Omnidirectional Video: Adaptive Coding based on Saliency

Guilherme Luz
Instituto Superior Técnico, Av. Rovisco Pais, 1049-001 Lisboa, Portugal
guilherme.luz@tecnico.ulisboa.pt

***Abstract*—Omnidirectional video requires very high bitrates due to its required high spatial resolution and, ideally, high frame rates to provide an immersive experience. However, the current video coding solutions were not designed for omnidirectional video and thus omnidirectional or spherical images and video are with the available standard codecs, such as H.264/AVC and HEVC, after applying a 2D rectangular projection. Therefore, these codecs should be adapted to become more efficient for this type of images and video. Motivated by this situation, this work aims to design, implement and assess an omnidirectional image and video coding solution with improved compression performance by using an adaptive coding solution where the more important omnidirectional video regions are coded with higher quality while less important regions are coded with lower quality with this process controlled by the appropriate control of the quantization parameter. To determine the importance of the various regions, a machine-learning, adapted omnidirectional image saliency detection model is proposed which is able to identify the regions standing out from their neighbors. The proposed solution achieves considerable overall compression rate gains with improved quality for the most important regions of the image when an appropriate quality metric is used for the assessment. This objective quality metric is validated by formal subjective assessments where very high correlations with the subjective tests scores are achieved.**

***Index Terms*— omnidirectional video; HEVC; saliency detection; video compression; adaptive coding.**

I. Introduction

In the recent years, omnidirectional video popularity has been dramatically increasing, motivated by the rising processing capacity of computers and mobile devices, 3D graphics capabilities and also the emergence of high-density displays. Currently, even a common user, with access to an omnidirectional camera, can produce omnidirectional images and video and broadcast this type of content through the Internet to be visualized in Head-Mounted Displays such as Google Cardboard and Oculus Rift.

Omnidirectional video has high resolution and consequently also demanding bitrate requirements. Moreover, due to the sudden growing popularity of omnidirectional video, there are no efficient coding tools specifically designed for omnidirectional video. For this reason, the currently available standard video codecs, e.g. H.264/Advanced Video Coding (AVC) [1] and High Efficiency Video Coding (HEVC) [2], have been used for omnidirectional video coding but this requires that the omnidirectional video is converted to a 2D rectangular projection before coding. As a consequence, the encoder input is a distorted representation of the omnidirectional video, which is not coding-friendly and at the same does not exploit specific omnidirectional video characteristics.

The proposed solution is a content adaptive codec that, by considering the output of a saliency detection model that intends to identify the regions where the viewer tends to fixate his/her attention [3], is able to reduce the bitrate for a target quality or increase the quality for a target rate by adaptively allocating the rate/quality within the image, notably penalizing the regions that the users are less likely to watch.

The remainder of this paper is organized as follows: Section II presents the omnidirectional video reference architecture with a brief description of the different modules. Section III presents a saliency detection model developed for omnidirectional video, including the overall architecture, description of its processing modules. Section IV describes the proposed omnidirectional video coding. Section V presents its performance using an appropriate assessment methodology and meaningful test conditions and metrics, confirmed by subjective evaluation results. Section VI presents the final conclusions and suggest future work directions.

II. Omnidirectional Video: Basics and Main Processing

Omnidirectional video, also known 360º video, is a video format where visual information is acquired in every scene direction at the same time. As it covers all the directions around a specific point, the user position, it is possible to have an experience where the viewer is able to navigate through any direction. This section present and describes the basic omnidirectional video system architecture.

A. Basic Omnidirectional Video Coding Processing Architecture

The adopted basic architecture to process omnidirectional video from acquisition to visualization, see Fig. 1, divided in a sequence of processing modules.

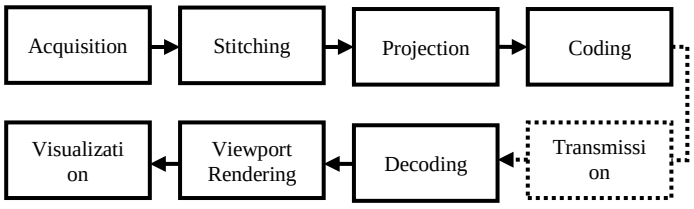


Fig. 1 - Omnidirectional video system architecture.

The modules in the presented architecture are briefly introduced as:

- **Acquisition -** Content is acquired with the use of a camera, and thus lenses and additional hardware and/or software to allow different, relevant capture modes and functionalities. Since a single lens is not able to capture a 360º horizontal and vertical view, more than one lens is normally necessary. Then, the views need to be stitched, this means appropriately glued together; video may have different Field of View (FOV) be stereoscopic or not and audio can be mono, stereo and multi-channel.
- **Projection -** The projection relates to the technique which is applied in order the acquired image data may be represented in a rectangular format which is more appropriate to be coded, notably with available standard codecs. After the projection, which is basically a transformation of the stitched image, the image is not spherical anymore and some areas may have been significantly stretched, thus creating distorted zones. This process may affect the coding efficiency obtained with the available standard video codecs which are adapted to non-distorted images.
- **Coding -** The coding regards the process where data is compressed to minimize the rate for a target quality; currently mostly available standard video codecs are used.
- **Rendering and Viewport Selection –** Before the coding stage, the video had to be converted using a certain map projection; however, the displayed format is not the same that is coded. Therefore, rendering is necessary to transform the decoded video into a proper format to be exhibited. The objective is not to watch the whole area at the same time but instead to navigate on a reduced viewing area, having an immersive sensation; thus, the viewing area selected by the user has to be extracted from the decoded data and only that area is displayed [4].
- **Visualization -** At the end, the user needs a display where the content is exhibited; the type and sophistication of this device will determine the quality of the experience provided to the user. For example, omnidirectional videos may be seen in a Head Mounted-Display (HMD) that allows the user to control the viewing angle, e.g. by simply turning around, while the video is being played; the user must be able to change the viewing angle whenever he/she wants without noticeable displaying delays as these delays are one of the main reasons for user sickness and dizziness.

B. Mapping Projections: from Spherical Video to 2D Rectangular Video

As most common video codecs (notably the standard ones) only code rectangular frames, it is first necessary to map the omnidirectional video into a 2D rectangular shaped video. The transformation from a spherical to a rectangular shape implies that the video information has to be stretched in some regions. The projections have been the main focus of study of omnidirectional video hence, the most common projections are presented:

- **Equirectangular projection –** this is the most used projection to convert spherical video into a rectangular format. This projection uses a constant spacing latitude $\phi \in [-\pi/2, \pi/2]$ and longitude $\theta \in [-\pi, \pi]$ and addresses the vertical and horizontal positions in a panorama using ϕ and θ, respectively. The projected image is more stretched in the horizontal direction, the closer a region is to the poles since the perimeter of a circumference in the sphere at latitude ϕ that has to be represented, gradually decreases as $\cos \phi$ from the Equator to the poles and still the full (and constant) rectangle width has to be filled
- **Lambert Cylindrical Equal-area:** Comparing with Equirectangular projection (ERP), this projection attempts to compensate the horizontal stretching with the shrinking in the vertical direction. This follows that same relation of $\cos \phi$. The area in equal-area projection is constant regarding the sphere.
- **Cube -** it starts by inserting the spherical video inside a cube and then stretching the sphere to cover the faces of the cube completely. This implies that the content has to be stretched close to the cube edges and corners but not as much as for the Equirectangular projection, notably near the poles. After this, the six faces of the cube contain video information have to be re-organized to obtain a rectangular layout.

Other projections following the same step as the Cube have studied such as pyramid [5], octahedron [6] or dodecahedron [7]. Also, this kind of projection may create with different dimensions allowing certain regions of the image have higher resolution.

III. Saliency Detection Model in Omnidirectional Images

Although projections have been the most studied solution to improve coding performance, the omnidirectional content characteristics are not taken into account, and for this reason, a content adaptive solution may be desired.

Thus, a saliency detection model for omnidirectional images is developed with the aim of defining regions-of-interest that are encoded with higher quality while less salient/important regions are encoded with lower quality, naturally targeting bitrate savings.

A. Saliency Detection in Omnidirectional Images: Architecture

Fig. 2 shows the architecture of the Saliency Detection Model (SDM) for omnidirectional images proposed. The input of architecture is based on the omnidirectional image dataset provided in the context of the *Salient360!: Visual attention modelling for 360ºImages Grand Challenge* organized at ICME'2017 [8]. This dataset includes both a set of omnidirectional images and their ground truth saliency maps with experimentally obtained fixation maps.

The architecture in Fig. 2 considers three branches:

- **Ground truth branch** - The left branch represents the ground truth in terms of saliency for each omnidirectional image, in this case expressed by means of the so-called Viewport integrated Head Direction Map (VHDM).
- **Proposed SDM** - The center branch represents the proposed SDM, and it is image-specific as it determines

the saliency considering the specific characteristics of a certain, n-th, omnidirectional image; its resulting saliency scores are represented in the Latitude biased Viewport integrated Saliency Map (LVSM).

- **Latitude driven SDM** - The right branch represents an alternative SDM which is NOT image-specific and only considers the latitude impact represented by means of the so-called Viewport based Latitude Importance Map (VLIM) which expressed the user relevance of different latitudes for human subjects as computed from a representative set of ground truth maps with experimentally measured latitude viewing intensities.

To adjust some algorithm parameters in the SDM designing process, the final output LVSM may be compared with the experimental VHDM using metrics that shall evaluate how close/correlated are the automatically computed and experimental maps. In a coding perspective, the idea is to use the LVSM as a quality impact map indicating the regions that should be coded at higher and lower qualities.

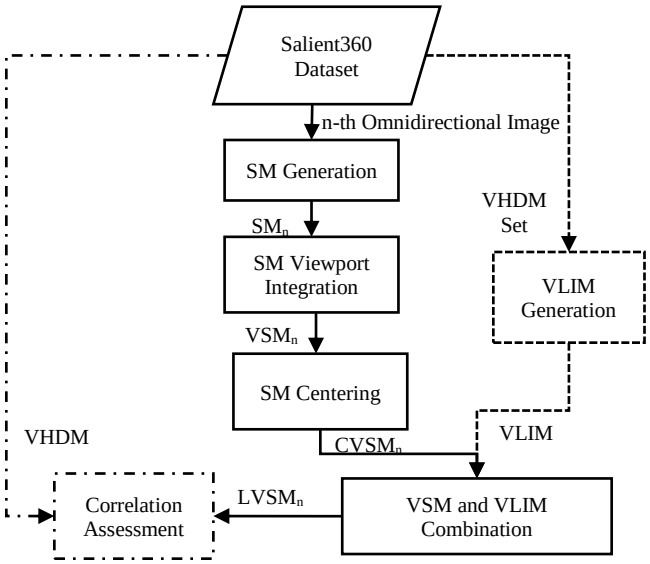


Fig. 2 - Architecture of the proposed Salient Detection Model for omnidirectional images.

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

1) Latitude driven SDM: VLIM Generation

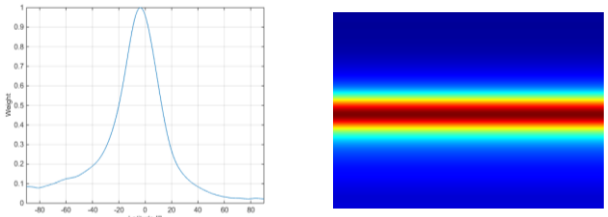
This module determines a saliency map only based on a set of experimental VHDMs from the *ICME 2017 Grand Challenge Salient360* [8] dataset.

The set of VHDM for each image in the dataset allow to extract some global statistics to globally characterize the importance of each latitude independently of a specific omnidirectional image. This map is computed through the average fixation intensity for each latitude (considering all longitudes) for the full set of the available VHDMs, computed as:

$$Q(i) = \frac{1}{N \times W} \sum_{n=1}^{N} \sum_{j=1}^{W} VHDM_n(i,j) \qquad (1)$$

where N is the number of available VHDMs (and thus images in the dataset) and W is the width of the VHDMs. This process is cumulative for the H rows in the VHDM. Finally, the $Q(i)$ score for each latitude expresses its corresponding user relevance, in this case independently of j-th longitude.

Fig. 3(a) shows the resulting weights depending on the latitude for the selected dataset. As shown in Fig. 3(b), the Q weights are converted into a 2D weight map that only varies in terms of the latitude.



(a) (b)

Fig. 3 - (a) Latitude weight as a function of the latitude; and (b) final VLIM.

2) Proposed SDM: Saliency Map Generation

The MLNET proposed by *Cornia et alli* [9] is one of the top scored SDM at the MIT Benchmark [10]. 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 and producing a temporary saliency map. Finally, a learned prior is considered to build the final saliency map that typically also considers the center bias, see Fig. 4.

The learning consists on a training process for the saliency map (SM) extraction algorithm 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.

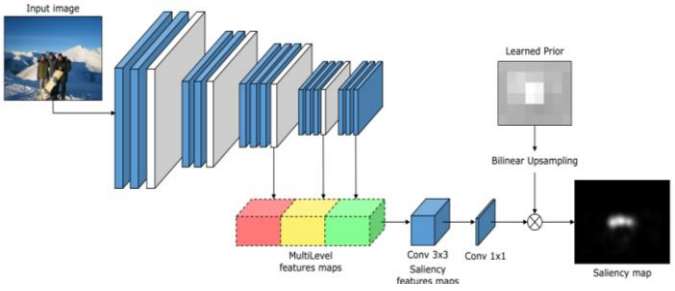


Fig. 4 - Architecture of the MLNET 2D SDM [9].

For the purposes of the omnidirectional image SDM, the MLNET pre-trained model is used. The MLNET input is an ERP omnidirectional image which has been zero-padded to fit a 4:3 aspect ratio and then resized to a 640x480 resolution which is the maximum size of the convolutional layers that are responsible for the feature extraction. As the SDM was trained for conventional images the saliency detection performance may be affected; however, the most affected regions are the poles which generally are not the most important/salient regions of the image.

3) Proposed SDM: Saliency Map Viewport Integration

Although usually the viewers look at the center of the viewport, eye and head directions are not always the same and this effect has to be taken into account (this means the users move the eyes around even when the head is fixed).
Thus, regions some processing has to be performed in this SM Viewport Integration module knowing the eyes move within the viewport.

Fig. 5 represents the architecture of the SM Viewport Integration algorithm which considers the following steps:

1) **Conversion to Spherical Domain** - For every pixel in the SM, its position is converted to spherical coordinates as longitude and latitude.
2) **Viewport Extraction** - Then, this position is assumed as the head direction and, using the same approach as for rendering, the viewport is computed.
3) **Weighted Saliency Computation** - After, the weighted average of the viewport saliency intensities is computed with weights defined by a foveation function, in this case a 2D Gaussian distribution with a standard deviation $\sigma = 4^{o}$ regarding the center of the viewport. The chosen standard deviation was determined as the value minimizing the RMSE between the VHDM and the LVSM.
4) **ERP Conversion** - The obtained result in the previous step is placed at the same position as in Step 1. At the end, when the process is repeated for every position, the resulting VSM is represented in a ERP.

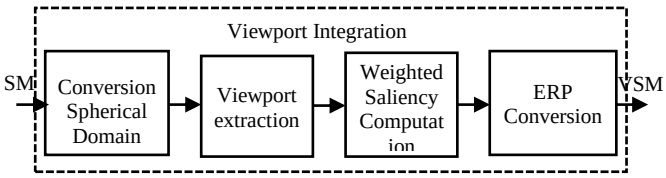


Fig. 5 - Viewport Integration module architecture.

4) Proposed SDM: 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, and instead the viewers complete the fixation by moving the eyes. For example, considering 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, to take this effect into account, the designed SDM should include a processing module that moves the VSM rows to a latitude closer to the Equator. It is important to refer that this creates a shrinking effect towards the Equator. Meanwhile, to keep the resolution and avoid creating any empty areas close to the poles, it is necessary to use some form of padding, copying the nearest valued pixel in the pole. To assure that every row is filled and there are no empty areas, each final Centered Viewport integrated Saliency Map (CVSM) coordinates are associated to some VSM coordinates implying that the corresponding saliency are allocated following this correspondence.

Assuming the coordinates $(i, j) = (0,0)$ at the image top left corner, the CVSM to VSM coordinates correspondence is defined by (naturally, only the i coordinates change as this is only a latitude move, not longitude):

$$i_{VSM} = \frac{1}{c}\, i_{CSM} \tag{2}$$

where c varies with the distance between the Equator ($i = H/2$ in the map representation) and the farthest very salient point with the relation given by

$$c = c_{max} - \frac{1}{H/2}\,\frac{d_{max}}{c_{max} - c_{min}} \tag{3}$$

while d_{max} is the distance between the Equator and the farthest very salient point expressed as

$$d_{max} = \max\left(\left|i_{VSM_{(i,j)>r}} - H/2\right|\right) \tag{4}$$

where $i_{VSM_{(i,j)>r}}$ represents the latitude of the farthest highly salient region to the Equator and $r = 0.8$, $c_{max} = 1.3$ and $c_{min} = 1$. The r value is here the threshold defining whether a region is very salient or not. The c_{min} and c_{max} values limit the effect of this centering effect.

5) Proposed SDM: CVSM and VLIM Combination

Despite the initially determined salient regions, the regions near the Equator are statistically the most viewed regions in an omnidirectional image, as shown in Fig. 3, and the computed SDM still does not considers this fact which is represented by the VLIM. The objective of this processing module is to introduce a latitude bias, where the saliency scores vary depending on the latitude.

As already explained, the VLIM expresses precisely the statistical latitude bias (experimentally computed), it is proposed here that the SM results from a combination between the computed CVSM and the experimental VLIM, more precisely by computing a weighted average of both maps:

$$LVSM = w \times CVSM + (1 - w) \times VLIM \tag{5}$$

where $w = 0.8$ is a previously determined value that minimizes the RMSE between the VHDM and the LVSM. This parameter setting should allow to decrease the difference between the LVSM and VHDM, as desired. This is also a safety measure for coding when the SDM is less accurate, as it increases the importance of the regions near the Equator, a fact that has strong experimental validation.

B. SDM Performance Assessment

The image dataset and experimental data provided in the context of the *Salient360!: Visual attention modelling for 360ºImages Grand Challenge* organized at ICME'2017 [8] has been adopted for the experiments. The images used from the dataset are: P2; P3; P4; P5; P6; P7; P10; P11; P12; P13; P14; P15; P17; P21;P22;P23;P24;P25; P27; and P28.

1) Performance Metric

The good performance of the proposed SDM should correspond to similar LVSM and VHDM, meaning that the computed saliency maps could replicate the experimental obtained saliency scores. As a performance reference/benchmark, the VHDM is also compared with the

VLIM which is only driven by the latitude bias; in theory, it is expected that the LVSM achieves better performance than the VLIM when both maps are compared with VHMD, since the VLIM statistically characterizes a whole set of images only in terms of head direction and not any specific image in particular.

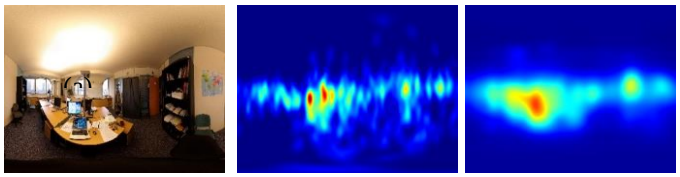
To compare the proposed LVSM and the VLIM, it is used the Root Mean Square Error (RMSE) metric. In this context, it measures the difference between two maps. Computing the RMSE for the two saliency maps is performed with the following expression:

$$RMSE = \sqrt{\frac{1}{N}\sum_{\theta=-\pi}^{\pi}\sum_{\varphi=-\pi/2}^{\pi/2}(LVSM(\theta,\varphi) - VHDM(\theta,\varphi))^2} \quad (6)$$

where θ is the longitude, φ is the latitude and N is the number of samples used to represent the spherical domain.

2) Performance Results and Analysis

Fig. 6 shows the example of image P3 and corresponding VHDM and LVSM. It shown the is a reasonable prediction of the salient regions as the LVSM presents the most salient regions in red, yellow and green in the same regions as the VHDM.



(a) (b) (c)

Fig. 6 Original equirectangular image P3 and corresponding (b) VHDM; and (c) LVSM;

Table 1 presents the average RMSE considering the entire set of images. It is shown that the average RMSE of VLIM is higher than RMSE of LVSM as it should be expected, meaning higher performance of LVSM, and the error decreases after each processing module.

TABLE 1
AVERAGE RMSE BETWEEN THE VHDM AND THE MAPS VLIM, VSM, CVSM AND LVSM.

	VLIM	VSM	CVSM	LVSM
Average RMSE	0.3298	0.1638	0.1538	0.1348

Fig. 7 presents the RMSE between the VLIM and the VHDM and between the LVSM and the VHDM for the full set of omnidirectional images. Comparing both cases, the LVSM RMSE is significantly lower than the VLIM RMSE for all the images in the dataset.

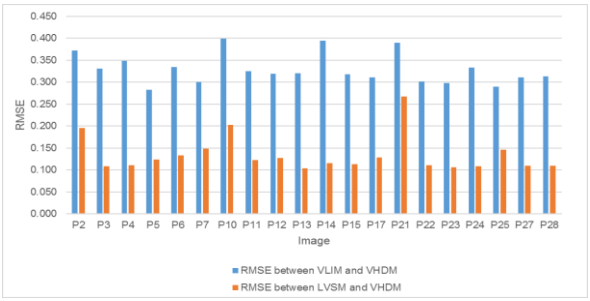


Fig. 7 - RMSE for the set of images in the dataset of VLIM and LVSM regarding VHDM.

IV. HEVC CODING

One of the most important parameters to control the rate and the obtained quality in HEVC coding is the quantization parameter (QP) as it defines how much distortion is inserted in the coding process. Naturally, using a quantization parameter which is adaptive to the image/video characteristics may allow achieving better compression at the price of some additional encoder complexity.

A. QP Selection based on the Saliency Map

The HEVC standard allows defining the QP at the Coding Unit (CU) level but 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. Moreover, it will be necessary to change the reference software encoder to compute the QP based on the saliency map intensity.

In [11], 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 same QP selection model is here applied to HEVC coding, notably to the Largest Coding Unit (LCU). It is applied at the LCU level and not at the CU level because the saliency map does not usually have drastic intensity variations within the LCU and thus it is not necessary to increase the system complexity. Simultaneously, it is necessary to consider that there are regions more sensitive to these QP variations, notably slow gradient regions. The QP for the i-th LCU is given by:

$$QP_i = round\left(\frac{QP_{slice}}{\sqrt{w_i}}\right) \quad (7)$$

where QP_{slice} is the default QP defined for the current slice (which defines the target quality). The w_i is a sigmoid function defined by

$$w_i = \begin{cases} a + \dfrac{b}{1+\exp(-c(S(X_i)/n - \bar{s})/\bar{s})}, & if\ l \le 10 \\ a + \dfrac{b}{1+\exp(-c(S(X_i) - \bar{s})/\bar{s}}, & if\ l > 10 \end{cases} \quad (8)$$

where $a = 0.7$, $b = 0.6$ and $c = 4$, the same parameters as in [11], $\bar{s}$ is the average saliency for all the set of LCU within the current frame, and $S(X_i)$ is the average saliency in the i-th LCU. The QP_i function depends on $\bar{s}$ (a frame level characteristic), which typically varies around 0.15, and $S(X_i)$ (a LCU level characteristic).

The normalized spatial activity n is computed as:

$$n = \frac{f \cdot l + t}{l + f \cdot t} \quad (9)$$

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

$$f = 2^{a/6} \quad (10)$$

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

$$l = 1 + \min(\sigma^2_{Y,k}), \text{where } k = 1, \dots, 4 \quad (11)$$

where $\sigma^2_{Y,k}$ denotes the spatial activity of the pixel values in sub-block k (of size N×N) in a luma CB (2Nx2N). Variable $\sigma^2_{Y,k}$ regards the luma samples variance, which is computed as

$$\sigma^2_{Y,k} = \frac{1}{z}\sum_{i=1}^{z}(w_i - \mu_Y)^2 \quad (12)$$

where z denotes the number of samples in the luma CB k, while w_i corresponds to the i_{th} sample in the luma CB k and μ_Y refers to the mean sample intensity of luma CB k, which is computed as:

$$\mu_Y = \frac{1}{Z}\sum_{i=1}^{z} w_i \quad (13)$$

Fig. 8 (a) represents w_i and Fig. 8(b) represents the QP_i depending on the specific $S(X_i)$ in the LCU, assuming $\bar{s} = 0.15$ and $QP_{slice} = 32$.

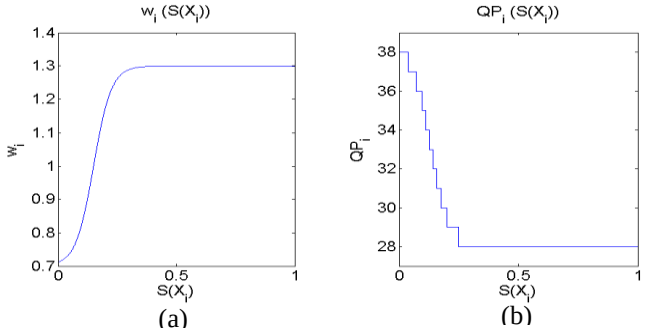


Fig. 8 - Variation of: (a) w_i ; (b) $1/\sqrt{w_i}$; and (c) QP_i depending on $S(X_i)$ for $\bar{s} = 0.15$ and $QP_f = 32$.

The spatial activity is also important, to determine which regions may be more sensitive for the Human Visual System (HVS). For this reason, for low spatial activity l, the normalized spatial activity is considered to compute the w_i and decrease the QP in these regions.

V. RD Performance Assessment

This section is divided into two main subsections. First, it is presented the objective evaluation with the test condition adopted and objective metrics, including the proposal of a new metric based on saliency. And second, the subjective evaluation is conducted to evaluate whether the objective metrics used are correlated with the quality perceived by human subjects.

A. Objective Evaluation

1) Test Conditions

The JVET Group has defined a set of assessment methodologies to conduct the omnidirectional video coding experiments [12]. 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 [13] define the basic QP values to be used, notably 22, 27, 32 and 37. The experiments should be performed for two different prediction configurations: Intra (only Intra coding is used) and Random Access (efficient temporal prediction structures are used).

The test material includes both omnidirectional images and videos. Regarding the images, the same images with 8192x4048 resolution used in Section IV from the Salient360 ICME 2017 Grand Challenge [8] dataset were selected to be Intra coded. The test video sequences have been selected from the JVET dataset and coded using the Random Access configuration for the first 128 frames.

TABLE 2
TEST SEQUENCES CHARACTERISTICS

Sequence name	Resolution	Frame Rate	Bit-depth
AerialCity	3840x1920	30	8
DrivingInCity	3840x1920	30	8
DrivingInCountry	3840x1920	30	8
GasLamp	8192x4096	30	8
Harbor	8192x4096	30	8
KiteFlite	8192x4096	30	8
PoleVault_le	3840x1920	30	8
SkateboardTrick	8192x4096	60	8
Trolley	8192x4096	30	8

2) Omnidirectional Image and Video Objective Quality Metrics

a) JVET Quality Metrics

The JVET Group has been very active in the area of omnidirectional video coding, notably studying performance assessment metrics, projections and new coding tools.
Therefore, a brief description is below:

- WS-PSNR 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.
- S-PSNR compares the input material before the projection with the output of the HEVC decoder after the conversion, both into the spherical domain.
- Viewport based PSNR (V-PSNR) compares the rendered viewports from the input material and the reconstructed/decoded image/video. Six viewports are used, identified as *Vx-PSNR*, with x between 0 and 5. The viewports are placed at specific longitude θ and latitude φ positions corresponding to (θ, φ), notably: viewport 0

(0,0); viewport 1 (90,0); viewport 2 (180,0); viewport 3 (-90,0); viewport 4 (0,90); and viewport 5 (0,-90).

b) Proposing a Saliency based Peak Signal-to-Noise Ratio Quality Metric

The appropriate quality assessment of omnidirectional image and video using content adaptive quantization based on visual saliency requires appropriate quality metrics which should take into account the applied saliency values, assuming that users concentrate their attention in the more salient regions.

With this purpose in mind, a saliency based PSNR (SAL-PSNR) is proposed here to evaluate the video quality according to the importance/saliency of each region in the omnidirectional image.

In this new metric, each pixel in the image has an importance associated to a weight defined by the multiplication of its saliency value by the distortion factor associated to the used projection as defined for the WS-PSNR metric, resulting in a weight $q(i,j)$ defined by

$$q(i,j) = \frac{W(i,j) \times SM(i,j)}{\sum_{i=0}^{m-1} \sum_{j=0}^{n-1} [W(i,j) \times SM(i,j])} \quad (14)$$

where $W(j,i)$ is the distortion factor for position (i,j) in the image and $SM(i,j)$ is the saliency map value for the same position.

The saliency based PSNR (SAL-PSNR) is thus defined as

$$SAL\text{-}PSNR = 10\, log\left(\frac{{MAX_f}^2}{SAL\text{-}MSE}\right) \quad (15)$$

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

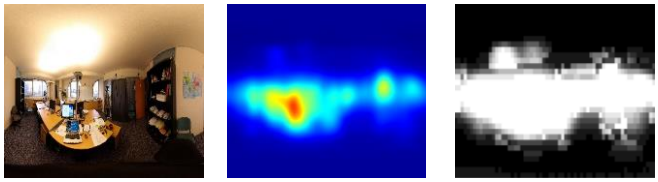
$$SAL\ MSE = \sum_{i=0}^{m-1} \sum_{j=0}^{n-1} ||f(i,j) - g(i,j)||^2 \times q(i,j) \quad (16)$$

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

3) Omnidirectional Image Compression Performance

In this section, the results are presented for the set of images following the test conditions previously described.

Fig. 9 (a) shows the image to be coded after ERP, Fig. 9 (b) is the LVSM, that is used to obtain the QP' values and Fig. 9 (c) represents the $\Delta QP_i'$ within the image, where the white regions represent the minimum $\Delta QP_i'$ (quality improved) while the black regions represent the regions where the $\Delta QP_i'$ is higher (quality reduced).



(a) (b) (c)

Fig. 9 - (a) Image P3; (b) LVSM used before coding; (c) Map with $\Delta QP_i'$.

Fig. 10 shows the RD-curves for the WS-PSNR, SAL-PSNR and V-PSNR metrics for image P3. The results allow deriving the following conclusions:

- For Y-WS-PSNR, the standard HEVC performs better than the Adaptive QP solution. This is expectable as the metric does not take into account the saliency map.
- For Y-SAL-PSNR and V0-PSNR, there is considerable rate-distortion (RD) gain for the Adaptive QP solution. This is expectable as the first metric takes into account the saliency map and the second metric corresponds to an area with high saliency values.
- V4-PSNR reveals a RD loss of Adaptive QP solution regarding the standard HEVC because the viewport is located in area with low saliency values.

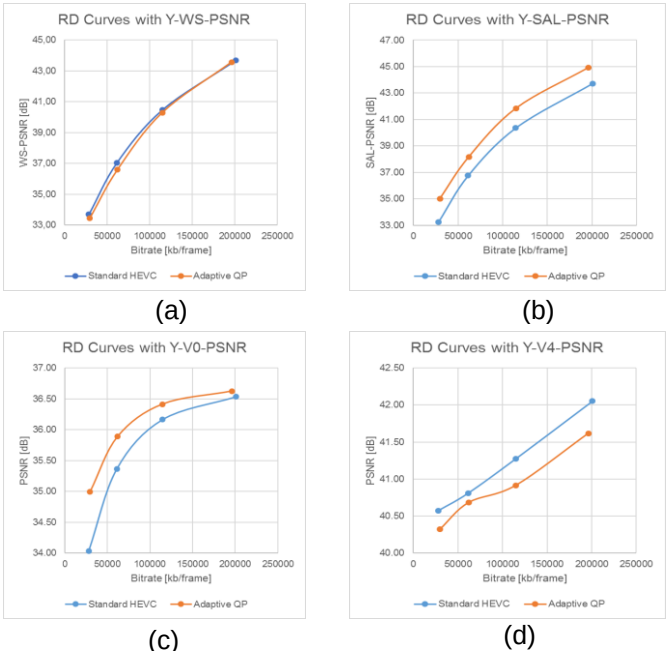


Fig. 10 - RD curves for image P3: (a) Y-WS-PSNR; (b) Y-SAL-PSNR; (c) Y-PSNR-VP0; and (d) Y-PSNR-VP4.

Table 3 presents the BD-Rate results for the Adaptive QP coding solution regarding HEVC standard coding for several objective metrics. Negative values indicate there is an improvement in terms of RD performance when Adaptive QP is used regarding the standard HEVC and vice-versa. The following conclusions may be derived:

- Y-WS-PSNR shows an average BD-Rate loss of 7.82%; this is expectable due to the non-consideration of the saliency map by the quality metric. On the other hand, there is a gain for the chrominance components of 8.81% and 10.73%, for the U and V components, respectively.
- As expected, there are significant BD-Rate gains for all components when using the SAL-PSNR metric as it gives higher importance to the most salient regions. The average BD-Rate gain is 30.33% for luminance and 35.13% and 35.44% for the U and V components, respectively.
- Looking at the V-PSNR, the viewport 0 is located at the center of the ERP image, the center of the viewport 4 is

located at the north pole. As expected, there are considerable quality differences for these two cases because the most salient regions are usually located at the Equator, whereas the poles typically correspond to the less salient regions.

- For viewport 0, Table 3 shows average BD-Rate gains of 23.67% for luminance, 33.60% for chrominance U and 35.69% for chrominance V, while, for viewport 4, the average BD-Rate losses are significant, notably 107.95% for luminance, 106.51% for chrominance U and 102.34% for chrominance V.

TABLE 3
AVERAGE BD-RATE FOR ADAPTIVE QP CODING REGARDING STANDARD HEVC FOR WS-PSNR, SAL-PSNR,V0-PSNR AND V4-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-V4-PSNR**	**U-V4-PSNR**	**V-V4-PSNR**
-23.67	-33.60	-35.69	107.95	106.51	102.34

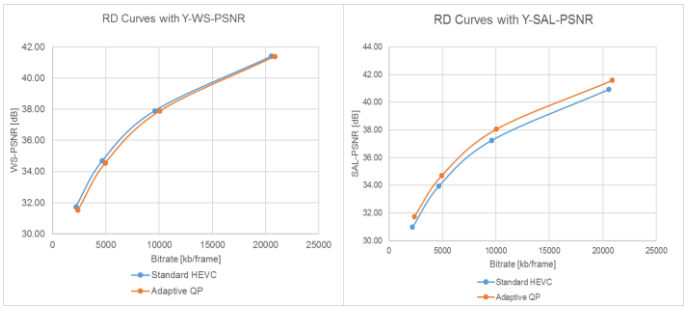
In summary, the most salient regions show an increase of quality regarding the standard HEVC, whereas the RD on the entire omnidirectional image is not strongly affected. Considering the WS-PSNR metric, which is saliency agnostic, only the luminance component suffers an average BD-Rate loss, while the chrominances actually show average BD-Rate gains. The V-PSNR also reaches BD-Rate gains when placed in regions around the Equator while the poles are penalized. These results seem to indicate that it is possible to obtain bitrate savings by reducing the quality of non-salient regions.

4) Omnidirectional Video Compression Performance

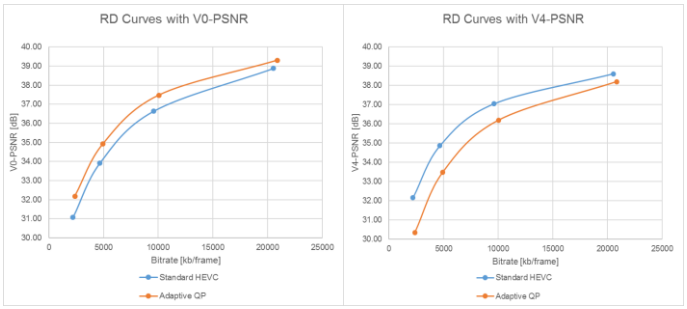
The study of the JVET video sequences follows the same methodology as for the Salient360 images; here the videos are coded using the Random Access prediction configuration, thus exploiting the temporal correlation, for the first 128 frames.

Fig. 11 shows the RD-curves for the WS-PSNR, SAL-PSNR and V-PSNR metrics for sequence *KiteFlite*. The results allow deriving the following conclusions:

- For Y-WS-PSNR, the standard HEVC performs better than the Adaptive QP solution. Once again, this is expectable as the metric does not take into account the saliency map.
- For Y-SAL-PSNR and V0-PSNR, there is a RD gain for the Adaptive QP solution. This is due to the fact that the first metric takes into account the saliency map and the second metric corresponds to an area with high saliency values.
- V4-PSNR reveals a RD loss of Adaptive QP solution regarding the standard HEVC because the viewport is located in area with low saliency values.



(a) (b)



(c) (d)

Fig. 11 - RD curves for test sequence KiteFlite: (a) Y-WS-PSNR; (b) Y-SAL-PSNR; (c) Y-PSNR-VP0; and (d) Y-PSNR-VP4.

To understand the compression performance behavior for several test sequences, Table 4 presents the BD-Rate results for the Adaptive QP coding solution regarding HEVC standard coding for the WS-PSNR and SAL-PSNR. The following conclusions may be derived:

- The Y-WS-PSNR results shows an average BD-Rate loss of 5.02% while the U-WS-PSNR results show in a loss of 0.52% and V-WS-PSNR has a gain of 0.61%. The chrominances present BD-Rates close to 0%.
- As expected, there are BD-Rate gains for all components when using the SAL-PSNR metric as it gives higher importance to the most salient regions. The average BD-Rate gains are 6.85% for luminance and 16.55% and 16.73% for the U and V components, respectively.

TABLE 4
AVERAGE BD-RATE FOR ADAPTIVE QP CODING REGARDING STANDARD HEVC FOR WS-PSNR AND SAL-PSNR.

	BD-Rate [%]					
	Y-WS-PSNR	**U-WS-PSNR**	**V-WS-PSNR**	**Y-SAL-PSNR**	**U-SAL-PSNR**	**V-SAL-PSNR**
AerialCity	5.51	13.42	13.07	-5.69	-12.08	-11.28
DrivingInCity	8.68	7.45	7.12	1.20	-14.70	-13.17
DrivingInCountry	-2.07	-10.93	-11.57	-13.71	-28.11	-27.56
GasLamp	3.50	-1.35	10.13	-8.37	-12.05	-8.74
Harbor	5.65	-2.97	-3.30	-5.90	-13.77	-13.36
KiteFlite	7.22	2.05	-3.04	-11.41	-20.09	-19.78
PoleVault	1.43	-3.84	-10.54	-11.55	-20.08	-24.75
SkateboardTrick	8.53	0.21	-1.57	3.62	-12.96	-15.04
Trolley	6.67	0.64	-5.76	-9.82	-15.10	-16.93
Average	**5.02**	**0.52**	**-0.61**	**-6.85**	**-16.55**	**-16.73**

Table 5 presents the BD-Rate results for the Adaptive QP coding solution regarding HEVC standard coding for the V0-

PSNR and V4-PSNR. The following conclusions may be derived:

- For viewport 0, shows average BD-rate gains of 8.47% for the luminance, 16.94% for chrominance U and 16.77% for chrominance V, while for viewport 4, the average BD-Rate losses are rather major, notably 39.94% for luminance, 52.88% for chrominance U and 51.41% for chrominance V.
- Once again, this was expected because the saliency is higher in the regions near the Equator, thus higher quality; opposite occurs for the viewport 4 at the pole.

TABLE 5
AVERAGE BD-RATE FOR ADAPTIVE QP CODING REGARDING STANDARD HEVC FOR V0-PSNR AND V4-PSNR.

	BD-Rate [%]					
	Y-V0-PSNR	**U-V0-PSNR**	**V-V0-PSNR**	**Y-V4-PSNR**	**U-V4-PSNR**	**V-V4-PSNR**
AerialCity	-1.21	2.82	0.65	60.46	67.29	59.73
DrivingInCity	0.61	-15.73	-13.99	65.99	73.32	75.83
DrivingInCountry	-9.29	-22.02	-24.61	34.58	64.23	58.15
GasLamp	-10.48	-14.73	-14.12	35.99	41.29	45.04
Harbor	-0.72	-12.40	-11.89	17.51	35.62	26.33
KiteFlite	-18.06	-26.97	-23.30	48.09	58.08	49.69
PoleVault	-14.91	-24.02	-27.55	46.05	44.15	52.05
SkateboardTrick	-0.41	-15.12	-11.67	36.28	68.86	74.47
Trolley	-21.81	-24.32	-24.41	14.48	23.05	21.42
Average	**-8.47**	**-16.94**	**-16.77**	**39.94**	**52.88**	**51.41**

B. *Subjective Evaluation*

1) *Test Conditions*

Four images from the *Salient360* dataset were selected to proceed with subjective tests such that different type of content is evaluated.

The subjective evaluation has followed the Absolute Category Rating with Hidden Reference methodology using a five-grade quality scale (1-Bad; 2-Poor; 3-Fair; 4-Good; 5-Excellent).

The images were coded following the JVET methodology described in Section V. The hidden references are images not coded but converted to the 4K (4096x2048) resolution while the coded images are coded at the same resolution as the reference. Two coding solution were evaluated: Standard HEVC and the Adaptive QP presented; both are coded using QP reference of 30, 35, 40 and 45. These values differ from the objective evaluation because using that range of values it was hard distinguish the different quality levels. For QPs below 30 are almost undistinguishable from the reference image.

The images were shown to the subjects using the Oculus Rift, which has a resolution of 1080x1200 per eye while they were sit on a rolling chair being free to rotate.

During the test session, 26 subjects visualized the image coded with Standard HEVC and Adaptive QP presented in a random order for each subject.

Subjects (21 males and 5 females) were between 22 and 54 years old and characterized by an average and median of 30.61 and 26.5 years old respectively.

Before the test session, written instructions were provided and a training session was conducted to adapt the subject to the assessment procedure while five training samples were shown, representing all the quality levels.

During the test session, each image was shown during 20 seconds and then the subject was asked to evaluate the overall quality of the omnidirectional image and report a score.

2) *Objective Metrics Validation*

After the experiment, outlier detection was performed according to the guidelines described in Section 2.3.1 of Annex 2 in [14] to remove subjects whose scores are highly deviated from others. In this experiment, two subjects were considered as outliers and removed from subsequent results. Then, the Differential Mean Viewer Scores (DMV) were determined for each quality level for each image, according to Section 6.2 in [15].

Then, the objective is to find a relationship between the mean score and each objective metric. A logistic function is fitted to DMV depending on the objective metrics. This function represents the predicted DMV depending on the objective metric.

Finally, it is possible to compare the experimental DMV with the predicted DMV using three performance indexes commonly used for this purpose. The Pearson Linear Correlation Coefficient (PLCC) evaluates linearity, the Spearman Rank Order Correlation (SROC) that evaluates the monotonicity and the RMSE which indicates the difference between two variables.

The sigmoid function is determined for the studied metrics considering the DMV obtained for the standard HEVC and adaptive QP conditions and the performance indexes may be computed.

It is seen in Table 6, that most of the metrics show very high correlation with the DMV values.

TABLE 6
PERFORMANCE INDEX OF PLCC, SROCC AND RMSE FOR THE DIFFERENT OMNIDIRECTIONAL METRICS.

	PLCC	**SROCC**	**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
V4-PSNR	0.466	0.373	1.021

VI. CONCLUSION AND FUTURE WORK

The compression performance results show considerable quality gains for the regions defined as important. Also, the quality of the entire omnidirectional video is not severely penalized according to the saliency independent metrics. In terms of rate, significant BD-Rate gains are obtained when considering the novel quality metric. This suggests that it may be possible to effectively reduce the bitrate without reducing the user quality of experience by assigning higher QP values to less salient regions and vice-versa without compromise the overall perceived quality and compression performance. However, QP variations within an image may create visual artifacts that affect the user quality of experience and, for this reason, subjective tests were conducted and show that the proposed coding solutions do not seem to present relevant

quality variations relative to the standard HEVC. It was possible to evaluate the correlation between the subjective and objective results. The proposed objective quality metric achieves very high correlation with the subjective testing results, notably better correlation performance than the available objective quality metrics used for omnidirectional content. Despite the encouraging results, there is certainly room for improvements:

- **Saliency Detection Model Improvements –** Although the proposed saliency detection model determines the salient regions rather reasonably, the saliency detection model may be improved. First, the original saliency detection model adopted was trained for non-omnidirectional images which means the resulting saliency is based on non-omnidirectional image characteristics. Therefore, a possible improvement is to train the model with omnidirectional images if large omnidirectional image datasets are made available. Also, the proposed saliency detection model only considers the spatial information; thus, it would be appropriate to improve it to consider the temporal information since it is known that moving objects and scenes are significant factors to catch the user attention in video. The omnidirectional content characteristics still need to be studied more carefully, however the latitude importance is currently one of the most important factors to consider. Moreover, it is necessary to find solutions to avoid the that the saliency maps are computed from distorted content from 2D projections without penalizing the saliency detection model performance.
- **Study of different projections** - Although the equirectangular projection is clearly the most widely used before coding omnidirectional images and videos, it has been shown that different projections may achieve better coding performances, as the content is also differently distorted. So, another possible direction could be the study the compression performance of the proposed coding solution when different projections are considered.

REFERENCES

[1] T. Wiegand, G. J. Sullivan, G. Bjøntegaard and A. Luthra, "Overview of the H.264/AVC Video Coding Standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560-576, July 2003.

[2] "High Efficiency Video Coding (HEVC) reference software HM," Fraunhofer Institute for Telecommunications, Heinrich Hertz Institute, 2016. [Online]. Available: https://hevc.hhi.fraunhofer.de/.

[3] A. Borji and L. Itti, "State-of-the-Art in Visual Attention Modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 185-207, January 2013.

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

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

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

[7] C.-W. Fu, L. Wan and T.-T. Wong, "The Rhombic Dodecahedron Map: An Efficient Scheme for Encoding Panoramic Video," *IEEE Transactions on Multimedia*, vol. 11, no. 4, pp. 634 - 644, April 2009.

[8] Y. Rai, L. Callet, Patrick and P. Guillotel, "Salient360 - The Training Dataset for the ICME Grand Challenge," in *IEEE International Conference on Multimedia & Expo*, Hong Kong, July 2017.

[9] M. Cornia, L. Baraldi, G. Serra and R. Cucchiara, "A Deep Multi-Level Network for Saliency Prediction," *arXiv:1609.01064 [cs.CV]*, September 2016.

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

[11] H. Hadizadeh and I. V. Bajic, "Saliency Aware VIdeo Compression," *IEEE Transactions on Image Processing*, vol. 23, no. 1, pp. 19-33, January 2014.

[12] J. Boyce, E. Alshina, A. Abaas e Y. Ye, "JVET Common Test Conditions and Evaluation Procedures for 360º video," Joint Video Exploration Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, JVET-E1030, 5th Meeting, Geneva, Switzerland, January 2017.

[13] K. Suehring and X. Li, "JVET Common Test Conditions and Software Reference Configurations," Joint Video Exploration Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, JVET-1010, 2nd Meeting, San Diego, CA, USA, February 2016.

[14] ITU-R BT.500, "Methodology for the subjective assessment of quality of television pictures," International Telecommunication Union, January 2012.

[15] ITU-T P.910, "Subjective video quality assessment methods for multimedia applications," International Telecommunication Union, April 2008.

[16] L. Prangnell, M. Hernández-Carbonero and V. Sanchez, "Cross-Color Channel Perceptually Adaptive Quantization for HEVC," *arXiv:1612.07893 [cs.MM]*, February 2017.