Visually Optimized Two-Pass Rate Control for Video Coding Using the Low-Complexity XPSNR Model

Christian R. Helmrich, Ivan Zupancic, Jens Brandenburg, Valeri George, Adam Wieckowski, and Benjamin Bross

Video Communication and Applications Group, Fraunhofer Heinrich Hertz Institute (HHI), Einsteinufer 37, 10587 Berlin, Germany

***Abstract*—Two-pass rate control (RC) schemes have proven useful for generating low-bitrate video-on-demand or streaming catalogs. Visually optimized encoding particularly using latest-generation coding standards like Versatile Video Coding (VVC), however, is still a subject of intensive study. This paper describes the two-pass RC method integrated into version 1 of VVenC, an open VVC encoding software. The RC design is based on a novel two-step rate-quantization parameter (*R-QP*) model to derive the second-pass coding parameters, and it uses the low-complexity XPSNR visual distortion measure to provide numerically as well as visually stable, perceptually R-D optimized encoding results. Random-access evaluation experiments confirm the improved objective as well as subjective performance of our RC solution.**

***Keywords*—QoE, rate control, video coding, VoD, VQA, VVC**

I. Introduction

Compressed high-resolution video content is increasingly distributed to consumers through video-on-demand (VoD) IP based streaming platforms operated by relatively new Internet technology companies, and video archive web-sites hosted by traditional TV broadcasters. Coded video catalogs are usually generated off-line according to certain average and maximum instantaneous bit-rate constraints, and it was found that two-pass rate control (RC) schemes are, in terms of the achievable level of visual coding quality, the preferred means to enforce these constraints [1]. Two-pass RC solutions, when encoding a certain motion picture sequence or subset thereof, employ a first (typically fast) *analysis* pass in which preliminary frame-wise coding statistics are collected for the full range of frames followed by a second (typically slower) *final* pass performing actual rate-distortion (R-D) optimized picture encoding, with imposed rate constraints derived from the first-pass data.

Since the ultimate receiver of the compressed video signal catalogs is the human visual system, perceptually motivated visual quality assessment (VQA) methods are generally used to quantify the visual coding quality of the bit streams resulting from RC assisted encoding runs. Moreover, to achieve a predefined level of quality of experience (QoE) of the videos given the bit-rate constraints determined a-priori, the psycho-visual VQA models are frequently incorporated as *perceptual distortion* measures into the RC video encoders [2] – [5]. The objective is to achieve, across the given set of video frames, a consistent VQA score with little variance between frames [1], [2] as well as a *visually optimal* tradeoff between mean and/or maximum bit rate and average VQA score. Finally, as with all two-pass and single-pass RC methods, the rate resulting from final-pass encoding should closely match the target rate.

A. Related Work

In most RC methods, statistical models are used to formulate the relationship between the encoding parameters, e. g., the quantization parameter (*QP*) and Lagrange multiplier (λ), and the actual bit rate R. Given a second-pass target bit count r_f, determined for some frame f based on first-pass statistics, and the overall rate requirements, these encoding parameters are chosen using the RC model, either per picture (*frame-level* RC) or sub-area (*block-level* RC). One of the best performing models, especially with modern video coding standards such as High Efficiency Video Coding (HEVC) [7], [8] or Versatile Video Coding (VVC) [9], [10], is the R-λ model, where Lagrange value λ_f is calculated from r_f using a hyperbolic model:

$$\lambda_f = \alpha \cdot {r_f}^{\beta}, \tag{1}$$

with α and β being the model parameters [3] – [5], which are often refined over time as the final-pass encoding progresses. A similarly behaving quadratic model, yielding slightly better R-D performance, has recently been devised for Intra frames:

$$\lambda_f = -\frac{a \cdot \ln r_f + b}{r_f}, \tag{2}$$

where a and b are the continuously updated parameters [6].

B. Motivation

The RC models of (1) and (2), although performing well on typical input video material, were found to be difficult to integrate into modern video encoders in terms of stable choice of their two parameters for input sequences with strongly or rapidly varying R-D statistics. This behavior can be attributed to the fact that two parameters must be updated and stabilized simultaneously. In addition, the combination of two-pass RC and *direct* perceptual encoding optimization controlled by a low-complexity psycho-visual model has rarely been studied. In fact, the authors are only aware of the work by Wang *et al.* [2], [11] and Yuan *et al.* [12], utilizing the structural similarity measure (SSIM) [13] as a visually motivated distortion model for "perceptual R-D optimization" during two-pass RC video coding. The other works previously referenced in this section only aim for *indirect* visual optimization by adjusting the final RC pass to result in relatively constant per-frame *VQA scores*.

C. Paper Outline

This paper proposes two new approaches for two-pass RC operation in fast HEVC or VVC encoders. The first, described in Sec. II, is the usage of a simple and implementation friendly two-step *R-QP* model instead of the R-λ models (1) or (2), in order to simplify the encoder integration and to stabilize the RC behavior on "difficult" videos. The second, introduced in Sec. III, is the adoption of the low-complexity psycho-visual model of the XPSNR metric for perceptual R-D optimization during first and second-pass encoding as well as for improved scene cut detection. Sec. IV outlines the conducted evaluation experiments and their results, and Sec. V concludes the paper.

Submitted to *IEEE VCIP*, Munich.

II. A Two-Step *R-QP* Model for Two-Pass RC

An alternative to the *R-λ* model for RC is described in the following. For reasons of brevity, the variable bit rate (VBR) use case adopted in streaming applications is emphasized and mostly broadcasting centric constant bit rate (CBR) use cases, with their need to enforce stricter rate limits, are ignored.

The motivation for the development of an alternative, yet simple RC model describing the relationship between R and the encoding parameters is based on the observation that, with modern codecs like HEVC or VVC, the overall rate resulting from encoding with a given overall QP and λ increases considerably when QP and λ fall below a certain threshold. Above this threshold, however, the logarithm of the rate change appears to be almost linearly related to QP and λ. Table I illustrates this observation on CTC coding results [14] collected using the VVC reference software encoder, VTM 12.1 [15]. Note that the resulting bit rates decrease with increasing base QP (since the QP value is proportional to the quantizer's step size) and that, for the All Intra case, UHD sequence *Campfire* was excluded from the UHD results due to outlier behavior.

From Table I it can be observed that, when reducing the base QP from 37 to 32 or from 32 to 27, the change in rate is a relatively consistent factor between 1.7 and 1.9 for All Intra (AI) and between 1.9 and 2.1 for Random Access (RA) coded content. When reducing the base QP from 27 to 22, however, the rate increases faster, especially for UHD and HD coding. This growth in the rate change factor can be attributed to the more prominent presence of film grain or camera sensor noise in high-resolution video recordings – the closer the resolution is to the physical limits of the camera optics and acquisition device, the greater the amount of noise is in a given pixel area. Below a base QP of roughly 25, this noise leads to a greater variance in the video coder's prediction error and, thereby, to many more residual transform coefficients being quantized to nonzero, even when reducing the QP in relatively small steps.

A. *First Part: QP Derivation for Low Rates*

The observed stronger rate change at high target rates than at lower target rates leads to the conclusion that a RC model comprising *two* parts, with the second part applied only when the result of the first part falls below a threshold, is desirable. Hence, a two-step *R-QP* model with a corrective second step, simplified to be easily implementable in fixed-point arithmetic, was devised. The first part of that *R-QP* model is given by

$$QP'_f = QP_f - c_{\text{low}} \cdot \sqrt{\max\left(1;\ QP_f\right)} \cdot \log_2\left(\frac{r'_f}{r_f}\right), \qquad (3)$$

where QP_f and QP'_f are the integer first-pass and preliminary second-pass QP values, respectively, while r_f and r'_f hold the resulting first-pass and target second-pass bit counts, respectively, for frame f. A very close fit to the low-rate $QP = 27, 32$ RA data of Table I is obtained using $c_{\text{low}} \approx 0.82$, as tabulated in Table II for first-pass QP_{base} values of 32 and 37. It is worth noting, in this regard, that the choice of the base QP value for the first coding pass, providing the frame-wise QP_f, r_f values required for the final bit allocation in the second coding pass, affects both accuracy and runtime (i. e., complexity overhead) of the RC encoding process, since higher QP values decrease the encoder runtime over lower values. In initial experiments, a good tradeoff was reached by the following QP assignment:

$$QP_{\text{base}} = \text{round}\left(40 - \sqrt{\frac{3840 \cdot 2160}{W \cdot H} \cdot \frac{R_{\text{target}}}{500000}}\right), \qquad (4)$$

where R_{target} is the overall user-specified target bit rate in bps and W and H are the input video width and height, respectively. For typical target rates between about 400 kbps and 40 Mbps for UHD, or a quarter of these rates for HD content, (4) results in reasonably accurate first-pass base QPs between 39 and 31.

Table I. Rate ratios $R(QP)/R(QP{+}5)$ resulting from VVC encoding with given base QP, geometrically averaged for each CTC class [16].

Base QP	All Intra			Random Access		
	UHD (A½)	HD (B)	SD (C)	UHD (A½)	HD (B)	SD (C)
22	3.148	2.252	1.724	2.584	2.832	2.217
27	1.787	1.893	1.747	2.007	2.138	2.067
32	1.744	1.856	1.828	1.902	2.001	1.972

Table II. Accuracy of second-pass QP estimator of (3) for Random Access, with $c_{\text{low}} = 105/128$ and base QPs from Tab. I as ground truth.

Ground Truth	First-Pass $QP_{\text{base}} = 32$			First-Pass $QP_{\text{base}} = 37$		
	UHD (A½)	HD (B)	SD (C)	UHD (A½)	HD (B)	SD (C)
22	20.98	19.94	21.81	20.52	19.04	21.15
27	27.34	26.91	27.14	27.36	26.54	26.88
32	32.00	32.00	32.00	32.37	32.01	32.11
37	36.30	36.64	36.55	37.00	37.00	37.00

B. *Second Part: QP Correction for High Rates*

Table II illustrates that, when rounding the output of (3) to integer, the estimated base QPs match the ground-truth QP_{base} data from Table I for 17 out of the 18 low/medium-rate cases (mismatches are colored blue) and are, thus, very accurate. At the very high rates (top row), however, model (3) consistently underestimates the actual QP because of the steeper increase in rate noted earlier. Therefore, a second *corrective* part to the *R-QP* model is required. An easily implementable solution is

$$QP''_f = \text{round}\left(QP'_f + c_{\text{high}} \cdot \max\left(0;\ QP_{\text{start}} - QP'_f\right)\right), \qquad (5)$$

where QP_{start} is the correction threshold and $0 \le c_{\text{high}} \le 1$ serves as a parameter to control the strength of the correction. Figure 1 shows the effect of varying c_{high} when QP_{start} is kept constant. Using $QP_{\text{start}} = 24$ and with $c_{\text{high}} = 0.5$ for (U)HD and 0.25 for SD, QP 22 is now reached for all 6 high-rate cases in Table II.

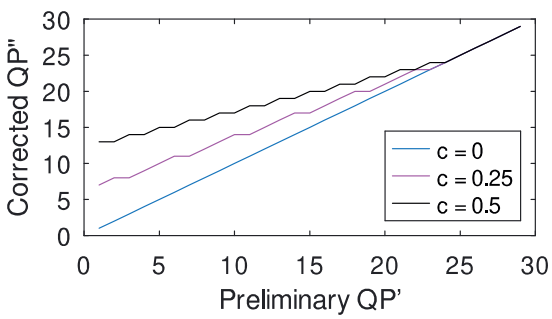

Figure 1. Effect of corrective *R-QP* step (5) on preliminary second-pass QP' obtained via (3). $QP_{\text{start}} = 24$ is fixed, c_{high} (abbr. c) is varied.

In summary, the above two-pass RC model is used as follows:

1. Perform a first-pass fixed-QP encoding, with a base QP of QP_{base} as specified by the empirically derived assignment rule of (4) and with frame-wise QP_f and λ_f values derived therefrom, e. g., as defined for RA encoding in [15], [16]. Store the resulting first-pass bit consumption r_f for each f.
2. Perform the second-pass varying-QP encoding with a preliminary QP'_{base} as specified in (6) which is then corrected similarly to QP'_f using (5), resulting in QP''_{base}. For each f, obtain QP''_f via (3) and (5) using the first-pass QP_f and r_f.
3. In the *R-λ* models, the frame-wise second-pass Lagrange values λ''_f are determined directly and the associated QP''_f values are derived therefrom. In the *R-QP* model, the QP''_f are determined directly, as in step 2, and the λ''_f values are derived, using the first-pass λ_f, as $\lambda''_f = \lambda_f \cdot 2^{(QP''_f - QP_f)/3}$. This *λ-QP* relationship is derived in detail in the appendix.

$$QP'_{\text{base}} = QP_{\text{base}} - c_{\text{low}} \cdot \sqrt{QP_{\text{base}}} \cdot \log_2\left(\frac{R_{\text{target}} \cdot F}{fps \cdot \sum_f r_f}\right), \quad (6)$$

where F is the total frame count and *fps* is the frame rate in Hz. The choice of r_f for each f is discussed further in Sec. IV.

III. Perceptual R-D Optimization Using XPSNR

In [17], a low-complexity VQA measure termed XPSNR, representing a generalization and extension of the traditional peak signal-to-noise ratio (PSNR), was proposed. The core of the XPSNR measure is a psycho-visually inspired, simplified spatiotemporal *sensitivity* model specified locally, for disjoint areas, or blocks B, of each input picture P of bit depth BD, as

$$w_k = \sqrt{\frac{\hat{a}_{\text{pic}}}{\hat{a}_k}} \quad \text{with} \quad \hat{a}_{\text{pic}} = 2^{2BD-9} \cdot \sqrt{\frac{3840 \cdot 2160}{W \cdot H}}, \quad (7)$$

where k is the block index and $\hat{a}_k$ is a *visual activity* given by

$$\hat{a}_k = \max\left(a^2_{\text{min}};\ \left(\frac{1}{4N^2}\sum_{[x,y]\in B_k} |h_{\text{s}}[x,y]| + 2|h_{\text{t}}[x,y]|\right)^2\right) \quad (8)$$

for each luminance-channel block $B_k \in P_f$. The definitions of a_{min}, h_{s}, and h_{t} are provided in [17] and omitted here for brevity and N^2 is the number of picture samples in each block. A low computational complexity is reached because the spatial high-pass operator h_{s} and temporal high-pass operator h_{t} use very simple fixed-point operations, and the square-root in the calculation of w_k cancels the squaring operations in (8) [18].

A. Perceptual R-D Optimization

The XPSNR model was explicitly specified to facilitate a simple implementation into modern image and video codecs.

$$N = \text{round}\left(128 \cdot \sqrt{\frac{W \cdot H}{3840 \cdot 2160}}\right), \quad (9)$$

in particular, was chosen to align the B with the largest coding blocks (coding tree block, CTB) so that all boundaries of each CTB are aligned with those of the collocated visual sensitivity blocks. In the case of VVC, this means that one and four w_k are calculated per CTB for UHD and HD input, respectively. This convenient parametrization allows for the w_k to be used as *perceptual weights* during bit allocation for R-D optimized encoding, as described in [19] and summarized hereafter.

Let distortion D_k be the sum of squared errors (SSE) or, as adopted more frequently, the mean squared error (MSE), with

$$\min_{\mathbf{p}_k} D_k(\mathbf{p}_k) + \lambda_f R_k(\mathbf{p}_k) \quad \text{for all } k \text{ in coding order} \quad (10)$$

constituting the (approximate) block-wise R-D encoding optimization problem regarding the block coding parameters $\mathbf{p}_k$, when any dependencies between blocks are ignored. Here, λ_f is an overall Lagrange multiplier for the given frame which is associated with that frame's quantization parameter QP_f. It is shown in [19] that (10) can be turned into a *perceptual* block-wise R-D optimization problem by local distortion weighting:

$$\min_{\mathbf{p}_k} w_k D_k(\mathbf{p}_k) + \lambda_f R_k(\mathbf{p}_k) \Leftrightarrow \min_{\mathbf{p}_k} D_k(\mathbf{p}_k) + \lambda_k R_k(\mathbf{p}_k) \quad (11)$$

with block-wise Lagrange multiplier $\lambda_k = \lambda_f / w_k$. Hence, (7) can be used *directly* to locally adapt λ_f and, thereby, QP_f, and this change is all that is required to achieve visually improved encoding; all other block coding operations can stay the same.

B. Combination with Two-Pass RC

Having shown how traditional block-wise R-D optimized encoding based on SSE or MSE can be generalized by means of block-wise weighting, the choice of Lagrange and quantization parameter for each k in each RC encoding pass remains to be made. Clearly, the XPSNR visual sensitivity weighting of (7), as a model of human vision, is a good choice for w_k in (11). It is, therefore, proposed to let the *first* RC encoding pass operate with the configuration outlined in [17], [19], namely:

- Perform first-pass R-D optimized encoding as in step 1 of Sec. II, but with the QP_f and λ_f parameters adapted per k (i. e., CTB, or quarter of CTB for HD or smaller input) as

$$QP_k = QP_f - \text{round}\big(3 \cdot \log_2(w_k)\big), \quad \lambda_k = \frac{\lambda_f}{w_k}. \quad (12)$$

The *second* RC coding pass can then be applied analogously:

- Perform the second-pass variable-QP encoding according to steps 2 and 3 of Sec. II, but with local adaptation of the QP''_f and λ''_f for each k. Since w_k is codec agnostic due to its sole dependence on the input images P, it follows that

$$QP''_k = QP''_f - \text{round}\big(3 \cdot \log_2(w_k)\big), \quad \lambda''_k = \frac{\lambda''_f}{w_k}. \quad (13)$$

Note that, aside from R-D optimization, w_k can be employed to detect sudden scene changes or camera switches, which is beneficial in RC coding since prior knowledge of changes in pixel value statistics can help stabilize the temporal RC parameter refinement (see also Sec. I and IV). The temporal high-pass component of the XPSNR model [17], [20], reflected by h_{t} in (8), causes an increase in visual activity on consecutive pictures P_{f-1}, P_f with significantly different content, such as at scene cuts. Defining a picture-wise mean luma visual activity

$$\hat{a}_f = \max\left(a^2_{\text{min}};\ \left(\frac{1}{4WH}\sum_{[x,y]\in P_f} |h_{\text{s}}[x,y]| + 2|h_{\text{t}}[x,y]|\right)^2\right) \quad (14)$$

for each frame f, instead of each block k, the simple condition

$$\hat{a}_f > 8 \cdot \hat{a}_{f-1} \tag{15}$$

accurately identifies cuts, i. e., f with changing characteristics.

IV. Implementation and Evaluation

The visually optimized two-pass RC method proposed in Secs. II and III was implemented into version 1.0.0 of VVenC, an open VVC encoder [21], and tested in RA configuration. A GOP[1] size of 32 and the non-normative temporal filtering tool was used, as in [15], [16]. XPSNR based visual QP adaptation (QPA), including the chroma extension of [19], was allowed.

A. RC Behavior in Second Coding Pass

After the first RC pass, the sets of collected frame coding statistics $\{QP, \lambda, r, \hat{a}\}_f$ are sorted in display order, and scene changes are detected according to Sec. III.B. Using a value of

$$c_{\text{high}} = \tfrac{1}{8} \cdot \max\big(0;\ \text{round}\big(\log_2 H\big) - 7\big), \tag{16}$$

which was found to be a better overall fit for HD input[2] than the constant mentioned in Sec. II.B, QP''_{base} is determined, and the second-pass per-frame target bit counts r'_f are initialized as

$$\hat{r}_f = \text{round}\left(r_f \cdot \frac{R_{\text{target}} \cdot F}{fps \cdot \sum_f r_f}\right), \tag{17}$$

where the ratio by which r_f is scaled is already available from (6). Then, before final encoding of each frame, $\hat{r}_f$ is refined to

$$r'_f = \max\left(1;\ \hat{r}_f + \left(\sum_{c \in C} \hat{r}_c - r''_c\right) \cdot d \cdot \frac{r_f}{g_f}\right) \tag{18}$$

to better match the target bit rate as the encoding progresses, where C is the set of all frames already encoded in the second pass and r''_c is the final bit consumption of each frame c in C. Constant d equals 1 for all f in the last encoded GOP, else 0.5. g_f is the sum of all bits r in the GOP to which f belongs. With (16)–(18), the combination of steps 2 and 3 of Sec. II and the visual optimization of Sec. III can now be applied; see function picInitRateControl() in file EncGOP.cpp of [21] for details.

B. Objective Evaluation (BD-Rate)

Bjøntegaard Delta-rate (BD-rate) statistics were compiled according to [22], using JVET's CTC sequences for SDR [16] (with the UHD coding extended to 10 seconds for more representative results) and Fraunhofer HHI's public *Berlin* test set [23]. Only VTM's single-pass RC could be tested for comparison since it is the only other open RC implementation compatible with GOP size 32 as of June 1, 2021 [24], [25]. VVenC was operated with and without perceptual optimization and in preset *slow* which, in terms of fixed-QP coding efficiency, is close to VTM [26]. For each of the three encoder conditions, the rates resulting from fixed-QP CTC-like coding were used as R_{target}, and rate accuracy was measured as in [6], [11].

The BD-rate results provided in Table III, averaged across YUV components and video class, illustrate that the two-pass RC proposal in VVenC approaches the efficiency of the fixed-QP reference quite closely, with or without QPA. The error in rate accuracy was just 0.5% on average. Using the single-pass RC method in VTM, in comparison, causes a notable BD-rate loss of 5–14 %. This, effectively, makes VVenC with RC outperform VTM with RC at a fraction of the runtime (9 – 10 %).

[1] group of pictures, [2] on larger training set excluding the CTC sequences

Table III. BD-rate results of RC methods. Timing relative to VTM.

Resolution	**VTM 12, no QPA**		**VVenC, no QPA**		**VVenC, vis. QPA**	
Class	PSNR	Runtime	PSNR	Runtime	XPSNR	Runtime
UHD A½	11.2%	97.7%	0.10%	9.60%	0.57%	9.61%
UHD HHI	9.13%	101%	2.30%	9.12%	4.30%	10.1%
HD B	6.31%	105%	0.54%	9.34%	1.36%	9.50%
HD HHI	14.0%	109%	1.72%	9.99%	2.47%	11.1%
SD C	4.66%	102%	0.23%	10.1%	0.51%	10.5%

C. Subjective Evaluation (Visual Testing)

According to formal visual tests conducted between September 2020 and May 2021, VVenC with activated perceptual optimization outperforms VTM in visual quality at moderate and high bit rates in RA configuration, while operating more than 100 times faster [27], [28]. An informal visual check was carried out by the present authors to assess whether updated VVenC encodings made with the two-pass RC proposal, rate matched to the VTM encodings, perform comparably. It was found that the visual quality advantage of VVenC over VTM could also be observed with the RC encodings, and no significant differences between VVenC with and without RC (when both variants exhibit the rates used in [27], [28]) were visible.

V. Summary and Conclusion

This paper introduced the two-pass rate control algorithm integrated into version 1.0 of VVenC, an open VVC encoder. It described the underlying two-step *R-QP* model, devised for simple yet numerically stable operation, as well as the combination with perceptually R-D optimized coding, for improved (i. e., visually stable) subjective quality. Experimental evaluation against VTM, the VVC reference encoder, confirmed the runtime and visual quality advantage of VVenC, without and with rate control enabled. Future work will focus on improvements to (16)–(18) and adaptations for single-pass operation.

Appendix: Relationship Between λ and QP

Using high-rate approximations, the relationship between λ and the respective quantization step size Δ can be derived as

$$\lambda \propto \Delta^2 \tag{A.1}$$

for all MPEG/ITU-T video codecs since AVC [29], which was verified experimentally. Having the approximate relationship

$$\Delta \propto 2^{QP/6} \tag{A.2}$$

and disregarding rounding, combining (A.1) and (A.2) yields

$$QP - 3 \cdot \log_2(\lambda) = \text{const}, \tag{A.3}$$

see also [19]. If the QP is varied by a delta dQP, it follows that

$$QP + dQP - 3 \cdot \log_2(\lambda \cdot d\lambda) = \text{const}, \tag{A.4}$$

i. e., additively changing the logarithm-domain QP requires a corresponding multiplicative change of the associated linear-domain Lagrange value. Solving (A.4) for $d\lambda$ via (A.3) yields

$$d\lambda = \tfrac{1}{\lambda} \cdot 2^{\frac{QP+dQP-\text{const}}{3}} = \tfrac{1}{\lambda} \cdot 2^{\frac{dQP}{3}+\log_2(\lambda)} = 2^{\frac{dQP}{3}}, \tag{A.5}$$

which completes the derivation of how λ must be scaled when QP is modified, in order for constraint (A.3) to stay enforced.

References

[1] V. P. Kumar M, K. C. Ravi, and S. Mahapatra, "A Novel Two Pass Rate Control Scheme for Variable Bit Rate Video Streaming," in *Proc. IEEE Int. Symposium Multimedia (ISM)*, Miami, pp. 140–143, Nov. 2015.

[2] S. Wang, A. Rehman, K. Zheng, and Z. Wang, "SSIM-Inspired Two-Pass Rate Control for High Efficiency Video Coding," in *Proc. IEEE Int. Workshop Multimedia Sig. Process. (MMSP)*, Xiamen, Oct. 2015.

[3] B. Li, H. Li, L. Li, and J. Zhang, "Lambda-Domain Rate Control Algorithm for High Efficiency Video Coding," *IEEE Trans. Image Process.*, vol. 23, no. 9, pp. 3841–3854, Sep. 2014.

[4] I. Zupancic, M. Naccari, M. Mrak, and E. Izquierdo, "Two-Pass Rate Control for Improved Quality of Experience in UHDTV Delivery," *IEEE J. Sel. Topics Sig. Process.*, vol. 11, no. 1, pp. 167–179, Feb. 2017.

[5] G. Cao, X. Pan, Y. Zhou, Y. Li, and Z. Chen, "Two-Pass Rate Control for Constant Quality in High Efficiency Video Coding," in *Proc. IEEE Visual Communic. and Image Process. (VCIP)*, Taichung, Dec. 2018.

[6] Y. Chen, S. Kwong, M. Zhou, S. Wang, G. Zhu, and Y. Wang, "Intra Frame Rate Control for Versatile Video Coding with Quadratic Rate-Distortion Modelling," in *Proc. IEEE Int. Conf. Acoustics, Speech, Sig. Process. (ICASSP)*, Barcelona/online, pp. 4422–4426, May 2020.

[7] ITU-T H.265 and ISO/IEC 23008-2, "High Efficiency Video Coding," Apr. 2013 (and subsequent ed.). https://www.itu.int/rec/T-REC-H.265

[8] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the High Efficiency Video Coding (HEVC) Standard," *IEEE Trans. Circuit Syst. for Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[9] ITU-T H.266 and ISO/IEC 23090-3, "Versatile Video Coding," Aug. 2020 (and subsequent editions). https://www.itu.int/rec/T-REC-H.266

[10] B. Bross, J. Chen, J.-R. Ohm, G. J. Sullivan, and Y.-K. Wang, "Developments in International Video Coding Standardization After AVC, With an Overview of Versatile Video Coding (VVC)," *Proc. IEEE*, Jan. 2021.

[11] Z. Wang, A. Rehman, K. Zheng, J. Wang, and Z. Wang, "SSIM-Motivated Two-Pass VBR Coding for HEVC," *IEEE Trans. Circuits Syst. for Video Technol.*, vol. 27, no. 10, pp. 2189–2203, Oct. 2017.

[12] H. Yuan, Q. Wang, Q. Liu, J. Huo, and P. Li, "Hybrid Distortion-Based Rate-Distortion Optimization and Rate Control for H.265/HEVC," *IEEE Trans. Consumer Electron.*, DOI TCE.2021.3065636, June 2021.

[13] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[14] F. Bossen, X. Li, K. Sühring, K. Sharman, and V. Seregin, "JVET AHG report: Test model software development," doc. JVET-V0003, Apr. '21.

[15] JVET and Fraunhofer HHI, "VVCSoftware_VTM," Gitlab repository, Apr. 2021. https://vcgit.hhi.fraunhofer.de/jvet/VVCSoftware_VTM

[16] F. Bossen, J. Boyce, X. Li, V. Seregin, and K. Sühring, "VTM common test conditions and software reference configurations for SDR video," doc. JVET-T2010, version 2. Nov. 2020. https://jvet-experts.org

[17] C. R. Helmrich, S. Bosse, H. Schwarz, D. Marpe, and T. Wiegand, "A Study of the Extended Perceptually Weighted Peak Signal-to-Noise Ratio (XPSNR) for Video Compression with Different Resolutions and Bit Depths," *ITU Journal: ICT Discoveries*, vol. 3, no. 1, May 2020. http://handle.itu.int/11.1002/pub/8153d78b-en

[18] J. Erfurt, C. R. Helmrich, S. Bosse, H. Schwarz, D. Marpe, and T. Wiegand, "A Study of the Perceptually Weighted Peak Signal-to-Noise Ratio (WPSNR) for Image Compression," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Taipei, pp. 2339–2343, Sep. 2019.

[19] C. R. Helmrich, S. Bosse, M. Siekmann, H. Schwarz, D. Marpe, and T. Wiegand, "Perceptually Optimized Bit-Allocation and Associated Distortion Measure for Block-Based Image or Video Coding," in *Proc. IEEE Data Compress. Conf. (DCC)*, Snowbird, pp. 172–181, Mar. 2019.

[20] C. R. Helmrich, M. Siekmann, S. Becker, S. Bosse, D. Marpe, and T. Wiegand, "XPSNR: A Low-Complexity Extension of the Perceptually Weighted Peak Signal-to-Noise Ratio for High-Resolution Video Quality Assessment," in *Proc. IEEE Int. Conf. Acoustics, Speech, Sig. Process. (ICASSP)*, Barcelona/online, pp. 2727–2731, May 2020.

[21] Fraunhofer HHI, "Fraunhofer Versatile Video Encoder (VVenC)," GitHub repository, May 2021. https://github.com/fraunhoferhhi/vvenc

[22] ITU-T HSTP-VID-WPOM and ISO/IEC TR 23002-8, "Working practices using objective metrics for evaluation of video coding efficiency experiments," 2021. https://www.itu.int/pub/T-TUT-ASC-2020-HSTP1

[23] B. Bross, H. Kirchhoffer, C. Bartnik, M. Palkow, and D. Marpe, "AHG4 Multiformat Berlin Test Sequences," doc. JVET-Q0791, Jan. 2020.

[24] Y. Li, Z. Liu, Z. Chen, and S. Liu, "Rate Control for Versatile Video Coding," in *Proc. IEEE Int. Conf. Image Process.*, Abu Dhabi, Sep. 2020.

[25] F. Liu, Z. Liu, Y. Li, and Z. Chen, "AHG10: Extension of RC to support RA configuration with GOP size of 32," doc. JVET-T0062, Oct 2020.

[26] J. Brandenburg, A. Wieckowski, A. Henkel, B. Bross, and D. Marpe, "Pareto-Optimized Coding Configurations for VVenC, a Fast and Efficient VVC Encoder," in *Proc. IEEE Int. Workshop on Multimedia Sig. Process. (MMSP)*, Tampere, Oct. 2021.

[27] M. Wien and V. Baroncini, "Report on VVC compression performance verification testing in the SDR UHD Random Access category," doc. JVET-T0097, Oct. 2020.

[28] V. Baroncini and M. Wien, "Report on VVC compression performance verification testing in the SDR HD Random Access, SDR HD Low Delay, and 360 category," doc. JVET-V0174, Apr. 2021.

[29] G. J. Sullivan and T. Wiegand, "Rate-Distortion Optimization for Video Compression," *IEEE Signal Process. Mag.*, vol. 15, no. 6, Nov. 1998.