FAST CONSTANT-QUALITY VIDEO ENCODING USING VVENC WITH RATE CAPPING BASED ON PRE-ANALYSIS STATISTICS

Christian R. Helmrich, Valeri George, Vignesh V Menon, Adam Wieckowski, Benjamin Bross, and Detlev Marpe

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

ABSTRACT

VVenC, an open Versatile Video Coding (VVC) encoder, has recently been equipped with rate capping functionality in its two-pass rate control modes, providing constrained variable bitrate coding governed by target rate and maximum rate parameters. This paper reports on implementations and evaluation results of straightforward extensions to VVenC which enable the use of the maximum rate parameter also in the single-pass fixed-QP modes, controlled by a base quantization parameter (QP) instead of a target rate. The rate capping in the fixed-QP mode is achieved, with sufficient accuracy, by evaluating only already calculated pre-processing statistics, thereby avoiding increases in encoder runtime. This encoding mode, given that it supports visual quality optimizations such as XPSNR based block-wise perceptual QP adaptation, can be considered a rate capped constant-quality mode, which was missing in VVenC and which is an interesting configuration for video streaming.

***Index Terms*—**QoE, rate control, VBR, video coding, VoD, VVC

1. INTRODUCTION

Encoding digital video content for online streaming or social media platforms, at different resolutions and bitrates to serve a large range of specific customer needs and achieve overall high quality of experience (QoE), is a frequent task. To satisfy the need for an openly available software encoder generating bitstreams compliant with the Versatile Video Coding (VVC) standard finalized in 2020 [1], Fraunhofer HHI developed and published VVenC, whose source code is available on GitHub [2]. During the four years of its development, VVenC has been equipped with a screen content detector, optimizing VVenC's encoding process for fast yet efficient performance on screen sharing or online gaming videos [3], a scene cut detector, and sequence or lookahead based two-pass rate control (RC) operation to allow the encoding process to be governed by a target rate R_{target} instead of the usual, and less intuitive, quantization parameter QP_{base} [4–6]. Note that the use of a QP_{base} parameter in conjunction with VVenC's block-wise perceptual QP adaptation based on the XPSNR psychovisual model [7] represents a "constant visual quality" encoding mode and, as such, QP_{base} may be considered a constant quality factor (CQF) for scaling of the resulting bitrate or, equivalently, overall level of visual quality for a video (or segment thereof) to be compressed.

In the QoE centric use case mentioned above, two-pass RC coding may not be necessary or desired since *1)* the additional pre-encoding pass increases the runtime of the encoding tasks [4, 5], *2)* consistent overall visual quality is regarded more important than precise matching of R_{target} per encoding task. The enforcement of an overall *rate limit* even during QP_{base} driven non-RC operation, however, remains a desirable feature since such "rate capped CQF" encoding can, with proper choices of the two parameters QP_{base} and maximum rate R_{max}, minimize the occurrence of playback stutter and data rebuffering phases on the decoder (i. e. consumer) side and, thus, maximize QoE.

1.1. Contribution of This Paper

In 2023, VVenC's two-pass RC modes have been enhanced to support the declaration of a maximum instantaneous rate parameter during encoder configuration, and constrained variable bitrate (CVBR) encoding is being triggered when such an (optional) instantaneous-rate limit is being specified [6]. The frequently updated, accurate rate estimation required for such an encoding process to function satisfactorily is based on the rate and QP statistics obtained during both the first (pre-encoding) and second (encoding) pass, i. e., a relatively large amount of actual coding information is available to take appropriate rate related decisions. In a fast single-pass "rate capped CQF" encoding mode, where an additional pre-encoding pass is neither present nor desired, this is not the case, and stronger compromises will need to be made. However, as noted earlier, precise rate matching is often deemed unnecessary in such scenarios.

In this paper, a CVBR CQF mode, based on a "fixed-QP" operation with XPSNR based perceptual QP adaptation on the CTU level [8], is described. The statistics required for the rate capping enforced before starting the *single* encoding process for a group of pictures (GOP) are exclusively derived from a set of picture statistics already determined in VVenC for other purposes, rendering the computational complexity of the rate capping related algorithm near-zero. In fact, as will be shown, with realistic parametrization of R_{max}, VVenC's encoding time actually decreases in capped CQF mode, due to restrictions to high QPs on high-activity scenes with large motion residuals.

1.2. Outline of This Document

The remainder of this paper is organized as follows. Related work on the subject is summarized in Sec. 2. The pre-analysis stage currently implemented in VVenC is examined in Sec. 3, focusing on the XPSNR [7] and MCTPF [9] data to be reused for rate capping. Sec. 4 describes the rate estimation process devised for the rate capping, separated into an Intra-frame and Inter-frame part. Sec. 5 continues with implementational details of the capped CQF mode, and the results of an evaluation experiment conducted to assess the performance of this mode in terms of Bjøntegaard delta-rate, encoder runtime, and rate accuracy are reported in Sec. 6. To conclude the paper, Sec. 7 discusses and summarizes the findings in the earlier sections.

2. REVIEW OF RELATED WORK

The individual components of the rate capped CQF proposal discussed herein—namely, the single-pass operation, reuse of previously calculated statistics, CVBR operation, constant video quality, and perceptual optimization—have already been discussed in the scientific literature but, to the authors' knowledge, only separately, not in full combination. Specifically, a review of recent publications related to random access (RA), instead of Intra-frame or still-picture, RC coding reveals that

- *Cao et al.* [10], for example, describe constant-quality encoding for HEVC based on GOP-level and frame-level QP allocation. However, their solution requires two encoding passes and does not enforce any maximum-rate constraint. *Tan et al.* [11] discuss a single-pass counterpart based on a "texture/non-texture" block classification in the context of AVC, with the same lack of any maximum-rate restriction.
- *Zhou et al.* [12], on the other hand, present a Lagrange multiplier method based single-pass quality control approach for the VVC standard (as a replacement for a RC solution) that, however, optimizes for mean squared error (MSE) or PSNR instead of a psychovisually motivated measure and, like [10], does not enforce any maximum-rate constraints. The same PSNR notion of "quality" is adopted by *Blestel et al.* [13], who propose a rate capped design wherein the rate capping aspect is, however, outlined only very briefly.
- *Xu et al.* [14] address the combination of RC and consistent visual quality by introducing a new objective visual quality metric for encoder control, but how to adopt their work from AVC to the modern VVC standard appears nontrivial to the present authors. More recent studies targeting VVC specifically, published in 2020 or later and including [15] or excluding [16] psychovisually motivated bit allocation, however, again do not enforce maximum-rate constraints.

Also, all of the above-mentioned prior works [10–16] have in common that they do not pay particular attention to, and thus do not evaluate, the computational complexity of the RC and rate capping related algorithms and their effect on the encoder runtime. The latter aspect, however, is of primary importance in the context of VVenC, especially with *fast* encoder presets.

To complete this section, further reports already published in [6] shall be repeated in the present context. In [13], *Blestel et al.* propose a constant quality control (CQC) scheme. However, that work controls an HEVC encoding run by enforcing GOP-wise average and maximum *distortion* constraints, upon which R_{target} and R_{max} then depend. A direct limitation of the encoding *rate* to R_{max} is, therefore, not the scope of that study, although, arguably, customers are, likely, much more familiar with the usage of maximum-rate values than distortion or QP values when configuring a typical video encoder. Considering CVBR, *Lin et al.* [17] present a RC for VVC, particularly for 360° video, but the description of the Intra-frame rate capping used therein remains vague and indirect: the proposed system "constrains the frame-level QP of each Intra frame." Similarly vague on rate limitation are the investigations of *Menon et al.* [18, 19], capping every GOP-level QP (or constant rate factor, CRF, as called therein) at a c_{max} or b_{max} during the second pass.

One may, therefore, conclude from the above review that the description of a rate capping, perceptually optimized CQF encoding solution with carefully designed efficiency-runtime tradeoff and single-pass operation has not been published yet.

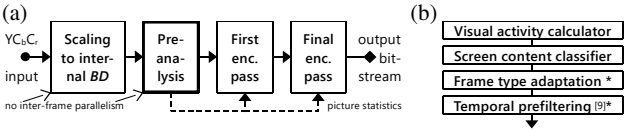

Figure 1. Processing stages in VVenC. (a) Location of pre-analysis stage in signal path, (b) components of pre-analysis stage in order of input signal processing, * possible restriction to low temporal levels.

3. VVENC'S PRE-ANALYSIS STAGE

The single-pass CVBR mode illuminated hereafter, as already established in Sec. 1, intends to utilize for all rate capping related estimates and calculations only such picture and coding data which are readily available in VVenC, i.e., determined in the context of other encoder decisions or optimizations during a regular non-RC encoding pass. Aside from *output* statistics such as mean QP, bit count, and MSE or PSNR, obtained for every *compressed* frame, at index f, written to the bitstream in *coding* order, these are *input* statistics such as spatiotemporal visual activity, motion strength, and noise power information, acquired from each *uncompressed* original frame picture P_f in *display* order. The common processing stage in VVenC where the input statistics are being calculated, for use by subsequent algorithms during rate-distortion optimized (RDO) encoding, is a pre-analysis (or pre-processing) stage, runtime-optimized by way of single-instruction multiple-data (SIMD) intrinsics and executed directly after picture data read-in before all other encoding stages. Figures 1(a) and 1(b) depict the location and consitution, respectively, of the pre-analysis stage in VVenC.

The picture-wise statistics determined during pre-analysis which are relevant to the present rate capping and CQF study are outlined in the following four paragraphs. A discussion of other obtained data, used primarily for scene cut detection and frame type adaptation [5] as well as screen content detection and coding tool selection [3], is omitted for reasons of brevity.

Visual activity calculation. For each component (Y and if available, Cb & Cr chroma) of each P_f, a mean spatiotemporal visual activity value is being obtained depending on the width W_c, height H_c, and bit depth D_c of the respective component c:

$$\hat{a}_{c,f} = \max\left(a_{\text{min}}^2; \left(\frac{1}{4W_cH_c}\sum_{[x,y]\in P_{c,f}}|h_{\text{s}}[x,y]| + 2|h_{\text{t}}[x,y]|\right)^2\right) \quad (1)$$

with $a_{\text{min}} = 2^{D_c-6}$ denoting a lower activity limit and h_{s}, h_{t} being the output sample values of spatial and temporal, respectively, high-pass filters operating on the input picture samples in $P_{c,f}$. Note that the constant factor 2 in front of term $|\,h_{\text{t}}[\,]\,|$ and limit a_{min} were chosen experimentally [7, 8] and that the individual averaged absolute high-pass outputs can be stored separately:

$$\hat{a}_{c,f} = \max\left(a_{\text{min}}^2; \left(s_{c,f} + t_{c,f}\right)^2\right) \text{ with } s_{c,f} = \frac{\sum_{[x,y]\in P_{c,f}}|h_{\text{s}}[x,y]|}{4W_cH_c} \quad \text{and } t_{c,f} = \frac{\sum_{[x,y]\in P_{c,f}}|h_{\text{t}}[x,y]|}{2W_cH_c} \quad (2)$$

and the high-pass filter kernels being resolution dependent, as in [7]. Details shall be omitted here for reasons of brevity; the only aspects worth mentioning are *1)* the spatial high-pass s is an approximation of a 2-dimensional, psychovisually inspired Laplacian-of-Gaussian filter with 9 taps [20] operating on 2×2 downsampled picture data in case of UHD resolution [7], and *2)* the temporal high-pass t is a sample-wise one-dimensional first or second-order filter without motion compensation [21], also applied on spatially downsampled input in case of UHD.

Temporal prefiltering. Again for each picture component c, motion compensated temporal prefiltering (MCTPF) [9] is applied for *denoising* purposes to those input pictures P_f being referenced the most in the motion compensated "inter-picture" prediction process during the RDO encoding pass. As shown in Fig. 1(b), this is the last step of the pre-analysis—or, in that case, pre-processing—stage since it modifies the input values of often-referenced, i.e., low-temporal-level, P_f while keeping the remaining P_f unchanged. It comprises a motion estimation (ME) *analysis* and a bilateral *filtering* part [22], where the ME step is of particular interest in the present study as it acquires motion information on the sequence of P_f. More specifically, a spatially hierarchical motion search is being conducted on a luma block level, starting at a coarse 1:4 downsampled picture level and ending at a fine, 16th-pel fractional-resolution stage. At each resolution level of the search "pyramid" and for each neighboring (in display order) reference picture P_r, a distance

$$d_{r,f} = \frac{1}{W_B H_B} \sum_{[x,y]\in(B\in P_{Y,f})} \left(p_f[x,y] - p_r[x + v_{r,x}, y + v_{r,y}]\right)^2, \quad (3)$$

representing the average of, in block B, the squared difference, or *motion error,* between the current-frame picture samples p_f and the "motion aligned" reference-frame samples p_r, is being calculated. The ultimate task of the motion search is to obtain, for each B and reference index r, a *motion vector* $[v_{r,x}, v_{r,y}]$ at the highest supported fractional resolution that minimizes (3). In VVenC, the block widths W_B and heights H_B depend on the resolution, being set to 16 on HD-or-larger videos and 8 otherwise and, if required, reduced at the bottom-right boundaries.

Having found all such motion vector data, the bilateral filter can then compensate for the block-wise P_f-to-P_r motion to improve the denoising performance, i. e., to minimize picture blurring and/or blocking artifacts during MCTPF. Reduction of random picture components especially in the luma channel without introduction of other types of distortion is essential to a successful, i. e., encoding performance improving, prefilter method (temporally uncorrelated signals exhibit high entropy and are, thus, hard to compress using hybrid predictive codecs like VVC). However, MCTPF is computationally quite complex, so to speed up especially VVenC's *fast*(*er*) presets, it is applied to fewer frames at low than high rates [9]. Moreover, the maximum distance N of a reference picture, at r, to frame f (i. e., $P_r = P_{f-N}, \ldots, P_{f-1}, P_{f+1}, \ldots, P_{f+N}$) is also preset dependent, with $N \le 4$ for the *fast*(*er*) and $N \le 6$ for the *slow*(*er*) presets.

In principle, when sufficiently large ME search spaces can be assumed, motion error $d_{r,f}$ in (3) represents a good estimate of the expected residual signal energy in block B during RDO encoding or, in other words, of the variance of the temporally uncorrelated—and, thus, unpredictable—quasi-random video content in B not related to motion, texture, or structure [3, 5]. Note that up to $2N$ motion errors are calculated per block, one for each reference picture P_r. Finding the smallest of these $d_{r,f}$ values per-block and saving that as a *minimum motion estimation error*, $MMEE_k$, where k equals the block index, was found to be a useful feature for screen content coding and RC [3, 5].

4. FAST RATE ESTIMATION FROM QP_{BASE}

The first objective of this work is to estimate the instantaneous bit consumption, when measured across each Intra-picture (I-frame) period, solely from the user-defined QP_{base} and a proper subset of the input statistics obtained in the pre-analysis stage.

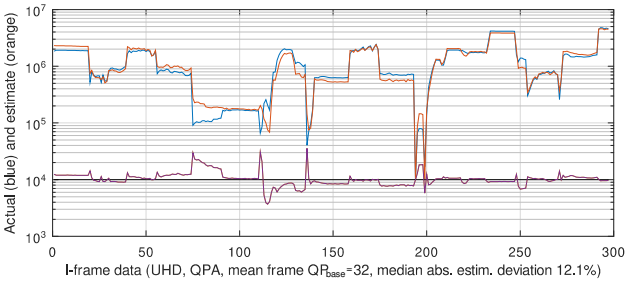

Figure 2. Comparison of (–) actual I-frame bit counts and (–) visual activity based estimates thereof for various concatenated UHD video sequences encoded using VVenC preset *medium* with QPA and QP_{base} 32. (–) Ratio between estimated and actual bit counts, scaled by 10^4.

In other words, the first (pre-)encoding pass used by VVenC's RC modes, as indicated in Fig. 1(a), is to be avoided, and no further computationally intensive operations shall be added to the pre-analysis stage, in order to minimize the runtime overhead during capped CQF encoding relative to fixed-QP (i. e., noncapped CQF) encoding. During preliminary experiments, it was concluded that *1)* a GOP-wise assessment of the coding bit consumption appears to be a good tradeoff between latency (due to frame lookahead needed for the picture analysis) and instantaneous-rate estimation accuracy, which is independent from the—possibly very large or small—I-frame period, *2)* a separate estimation of the bit counts for I-frames and all other (non-Intra-only) frames in every GOP, with subsequent accumulation, improves the overall performance of the estimation and, thereby, of the rate accuracy during CQF capping. Thus, individual estimators were devised for Intra and Inter (non-I) coded pictures within each GOP g, as described hereafter.

4.1. Intra-Picture Rate Prediction

Let $P_g = [P_{G\cdot g}, P_{G\cdot g+1}, \ldots, P_{G\cdot g+G-1}]$ be the set of pictures belonging to GOP g of size G, where $G = 32$ is adopted in this study, and let $P_{G\cdot g}$ be the *key frame* having the lowest temporal hierarchy layer $l_{G\cdot g} = 0$ in g and, as such, being encoded first in g. When $P_{G\cdot g}$ is an Intra-only coded I-frame, a typical hierarchical frame QP cascade, as used by the Joint Video Experts Team (JVET) during standardization work, assigns to $QP_{G\cdot g}$ some $QP_I = QP_{base} + q_I$ where, e. g., $q_I = -2$ with HEVC's $G = 16$ and $q_I = -3$ with VVC's $G = 32$ [23]. Thus, given a user-defined overall QP_{base}, the task of the proposed I-frame rate estimator is to obtain an approximation $\bar{b}_I$ of the total frame bit count b_I resulting from Intra-only RDO encoding of all components of $P_{G\cdot g}$ using $QP_{G\cdot g}$.

Recently, estimates of bit consumptions or rates produced by video encoders, given specific encoding parameters, have focused on machine learning solutions; see, e. g., the investigations by *Menon et al.* [18, 19, 24] and the literature overview therein. Pursuing a minimum-complexity estimator, this work resorts to traditional polynomial-function based prediction of $b_{G\cdot g}$ from $QP_{G\cdot g} = QP_I$ and the component-wise picture-average visual activity data $\hat{a}_{c,G\cdot g}$ determined in VVenC's pre-analysis step. In fact, it was found that a polynomial fitting of the form

$$\bar{b}_f = \max\left(10^4;\ \mathrm{round}\left(\delta + \beta_Y \cdot s_{Y,f}^{\alpha_Y} + \beta_C \cdot \left(\frac{s_{Cb,f}+s_{Cr,f}}{2}\right)^{\alpha_C}\right)\right), \quad (4)$$

where both exponents can be kept constant at $\alpha_Y = \alpha_C = \frac{4}{3}$ and the offset δ and scalars β_Y, β_C vary with $QP_{G\cdot g}$ and video resolution, yields satisfactory estimates of $b_{G\cdot g}$. Figure 2 illustrates the performance of bit count approximator (4) on UHD input.

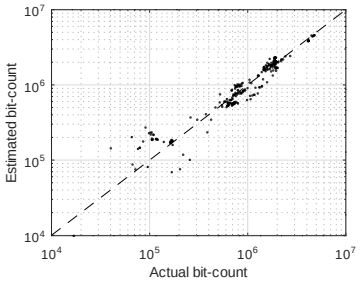

Figure 3. Scatter-plot representation of I-frame data shown in Fig. 2.

Note that, in (4), subscripts Y and C identify the luma and chroma channels, respectively, and s is the nonsquared spatial activity for each channel c and (I-)frame $f = G{\cdot}g$, as defined in (2). The $\bar{b}_{G\cdot g} = \bar{b}_{\mathrm{I}}$ estimates in Fig. 2 appear to be more accurate in the upper (high-rate) than in the lower (low-rate) data range which is sufficient since rate capping will be triggered only on hard-to-compress input typically requiring lots of coding bits. This observation is confirmed by a scatter-plot representation of the data in Fig. 2, depicted in Fig. 3. For the $QP_{\mathrm{base}} = 32$ data shown in both figures, the coefficient of determination equals $R^2 \approx 0.9$, which is in-line with the machine-learning results for low *max_depth*s in [24] obtained for the same estimation task and which is considered a very acceptable prediction accuracy given that (4) exhibits virtually zero complexity overhead.

Note also that, in commonly employed QP ranges between 24 and 40, VVC I-frame (non-RA) encodings roughly double in bitrate when reducing QP_{I} by about 5.8 [4][1]. Furthermore, a proportionality could be identified between the bit counts for different resolutions, which can be approximated quite closely by the 3rd root of the squared ratio of pixel counts of the given (e. g., HD) and of UHD resolution (3840·2160). Thus, the calculation of variables δ, β_{Y}, β_{C} in (4) from empirical data can be simplified by *1)* functional fitting to UHD bit counts resulting from a single QP, e. g. $QP_{\mathrm{base}} = 32$ and *2)* scaling, during a particular evaluation of (4), all above-noted three variables by

$$factor_{\mathrm{I}} = 2^{\frac{QP_{\mathrm{base}} - QP_{\mathrm{I}}}{5.8}} \cdot \left(\frac{W_{\mathrm{Y}} \cdot H_{\mathrm{Y}}}{3840 \cdot 2160}\right)^{\frac{2}{3}} \approx 2^{\frac{(32 - QP_{\mathrm{I}}) \cdot 11}{64}} \cdot \left(\frac{W_{\mathrm{Y}} \cdot H_{\mathrm{Y}}}{3840 \cdot 2160}\right)^{\frac{2}{3}} \quad (5)$$

when $QP_{\mathrm{I}} \neq 32$, with $W_{\mathrm{Y}} \cdot H_{\mathrm{Y}}$ specifying the luma video size as already applied in (1, 2) and the resolution ratio resembling the inverse of ratio D_{I} already in use in VVenC's two-pass RC [5].

With (5) realizing the desired dependency of (4) on the QP_{I} and video resolution, an accurate parametrization of the three variables in (4), obtained via polynomial regression fitting of I-frame data of a diverse set of UHD video input [23, 25–27] encoded with $QP_{\mathrm{base}} = 32$, $q_{\mathrm{I}} = 0$ (i.e. $QP_{\mathrm{I}} = QP_{\mathrm{base}}$), is given by

$$\delta = 1.5 \cdot 2^{18} \cdot factor_{\mathrm{I}},\ \beta_{\mathrm{Y}} = 173 \cdot factor_{\mathrm{I}},\ \beta_{\mathrm{C}} = 35 \cdot factor_{\mathrm{I}}. \quad (6)$$

Note that, since only three parameters are fitted to billions of video pixels, the authors are convinced that data *overfitting*, a key aspect in machine learning applications, is impossible to occur in the present study. For this reason and to increase the amount of high-quality video content available for data fitting via (4), the six UHD sequences defined in the JVET common test conditions (CTC) [23] and Fraunhofer's *Berlin* sequences [26], downsampled to UHD, were included in the calculations (i. e., the "training" set), despite the fact that these videos are part of the experimental evaluation (i.e., the "test" set) as well.

[1] In Tab. I of [4], $R(QP)/R(QP{+}5)$ averages at 1.81 for All Intra, $QP > 22$.

4.2. Inter-Picture Rate Prediction

Having predicted the bit consumption of the Intra-only coded, independently decodable I-frame (i. e., RA point) in each Intra period, an estimator for the remaining bits in each Intra period (i. e., the total bit count of all frames at $f \neq G{\cdot}g$ supporting Inter coding using motion compensation) remains to be developed. During preliminary experiments, the following was observed:

- On most rate capped video content requiring high bitrates, the I-frame bit count b_{I} is lower than the sum of the non-I frame bit counts per Intra period. Hence, good Inter-frame rate prediction is as crucial as good Intra-frame estimation.
- Inaccurate GOP-wise Inter-frame rate predictions could be improved fast using actual past-GOP bit allocation results, since at least $G{-}1$ non-I-frames are encoded in every GOP.
- The per-scene rate consumption of non-I-frames correlates with b_{I}, i. e., scenes with high spatial visual activity s^2 often consume relatively many bits in Inter-frames as well. This can be explained by the fact that, statistically, picture input with substantial high-frequency (structure, texture, noise) content likely results in significant motion residual energy and, thereby, time and rate intensive residual quantization.

Given these findings, a simple and minimum-complexity, QP_{I} driven estimator of the Inter-frame bit counts $b_{\text{non-I}} = b_{f \neq G\cdot g}$ resulting from Inter-predicted RDO encoding in each GOP g was devised which *1)* averages all available block-wise $MMEE_k$ of all frames f associated with g, as produced by the MCTPF (see Sec. 3), *2)* scales the square of the resulting $MMEE_g$ by $2^{-3D_{\mathrm{Y}}/2}$ to attain a certain value range, and *3)* multiplies that result by

$$factor_{\text{non-I},g} = m_{\mathrm{I}} \cdot m_{\mathrm{a},g} = \frac{4 \cdot \bar{b}_{\mathrm{I}} \cdot (I{-}1)}{\sqrt{\max_c(s_{c,G\cdot g})}} \cdot \left(\frac{(I{-}1) \cdot \sum_{i=1}^{N_g} b_{\text{non-I},g-i}}{(G{-}1) \cdot \sum_{i=1}^{N_g} \frac{\bar{b}_{\text{non-I},g-i}}{m_{\mathrm{a},g-i}}}\right)^{\varepsilon} \quad (7)$$

to reach the desired dependency on spatial activity and QP_{I} in g and the number I of frames in the Intra period. Note that the dependency on QP_{base} is achieved through QP_{I}. To reduce the effect of scene cuts on the output of (7) and, thereby, improve the Inter-rate estimator, $MMEE_g$ is attenuated by ½ in each g in which a scene cut occurs on one of the non-I-frames. Constant 4 and the square-root in (7) were determined experimentally, and N_g used in adaptation factor $m_{\mathrm{a},g}$ counts the number of previously encoded GOPs. For stability, N_g and the two accumulators in $m_{\mathrm{a},g}$ are zeroed out in GOPs with a scene cut, and exponent ε, instead of being fixed at $\varepsilon = 1$, set to $\varepsilon = N_g/(1 + N_g)$.

The result of the multiplication of the $MMEE_g$ data by (7) is the per-GOP approximation $\bar{b}_{\text{non-I},g}$ of the total Inter-coding bit count of the Intra period containing g. On the typical 32 ± 8 QP range noted in Sec. 4.1, the coefficient of determination is, for such an estimator, lower than that for the I-frame estimator when $N_g = 0$ ($R^2 \approx 0.8$) but reaches a similar R^2 level when N_g increases. Hence, its performance is regarded sufficient for an adequately reliable anticipation of rate spikes in most videos.

5. IMPLEMENTATION INTO VVENC

Having determined, before RDO encoding a GOP g, the Intra-period bit consumption estimate $\bar{b}_g = \bar{b}_{\mathrm{I}} + \bar{b}_{\text{non-I},g}$ and maximum allowed per-I bit count $b_{\max} = R_{\max} \cdot I/fps$, where fps is the video frame rate, rate capping in VVenC's CQF encoding mode can be realized using the *R-QP* model applied in the RC modes [4]:

$$q_{\mathrm{a},g} = \min\left(63; \mathrm{round}\left(\begin{matrix}\text{model}\\ \text{result for}\\ QP_{\mathrm{I}} - q_{\mathrm{I}}\\ b_{\max}/\bar{b}_g\end{matrix}\right) - QP_{\mathrm{I}} + q_{\mathrm{I}}\right) \text{ if } \bar{b}_g > b_{\max},\ 0 \text{ otherwise}, \quad (8)$$

where offset $q_{\mathrm{a},g}$ is added to all block QPs in g before RDO encoding.

Table 1. $R_{\max}$ in Mbps resulting from (11) for experimental values of QP_{base} and different video resolutions. Results for QP_{base} = 23, 28, typical defaults [30], are listed for information only and not used herein.

Resolution	QP_{base} **22**	QP_{base} 23	QP_{base} **27**	QP_{base} 28	QP_{base} **32**	QP_{base} **37**
4K, 2160p	42.01	36.03	19.50	16.72	9.051	4.201
HD, 1080p	14.00	12.01	6.499	5.574	3.017	1.400
SD, <1080p	3.500	3.002	1.624	1.393	0.754	0.350

The combination of rate capping and perceptual QPA can be realized by *1)* obtaining an initial QPA adjusted I-frame QP

$$QP'_{\text{I}} = QP'_{G\cdot g} = QP_{\text{base}} + q_{\text{I}} + q_{\text{aux}} + \text{round}\left(3 \cdot \log_2 \sqrt{\frac{\hat{a}_{\text{Y},G\cdot g}}{\hat{a}_{\text{pic}}}}\right), \quad (9)$$

where q_{aux} is an auxiliary, coding tool (e. g. BIM) induced offset, $\hat{a}_{\text{Y},G\cdot g}$ is given by (1, 2), $\hat{a}_{\text{pic}} = 2^{2D_{\text{Y}}-9} \cdot \sqrt{(3840 \cdot 2160)/(W_{\text{Y}} \cdot H_{\text{Y}})}$ [7], *2)* estimating therefrom QPA related GOP bit count $\bar{b}'_g$, *3)* obtaining $b_{\max}$ as well as $QP'_{\text{I}} - q_{\text{I}}$, and *4)* proceeding as above.

6. EXPERIMENTAL EVALUATION

The rate capping method devised in Secs. 4 and 5 was implemented into the mid-May 2024 source revision of VVenC [28] and its effect on the encoding performance and rate allocation was assessed by means of Bjøntegaard delta-rate (BDR) [29] and bitstream size comparisons, respectively. In addition, the runtimes of CQF encoding tasks with vs. without rate capping were evaluated. To this end, the JVET videos of classes A–C [23], extended to 10 sec duration in case of UHD resolution to reflect typical use cases, and the HHI *Berlin* videos [26] were encoded using VVenC's preset *slow* and perceptual QPA as in Sec. 5, taking the values 22, 27, 32, 37 listed in [23] as QP_{base}. All other encoder settings were left at their defaults or, when applicable, set equal to the CTC options for RA, except for *I*, which was changed from ≈1 sec to 4 sec, a more popular value in video streaming applications. Note that in the class-B CTC sequences with scene changes, intermediate I-frames may be added by VVenC's frame type adaptation [5]. It is also worth repeating that two-pass RC is not used in this evaluation and that, due to activated perceptual QPA, QP_{base} represents a CQF.

For the BDR measurements, sequence-wise XPSNR values [7] were used since all encoders under test employ perceptual optimization. Changes in encoder runtime were quantified by calculating per-bitstream encoding time ratios *ETR*, defined as

$$ETR = \frac{\text{duration of CVBR encoding at } QP_{\text{base}} \text{ with rate capping}}{\text{duration of VBR encoding at } QP_{\text{base}} \text{ without rate capping}}, \quad (10)$$

and averaging the four QP_{base}-wise time ratios for every video. The rate capping accuracy was assessed by deriving the actual rate R_{bs} from a bitstream's size and *fps*, obtaining ratio $R_{\text{bs}}/R_{\max}$ per-bitstream, and averaging the four $R_{\dots}$ ratios for each video.

6.1. Selection of $R_{\max}$ and Reference Encoder

In order to avoid continuous rate capping—and, thereby, near-constant bitrate (CBR) behavior—for certain combinations of QP_{base} and $R_{\max}$, the following QP_{base} and sequence resolution dependent, empirical definition of $R_{\max}$ (in kbps) was devised:

$$R_{\max} = \left\lfloor m_{\text{res}} \cdot 2^{15.323 \,-\, QP_{\text{base}}/4.516} \right\rfloor, \; m_{\text{res}} \in \{2.5, 10, 30\} \quad (11)$$

with resolution factor $m_{\text{res}} = 2.5$ for SD ($H_{\text{Y}} \leq 540$), 10 for HD ($H_{\text{Y}} \leq 1080$), and 30 for UHD videos ($H_{\text{Y}} > 1080$). This yields, for each sequence of the test set, a reasonable combination of the CQF and maximum-rate parameter, as listed in Table 1.

Table 2. XPSNR based BDR and rate results for UHD test content.

UHD Class	Luma	Chroma		Average	Mean *ETR*, $R_{\text{bs}}/R_{\max}$	
Sequence	BDR_{Y}	BDR_{U}	BDR_{V}	BDR_{YUV}	VVenC [28]	*x265 3.5*
CTC, A1 Tango4K	–0.17%	0.13%	0.09%	–0.08%	98, 60%	*95, 65%*
CTC, A1 FoodMarket	0.00%	0.00%	0.00%	0.00%	99, 45%	*98, 49%*
CTC, A1 Campfire	0.38%	–0.44%	0.53%	0.20%	91, 53%	*96, 72%*
CTC, A2 CatRobot	0.00%	0.00%	0.00%	0.00%	96, 52%	*98, 58%*
CTC, A2 DaylightRd.	0.12%	0.87%	0.72%	0.29%	94, 38%	*97, 41%*
CTC, A2 ParkRunning	2.77%	0.86%	0.96%	2.51%	**84**, 96%	*93, 99%*
Fraunhofer HHI, *Berlin* BCrossroads	0.40%	0.04%	0.15%	0.39%	94, 34%	*97, 30%*
Fraunhofer HHI, *Berlin* ChestnutTr.	**4.12%**	–0.30%	1.19%	**3.45%**	**74**, **103%**	*88, **104%***
Fraunhofer HHI, *Berlin* March18thS.	0.45%	0.28%	0.58%	0.53%	95, 62%	*96, 63%*
Fraunhofer HHI, *Berlin* NeptuneFnt.	3.30%	0.31%	–0.48%	2.99%	89, **103%**	*93, **104%***
Fraunhofer HHI, *Berlin* Oberbaum	0.00%	0.00%	0.00%	0.00%	101, 21%	*100, 19%*
Fraunhofer HHI, *Berlin* Quadriga	0.95%	0.24%	0.42%	0.82%	99, 35%	*98, 32%*
Fraunhofer HHI, *Berlin* ReichstagTr.	0.79%	0.01%	–0.50%	0.88%	97, 90%	*94, 91%*
Fraunhofer HHI, *Berlin* Spree	2.34%	–1.26%	–1.23%	1.90%	97, **113%**	*86, **104%***
Overall UHD	**1.10%**	**0.05%**	**0.17%**	**0.99%**	**93**, **64%**	***95**, **67%***

The rate capping performance is judged best by comparing it against the behavior of a popular alternative realization in a different encoder because, as noted earlier, the implementation in VVenC described here represents a certain low-complexity compromise from which highest accuracy operation cannot be expected. As such alternative rate capping "reference", x265 version 3.5, an open HEVC encoder in FFmpeg [30], was used in this study, with option *-tune ssim -crf* $QP_{\text{base}} + q_{\text{vid}}$ to produce CQF-like, visually optimized encodings with bitrates similar to those of the respective VVC encodings created by VVenC. Base QP modifier $-3 \leq q_{\text{vid}} \leq 3$ was determined experimentally for each video, and as with VVenC, preset *slow*, a max. I-frame interval (*-keyint*) of 4 sec, and the $R_{\max}$ of Tab. 1 were utilized.

6.2. Results of the Experimental Evaluation

Comparisons of the CVBR encodings to x265 or to same-QP VVenC encodings without rate capping, summarized in Table 2 for UHD and Table 3 for HD and SD resolution, reveal that

- BDR losses due to the rate capping remain small and confined to those sequences being rate constrained most (e.g., *Berlin* HD and UHD sequences *ChestnutTree* and *Spree*). This is especially the case when applying 6:1:1 averaging of the per-component XPSNR values [29], as in BDR_{YUV}.
- the VVenC rate capping accuracy, reflected by the number of videos for which the mean $R_{\text{bs}}/R_{\max}$ exceeds 1, matches that of x265, except on *Berlin* video *Spree*, showing a wavy river. Future work will focus on such content in particular.
- the VVenC runtime does not increase due to rate capping; it actually *decreases* overall since a higher frame QP (and, thereby, coarser and faster block-residual quantization) is often used on time consuming "difficult to encode" scenes with large motion residuals. Prominent examples are UHD video *ChestnutTr.* and HD video *Spree*, encoded ≈¼ faster. On HD or SD input, x265 is slowed down by rate capping.

Given the above observations and the fact that the devised $\bar{b}_g$ estimator of Sec. 4 avoids an additional pre-encoding pass, it can be concluded that the desired fast, single-pass, CQF-like, rate capped RA encoding mode for VVenC has been achieved. Note that recent versions of x265 and most AV1 encoders also utilize MCTPF-like pre-filters, so the present work can likely be integrated in and evaluated on top of these encoders as well.

7. SUMMARY AND CONCLUSION

This paper presented a low-complexity extension of VVenC's visually optimized, QPA enhanced single-pass mode allowing constant-quality encoding in a CVBR fashion by introducing a GOP-wise rate capping procedure similar to the one recently integrated into VVenC's two-pass rate control modes [6]. The rate capping leverages spatial visual activity and motion error related picture statistics already calculated as part of VVenC's pre-analysis/pre-processing stage to anticipate with sufficient accuracy, as shown in Secs. 4–6, the GOP-wise instantaneous bit consumption resulting from RD optimal encoding with the initially determined perceptually motivated frame QPs, which is required for the derivation of the GOP-wise QP constraints.

The results of an experimental evaluation indicate that the developed *capped CQF* mode exhibits a rate limitation performance similar to that of x265, a popular HEVC encoder, while avoiding significant BD-rate losses and simultaneously speeding up VVenC by raising the frame QPs in "difficult" scenes.

Table 3. XPSNR based BDR and rate results for HD and SD content.

HD/SD Class		Luma	Chroma		Average	Mean *ETR*, R_{bs}/R_{max}	
	Sequence	BDR_Y	BDR_U	BDR_V	BDR_{YUV}	VVenC [28]	*x265 3.5*
CTC, B	MarketPlace	0.60%	–1.61%	–2.34%	0.13%	93, 72%	*106, 64%*
	RitualDance	0.68%	–0.95%	–0.85%	0.37%	95, 81%	***113**, 81%*
	Cactus	0.00%	0.00%	0.00%	0.00%	96, 49%	*104, 56%*
	BasketballD.	0.00%	0.00%	0.00%	0.00%	101, 60%	*107, 63%*
	BQTerrace	0.00%	0.00%	0.00%	0.00%	101, 38%	*105, 45%*
Fraunhofer HHI, *Berlin*	BCrossroads	0.00%	0.00%	0.00%	0.00%	98, 20%	*102, 18%*
	ChestnutTr.	**4.01%**	1.38%	0.20%	**3.49%**	**81**, **109%**	***84**, **104%***
	March18thS.	0.00%	0.00%	0.00%	0.00%	101, 56%	*104, 55%*
	NeptuneFnt.	0.98%	–0.26%	–0.57%	0.85%	91, **103%**	*102, **103%***
	Oberbaum	0.00%	0.00%	0.00%	0.00%	98, 20%	*103, 21%*
	Quadriga	0.00%	0.00%	0.00%	0.00%	97, 13%	*99, 14%*
	ReichstagTr.	0.00%	0.00%	0.00%	0.00%	97, 88%	*100, 89%*
	Spree	**8.89%**	2.76%	3.89%	**8.17%**	**74**, **129%**	***80**, **104%***
C	CTC, C (SD)	1.10%	–0.22%	–0.24%	0.81%	99, 89%	***152**, 90%*
Overall HD		**1.16%**	**0.10%**	**0.03%**	**1.00%**	**94**, **64%**	***101**, **63%***

8. REFERENCES

[1] B. Bross, Y.-K. Wang, Y. Ye, S. Liu, J. Chen, G. J. Sullivan, and J.-R. Ohm, "Overview of the Versatile Video Coding (VVC) Standard and Its Applications," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 3736–3764, Oct. 2021.

[2] A. Wieckowski, J. Brandenburg, T. Hinz, C. Bartnik, V. George, G. Hege, C. R. Helmrich, A. Henkel, C. Lehmann, C. Stoffers, I. Zupancic, B. Bross, and D. Marpe, "VVenC: An Open and Optimized VVC Encoder Implementation," in *Proc. IEEE ICME*, virtual/online, July 2021.

[3] C. R. Helmrich, A. Henkel, T. Hinz, A. Wieckowski, B. Bross, and D. Marpe, "Finalization of VVenC's Screen Content Detector and Two-Pass Rate Control Using Pre-Filtering Statistics," in *Proc. IEEE Int. Conf. Image Process (ICIP)*, Kuala Lumpur, Oct. 2023.

[4] C. R. Helmrich, I. Zupancic, J. Brandenburg, V. George, A. Wieckowski, and B. Bross, "Visually Optimized Two-Pass Rate Control for Video Coding Using the Low-Complexity XPSNR Model," in *Proc. IEEE Int. Conf. Visual Communic. & Image Process. (VCIP)*, Munich, Dec. 2021.

[5] C. R. Helmrich, C. Bartnik, J. Brandenburg, V. George, T. Hinz, C. Lehmann, I. Zupancic, A. Wieckowski, B. Bross, and D. Marpe, "A Scene Change and Noise Aware Rate Control Method for VVenC, an Open VVC Encoder Implementation," in *Proc. IEEE PCS*, San Jose, Dec.'22.

[6] C. R. Helmrich, C. Bartnik, J. Brandenburg, A. Wieckowski, B. Bross, and D. Marpe, "A Constrained Variable Bitrate (CVBR) Algorithm for VVenC, an Open VVC Encoder Implementation," in *Proc. IEEE Int. Conf. Visual Communic. & Image Process. (VCIP)*, Jeju, Dec. 2023.

[7] 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.

[8] 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.

[9] A. Wieckowski, T. Hinz, C. R. Helmrich, B. Bross, and D. Marpe, "An Optimized Temporal Filter Implementation for Practical Applications," in *Proc. IEEE Picture Coding Symposium (PCS)*, San Jose, Dec. 2022.

[10] 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 Int. Conf. Visual Commun. & Image Process. (VCIP)*, Taichung, Dec. 2018.

[11] Y. H. Tan, C. Yeo, and Z. Li, "Single-Pass Rate Control With Texture and Non-Texture Rate-Distortion Models," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 8, pp. 1236–1245, Aug. 2012.

[12] M. Zhou, X. Wei, C. Ji, T. Xiang, and B. Fang, "Optimum Quality Control for Versatile Video Coding," *IEEE Trans. Broadcast.*, vol. 68, no. 3, pp. 582–593, Sep. 2022.

[13] M. Blestel, J. Le Tanou, and M. Ropert, "Constant Quality Control Based on Temporal Distortion Backpropagation in HEVC," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Athens, Oct. 2018.

[14] L. Xu, S. Li, K. N. Ngan, and L. Ma, "Consistent Visual Quality Control in Video Coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 6, pp. 975–989, June 2013.

[15] M. Wang, S. Wang, J. Li, L. Zhang, Y. Wang, and S. Ma, "SSIM Motivated Quality Control for Versatile Video Coding," in *Proc. APSIPA Ann. Summit & Conf.*, Auckland, pp. 1122–1127, Dec. 2020.

[16] Y. Li, Z. Liu, Z. Chen, and S. Liu, "Rate Control for Versatile Video Coding," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Abu Dhabi, pp. 1176–1180, Sep. 2020.

[17] Y-H. Lin, C.-Y. Chen, and C.-W. Tang, "VVC Based Rate Control Using SKIP CTU Predictor," in *Proc. IEEE ICCE-Asia*, Yeosu, Oct. 2022.

[18] V. V Menon, H. Amirpour, M. Ghanbari, and C. Timmerer, "ETPS: Efficient Two-Pass Encoding Scheme for Adaptive Live Streaming," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Bordeaux, Oct. 2022.

[19] V. V Menon, P. T Rajendran, C. Feldmann, K. Schoeffmann, M. Ghanbari, and C. Timmerer, "JND-aware Two-pass Per-title Encoding Scheme for Adaptive Live Streaming," *IEEE Trans. Circuits Syst. Video Technol.*, vol. TBD, July 1, 2023. DOI: 10.1109/TCSVT.2023.3290725.

[20] C. R. Helmrich, H. Schwarz, D. Marpe, and T. Wiegand, "Improved perceptually optimized QP adaptation and associated distortion measure," *document JVET-K0206*, Ljubljana, July 2018. https://jvet-experts.org.

[21] 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. Acoust., Speech, Sig. Process. (ICASSP)*, virtual/online, May 2020. http://www.ecodis.de/xpsnr.htm.

[22] J. Enhorn, R. Sjöberg, and P. Wennersten, "A Temporal Pre-Filter for Video Coding Based on Bilateral Filtering," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Abu Dhabi, pp. 1161–1165, Sep. 2020.

[23] F. Bossen, X. Li, K. Sharman, V. Seregin, and K. Sühring, "VTM and HM common test conditions and software reference configurations for SDR 4:2:0 10-bit video," *document JVET-AB2010*, teleconf., Dec. 2022.

[24] V. V Menon, A. Henkel, P. T Rajendran, C. R. Helmrich, A. Wieckowski, B. Bross, C. Timmerer, and D. Marpe, "All-Intra Rate Control Using Low Complexity Video Features for Versatile Video Coding," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Kuala Lumpur, Oct. 2023.

[25] Netflix, "Netflix Open Content," 2022. https://opencontent.netflix.com.

[26] B. Bross, H. Kirchhoffer, C. Bartnik, M. Palkow, and D. Marpe, "AHG4 Multiformat Berlin Test Sequences," *document JVET-Q0791*, Brussels, Jan. 2020. https://jvet-experts.org/ or https://hhi.fraunhofer.de/8kberlin.

[27] Blender Foundation, "Tears of Steel", open movie, 2.35:1 aspect ratio, Creative Commons Attrib., 2012. https://mango.blender.org/download.

[28] Fraunhofer HHI, "Fraunhofer Versatile Video Encoder (VVenC)," commit https://github.com/fraunhoferhhi/vvenc/commit/e4bb8, May 2024.

[29] 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

[30] MulticoreWare, "x265 HEVC Encoder," 2024. https://www.x265.org or libx265 as part of FFmpeg, https://trac.ffmpeg.org/wiki/Encode/H.265.