A VVC anchor for the CVPR 2022 CLIC video track

Pierrick Philippe and Théo Ladune

Abstract

In 2022 the CVPR Challenge for Learned Image Compression includes a video track which targets to explore technologies for the compression of HD video sequences. The proposed technologies are evaluated through a subjective test at two operating points: 100 kb/s and 1 Mb/s.

This contribution proposes to generate coded videos compliant with the latest standardized video coder, Versatile Video Coding (VVC). The primary objective of this candidate is to assess the recent developments in video coding with respect to this standard to measure the progress made by learning based techniques. To this end, this paper explains how to generate video sequences fulfilling the requirements of this challenge, in a reproducible way, targeting the maximum performance for VVC.

1. Introduction

From the 1990s standardization bodies, ISO and ITU-T, have defined several video coding standards [1]. Advanced Video Coding (AVC) was finalized in 2003 followed by HEVC (High Efficiency Video Coding) in 2013 and VVC (Versatile Video Coding) was recently released in 2020.

From a generation to an other it is targeted, among additional functionalities, to reduce the bit-rate by a factor of two for an equivalent subjective quality. HEVC has effectively proven to halve the bit-rate compared to AVC. VVC also demonstrates 50% bit-rate savings compared to HEVC [2]. ITU/MPEG standards have consistency shown that they represent the state of the art in terms of image quality. VVC, its latest technology is therefore considered as the flagship of the standardized solutions. In the context of the Challenge on Learned Image Compression (CLIC) it is therefore important to consider the level of performance of this last iteration of video coding standards.

The ITU/MPEG Joint Video Experts Team (JVET) is currently pursuing its technology investigations. Currently, neural approaches are explored to further enhance VVC compression capability. Three major areas are addressed in JVET:

- Enhancement or replacement of VVC coding tools with neural networks (intra prediction, transforms, frame prediction...)
- Introduction of filtering operations such as inloop filters or addition of super resolution as a post processing
- End-to-end solutions that do not rely on VVC

For those areas, VVC is used as a reference to quantify the amount of improvement provided by the proposed technologies. Currently, the range of improvement in the course of this activity is around 10% of bit-rate savings.

In order to monitor the performance of emerging technologies, it is also important for the industry to have appropriate anchors in the course of the CLIC challenges. In this paper, we strive to provide adapted VVC configurations for the CLIC video track at both target rates: 100 kb/s and 1 Mb/s.

In order to do so, the first section provides a brief overview of the general strategy used to generate anchors using the VVC coding standard.

As the challenge targets two bit-rates with a significant gap, the adopted coding strategy is made adaptive. The approach used is presented in a second section.

A last section summarizes the coding results and provides some additional statistics.

2. Adaptation of the VVC coding configuration to the challenge requirements

This section explains the coding strategy used to provide VVC coded items for the CLIC challenge. For an introduction on the Versatile Video Coding standard, the reader should refer on [1] to have an overview of VVC and its development phase. Also, in 2021, VVC was also contributed as anchors for the CLIC challenge, more details can be found in [3].

This year, the challenge addresses the compression of HD sequences (mostly 1920x1080 pixels) at 100 kb/s and 1 Mb/s. In the validation phase, 30 sequences, 10 seconds each, are proposed to exercise the coding tools in the anticipation of the test set (not know at the time of writing this paper).

The limit for the submission size is set to 300 seconds times the rate target which translates to respectively to 3,750,000 and 37,500,000 bytes for 100 kb/s and 1 Mb/s.

Given this overall limit, the objective of the challenge is to maximize the quality of the videos, in a subjective fashion.

2.1. Subjective optimization

The winners will be chosen based on a human rating task for this challenge. A subjective assessment is to be organized for this challenge to rate the candidate submissions. This is a significant difference compared to the CLIC 2021 challenge, as last year the MS-SSIM was used. Since there is no recommended metric this year, it is up to the proponents to select the metric they feel appropriate toward the subjective assessment.

In the course of the development of the VVC reference software (referred to as VTM, for VVC Test Model) a perceptually driven optimization method, quantization parameter adaptation (QPA), was proposed [4, 5]. The quantization adaptation strives to optimize a weighted PSNR metric, called XPSNR. Given the correlation of the XPSNR metric with numerous subjective assessments [6], this metric is selected in this paper in order to optimize the perceptual quality.

Note that this optimization method was already chosen last year for the generation of VVC anchors [3] : the results of last year's challenge using this metric was shown consistent with the MS-SSIM objective.

XPSNR is a weighted PSNR motivated by the fact that in compressed videos coding artifacts are often only perceivable in specific regions. Distorsion in highly textured areas are less visible than in low-contrast regions. Consequently, the mean squared error (MSE) is weighted in a block-based manner to take into account the local contrast of the image. Due to the block-based approach, XPSNR can directly be turned into a rate distorsion cost in a block-based coding scheme. This has been done for VVC for the VTM and also for a faster VVC implementation called VVEnC [7]. The XPSNR can directly be chosen as the optimisation metric when encoding a video sequence.

Thanks to the XPSNR, the challenge objective is turned into a classic rate distorsion optimization problem. This is commonly solved using a Lagrangian optimization method in which the distorsion and bit-rate are combined into a single metric $J(\lambda)$:

$$J(\lambda) = \text{XMSE} + \lambda \cdot \text{Rate} \tag{1}$$

Where XMSE is the overall weighted MSE for the video sequences : that is the XPSNR converted on the linear scale. Rate is the bitstream size and λ is a multiplier that aims at balancing those two quantities. XPSNR $= 10 \cdot log(\text{XMSE})$. And XMSE is the sum of individual sequence errors :

$$\text{XMSE} = \sum_s Xmse_s \tag{2}$$

The linear scale is chosen to prevent from wasting unnecessary bits for sequences which can easily have a high XPSNR on the decibel scale. As a consequence, the quality range is reduced and the critical items are allocated more bit-rate.

As the XPSNR and the submission size are additive, the optimization is solved, for a given λ value, by finding the optimal rate distorsion point sequence-wise. The size constraint λ is to be selected to match the submission size either for the 100 kb/s or the 1 Mb/s target.

To further improve the coding efficiency, the coding structure is relaxed to avoid unnecessary constraints. For example, only a single intra frame is needed in the context of the challenge as no periodical random access point is needed.

2.2. Subsampling

In order to address low bit-rates, it is common to reduce the video resolution is order to have a proper number of bits per pixel range. Especially, at 100 kb/s, retaining the initial HD resolution seems out of reach.

For this candidate we propose to pre-process the input sequences in order to provide lower resolutions, e.g. 1920x1080 sequences are converted to 960x540 pixels (one quarter of the initial number of pixels). The low-resolution sequence is fed to the VVC encoder, and after decoding the original resolution is retrieved using a simple up-sampling process.

In this work the down and up-sampling processes use the Lanczos filter implementation of ffmpeg. The Lanczos filter is recommended as it better preserves details and edges compared to the bicubic or bilinear sampling filters. The option `-vf scale=960x540:flags=lanczos` in invoked in the ffmpeg command line.

Four different conversions are proposed in this candidate: 2/3, 1/2 and 1/3 downscaling are possible. This leads, in the case of a 1920x1080 source to intermediate resolutions of 1280x720, 960x540 and 640x360. This 3 down sampling candidates provide a rich level of flexibility in a way similar to the ladder resolutions offered in adaptive streaming [8].

To handle odd resolutions during downscaling, padding is applied when appropriate. For example, 1280x720 sequences are padded to 1284x720 in order to be processed with the 2/3 downscaling. After decoding, and upsampling that additional area was cropped in the inverse conversion.

2.3. VVC encoder selection and parameterization

To summarize, the desired coding configuration should include :

- Perceptual quality optimization, targetting XPSNR maximization ;

- One single Intra frame insertion at the beginning of the sequence ;
- Adaptive use of resampling to match the bit-rate range.

The VVC standard includes a reference encoder [9] that contains selectable options to accommodate most of these desired features. It also features a perceptual optimization strategy [10]. An other VVC software candidate is the open source VVEnC software (https://github.com/fraunhoferhhi/vvenc) which is significantly faster than the VTM and provides the same perceptual optimisation mechanism.

As the VTM and VVenC provide roughly the same coding performance, the later was used for this candidate. In our experience, VVenC in its slower mode is 11 times faster than the VTM and adds less than 2% of bit-rate overhead. As the CLIC sequences last 10 seconds and since distributed coding of chunks could not be massively used to the longer intra period, VVenC is an obvious choice.

The rate distorsion point is selected using the Quantization Parameter (QP) in the command line. A large QP indicates a larger quantization step leading to a smaller bit-rate. In contrast, smaller QPs increase the quality. When the encoder is driven by a QP parameter, the encoding quality is mostly constant as the coding noise level is directly related to the quantization step.

Each file in the validation set is encoded with a set of QPs : in practice, here the QP ranks from 22 to 50 to address a sufficient bit-rate range for the challenge objectives.

3. Coding Results

The rate distorsion optimization process selects the best coding resolution and the appropriate QP for each sequence. Figure 1 reveals the rate distorsion characteristic for the adaptive resolution using the 4 possibles scales, and for the Full (1/1) and Half (1/2) resolutions.

This figure confirms the benefits of the adaptive down sampling selection for the lower rate range : at 100 kb/s the adaptive scaling outperforms both the full and half resolutions by roughly 0.5 dB on the XPSNR scale. This translates into roughly 25% bit-rate savings for the same quality.

At 1 Mb/s however,the full and adaptive resolution characteristics superimpose : according to the XPSNR metric, there is no need to sub-sample at this rate and the source resolution is kept.

Table 2 reveals the resolutions chosen at the two bit-rates. At 100 kb/s the 4 scalings are evenly used.

The HEVC characteristic, using the HEVC reference software (HM16.24) using the same VVC set-up, is also drawn Figure 1. This illustrates the gap between the two coding scheme generations, in the range of 50% of bit-rate reduction according to this characteristic.

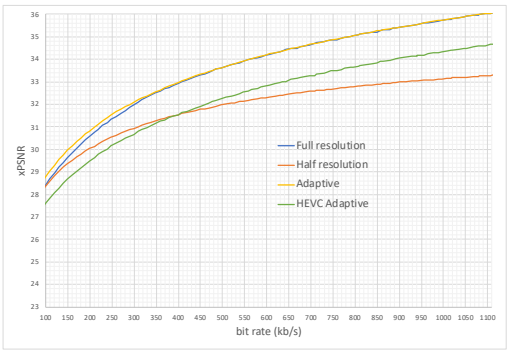

Figure 1. Rate distorsion characteristic

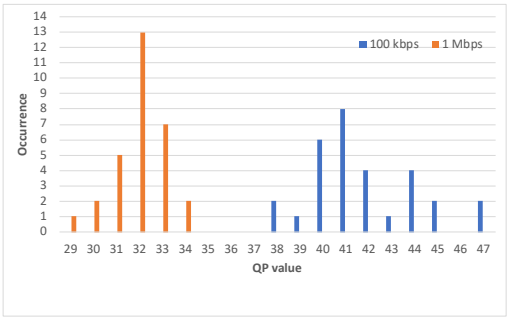

Figure 2. Selected QP values

Figure 2 illustrates the repartition of the QP values. The median QP value is 32 at 1 Mb/s and 41 at 100 kb/s. It anticipates a significant amount of visible distorsion at the lower rate during the subjective testing.

4. Acknowledgment

The authors thank Christian Helmrich and Adam Wieckowski from the Fraunhofer Heinrich Hertz Institute (HHI) for their support.

5. Conclusion

This paper reports the design of a VVC compliant candidate for the CLIC22 video track. A rate distorsion process is described to provide a set of encoded sequences for which the toolset and the quality is adjusted to match the challenge requirements.

The XPSNR metric is selected to optimize the visual quality, in the anticipation of the test phase subjective assessment.

This paper attempts to make this anchor generation as reproducible as possible. The video bitstreams are available upon request by contacting the first author.

Option	Description
`--input`	Selects the input file
`--output`	Indicates the bistream file
`--size`	Selects the video width × height, e.g. 1920x1080
`--framerate`	Selects the video frame rate, e.g. 30 fps
`--refreshsec`	Selects the video refreshing rate in seconds, e.g 20 s
`--qp`	Specifies the base value of the quantization parameter e.g. 32

Table 1. Parameters and command line used for the VVenC software.

Sampling	100 kb/s	1 Mb/s
1/3	6	0
1/2	9	0
2/3	8	0
1/1	7	30
total	30	30

Table 2. QP selection for the 100 kb/s and 1 Mb/s operating points

References

[1] 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). *Proceedings of the IEEE*, pages 1–31, 2021. 1

[2] N. Sidaty, W. Hamidouche, O. Déforges, P. Philippe, and J. Fournier. Compression performance of the versatile video coding: Hd and uhd visual quality monitoring. In *2019 Picture Coding Symposium (PCS)*, pages 1–5, 2019. 1

[3] Théo Ladune and Pierrick Philippe. Coding standards as anchors for the CVPR CLIC video track. In *2021 CVPR Workshop on Learned Image Compression*, 2021. 1, 2

[4] Christian Helmrich et al. AHG10: Improved perceptually optimized QP adaptation and associated distortion measure. In *doc. JVET-K0206, Ljubljana, July 2018*, 2018. 2

[5] Johannes Erfurt and Christian Helmrich et al. A study of the perceptually weighted peak signal-to-noise ratio (wpsnr) for image compression. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 2339–2343, 2019. 2

[6] Christian Helmrich et al. XPSNR: A low-complexity extension of the perceptually weighted peak signal-to-noise ratio for high-resolution video quality assessment. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2727–2731, 2020. 2

[7] Adam Wieckowski, J. Brandenburg, T. Hinz, C. Bartnik, V. George, G. Hege, C. Helmrich, A. Henkel, C. Lehmann, C. Stoffers, I. Zupancic, B. Bross, and D. Marpe. Vvenc: An open and optimized VVC encoder implementation. In *2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 1–2. IEEE Computer Society, jul 2021. 2

[8] Adam Wieckowski and Gabriel Hege Christian Lehmann Benjamin Bross Detlev Marpe Christian Feldmann Martin Smole. VVC in the cloud and browser playback – it works. In *Proceedings of ACM MHV 2022. ACM, New York, NY, USA*, 2022. 2

[9] JVET. https://vcgit.hhi.fraunhofer.de/jvet/vvcsoftware_vtm, 2021. 3

[10] Christian Helmrich, Sebastien 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 *2019 Data Compression Conference (DCC)*, pages 172–181, 2019. 3