PERCEPTUALLY-ADAPTIVE QUANTIZATION FOR STEREOSCOPIC VIDEO CODING

Sami Jaballah[1,3], *Mohamed-Chaker Larabi*[2], *Jamel Belhadj Tahar*[1]

[1] InnovCom Laboratory, SupCom, University of Carthage,Tunisia
[2] XLIM Laboratory, CNRS, University of Poitiers, France
[3] National School of Engineering of Tunis, University of Tunis-El-Manar, Tunisia

ABSTRACT

In this paper, we present a novel perceptually-based optimization for the improvement of stereoscopic video coding efficiency. The main idea of this proposed scheme is to adaptively adjust the quantization parameter by taking into account the Human Visual System perceptual characteristics. For this, a saliency map is generated from both views and then segmented into salient and non-salient regions. To make the proposed scheme effective, and inspired from the binocular suppression theory, the asymmetry is ensured by altering the saliency map and not the view . As a result, the proposed perceptual coding scheme effectively reduces the bit-budget without affecting the perceptual quality based on an optimization approach with asymmetric video coding taking into account the saliency map of each view. Experimental results on HEVC-MV show that the proposed algorithm can achieve over 20% bit-rate saving while preserving the perceived image quality.

Index Terms— Perceptual optimization, 3D-HEVC, Quality, MS-SSIM

1. INTRODUCTION

In the recent years, the rapid expansion of three-dimensional (3D) technologies in the television and the cinema industry is increased by the recent advances in compression and display of 3D video technologies. The stereoscopic 3D video (S3D) is considered as the widely used video format due to its coding, transmission and display simplicity among other video formats. In the same vein, stereoscopic video can be easily adapted in communication applications with the support of existing video technologies. However, to answer the enormous requirements in storage and bandwidth, many efforts have been made to further improve the coding performance of the 3D videos coding schemes. The 3D High Efficiency Video Coding (3D-HEVC) [1, 2] is the latest 3D extension of the video coding standard developed by the Joint Collaborative Team on Video Coding (JCT-VC) in July 2012 by ITU-T Video Coding Experts Group (VCEG) and ISO/IEC Moving Picture Experts Group (MPEG). Like the H.264/MVC and taking advantage of the new coding tools introduced in the High Efficiency Video Coding (HEVC/H.265) [3] and the inter-view redundancies, The 3D-HEVC enables coding different views of the same scene acquired from multiple cameras. It also offers a coding efficiency improvement of about 50% bit-rate reduction compared to the previous standard H.264/MVC. The main objective of a video codec is to reduce the bit-budget by taking into account the quality of the video after decoding. To improve encoding performance, video compression standards use various tools to remove the statistical redundancy of pixels [4]. Some of these tools can cause distortions that are not sensed by the metrics used by these standards, but perceived by the Human Visual System (HVS). Basically, HVS models are used to better interpret the large amount of visual information and reduce the complexity of the scene [5]. Perceptual based video coding approaches provide further coding efficiency, by exploiting the properties of HVS. In Psychology, the concept of visual attention is considered as the process of selecting a visual objects from among other competitors objects. Furthermore, Visual attention is an important mechanism of the HVS and one of the most concrete cognitive processes aiming at detecting the notable (salient) objects in a particular scene. We distinguish two different approaches of visual attention [6]: bottom-up and top-down. In bottom-up approaches, the attention is led involuntarily, and the HVS is attracted unconsciously to salient visual area. Top-down approaches are task-driven, where the visual attention depend on the performed task and the semantic information of the scene. Visual attention models have been deployed in video coding schemes to increase coding performance by reducing the bit-budget with almost no-noticeable artifact. Thus, a variety of sophisticated saliency computational models have been proposed in the literature in order to generate saliency maps showing important areas of the scene. Most of the proposed models are based on the feature integration theory introduced in [7]. The extracted features from the scene are generally color, intensity and orientation. Several conspicuity maps are calculated from each feature and then integrated in competitive way to form the saliency map. Over the last decades, a few studies have been proposed for the perceptual optimization of 3D video coding. To ensure unaltered binocular perception, authors in [8], have proposed a non-uniform asymmetric stereoscopic coding. Based on the Binocular Just Noticeable Difference (BJND)

combined with visual saliency and depth information, this approach determines the best level of blur to be introduced for each region of the image, allowing to decrease the required bandwidth for S3D delivery. In [9], authors have investigated the impact of using visual attention-aided ROI coding within the framework of asymmetric stereoscopic video coding. The main view is coded with a uniform QP while the other is using a non-uniform QP, with a ΔQP increase at different saliency layers. Authors of this work have also evaluated the impact of incorporating structural sensitivity to compression errors in the formulation of the visual attention model. A stereoscopic visual attention (SVA) for regional bit allocation optimization is proposed in [10]. Multiple perceptual features including depth, motion, intensity, color, and orientation contrast are used, to mimic the visual attention mechanisms of stereoscopic perception. The semantic region-of-interest (ROI) is extracted based on the saliency maps of SVA. In this paper, we propose a perceptually-based optimization of stereoscopic video coding scheme. By accounting for the properties of the HVS, an adaptive QP adjustment is introduced. As a result, the proposed perceptual coding scheme effectively reduces the bit-budget without affecting the perceptual quality thanks to an asymmetric optimization approach. Furthermore, we take advantage of the binocular suppression theory by only altering the saliency map instead of the texture view. For this, we use different thresholds to segment the saliency map into salient and non-salient regions. The QP adjustment process remains the same for both view, so the QP of each coding block is adjusted by a ΔQP according to its saliency, activity and size. The adjustment by ΔQP is limited by an adaptive ΔQP_{max} in order to limit the visual distortion between neighbouring coding blocks.

The remainder of this paper is organized as follows. In Section 2, we introduce our proposed perceptual scheme for stereoscopic video coding. The experimental results and comparisons are provided in section 4. Finally, the contributions of this paper are summarized and the future work is outlined in section 5.

2. PROPOSED APPROACH

2.1. Overview of the proposed solution

In the standard-compliant approaches, no a-priori information is known on the content of the video sequence. However, an adaptive compression scheme could require information concerning the spatial arrangement of colors or intensities, saliency of the important regions and edges of the scene. The main purpose of the proposed perceptual stereoscopic video coding is to remove perceptual redundancy and reduce the the bit-budget without any noticeable distortion. Our approach is based on the fact that coding non-salient regions with higher QP will not affect the perceptual quality of the video. Two stages are presented in this approach, they are performed before and while coding. First, the offline stage is performed before the coding process and is dedicated to the extraction of information about to the scene. In this stage, a saliency map is generated from both views and then segmented into salient and non-salient regions. For the on-line stage, we propose a new method to adaptively adjust the QP of each coding block based on the a-priori extracted information from the scene: saliency information, localization, size. The perceptual adjustment of QP is performed for each Coding Unit (CU). The diagram of the proposed model is depicted on Figure 1

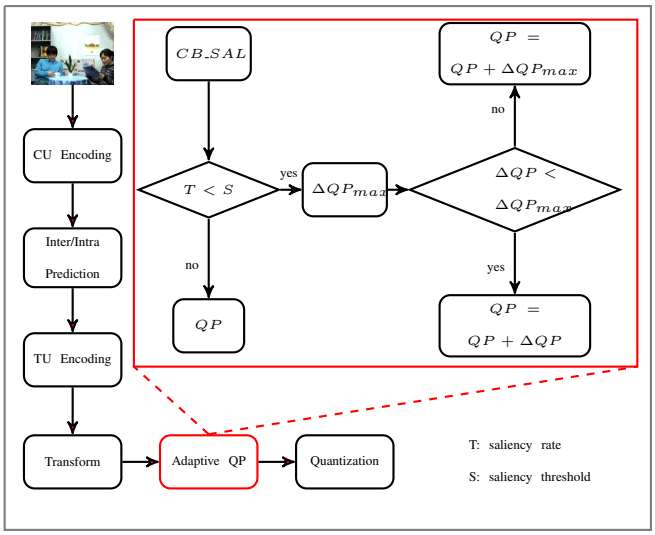
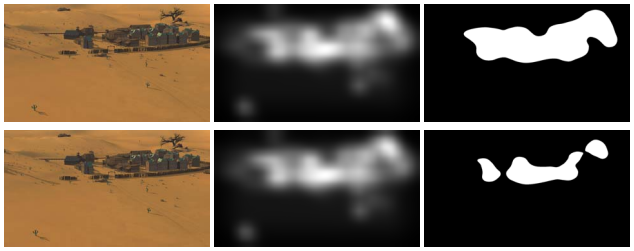


Fig. 1. Block diagram of the proposed approach.

2.2. Saliency map

For the application of an adaptive strategy for QP selection, based on perceptual cues of the visual scene, we use saliency information showing how conspicuous each region is in the given scene. In the proposed approach, we opted for the well-known Itti-Koch saliency model [11], where saliency is derived from low-level visual features. Other models can be applied without changing the online part of the proposed approach. First, several pre-attentive visual features are extracted and form multiple low-level feature maps. These features are sensitive to intensity contrast, four orientations (0, 45,90, and 135) and color contrast. Center-surround differences are determined using Gaussian pyramids with 9 scales from scale 0 to scale 8. Six center-surround difference are computed as the pointwise differences across pyramid scales. Then six feature maps are computed for each feature. Finally, all feature maps are combined to form the unique scalar saliency map.

In order to be exploited in the video coding process, the saliency map is segmented into salient and non-salient regions. Indeed, instead of using a simple threshold to obtain binary maps of salient regions, we use a content-driven thresh-



(a) (b) (c)

Fig. 2. Example images from GhostTownFly sequence, (a) color texture image,(b) Predicted saliency map, (c) segmented saliency map for the left and right views, with $Th_{main} = 0.4390$ and $Th_{aux} = 0.6557$.

old. In this model we propose to segment the saliency maps of the left and right views in a different fashion. One map (denoted as main) is segmented with a low threshold gathering most of the salient information. The other map (auxiliary) is segmented with higher threshold in order to keep the most salient regions. The thresholds for obtaining the main 1 and auxiliary 2 maps for every Group of Picture (GoP) are obtained as function of the average of saliency pixels values in the given GoP as described by the following equation.

$$Th_{main} = 2 * \frac{1}{n\_GoP} \sum_{g=0}^{n\_GoP-1} \left(\frac{1}{w * h} \sum_{x=0}^{w} \sum_{y=0}^{h} S_g(x,y) \right) \quad (1)$$

$$Th_{aux} = 3 * \frac{1}{n\_GoP} \sum_{g=0}^{n\_GoP-1} \left(\frac{1}{w * h} \sum_{x=0}^{w} \sum_{y=0}^{h} S_g(x,y) \right) \quad (2)$$

where Th_{main} and Th_{aux} are the threshold of the main and the auxiliary maps, respectively. n_GoP is the number of frames within a GoP, g is the frame index. w and h are the width and height of the saliency map, respectively, and $S(x,y)$ is the saliency value at position (x,y)

3. ADAPTIVE ADJUSTMENT OF QUANTIZATION PARAMETER

In the proposed model, the QP of each coding block is adjusted based on its saliency level computed offline. Taking into account the human perception characteristics, every block with a saliency rate lower than a given threshold is considered as less important region for the viewer. Therefore, in the encoding process, these blocks could be coded with lower quality (higher QP). In other words, less amount of bits are allocated to regions with low saliency rate. Such an encoding technique could cause a noticeable visual quality variation between neighbouring salient and non-salient regions. To deal with this issue, we limit the QP adjustment (ΔQP) by the maximum variation ΔQP_{max}.

Our model outlined in figure 1 is detailed in the following steps:

- Compute the saliency rate T of the current block:

$$CB\_SAL = \sum_{i=0}^{W_B} \sum_{j=0}^{H_B} A_{i,j} \quad \begin{cases} A_{i,j} = 1 & if P_{i,j} > Th \\ 0 & else \end{cases}$$

$$T = \frac{CB\_SAL}{W * H} \quad (3)$$

  where W_B and H_B are the size of the coding block and $P_{i,j}$ is the pixel value in the coding block at i,j position.

- A saliency threshold S is calculated for each block based on the previously calculated thresholds Th (Th_{main} or Th_{aux}) and the block size, as follows:

$$S = 1 - \alpha \left(\frac{size\,(CB)}{SIZE\_MAX} + Th \right) \quad (4)$$

  where α is a weighting factor used to limit S between $[0,1]$ $\alpha \leq 0.5$ (for the experimental part α is set to 0.5).

- Before adjusting the QP value, we have to define its maximum variation ΔQP_{max} as shown by the following equations:

$$\Delta QP_{max} = T * QP + \beta,$$
$$where \quad \begin{cases} \beta = 0 & if \quad I\_frame \\ \beta = 1 & if \quad P\_frame \\ \beta = 2 & if \quad B\_frame \end{cases} \quad (5)$$

- The adjustment of ΔQP is controlled by:

$$\Delta QP = \left(1 - \frac{S}{T}\right) * \Delta QP_{max} \quad (6)$$

- The Final QP value of the coding block is defined as :

$$QP_{adj} = min\left\{51, max\left\{QP + \Delta QP, 1\right\}\right\} \quad (7)$$

4. EXPERIMENTAL RESULTS

In order to assess the performance of the perceptually-based QP adjustment model, the proposed approach is compared to 3D-HTM ver.14.1 reference software MV-HEVC version. Comprehensive experiments were conducted using seven well-known video sequences with two different resolutions (1024×768 and 1920×1088) under the common test conditions (CTC), with the quantization parameters 22, 27, 32 and 37. The coding performances are evaluated using the Bjontegaard metrics providing Δ-PSNR and Δ-Rate [12]. In addition to these metrics, we used the average Multi-Scale Structural SIMilarity (MS-SSIM) index which is considered as being well correlated with the human judgment. This metric has been chosen in order to analyze the perceptual impact generated by our approach and the inherent bitrate reduction.

Unfortunately, it was impossible to compare the proposed model with other similar approaches from the literature because we could not find such works applied to 3D-HEVC optimization process.

Table 1 illustrates the results obtained with the proposed perceptual optimization approach in terms of bitrate-saving and PSNR variation. The proposed model provides significant bitrate savings depending on the used content and achieving up to 38% for the *Balloons* sequence. Of course these important saving comes with the cost of PSNR losses that are relatively for most of the sequences ($< 0.30\ dB$) except for *Balloons* which presents a decrease of $1.05\ dB$. However, this decrease for *Balloons* is negligible when compared to the huge saving achieved with the proposed approach. Moreover, a PSNR loss does not necessarily imply a decrease of perceived quality. Moreover, one needs to take into account that thanks to the binocular suppression phenomenon, the perceived 3D image is image is close to the sharpest view.

Table 1. Coding performance of the proposed model compared to the reference software.

Sequences	Δ-BR	Δ-PSNR
Balloons	-38.88%	1.05
GhostTownFly	-16.07%	0.38
ici Kendo	-5.86%	0.20
Newspaper	-12.56%	0.34
PoznanHall2	-10.04%	0.13
PoznanStreet	-12.97%	0.39
UndoDancer	-7.72%	0.21

These results of Table 1 are confirmed by Figure 4 which illustrate the RD performance of $Balloon$, $PoznanHall$ and $Newspaper$ sequences in term of MS-SSIM. This figure shows that the proposed model does not alter the perceived quality of the sequences and can outperform, in some cases, the reference software encoder. The perceptual selection of the QP value and its maximum variation based on the saliency map has limited the perceived quality distortion. To illustrate this purpose, Fig. 3 compares the visual quality of the left-eye images of "PoznanHall2" and "balloons" sequences encoded with $QP = \{27, 32, 37\}$, when the proposed approach and HTM ver.14.1 reference software are applied.

5. CONCLUSION

In this paper, a perceptually driven approach is proposed for the latest 3D video coding standard. It performs a perceptual optimization of the stereoscopic video coding based on the concept visual attention using a generated saliency. Experiments were carried out on seven stereoscopic video sequences, objective and subjective results demonstrated that the proposed method can provide important bitrate saving without perceptible visual quality distortion. It is important to note that the proposed perceptual optimization model provides a better tradeoff between bitrate and perceived quality. An extension of this work consists in reducing blocking effect caused by the QP variation at the salient regions edges..

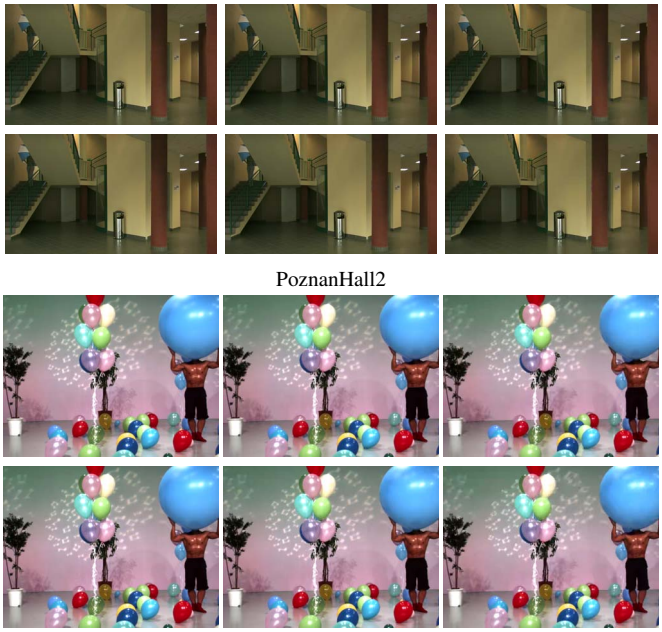
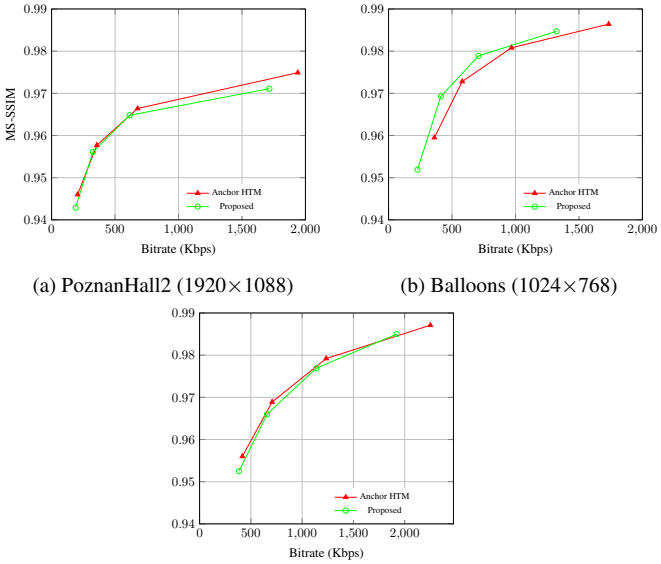


Fig. 3. Subjective quality comparison of the left-eye images of PoznanHall2 and Balloons sequences for three QP values QP=27 (left), QP=32 (middle) and QP=37 (right). For each sequence, the first line represents reference software HTM results and the second line the proposed method.



(a) PoznanHall2 (1920×1088) (b) Balloons (1024×768)

(c) Newspaper (1920×1088)

Fig. 4. RD-curves of of three different sequences with different resolutions using the widely used MS-SSIM.

6. REFERENCES

[1] G. Tech, K. Wegner, Y. Chen, and S. Yea, "3d-hevc draft text 1," *Joint Collaborative Team on 3D Video Coding Extension Development (JCT-3V) document JCT3V-E1001*, 2013.

[2] Miska M Hannuksela, Ye Yan, Xuehui Huang, and Houqiang Li, "Overview of the multiview high efficiency video coding (mv-hevc) standard," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2154–2158.

[3] Gary J Sullivan, J-R Ohm, Woo-Jin Han, and Thomas Wiegand, "Overview of the high efficiency video coding (hevc) standard," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 22, no. 12, pp. 1649–1668, 2012.

[4] Rafał K Mantiuk and Karol Myszkowski, "Perception-inspired high dynamic range video coding and compression," in *CHIPS 2020 VOL. 2*, pp. 211–220. Springer, 2016.

[5] Alexandra Frischen, Andrew P Bayliss, and Steven P Tipper, "Gaze cueing of attention: visual attention, social cognition, and individual differences.," *Psychological bulletin*, vol. 133, no. 4, pp. 694, 2007.

[6] Laurent Itti and Christof Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision research*, vol. 40, no. 10, pp. 1489–1506, 2000.

[7] Anne M Treisman and Garry Gelade, "A feature-integration theory of attention," *Cognitive psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[8] Sid Ahmed Fezza, Mohamed-Chaker Larabi, and Kamel Mohamed Faraoun, "Asymmetric coding using binocular just noticeable difference and depth information for stereoscopic 3d," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 880–884.

[9] Erhan Ekmekcioglu, Varuna De Silva, Peter Tho Pesch, and Ahme Kondoz, "Visual attention model aided non-uniform asymmetric coding of stereoscopic video," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 8, no. 3, pp. 402–414, 2014.

[10] Yun Zhang, Gangyi Jiang, Mei Yu, Ken Chen, and Qionghai Dai, "Stereoscopic visual attention-based regional bit allocation optimization for multiview video coding," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, pp. 60, 2010.

[11] Laurent Itti, Christof Koch, and Ernst Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, , no. 11, pp. 1254–1259, 1998.

[12] Gisle Bjontegaard, "Calculation of average psnr differences between rd-curves," *VCEG-M33*, April 2001.