Depth Perceptual Video Coding for Free Viewpoint Video based on H.264/AVC

G. Cernigliaro , M. Naccari , F. Jaureguizar' , J. Cabrera , N. García

Abstract—**A novel scheme for depth sequences compression, based on a perceptual coding algorithm, is proposed. A depth sequence describes the object position in the 3D scene, and is used, in Free Viewpoint Video, for the generation of synthetic video sequences. In perceptual video coding the human visual system characteristics are exploited to improve the compression efficiency. As depth sequences are never shown, the perceptual video coding, assessed over them, is not effective. The proposed algorithm is based on a novel perceptual rate distortion optimization process, assessed over the perceptual distortion of the rendered views generated through the encoded depth sequences. The experimental results show the effectiveness of the proposed method, able to obtain a very considerable improvement of the rendered view perceptual quality.**

Index Terms—**Free Viewpoint Video, H.264/AVC, Depth Maps, Rate Distortion Optimization, JND, HVS.**

I. Introduction

Free Viewpoint Video (FVV) and 3DTV [1] represent the next generation of video paradigms whose goal is the observer involvement thanks to the 3D scene reconstruction or the depth perception without stereoscopic glasses or other additional devices. Immersion performance depends on the number of viewpoints which, when high, improves the 3D sensation. However, the increasing of the cameras entails a considerable amount of information to be transmitted or recorded. To limit the viewpoint number, the virtual view rendering concept has been introduced. Virtual video sequences can be generated and placed in any point of the scene. The virtual view generation is set in a well defined environment where every texture sequence has a corresponding depth sequence. This setting is called Multi View plus Depth (MVD) environment [2]. Virtual sequences are generated by interpolating the warped reference textures to the location of the virtual one, using the depth to locate objects in the 3D space. One of the first synthetic view generation methods is the widely used Depth-Image-Based Rendering (DIBR) [3].

For the depth encoding, research started focusing on novel approaches adapted on the characteristics of this content. Morvan *et al.* [4], for example, proposed a platelet based depth encoder which geometrically adapts the blocks area according to the boundaries, refining the prediction in the depth areas which are more relevant in view synthesis. As, in FVV, depth sequences are never displayed, a coding approach, based on the depth quality, is not effective because it is not possible to predict how depth compression errors can affect the virtual sequences. As consequence, other researchers focused on a depth coding optimization oriented to the view synthesis. Following this idea, Woo-Shik Kim *et al.* [5] introduced a Rate Distortion Optimization (RDO) algorithm where the distortion is evaluated on the virtual view rendered through the encoded depth sequence.

In this work we propose to manage the depth information in a perceptual video coding environment based on the characteristics of the Human Visual System (HVS) which is not able to perceive all the distortion introduced by the compression. The HVS sensitivity can be modeled by means of the Just Noticeable Difference (JND) which represents the minimum visibility threshold below which no change can be perceived by a human observer [6]. The JND concept has been already applied on the depth information. In the work of De Silva *et al.* [7], [8], exploiting the human perception, a Just Noticeable Differences in Depth (JNDD) model has been derived. The JNDD model described how the HVS perceives the object position, in depth, with respect to the others in the same scene. This approach is useful to model the HVS depth perception. However, in view synthesis, depth maps are used to generate synthetics sequences, hence the depth compression quality does not affect the depth visual quality, but rather the pixel positions in the virtual sequences. In this paper we propose to perceptually manage these position errors through a depth encoder, based on a perceptual RDO, oriented to the view synthesis. The distortion, in this case, is evaluated as perceptual difference between the virtual view, rendered with the original depth map, and the virtual view rendered through the depth encoded with the novel Depth Perceptual Encoder (DPE). The DPE is able to optimize the depth encoding according to the perceived rendered view quality.

The paper is structured as follows. Section II shows the JND concept, focusing on the Spatial JND. In Section III the Depth Perceptual Encoder is presented. Section IV shows the experimental results. Finally, in Section V, the conclusions and the future work are illustrated.

II. Adopted JND model

As mentioned above, the HVS is considered in the design of the proposed encoder. In this work, we take into account the luminance sensitivity of the human eye. Luminance differences are perceived depending on the surrounding luminance level which makes some differences more visible than others. The JND model has been derived according to the visual sensitivity masking effects which can depend on the spatial, temporal and spectral signal characteristics [9].

In this work we have simplified as much as possible the conditions, taking into account only the spatial masking effects, modeled on the pixel domain. The considered JND model is based on part of the work of Chou and Li [10], where the JND spatial masking effects have been analyzed developing an

evaluation of spatial JND (SJND) thresholds. The thresholds have been modeled with a function of luminance contrast and spatial masking (Eq. 1).

$$S_{JND}(i,j) = \max\{f_1(bg(i,j), mg(i,j)), f_2(bg(i,j))\} \quad (1)$$

where $f_1(bg(i,j), mg(i,j))$ and $f_2(bg(i,j))$ are respectively used to estimate the spatial masking and luminance contrast. The function $f_1(bg(i,j), mg(i,j))$ is used to evaluate the relationship between the visibility threshold and the luminance difference, where the function $bg(i,j)$ is the average background luminance and $mg(i,j)$ is the maximum weighted average of luminance differences. The function $f_2(bg(i,j))$ is used to measure the relationship between the visibility threshold and luminance background.

Focusing on the depth encoding, the mentioned SJND model is applied on the virtual view rendered through the original depth maps. The sequence so rendered is used as reference and, applying the SJND model on it, it is possible to know which distortions, produced by the encoder on the depth, are not perceived by a human observer.

III. Depth Perceptual Encoder design

In the H.264/AVC based video compression, the sequence encoding is made through various processes based on different prediction strategies. The video frames are divided into 16×16 blocks called Macro Blocks (MBs) and the prediction is applied on every MB. Every possible strategy produces a distortion with respect to the original signal, and a certain amount of bits, used for the representation. To improve the encoding efficiency, the encoder selects the one with the minimum rate-distortion cost, between all the possible procedures. The best prediction strategy is chosen through the RDO process [11] and the RD cost is evaluated as in Equation 2.

$$J(QP, M) = D(QP, M) + \lambda_{QP} R(QP, M) \quad (2)$$

where QP is the quantization parameter used to quantize the signal and M is the mode used for the prediction. The value λ_{QP} is the Lagrangian multiplier evaluated, for every QP, minimizing the RD cost function (Eq. 3).

$$\lambda_{QP} = 0.85 \times 2^{\frac{(QP-12)}{3}} \quad (3)$$

A traditional encoding strategy, in this case, would optimize the RD cost according to the reconstructed depth frames quality. However, as mentioned in the introduction, the research has been oriented to a rendered view based RDO. In a perceptual video coding environment this approach is even more effective due to its characteristics which are completely oriented to the human observer. Hence, in this novel DPE, we propose a new perceptual RDO process, where the distortion is evaluated on the rendered view, taking into account the HVS characteristics modeled by the JND concept. Equation 4 shows how the perceptual RD cost is evaluated.

$$PJ_s(QP, M) = PD_s(QP, M) + \lambda_{QP} R(QP, M) \quad (4)$$

where PJ_s denotes the perceptual RD cost evaluated as in the standard approach (Eq. 2) but with a different distortion metric PD_s which represents the perceptual distortion measured between the synthetic view rendered through the original depth and the synthetic view rendered through the compressed depth taking into account the HVS sensitivity modeled by the SJND. For the perceptual distortion evaluation it has been used a Mean Squared Perceptual Error (MSPE) method derived by the work of Chou and Chen [12], as shown in Equation 5.

$$PD_s(QP, M) = \frac{1}{256} \sum_{i=1}^{16} \sum_{j=1}^{16} \{[|p_r(i,j) - \tilde{p_r}(i,j)| - S_{JND}(i,j)]^2 \delta(i,j)\} \quad (5)$$

where $p_r(i,j)$ is the luminance value of the pixel located in the position (i,j) of the virtual MB rendered by using the original depth MB and $\tilde{p_r}(i,j)$ is the same value of the virtual view rendered though the depth MB encoded using the mode M and the quantization parameter QP. The value $S_{JND}(i,j)$ represents the visibility luminance threshold evaluated for the pixel $p_r(i,j)$. The function $\delta(i,j)$ is used to force the difference to 0 when the distortion does not reach the SJND threshold as it is depicted in Equation 6.

$$\delta(i,j) = \begin{cases} 1, & |p_r(i,j) - \tilde{p_r}(i,j)| \geq S_{JND}(i,j) \\ 0, & |p_r(i,j) - \tilde{p_r}(i,j)| < S_{JND}(i,j) \end{cases} \quad (6)$$

The Lagrangian multiplier λ_{QP} is evaluated as in Eq. 3, where a traditional RD cost function is considered. In the DPE, the distortion is evaluated as in Eq. 5, where the MSE metric is modified in order to take into account the HVS characteristics. In the proposed method we consider Lagrangian multipliers evaluated in a traditional fashion which, due to the similar characteristics of the used distortion metric with respect to a traditional one, can well approximate the perceptual RD cost function minimization.

In some perceptual video encoders the strategy used is to change the QP in function of the signal JND characteristics. Normally higher QPs (which mean worst quality) are used in areas less sensible to errors in order to save bit-rate when the perceptual distortion does not increase. In this work the perceptual RDO process is applied on all the possible modes available in H.264/AVC and between various QPs which range around a fixed value. The perceptual RDO process decides the optimum QP and mode (QP_{opt} and M_{opt}) which minimize the PJ_S cost, according to Equation 7.

$$(QP_{opt}, M_{opt}) = \arg\min_{QP_q}\{\arg\min_{M_m}\{PJ_S(QP_q, M_m)\}\} \quad (7)$$

where QP_q represents all the allowed QPs and M_m represents all the available modes for every QP. The operations needed for the DPE application on a depth frame are depicted in Algorithm III.1.

Algorithm III.1: DPE($PerceptualRDO$)

for *each frame* $\mathbf{f(n)}$
  for *each frame pixel* $\mathbf{(i,j)}$
    $S_{JND}(i,j)$ *evaluated as in Eq.* 1
  for *each* **MB** of $\mathbf{f(n)}$
   while $QP_{min} \leq QP_q \leq QP_{max}$
    for *each available mode* $\mathbf{M_m}$
      $PJ_s(QP_q, M_m)$ *evaluated as in Eq.* 4
      if $PJ_s(QP_q, M_q) < \min(PJ_s)$
        $QP_{opt} = QP_q, \quad M_{opt} = M_m$

As shown in the pseudo-code, the algorithm starts by evaluating the S_{jnd} for every pixel of a frame $f(n)$. Then, for every frame MB, it evaluates the QP_{opt} and M_{opt} by minimizing the perceptual RD cost evaluated on all the QPs in the range and all the available modes.

For the evaluation of the virtual view perceptual distortion, the implementation in the encoder of a view synthesis algorithm is needed. According to the proposed DPE, there are not restrictions about the algorithm used but, in order to obtain reasonable results, the method implemented has to be as similar as possible to the method used for the view synthesis after the stream decoding. The view synthesis method used in this work is specified below in Section IV.

IV. Experimental Results

In this section the performance obtained by the DPE is shown and compared with the performance obtained by a traditional H.264/AVC based encoder. The section is divided into three parts: in the first the used sequences and the coding settings are depicted, in the second part the perceptual performance assessment methodology is proposed, the experimental results are provided in the third one.

A. Experimental setup

In FVV depth maps are never shown, hence the performance of a depth coding algorithm has to be evaluated on the virtual sequence rendered though the compressed depth. For the virtual view generation, the used video sequences are MVD sequences captured with calibrated cameras. For the experiments three MVD sequences have been used. In order to test the algorithm in different situations, the sequences have three different resolutions and one of them has been captured by moving cameras. Table I shows the used sequences with the corresponding resolution, frame-rate, real camera used and virtual camera rendered.

The DPE algorithm has been implemented on the JM Reference Software (RS) version 18 [13]. The traditional H.264/AVC based depth coding has been made through the same RS version without any change. In both cases the encoder settings are depicted as follows: the profile used is the Main Profile with Inter and Intra modes enabled; the QP range has been made vary in a range of $\pm$ 1 QPs around a fixed QP value. All the experiments have been made on a GOP structured as IPPP.... The number of frames used in the experiments corresponds to one second of video.

TABLE I
MVD SEQUENCES USED FOR THE EXPERIMENTS.

Sequence	Resolution	Orig. View	Sinth. View	fps
Mobile	720×540	4	4.5	25
Beergarden	1920×1080	5	5.5	25
Kendo	1024×768	1	2	30

For the virtual view rendering, the algorithm proposed by Fehn in [3] has been used. The same algorithm has been implemented in the DPE to obtain the synthetic view perceptual distortion $PD_s(QP, M)$. In the synthetic sequences there are some frames areas which information is not available because it is occluded, hence, warping to the location of a virtual view, some de-occlusions can appear. To provide good visual results, the rendering algorithms are followed by post-processing techniques for the de-occluded areas filling. In the rendering method implemented in the proposed algorithm, the rendered view de-occlusions are important areas where it is easier to perceptually notice a pixel displacement due to depth compression errors, hence, at this stage of the study, the perceptual distortion (Eq. 5) have been evaluated on the rendered frame without considering any filling. However, as the experimental results have to show the perceptual quality of the sequences perceived by the observer, the performance has been evaluated on the rendered views after a simple de-occlusion filling made by repeating the values of the nearest available pixels with the higher depth value [14]. There are not limitations about the use of other filling techniques.

B. Performance indicators

This work has been conceived for the depth encoding optimization, with the goal to improve the perceptual quality of the rendered views. For the evaluation, a perceptual quality metric, called Peak Signal-to-Perceptible-Noise Ratio (PSPNR), proposed in [12], is used. The PSPNR is evaluated as shown in Equation 8.

$$PSPNR = 10 \log_{10} \frac{255^2}{MSPE} \tag{8}$$

where $MSPE$ is the Mean Squared Perceptual Error already used in Eq. 5 for the perceptual distortion evaluation (Eq. 9).

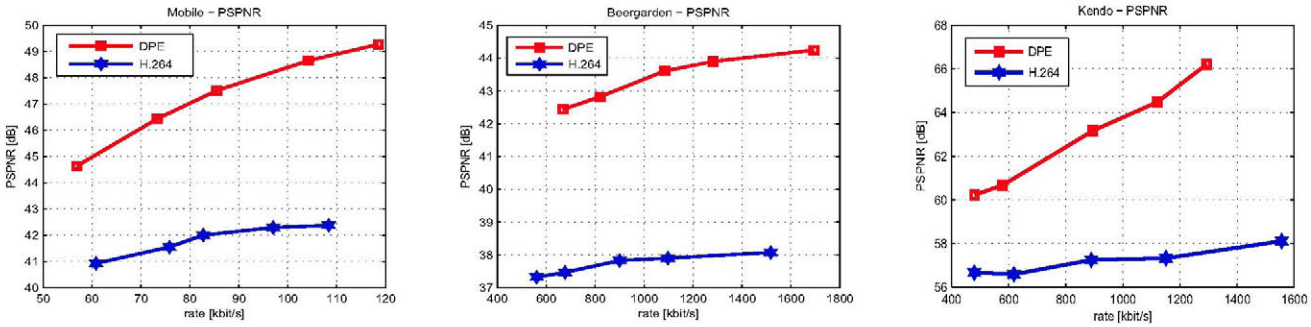
$$MSPE = \frac{1}{256} \sum_{i=1}^{16} \sum_{j=1}^{16} [|p_r(i,j) - \tilde{p_r}(i,j)| - S_{JND}(i,j)]^2 \delta(i,j)\} \tag{9}$$

The variables used in Equation 9 are the same already depicted for the Equations 5 and 6.

C. RD Performance Evaluation

In this section, the performance evaluation measured by mean of the PSPNR is shown. The curves are obtained measuring the quality of the virtual view rendered, first through the depth maps encoded with the proposed algorithm and, second, with a traditional H.264/AVC based encoder.

Figure 1 shows, in terms of PSPNR and bit-rate for all the considered MVD sequences, the gain the DPE algorithm obtains in perceptual performance with respect to the traditional H.264/AVC encoder.



(a) Sequence Mobile (b) Sequence Beergarden (c) Sequence Kendo

Fig. 1. Comparison of the perceptual performance of the DPE vs. a traditional H.264/AVC encoder (Mobile, Beergarden, Kendo)

The results of the traditional H.264/AVC encoders are optimized on the traditional MSE evaluated on the depth maps, therefore the quality of the rendering and of the de-occlusion filling is unpredictable. This is the reason why the perceptual RD curves, in this case, do not follow the typical logarithmic trend. On the contrary, observing the curves obtained using the DPE, the logarithmic trend is better approximated thanks to the perceptual RDO oriented compression. The approximation is not yet perfect due to the presence of a de-occlusion filling algorithm at the post-processing stage which can change the results predicted by the encoder. A perfect logarithmic approximation should be obtained with the implementation of a de-occlusion filling algorithm inside the DPE.

The improvement obtained by the DPE is noticeably high in all the cases for all the considerate rates: 3.5 dB for very low bit rates, 6 dB for rates around 1 Mb/s and up to 9 dB for higher rates. These results indicate that the depth maps encoding could be set in a coding environment different to the existing one and oriented to the real function they have, which is the virtual view generation. Virtual views are often affected by the presence of artifacts so the use of a perceptually oriented encoding of the depth maps can be the best option for the depth compression.

V. Conclusions and Future Work

In this paper, to the best of our knowledge, the first depth perceptual encoder has been proposed. The encoding of the depth sequences is optimized through a perceptual based RDO process which takes into account the HVS characteristics modeled by the SJND. The results obtained at this stage of the study show how the virtual view perceptual quality obtained with the DPE is always better than the quality obtained with a traditional H.264/AVC based algorithm, considerably improving the representation of the synthetic sequences (up to 9 dB).

In addition, this work only represents the first stage of a study on the depth perceptual coding. The next steps of this research have a threefold orientation. First, also the JND modeled on the temporal masking can be considered, in order obtain best perceptual results. Second, the use of de-occlusions filling algorithm can be included in the DPE. Third, the Lagrangian multipliers for the minimization of the RD cost function have to be analyzed more in depth and set in the depth perceptual coding environment.

According to the results and to the further improving possibilities, the perceptual depth encoding, oriented to the rendered view perceptual quality, can entail an important advance for the depth compression.

Acknowledgment

This work has been partially supported by the Ministerio de Ciencia e Innovación of the Spanish Government under project TEC2010-20412 (Enhanced 3DTV).

References

[1] A. Smolic, "An overview of 3D video and free viewpoint video," *Lecture Notes in Computer Science*, vol. 5702, pp. 1–8, 2009.

[2] P. Merkle, A. Smolic, K. Muller, and T. Wiegand, "Multi-view video plus depth representation and coding," in *ICIP 2007*, vol. 1, Oct. 2007, pp. I –201 –I –204.

[3] C. Fehn, "Depth-image-based rendering (DIBR), compression and transmission for a new approach on 3D-TV," *SPIE*, vol. 5291, pp. 93–104, 2004.

[4] Y. Morvan, P. H. N. de With, and D. Farin, "Platelet-based coding of depth maps for the transmission of multiview images," vol. 6055, no. 1. SPIE, 2006, p. 60550K.

[5] W.-S. Kim, A. Ortega, P. Lai, D. Tian, and C. Gomila, "Depth map coding with distortion estimation of rendered view," vol. 7543, no. 1, p. 75430B, 2010.

[6] H. R. Wu and K. R. Rao, *Digital video image quality and perceptual coding*. CRC Press, 2010.

[7] D. De Silva, W. Fernando, S. Worrall, S. Yasakethu, and A. Kondoz, "Just noticeable difference in depth model for stereoscopic 3D displays," in *ICME 2010*, July 2010, pp. 1219 –1224.

[8] D. De Silva, W. Fernando, G. Nur, E. Ekmekcioglu, and S. Worrall, "3d video assessment with just noticeable difference in depth evaluation," in *ICIP 2010*, Sept. 2010, pp. 4013 –4016.

[9] N. Jayant, J. Johnston, and R. Safranek, "Signal compression based on models of human perception," *Proceedings of the IEEE*, vol. 81, no. 10, pp. 1385 –1422, Oct. 1993.

[10] C.-H. Chou and Y.-C. Li, "A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile," *IEEE Trans. on Circ. and Sys. for Video Tech.*, vol. 5, no. 6, pp. 467 –476, Dec. 1995.

[11] G. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *Signal Processing Magazine, IEEE*, vol. 15, no. 6, pp. 74 –90, Nov. 1998.

[12] C.-H. Chou and C.-W. Chen, "A perceptually optimized 3-D subband codec for video communication over wireless channels," *IEEE Trans. on Circ. and Sys. for Video Tech.*, vol. 6, no. 2, pp. 143 –156, Apr. 1996.

[13] F.-I. HHI, "H.264/AVC JM reference software (JM 18)," 2011.

[14] W. Bruls, C. Varekamp, R. Gunnewiek, B. Barenbrug, and A. Bourge, "Enabling introduction of stereoscopic (3D) video: Formats and compression standards," in *ICIP 2007*, vol. 1, 16 2007-oct. 19 2007, pp. I –89 –I –92.