here, motivated by the fact that the viewers tend to concentrate their attention in the most salient regions of the omnidirectional image. The new saliency (SAL) biased metric is called SAL-PSNR and has the purpose of evaluating the image/video quality while taking into account the importance/saliency of each region in the omnidirectional image. In this new metric, each pixel in the omnidirectional image has a weight defined as the product of its saliency score by a distortion factor associated to the used projection, thus resulting in a weight $q(i,j)$ defined as:

$$q(i,j) = \frac{W(i,j) \times S(i,j)}{\sum_{i=0}^{m-1} \sum_{j=0}^{p-1} [W(i,j) \times S(i,j)]} \tag{7}$$

where $W(j,i)$ is the distortion factor for position (i,j) in the image and $S(i,j)$ is the saliency map score for the same position. The saliency based PSNR (SAL-PSNR) objective quality metric is thus defined as:

$$SAL\text{-}PSNR = 10\, log \left(\frac{{MAX_f}^2}{SAL\text{-}MSE} \right) \tag{8}$$

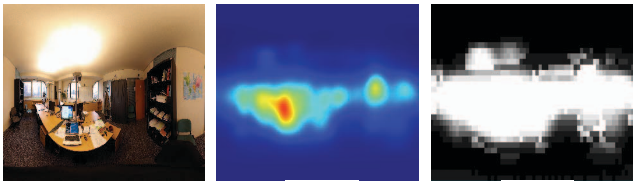
where MAX_f is the maximum signal value, e.g. 255 for 8-bit representations, and

$$SAL\text{-}MSE = \sum_{i=0}^{m-1} \sum_{j=0}^{p-1} \left[\left(f(i,j) - g(i,j) \right)^2 \times q(i,j) \right] \tag{9}$$

where f is the codec input image (after the projection) and g the decoded image at the codec output.

3) RD Performance

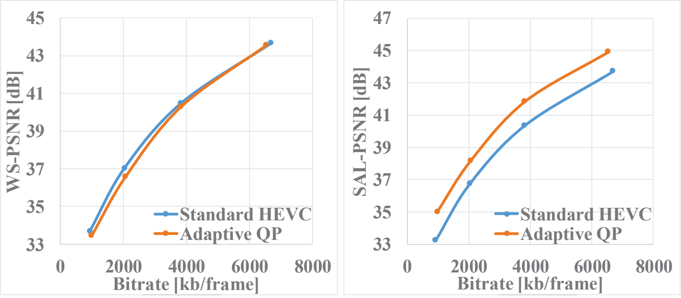
This section presents and analyzes the RD performance results for the set of 20 omnidirectional images following the JVET test conditions previously described. Fig. 2(a) shows one of the ERP images to be coded, Fig. 2(b) the LVSM obtained for that image, and Fig. 2(c) the QP variations around a reference QP equal to 32; in Fig. 2(c), the white regions are associated to QP reductions (thus quality increases), while the darker regions are associated to QP increases relative to the reference QP (thus quality reductions).



(a) (b) (c)

Fig. 2. (a) Image P3; (b) LVSM; (c) QP variation relative to the reference QP.

Fig. 3 shows the RD performance both for the WS-PSNR and SAL-PSNR metrics for image P3 (luminance component, Y, only).



(a) (b)

Fig. 3. RD performance for image P3: (a) Y WS-PSNR; and (b) Y SAL-PSNR.

The results in Fig. 3 allow concluding that, according to WS-PSNR, the standard HEVC coding performs better than the Adaptive QP solution; this is expected as this metric does not take into account the saliency map. However, according to SAL-PSNR, there is a considerable RD gain for the Adaptive QP solution, which is expected as this metric takes into account the saliency map.

Table II presents the average BD-Rate results for the Adaptive QP coding solution regarding HEVC standard coding for several objective quality metrics and considering all the 20 coded images. Negative values indicate there is an RD performance improvement, this means a rate reduction for the same quality, when Adaptive QP coding is used regarding standard HEVC and vice-versa. The results confirm the trend from Fig. 3 as Y WS-PSNR shows a BD-Rate loss (although gains are shown for the chrominances). However, the proposed SAL-PSNR metric presents considerable average BD-Rate gains for the luminance and chrominances. It is worth to mention the considerable BD-Rate gains for V0-PSNR, which are expected due to the typical higher saliency/importance around the Equator. The opposite occurs for V1-PSNR, as this viewport is located at the north pole, which is typically a region with lower saliency, and thus the quality was worsened with the adaptive QP coding.

TABLE II. AVERAGE BD-RATE FOR ADAPTIVE QP CODING REGARDING STANDARD HEVC FOR WS-PSNR, SAL-PSNR, V0-PSNR AND V1-PSNR.

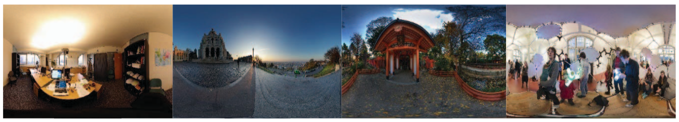
Y WS-PSNR	**U WS-PSNR**	**V WS-PSNR**	**Y SAL-PSNR**	**U SAL-PSNR**	**V SAL-PSNR**
7.82	-8.81	-10.73	-30.33	-35.13	-35.44
Y V0-PSNR	**U V0-PSNR**	**V V0-PSNR**	**Y V1-PSNR**	**U V1-PSNR**	**V V1-PSNR**
-23.67	-33.60	-35.69	107.95	106.51	102.34

B. *SAL-PSNR Subjective Testing Validation*

While the proposed SAL-PSNR objective quality metric shows significant quality gains when using the Adaptive QP tool, this may be considered expectable as this metric gives more importance to the areas where the saliency scores are higher and thus the QP is made lower. Naturally, to support these RD performance gains claims in a solid way, it is critical to validate the proposed SAL-PSNR in terms of subjective assessment, which requires to perform formal subjective tests and assess the correlation performance of SAL-PSNR with subjective scores.

1) *Test Conditions*

Four omnidirectional images (P3, P4, P13 and P22) from the Salient360! Grand Challenge dataset were selected to perform the subjective tests with the purpose of including rather different types of content, see Fig. 4. The set of images includes indoor and outdoor images, images with people and images with different types of textures.



(a) (b) (c) (d)

Fig. 4. Images for subjective testing: (a) P3; (b) P4; (c) P13; (d) P22.

The coding process for the selected images has followed the JVET test methodology described in Section IV.A.1). The standard HEVC and Adaptive QP coding solutions are evaluated, both using the same set of reference QPs, notably 30, 35, 40 and 45. These QPs differ from the QP values adopted for the objective evaluation, notably by eliminating QPs below 30, since for that QP range, the coded images are almost undistinguishable in terms of quality from the reference image.

The subjective evaluation has followed the Absolute Category Rating with Hidden Reference (ACR-HR) methodology using a five-grade quality scale (1-Bad; 2-Poor; 3-Fair; 4-Good; 5-Excellent). In this method, the user scores not only some coded images but also a reference image, in this case not coded, in terms of overall quality. The omnidirectional images were shown to the subjects using the Oculus Rift HMD, which has a resolution of 1080×1200 per eye, while they were sit on a rolling chair and free to rotate in any direction. A total of 26 subjects (21 males and 5 females) were used with ages between 22 and 54 years old. Before the test session, written instructions were provided to the subjects and a training session was conducted to adapt the subject to the assessment procedure and quality levels. More specifically, five training samples were shown, each associated to each possible quality level. During the test session, each image to be scored was shown during 20 seconds and after, the subject was asked to evaluate the overall quality of the omnidirectional image by orally reporting a quality score that was registered by a nearby operator. Each subject has visualized the four images coded with both the standard HEVC and Adaptive QP coding solutions with the selected reference QPs as well as the (hidden) reference images, shown in a random order.

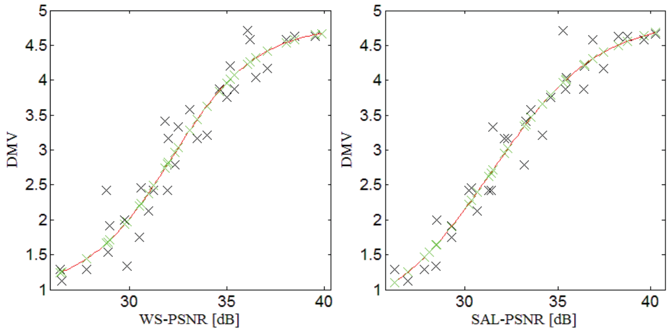
2) *Objective Metric Correlation Performance*

After the subjective tests, outlier detection was performed according to the guidelines described in ITU-R Recommendation BT.500 [18] to exclude subjects whose scores are highly deviating from the others. Two subjects were considered outliers and removed from subsequent results. After, the so-called Differential Viewer Score (DV) is computed for each collected score (V) as the difference between the score for each coded image (CI) and the score for its corresponding hidden reference (HR) given by subject i according to Recommendation ITU-T P.910 [19]. DV is given by

$$DV_i(CI) = V_i(CI) + V_i(HR) + 5 \qquad (10)$$

Next, the so-called Differential Mean Viewer Score (DMV) is computed per coded image as the average of the DV scores of all the subjects.

The next data processing step is to find the relationship between the experimental DMV scores and the objective quality metric scores. As the DMV scores present a non-linear distribution, a logistic function fitting of the DMV scores to the objective metric values is performed for each objective metric according to [18], considering both coding solutions for all the images. The logistic function has the intent to represent the predicted DMV scores as a function of the objective quality metric scores, see Fig. 5. This allows comparing the experimental DMV scores represented by the black crosses, with the predicted DMV scores values determined using the objective metric results represented by the green crosses.



(a) (b)

Fig. 5. Fitting the experimental DMV to the objective metric score using a logistic function: (a) WS-PSNR; and (b) SAL-PSNR.

Finally, the similarity between the experimental and predicted DMV scores is evaluated using three performance indexes commonly used for this purpose: the Pearson Linear Correlation Coefficient (PLCC), which evaluates if there is a linear correlation; the Spearman Rank Order Correlation (SROC), which evaluates the monotonicity; and the RMSE, which evaluates the accuracy by measuring the error between the two sets of values.

The results in Table III show that the assessed objective quality metrics are highly correlated with the experimental DMV scores. However, SAL-PSNR presents the highest correlation coefficients and the lowest RMSE value, which means that SAL-PSNR is the objective quality metric that is able to more reliably assess the subjective quality of omnidirectional images. On the other hand, V1-PSNR presents a rather low correlation performance since this metric only evaluates the quality of the viewport located at the upper polar region. As the upper polar region is a region that the user visualizes less often, it is just natural that V1-PSNR is not a good objective metric, which is confirmed by the correlation performance results.

TABLE III. PLCC, SROC AND RMSE FOR THE OBJECTIVE QUALITY METRICS.

	PLCC	SROC	RMSE
WS-PSNR	0.962	0.953	0.315
SAL-PSNR	**0.972**	**0.963**	**0.273**
PSNR	0.957	0.949	0.334
V0-PSNR	0.901	0.893	0.501
V1-PSNR	0.466	0.373	1.021

V. FINAL REMARKS

This paper proposes a saliency-driven adaptive coding solution for omnidirectional imaging using an adaptive QP selection strategy where the most salient regions are coded with higher quality and vice-versa. This solution achieved considerable BD-Rate gains as assessed by a novel, proposed objective quality metric biased by saliency. To confirm this conclusion, subjective tests were performed to validate the new metric and it was shown that it achieves higher correlation with the subjective tests results than any other currently used objective quality metric for omnidirectional imaging.

REFERENCES

[1] T. Wiegand *et al.*, "Overview of the H.264/AVC Video Coding Standard," *IEEE TCSVT*, vol. 13, no. 7, pp. 560-576, July 2003.

[2] G. J. Sullivan *et al.*, "Overview of the High Efficiency Video Coding (HEVC) Standard," *IEEE TCSVT*, vol. 22, no. 12, pp. 1649-1668, December 2012.

[3] M. Yu, H. Lakshman and B. Girod, "A Framework to Evaluate Omnidirectional Video Coding Schemes," in *IEEE Int. Symposium on Mixed and Augmented Reality*, Fukuoka, Japan, October 2015.

[4] G. Van der Auwera, M. Coban and M. Karczewicz, "AHG8: Truncated Square Pyramid Projection (TSP) For 360 Video," JVET of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, JVET-0071, Chengdu, China, October 2016.

[5] T. Engelhardt and C. Dachbacher, "Octahedron Environment Maps," *Proceedings of Vision Modelling and Visualization,* pp. 383-388, 2008.

[6] C.-W. Fu *et al.*, "The Rhombic Dodecahedron Map: An Efficient Scheme for Encoding Panoramic Video," *IEEE TMM,* vol. 11, no. 4, pp. 634-644, April 2009.

[7] R. G. Youvalari *et al.*, "Efficient Coding of 360-Degree Pseudo-Cylindrical Panoramic Video for Virtual Reality Applications," in *IEEE Int. Symposium on Multimedia*, San Jose, CA, USA, December 2016.

[8] M. Yu, H. Lakshman and B. Girod, "Content Adaptive Representations of Omnidirectional Videos for Cinematic Virtual Reality," in *ACM Multimedia Workshop on Immersive Media Experiences*, Brisbane, Australia, November 2015.

[9] J. Li *et al.*, "Novel Tile Segmentation Scheme for Omnidirectional Video," in *IEEE Int. Conf. on Image Processing*, Phoenix, AZ, USA, September, 2016.

[10] E. Upenik, M. Rerabek and T. Ebrahimi, "A Testbed for Subjective Evaluation of Omnidirectional Visual Content," in *32nd Picture Coding Symposium*, Nuremberg, Germany, December 2016.

[11] J. Boyce *et al.*, "JVET Common Test Conditions and Evaluation Procedures for 360º Video," JVET of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, JVET-E1030, Geneva, Switzerland, January 2017.

[12] A. Borji and L. Itti, "State-of-the-Art in Visual Attention Modeling," *IEEE TPAMI*, vol. 35, no. 1, pp. 185-207, January 2013.

[13] Y. Rai, P. Le Callet and P. Guillotel, "Salient360 - The Training Dataset for the ICME Grand Challenge," in *IEEE Int. Conf. on Multimedia & Expo*, Hong Kong, July 2017.

[14] M. Cornia *et al.*, "A Deep Multi-Level Network for Saliency Prediction," *arXiv:1609.01064 [cs.CV],* September 2016.

[15] "MIT Saliency Benchmark," October 2016. [Online]. Available: http://saliency.mit.edu/. [Accessed December 21, 2016].

[16] H. Hadizadeh and I. V. Bajic, "Saliency Aware VIdeo Compression," *IEEE TIP*, vol. 23, no. 1, pp. 19-33, January 2014.

[17] K. Suehring and X. Li, "JVET Common Test Conditions and Software Reference Configurations," JVET of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, JVET-1010, San Diego, CA, USA, February 2016.

[18] ITU-R BT.500, "Methodology for the Subjective Assessment of Quality of Television Pictures," ITU, January 2012.

[19] ITU-T P.910, "Subjective Video Quality Assessment Methods for Multimedia Applications," ITU, April 2008.