Perceptually Optimized Bit-Allocation and Associated Distortion Measure for Block-Based Image or Video Coding

Christian R. Helmrich[*], Sebastian Bosse[*], Mischa Siekmann[*], Heiko Schwarz[*],
Detlev Marpe[*], and Thomas Wiegand[*+]

[*]Fraunhofer HHI
Video Coding and Analytics Department
Einsteinufer 37, 10587 Berlin, Germany

[+]Technical University of Berlin
Telecommunication Systems Department
Einsteinufer 17d, 10587 Berlin, Germany

Abstract

It is well known that input-invariant quantization in perceptual image or video coding often leads to visually suboptimal results and that quantization parameter adaptation (QPA) based on a model of the human visual system can improve subjective coding quality. This paper introduces a simple low-complexity QPA algorithm, controlled using a block-wise perceptually weighted distortion measure representing a generalization of the PSNR metric. The weighting scheme of this WPSNR metric is based on a psychovisual model. It directly leads to a perceptually adapted scaling of the block-wise Lagrange parameter used in the bit-allocation process in the encoder and, consequently, to a block-wise QPA. Unlike prior QPA approaches, the proposal avoids classifications of picture regions and easily extends from still-image or grayscale to video or chromatic coding. The WPSNR metric also uses fewer algorithmic operations than e. g. the multiscale structural similarity measure (MS-SSIM). Due to the results of two formal subjective tests indicating its visual benefit, the QPA proposal has been adopted into VTM, the currently developed Versatile Video Coding (VVC) reference software.

1. Introduction

Perceptual transform coding of still images or videos is known to, for certain input, benefit from visual quantization models adhering to the characteristics of the human visual system (HVS) [1, 2]. Especially in rate-constrained applications where the encoder must adjust the bit-rate of the coded stream dynamically [3], HVS motivated quantization usually provides significantly increased subjective quality of the decoded content compared with objectively optimized quantization based on, e. g., peak signal-to-noise ratio (PSNR) or mean squared error (MSE) measures. Recently, various subjectively optimized quantization methods have been presented. For example, HVS-based rate control based on [1] for video compression is described in [4], input adaptive quantization optimizing for a structural similarity metric (SSIM) is proposed in [5, 6], a just-noticeable difference model simultaneously considering several psychovisual effects is applied in [7], and subjectively tuned quantization is derived from a texture mask model in [8].

It is worth noting that the abovementioned perceptual quantization approaches, as well as the HVS inspired distortion metrics which they employ, tend to increase in algorithmic complexity over the years since their underlying models become more and more elaborate. As a result, their implementations into image or video encoder software require increasing amounts of computational complexity, which may slow down the encoding process quite considerably. A low-complexity quantizer adaptation method is, therefore, a desirable objective. Moreover, the present authors observed that, of the psychovisual effects commonly exploited in the literature, namely,

- frequency sensitivity, defined by a contrast sensitivity function,
- spatiotemporal masking effects, particularly luminance masking, and
- higher-level perceptual factors such as attention or eye movement,

only the second category is viewer and environment invariant. Specifically, the frequency or contrast sensitivity depends on factors like viewing distance to, and resolution of, the display reproducing the decoded image or video, whereas higher-level HVS factors highly depend on the picture region the observer is looking at [1, 2] (humans only possess high visual acuity over a small viewing area known as the fovea). Furthermore, a viewer is often given control over the video playback and is often allowed to look at the content repeatedly and to focus on any spatiotemporal region thereof. Therefore, it can be concluded that such HVS characteristics cannot be exploited reliably in image or video coding.

Given the desired requirements of low computational complexity and high viewer and environment invariance in visually optimized quantization, a simple perceptually motivated quantization parameter adaptation (QPA), along with a corresponding equally straightforward, psychovisually inspired distortion metric, is proposed in this paper. As discussed in Section 2, the distortion measure is a generalization of the PSNR metric and loosely based on the spatial activity model devised in [1] and its extension described in [9], both of which are luma-only. However, unlike these two models, which classify picture blocks into either *smooth/flat*, *edge*, or *texture* regions, an alternative design avoiding classification is used.

Section 4 continues with a description of the QPA algorithm implementing the desired bit-allocation according to the visual rate-distortion metric of Section 3, along with a simple extension to chromatic picture components. Section 5 then outlines the basic parametrization of the QPA and generalized PSNR for modern block-based image or video coding and reports on the results of objective and subjective experiments conducted to assess the visual merit of our work. Section 6, finally, summarizes and concludes the paper.

2. Perceptually Motivated Distortion Measures

2.1 Block-Based Weighted Squared-Error Distortion Measures

Let D_k^{SSE} denote the sum of squared errors (SSE) for a picture block B_k at index k. With s and $\hat{s}$ being the original and reconstructed signal for the considered picture, it is given by

$$D_k^{\text{SSE}} = \sum_{(x,y)\in B_k} (s[x,y] - \hat{s}[x,y])^2. \tag{1}$$

We assume that a distortion measure D_k^{wSSE} which better reflects the subjective quality for a block B_k (in comparison to other blocks) can be defined according to

$$D_k^{\text{wSSE}} = w_k \cdot D_k^{\text{SSE}}, \tag{2}$$

where the weighting factors w for the blocks B of a picture can, basically, be any measures which only depend on the original (i. e., uncoded) picture samples. These weights w can be interpreted as visual subjective sensitivity measures for the blocks. For B_k with large visual sensitivity measure w_k, a given SSE distortion has a higher impact on perceptual quality than for blocks with small w_k. The overall distortion D_{pic} for a picture shall be given by the sum of the distortions for the individual blocks. Hence, for the SSE distortion, we obtain

$$D_{\text{pic}}^{\text{SSE}} = \sum_k D_k^{\text{SSE}} = \sum_{(x,y)} \left(s[x,y] - \hat{s}[x,y]\right)^2, \tag{3}$$

whereas for the weighted SSE distortion, we have

$$D_{\text{pic}}^{\text{wSSE}} = \sum_k D_k^{\text{wSSE}} = \sum_k \Bigg(w_k \cdot \sum_{(x,y)\in B_k} \left(s[x,y] - \hat{s}[x,y]\right)^2 \Bigg). \tag{4}$$

Note that for $w_k = 1$, the weighted $D_{\text{pic}}^{\text{wSSE}}$ equals the unweighted $D_{\text{pic}}^{\text{SSE}}$. Therefore, a simple weighted measure according to Eq. 2 and Eq. 4 can be considered a generalization of Eq. 3.

2.2 The PSNR and its Generalization, the Weighted PSNR

To approximate the logarithmic sensitivity of the HVS [10], subjective quality is typically reported in terms of the peak signal-to-noise ratio (PSNR) rather than SSE. Following this widely adopted convention, a perceptually weighted PSNR (WPSNR) can be defined as

$$\text{WPSNR} = 10 \cdot \log_{10} \left(\frac{W \cdot H \cdot 255^2 \cdot 2^{2BD-16}}{D_{\text{pic}}^{\text{wSSE}}} \right), \tag{5}$$

where W and H specify the number of samples in horizontal and vertical direction, respectively, and BD is the bit-depth for the samples of the color component at hand. Note that the numerator in Eq. 5 can be simplified to $W \cdot H \cdot (2^{BD} - 1)^2$ without much effect on the output. Again, if all blocks exhibit weight $w_k = 1$, the WPSNR is identical to the PSNR.

Following the common practice in codec standardization and visual quality research, with N being the number of frames in a video sequence and $\text{PSNR}[n]$ indicating the PSNR of the given color channel for the n-th frame, the average WPSNR for the video is given by

$$\text{WPSNR}_{\text{seq}} = \frac{1}{N} \cdot \sum_n \text{WPSNR}[n]. \tag{6}$$

To guarantee that the WPSNR values lie in the range familiar from PSNR, the mean of the block-wise weights w_k for an image, or a set of images or video frames, must be close to 1.

3. Visually Optimized Encoder Control

3.1 Review of Lagrangian Bit-Allocation

Having outlined a generalization of the PSNR distortion metric in the previous section, we continue with a brief review of the concept of Lagrangian bit-allocation applied in modern image or video coding. At this point, we do not make any particular assumptions about the actual distortion measure used; we only assume that it is *block-additive*, i. e., the distortion for a full picture equals the sum of the distortions for the individual blocks. This property is valid for both the conventional SSE and the above-introduced weighted SSE distortion.

Let $\mathbf{p}_k$ be the vector of encoding decisions at index k. The main objective of an encoder control is to determine the block coding parameters $\{\mathbf{p}_k\}$ in a way that the overall distortion D_{pic} is minimized while the associated rate R_{pic} does not exceed a given rate budget. Using the concept of Lagrangian multipliers, we obtain the optimization problem [11, 12]

$$\min_{\{\mathbf{p}_k\}} D_{\text{pic}}(\{\mathbf{p}_k\}) + \lambda \cdot R_{\text{pic}}(\{\mathbf{p}_k\}). \tag{7}$$

If we use any block-additive distortion measure, we can rewrite the minimization as

$$\min_{\{\mathbf{p}_k\}} \sum_k D_k(\mathbf{p}_k) + \lambda \cdot R_k(\mathbf{p}_k), \tag{8}$$

where $R_k(\mathbf{p}_k)$ represents the rate that is required for conveying all decisions $\mathbf{p}_k$ for a block B_k and $D_k(\mathbf{p}_k)$ is the resulting distortion for B_k. To achieve a feasible encoding algorithm, dependencies between blocks may be ignored. Then, the overall optimization problem of Eq. 8 can be solved approximately via separate optimizations (in coding order) for every k:

$$\forall k \quad \min_{\mathbf{p}_k} D_k(\mathbf{p}_k) + \lambda \cdot R_k(\mathbf{p}_k). \tag{9}$$

In case of independent blocks, the solution of the overall optimization problem of Eq. 7 is obtained when all decisions for the individual blocks are made with the same value of the Lagrange multiplier λ. For non-independent blocks (which we have in predictive image or video coding), still the same Lagrange multiplier λ should be used for all block decisions.

3.2 Lagrangian Bit-Allocation for Weighted SSE Distortion

If we use the conventional SSE distortion measure $D_k = D_k^{\text{SSE}}$, the coding decision for each block can be written as

$$\min_{\mathbf{p}_k} D_k(\mathbf{p}_k) + \lambda \cdot R_k(\mathbf{p}_k). \tag{10}$$

If, instead, we employ the weighted SSE distortion measure $D_k^{\text{wSSE}} = w_k \cdot D_k^{\text{SSE}}$, we obtain

$$\min_{\mathbf{p}_k} w_k \cdot D_k(\mathbf{p}_k) + \lambda \cdot R_k(\mathbf{p}_k) \quad \Leftrightarrow \quad \min_{\mathbf{p}_k} D_k(\mathbf{p}_k) + \lambda_k \cdot R_k(\mathbf{p}_k) \quad \text{with} \quad \lambda_k = \frac{\lambda}{w_k}. \tag{11}$$

In other words, for each individual block B_k, we get the same optimization problem as with the conventional SSE distortion; only the Lagrange multiplier λ_k is altered block-by-block. Still, we can exploit the main advantage of using the SSE distortion in the encoder control.

3.3 Selection of Quantization Parameter (QP)

In general, the quantization parameter QP_k for a block B_k can be considered as an element of the block decision parameter vector $\mathbf{p}_k$. However, for speeding up the encoding process, it is common practice to determine the quantization parameter QP_k for a block B_k a-priori. Then, the block decision process can be conducted given the Lagrange multiplier λ_k and the pre-selected quantization parameter QP_k. Using high-rate approximations for the SSE distortion and the operational rate-distortion curve, the relationship between the Lagrange multiplier λ_k and the associated quantization step size Δ_k can be derived as [13, 14]

$$\lambda_k \propto \Delta_k. \tag{12}$$

This relationship has also been verified experimentally [13, 14] and is used in virtually all modern image and video encoding algorithms. Since we have the approximate relationship

$$\Delta_k \propto 2^{\frac{QP_k}{6}} \tag{13}$$

in Advanced Video Coding (AVC) and its recent successors, High-Efficiency Video Coding (HEVC) [15] and Versatile Video Coding (VVC) [16], combining Eq. 12 and Eq. 13 yields

$$QP_k - \left\lfloor 3 \cdot \log_2 \lambda_k \right\rceil = \text{const}, \tag{14}$$

where $\lfloor \cdot \rceil$ means rounding (QP_k is an integer). Assuming that the overall multiplier λ for a picture is associated with the overall QP for said picture, we obtain the QP assignment rule

$$QP_k = QP' - \left\lfloor 3 \cdot \log_2 \frac{\lambda}{\lambda_k} \right\rceil = QP' - \left\lfloor 3 \cdot \log_2 w_k \right\rceil, \tag{15}$$

with QP' denoting the predefined overall QP. At this point, we note again that it is preferable to scale the weighting factors w_k in a way that their average across a picture, or a set of pictures or video frames, is close to 1. Then, the same relationship between the picture/set Lagrange parameter λ and the picture/set QP as for unweighted SSE distortion can be used.

3.4 Example for the Case of High-Rate Approximations

A simple experiment shall demonstrate the effect of weighted distortion metrics and scaled Lagrangian multiplies on a block basis. We consider five different quantizers, which should represent five independent entities (such as blocks B_k in a picture). Each quantizer provides six operation points (i. e., six rate-distortion points available for selection by an encoding algorithm). The corresponding rate-distortion curves for SSE and weighted SSE distortion are depicted in Fig. 1a and Fig. 1b, respectively. Further given are the sensitivity weighting factors $\{w_k\} = \{1, 0.3, 1.5, 0.4, 3.1\}$. Finally, overall rate-distortion plots for the measures D^m and D^w are provided in Fig. 1c and Fig. 1d, respectively.

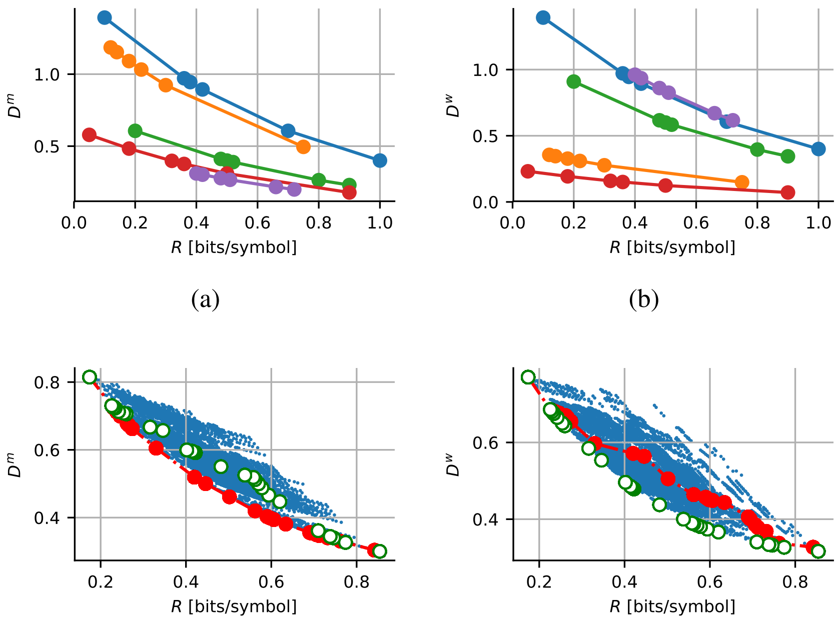
The blue points represent all rate-distortion points that can be achieved by selecting one of the available rate-distortion points for each of the five quantizers, resulting in $6^5 = 7776$ possibilities. Now, let us consider two variants of Lagrangian optimization: 1. Distortion measure D^m: $\min D_k + \lambda \cdot R_k$, and 2. Distortion measure D^w: $\min D_k + \lambda_k \cdot R_k$ with $\lambda_k = \frac{\lambda}{w_k}$. For varying the Lagrange multiplier λ, the first optimization (for the distortion measure D^m) yields the rate-distortion points that are marked with red circles. The second optimization (weighted distortion D^w) results in the rate-distortion points marked with green circles. The example clearly demonstrates that Lagrangian optimization based on D^w leads to a set of encoder decisions (and, thus, convex hull) that is distinct to an optimization based on D^m. It also verifies that unweighted D^m and weighted D^w are equally applicable for bit-allocation.

4. Implementation into Block-Based Codec

Section 3 outlined the relationship between the visual sensitivity measure w_k, the Lagrange parameter λ_k, and the quantization parameter QP_k for each picture block B_k at index k. In the following, we propose a simple measure for characterizing w_k which can be applied to calculate WPSNR values and to adapt the encoder control as described previously.

4.1 Luminance-Based Block Weighting Algorithm

It can frequently be observed that, when using the same QP for a complete picture, regions with rather smooth (low-frequency) content, i. e., low visual activity, are perceptually more



(a) (b)

(c) (d)

Figure 1: *Top row: Rate-distortion curves for five quantizers, each with six selectable rate-distortion points for (a) conventional SSE distortion measure, (b) weighted SSE distortion measure. Bottom row: Obtained rate-distortion points for (c) SSE, (d) weighted SSE. Small blue dots denote the set of all avilable rate-distortion points, while red points indicate rate-distortion points obtained by a fixed Lagrange parameter with SSE and green circles depict rate-distortion points obtained by a weighted Lagrange parameter and weighted SSE.*

degraded than regions with highly varying (high-frequency) content, i. e., high activity. To address this effect, we devise a very simple w_k based on a local measure of the mean visual activity of a range of uncoded initial picture samples $s[x,y]$ with $0 \le x < W$ and $0 \le y < H$, where W and H are the picture's width and height, respectively. Specifically, we first obtain a high-pass filtered version of the picture's luma channel using a discrete Laplace operator

$$\begin{aligned} h[x,y] = \tfrac{1}{4} \cdot (12 \cdot s[x,y] - 2 \cdot s[x-1,y] - 2 \cdot s[x+1,y] - 2 \cdot s[x,y-1] - 2 \cdot s[x,y+1] \\ - s[x-1,y-1] - s[x+1,y-1] - s[x-1,y+1] - s[x+1,y+1]) \end{aligned} \quad (16)$$

for $0 < x < W-1$ and $0 < y < H-1$, which can be implemented efficiently using a 9-tap fixed-point filter algorithm. For $x = 0$, $x = W-1$, $y = 0$, and/or $y = H-1$, the unavailable neighboring samples of s used in Eq. 16 are assumed to equal $s[x,y]$ for simplicity. Such a high-pass operation serves as a low-complexity approximation of the *difference of Gaussians* behavior of the human retina [2]. Then, the local visual activity a_k can be determined:

$$a_k = \max\left(a_{\min}^2; \left(\frac{1}{|B_k|} \cdot \sum_{(x,y) \in B_k} |h[x,y]|\right)^2\right), \quad a_{\min} = 2^{BD-6}, \quad (17)$$

where the bit-depth dependent $a_{\min}$ models the lower sensitivity limit of the HVS and $|B_k|$ holds the number of samples in block region B_k. The squared-mean in Eq. 17 was preferred over a mean-square design to simplify fixed-point implementations. Let us define w_k as the normalized inverse of a_k raised to the β-th power. Choosing $\beta = 0.5$ empirically, we obtain

$$w_k = \left(\frac{a_{\text{pic}}}{a_k}\right)^{0.5}, \quad a_{\text{pic}} = 2^{BD} \cdot \sqrt{\frac{3840 \cdot 2160}{W \cdot H}}, \tag{18}$$

where the picture size dependent normalization factor a_{pic} was determined experimentally for minimal differences between average PSNR and WPSNR values (or, in other words, an average of $w_k \approx 1$) on a large set of image and video material [17]. This implies that, when $a_k < a_{\text{pic}}$ (low activity), then $w_k > 1$ (high weight) and vice versa, as desired.

Note that the squares in Eq. 17 and the power-of-0.5 in Eq. 18 effectively cancel out, so when adjusting a_{pic} appropriately, low-complexity implemenations are possible. It is also worth repeating that with $\beta = 0$, we acquire $w_k = 1$, in which case the WPSNR reduces to the conventional PSNR and the encoder employs *fixed* QP values inside each picture block.

4.2 Extension to Chromatic Image Components

The block-wise sensitivity weights w_k obtained for the luminance channel according to the preceding subsection could also be applied to any chromatic channels during bit-allocation or WPSNR measurements. However, the authors observed that this approach can result in degraded subjective quality when the QPA algorithm is applied to images in which the a_k in the chroma components tend to be higher than those in the luma component. In this case, some chromatic blocks tend to be *overcoded* while an excessive amount of bit-rate is taken away from visually more salient luminance blocks, thus causing *undercoding*.

One solution to the aforementioned issue is the introduction of a perceptually adapted *luma-to-chroma* QP offset O_c for each available chromatic channel $c \in$ [Cb, Cr], defined as

$$O_c = \begin{cases} \left\lfloor 3\beta \cdot \log_2 \dfrac{\alpha \cdot a_c}{a_{\text{Y}}} \right\rceil, & \alpha \cdot a_c > a_{\text{Y}}, \\ 0, & \text{otherwise}, \end{cases} \qquad a_p = a_k \text{ of (17) with } B_p \text{ instead of } B_k, \tag{19}$$

for $QP_c = QP_{\text{Y}} + \min(O_{\max}; O_c)$, where Y specifies the luma component, $\beta = 0.5$ as before, α and $O_{\max}$ can be selected empirically (value 4 works well), $p \in$ [Y, Cb, Cr], and B_p stands for the component's complete picture of size $W_p \cdot H_p$ samples. In other words, O_c represents a (positive) delta-QP offset which, effectively, lowers the coding rate for the given c when its mean visual activity a_c is high in comparison to the luma channel's mean activity a_{Y}. Note that O_c could also be specified *block-wise*, but due to the limited acuity of the HVS to chromatic stimuli [2] and according to the results of our own informal experiments, the definition of a single (coarse) *picture-wise* O_c appears to be sufficient (as a welcome side-effect, it is also compliant with the QP signaling syntax in the HEVC and VVC standards). To summarize, applying Eq. 15 and Eq. 19 with $QP_{\text{Y}} = QP_k$ leads to a polychromatic QPA.

5. Objective and Subjective Evaluation

Objective and subjective experiments were conducted to assess the benefits of w_k-weighted distortion measurement and corresponding bit-allocation in the context of image and video coding. To this end, the block size $|B_k| = W_k \cdot H_k$ was set to a square of size $d/a_{\text{pic}} \cdot d/a_{\text{pic}}$ with $d = 2^{17}$, i. e, $64 \cdot 64$ for high-definition (HD) and $128 \cdot 128$ for ultra-high-definition (UHD) content. The other model parameters ($a_{\min}$, α, β, and $O_{\max}$) are constants as above.

5.1 Objective Evaluation of WPSNR Algorithm

Our perceptually weighted distortion metric, in the form of the WPSNR specification of Eq. 5 with $\beta = 0.5$, was compared with other well-known psychovisually motivated measures. To be specific, we determined the overall Spearman rank-order correlations between several measurement results and the mean opinion score (MOS) data collected in formal visual tests on the JPEG and JPEG 2000 distorted still-image sets of the LIVE database [18].

Analyzing the results of this study, we observe that the WPSNR metric achieves higher correlation with the subjective scores than the traditional PSNR measure (0.962 vs. 0.888, respectively, when averaged across the two distortion types), approaching the performance of the widely employed multiscale structural similarity measure (MS-SSIM, 0.972) [19]. Details on this investigation are published in [20]. In addition, the run-time of the WPSNR method was found to be an order of magnitude lower than that of a C-code MS-SSIM software implementation, thus indicating that the former exhibits a much lower computational complexity (the WPSNR run-time is only 2–3 times higher than that of PSNR).

5.2 Subjective Evaluation of QP Adaptation

The perceptual QPA design devised herein was evaluated via two comparative subjective A/B tests. The first test intended to assess the overall video coding quality of the proposal (called "WPSNR" in the following) in comparison to an alternative, MS-SSIM optimized QP adaptation (named "MS-SSIM" below) employing the same block sizes $|B_k|$ and spatio-temporal QP granularity. The second test assessed the visual advantage of the *WPSNR* QPA relative to coding optimized for unweighted PSNR-like distortion (i. e., "fixed-QP" coding). The bit-stream sizes of the *WPSNR* and *MS-SSIM* conditions were matched per-sequence, via QP', to those of the *fixed-QP* encodings, where a global base-QP of 37 was employed. The QP step-size for this rate matching was 0.5, meaning that a base-QP increase by 1 was allowed halfway through the encoding process. The underlying software selected for this investigation is version 1 of VTM [16], the reference software developed in the course of the current VVC standardization. Its maximum coding block size (the coding tree unit) is $128 \cdot 128$ luma samples, i. e., an integer multiple of $|B_k|$ for HD and UHD input.

The subjective evaluation was done using direct A/B comparisons largely following the ITU-R BT.500 methodology [21]. Twelve subjects experienced in detecting video coding artifacts participated in each test. Each subject claimed to possess normal color perception and took part in only one of the tests, but not both. The decoded bit-stream stimuli of codec configurations A and B (randomly assigned) were presented in a temporally concatenated "$G_A - A - G_B - B - G_A - A - G_B - B - G_{Vote}$" manner, where G stands for a one-second mid-gray separating picture labeled with a black letter "A" or "B" at its center, depending on which coding conditions follows, or in case of the G_{Vote}, the word "VOTE" to indicate the 10-second voting period. A color-calibrated 65-inch Sony XE93 (KD-65XE9305) TV set with UHD panel resolution was used for presentation, on which the HD sequences were shown centered in 1:1 (i. e., not upscaled) resolution on a mid-gray background.

The participants were asked to vote whether A or B had better overall coding quality (A better: –1, B better: 1) or whether both exhibited similar quality (0). Each subject's verdict was translated to the number in parentheses which, given the expected subtle differences between the coding variants, is interpreted directly as a difference opinion score (DOS). Averaging the DOS values results in a DMOS number for codec B relative to codec A [17].

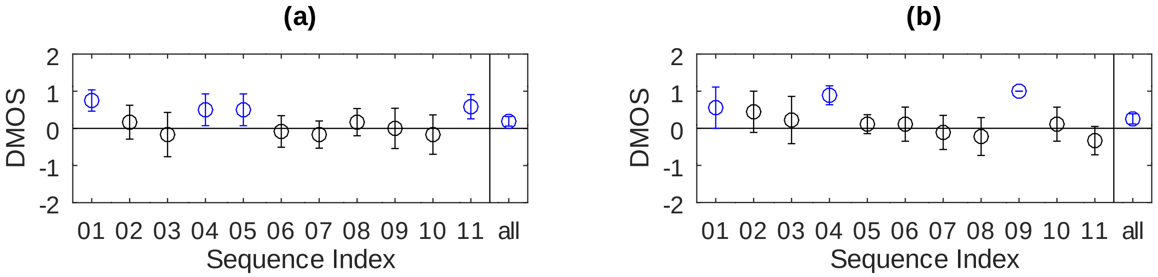


Figure 2: *Results of the subjective tests of QPA in the VTM-1 encoder: (a) WPSNR vs. MS-SSIM based, (b) WPSNR vs. PSNR based (12 experienced viewers, including three females). The video material used in this evaluation comprises five 10-second HD sequences and six 5–10-second UHD sequences, coded at a mean bit-rate of 2.8 Mbit/s (median: 2.3 Mbit/s). The PSNR reference bit-streams were encoded with a base-QP of 37. See [17] for details.*

Fig. 2 illustrates the outcome of the two subjective experiments. For each sequence and the overall average, the horizontal bars indicate the limits of the 95% confidence intervals associated with the given DMOS (circle), while colors indicate statistical significance. Surprisingly, our QPA based on the perceptual distortion weighting outperforms the reference QPA, controlled by the MS-SSIM algorithm, on four of the 11 video sequences (blue color) in terms of visual coding quality, while never yielding worse quality than the latter scheme. Compared with the *fixed-QP* coding modeled by the unweighted PSNR-like distortion, the QPA method improves the visual coding quality on three of the 11 sequences (particularly on sequence 9, the *ParkRunning* recording currently used in the VVC standardization work [17]), again without causing significant quality degradations. This leads us to conclude that the parametrization of the QPA and WPSNR algorithm is, psychovisually, well chosen.

6. Summary and Conclusion

This paper investigated the issue of objectively and subjectively motivated bit-allocation in modern image and video coding. It then presented a generalization of the underlying PSNR distortion measure by extending it with input-adaptive, psychovisually derived block-wise weights $w_k \neq 1$. The resulting weighted PSNR (WPSNR) was demonstrated to be equally applicable for encoder control as the conventional PSNR. In fact, the reported experimental results indicate that, with proper parameter selection, a quantization parameter adaptation (QPA) inside the encoder can lead to substantial improvements of visual coding quality on some content, while the weighted distortion metric can match (or even exceed, when used to control the QPA) the performance of other state-of-the-art perceptual distortion metrics such as MS-SSIM at a fraction of their algorithmic complexity. Thanks to these results, the WPSNR-based QPA algorithm was adopted into VTM, the currently developed reference software for the Versatile Video Coding (VVC) standard [16], in 2018.

The proposed QPA-controlled bit-allocation was shown to easily extend from luminant (grayscale) to chromatic (color) image and video coding. Further research could focus on an optimized parametrization of the weighted polychromatic distortion model, possibly by considering more spatiotemporal and/or data-driven aspects [22], as well as more thorough evaluations of the model on MOS-annotated image or video databases.

References

[1] W. Osberger, S. Hammond, and N. Bergmann, “An MPEG Encoder Incorporating Perceptually Based Quantisation,” in *Proc. IEEE Annual Conf. Speech and Image Technol. for Comput. and Telecomm. (TENCON)*, Brisbane, Dec. 1997, vol. 2, pp. 731–734.

[2] A. Valberg, *Light Vision Color*, Wiley, Mar. 2005.

[3] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, and G. J. Sullivan, “Rate-Constrained Coder Control and Comparison of Video Coding Standards,” *IEEE Trans. Circuits and Systems for Video Technol.*, vol. 13, no. 7, pp. 688–703, July 2003.

[4] V. V. Gohokar and V. N. Gohokar, “Optimum Bit Allocation Using Human Visual System for Video Compression,” in *Proc. IEEE Int. Conf. Comput. Intell. and Multim. Applic. (ICCIMA)*, Sivakasi, Dec. 2007, vol. 3, pp. 84–88.

[5] T. S. Ou, Y. Huang, and H. H. Chen, “SSIM-Based Perceptual Rate Control for Video Coding,” *IEEE Trans. Circuits and Systems for Video Technol.*, vol. 21, no. 5, pp. 682–691, May 2011.

[6] C. Yeo, H. L. Tan, and Y. H. Tan, “SSIM-Based Adaptive Quantization in HEVC,” in *Proc. IEEE Int. Conf. Acoust., Speech, Sig. Proc. (ICASSP)*, Vancouver, May 2013, pp. 1690–1694.

[7] W. W. Chao, Y. Y. Chen, and S. Y. Chien, “Perceptual HEVC/H.265 System with Local Just-Noticeable Difference Model,” in *Proc. IEEE Int. Symp. Circuits and Systems (ISCAS)*, Montreal, May 2016, pp. 2679–2682.

[8] F. Zhang and D. R. Bull, “HEVC Enhancement using Content-Based Local QP Selection,” in *Proc. IEEE Int. Conf. Image Proc. (ICIP)*, Phoenix, Sep. 2016, pp. 4215–4219.

[9] Z. Wei and K. N. Ngan, “Spatio-Temporal Just Noticeable Distortion Profile for Grey Scale Image/Video in DCT Domain,” *IEEE Trans. Circuits and Systems for Video Technol.*, vol. 19, no. 3, pp. 337–346, Mar. 2009.

[10] G. T. Fechner, *Elemente der Psychophysik. 2*, Breitkopf & Härtel, 1907.

[11] H. Everett, “Generalized Lagrange Multiplier Method for Solving Problems of Optimum Allocation of Resources,” *Operations Research*, vol. 11, no. 3, pp. 399–417, June 1963.

[12] Y. Shoham and A. Gersho, “Efficient Bit Allocation for an Arbitrary Set of Quantizers,” *IEEE Trans. Acoustics, Speech, and Signal Proc.*, vol. 36, no. 9, pp. 1445–1453, Sep. 1988.

[13] G. J. Sullivan and T. Wiegand, “Rate-Distortion Optimization for Video Compression,” *IEEE Signal Proc. Mag.*, vol. 15, no. 6, pp. 74–90, Nov. 1998.

[14] T. Wiegand and B. Girod, “Lagrange Multiplier Selection in Hybrid Video Coder Control,” in *Proc. IEEE Int. Conf. Image Proc. (ICIP)*, Thessaloniki, Oct. 2001, vol. 3, pp. 542–545.

[15] ITU, “Rec. H.265: High efficiency video coding,” 2018, `www.itu.int/rec/T-REC-H.265`.

[16] JVET, “VVCsoftware_vtm,” 2019, `vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM`.

[17] C. Helmrich, H. Schwarz, D. Marpe, and T. Wiegand, “Improved Perceptually Optimized QP Adaptation and Associated Distortion Measure,” *JVET document K0206*, July 2018, online at `phenix.it-sudparis.eu/jvet/doc_end_user/current_document.php?id=3715`.

[18] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, “LIVE Image Quality Assessment Database,” *rel. 2*, 2005, `live.ece.utexas.edu/research/Quality/subjective.htm`.

[19] Z. Wang, E. Simoncelli, and A. C. Bovik, “Multi-Scale Structural Similarity for Image Quality Assessment,” in *Proc. IEEE Asilomar Conf. Signals, Systems, and Comp.*, Pacific Grove, 2003.

[20] S. Bosse, C. Helmrich, H. Schwarz, D. Marpe, and T. Wiegand, “Perceptually Optimized QP Adaptation and Associated Distortion Measure,” *JVET document H0047*, Oct. 2017, online at `phenix.it-sudparis.eu/jvet/doc_end_user/current_document.php?id=3319`.

[21] ITU, “Methodology for the subjective assessment of the quality of television pictures,” 2012.

[22] S. Bosse, S. Becker, Z. V. Fisches, W. Samek, and T. Wiegand, “Neural Network-Based Estimation of Distortion Sensitivity for Image Quality Prediction,” in *Proc. IEEE Int. Conf. Image Proc. (ICIP)*, Athens, Oct. 2018, pp. 629–633.