TECHNICAL REPORT

ISO/IEC TR 23002-8

First edition
2021-05

Information technology — MPEG video technologies —

Part 8: Working practices using objective metrics for evaluation of video coding efficiency experiments

Reference number
ISO/IEC TR 23002-8:2021(E)

Contents

Page

Foreword

ISO (the International Organization for Standardization) and IEC (the International Electrotechnical Commission) form the specialized system for worldwide standardization. National bodies that are members of ISO or IEC participate in the development of International Standards through technical committees established by the respective organization to deal with particular fields of technical activity. ISO and IEC technical committees collaborate in fields of mutual interest. Other international organizations, governmental and non-governmental, in liaison with ISO and IEC, also take part in the work.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives or www.iec.ch/members_experts/refdocs).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO and IEC shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents) or the IEC list of patent declarations received (see patents.iec.ch).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT) see www.iso.org/iso/foreword.html. In the IEC, see www.iec.ch/understanding-standards.

This document was prepared by Technical Committee ISO/IEC JTC1, *Information technology*, Subcommittee SC 29, *Coding of audio, picture, multimedia and hypermedia information*, in collaboration with ITU-T (as ITU-T HSTP-VID-WPOM (07/2020)).

A list of all parts in the ISO/IEC 23002 series can be found on the ISO and IEC websites.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html and www.iec.ch/national-committees.

Information technology — MPEG video technologies —

Part 8: Working practices using objective metrics for evaluation of video coding efficiency experiments

1 Scope

This document provides general information about coding efficiency measurement practices for video coding. This document does not provide recommendations for evaluating video quality; it describes the practices that have recently been followed for coding efficiency experiments conducted during work to develop video coding standards.

2 Normative references

There are no normative references in this document.

3 Terms and definitions

For the purposes of this document, the following term and definition apply.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at http://www.electropedia.org/

3.1
Bjøntegaard delta bit rate
BD-rate
average percentage bit rate difference at equal measured distortion, integrated across a range of bit rates in the log domain

Note 1 to entry: The Bjøntegaard delta bit rate measurement method was originally specified in VCEG-M33[1].

4 Abbreviated terms

AVC	advanced video coding (Rec. ITU-T H.264 | ISO/IEC 14496-10)
BD-rate	Bjøntegaard delta bit rate
HDR	high dynamic range
HEVC	high efficiency video coding (Rec. ITU-T H.265 | ISO/IEC 23008-2)
HLG	hybrid log gamma
JCT-VC	joint collaborative team on video coding (for development of HEVC)
JVET	joint video experts team (for development of VVC)

MPEG	moving picture experts group
MS-SSIM	multi-scale structural similarity
MSE	mean square error
PQ	perceptual quantizer (as defined in SMPTE ST 2084[10] and Rec. ITU-R BT.2100[11])
PSNR	peak signal-to-noise ratio
QP	quantization parameter
SDR	standard dynamic range
SSIM	structural similarity
VCEG	visual coding experts group
VMAF	video multimethod assessment fusion
VVC	versatile video coding (Rec. ITU-T H.266 | ISO/IEC 23090-3)
WCG	wide colour gamut
WVGA	wide video graphics array
$Y'C_BC_R$	colour space representation commonly used for video/image distribution, also written as *YUV*
YUV	colour space representation commonly used for video/image distribution, also written as $Y'C_BC_R$

5 Video coding experiments using Bjøntegaard delta bit rate (BD-rate) measurements

This document provides general information about coding efficiency measurement practices that have been used for video coding experiments for the development of video coding standards in the ITU-T SG 16 VCEG and ISO/IEC JTC 1/SC 29 MPEG communities. Such work has often been conducted together in the JVET and JCT-VC joint collaborative teams. In particular, the document describes the use of Bjøntegaard delta bit rate (BD-rate) measurements. It provides a concept-level overview of recent practices and provides references to other works that describe further details. It includes comments on why some of the choices were made and indicates situations where caution is needed when interpreting the results.

For comparing different encodings, often it is helpful to control the encodings so that similar types and degrees of encoder optimization are applied, except for the aspects to be tested.

When there are large differences between the coding technologies being tested, and especially when there can be a substantial difference between the resulting subjective quality, subjective testing (i.e. using humans to measure the visual quality) is the appropriate action. There are also cases where the quality difference is expected to be primarily a matter of subjective effect – for example, when measuring the effects of deblocking filters.

The video coding community has typically used formal subjective testing at the call for proposals and verification testing stages of projects for standardization (i.e. at the beginning and the end of the work). For measuring smaller effects and where formal subjective testing is not feasible, it is necessary to use objective measurements. Since objective measurements are collected at multiple operational points, and to better understand coding behaviour across all these points, what has commonly been used in this community is the technique known as the BD-rate (Bjøntegaard delta bit rate) comparison.[1]

Encoding for video distribution is ordinarily performed in the $Y'C_BC_R$ domain (YUV). For typical multimedia applications, it is well known that the human visual system is most sensitive to the fidelity of the Y component. The Y component also tends to use most of the bit rate, so it is natural to focus primarily on the Y component. However, it is advised to measure and report the fidelity of all three components and review the balance between luma and chroma fidelity when interpreting the results. This practice can help avoid situations where luma gain can be achieved at a significant cost of chroma fidelity.

To calculate the Bjøntegaard delta bit rate, a distortion metric needs to be used. For standard-dynamic range video, the distortion metric primarily used in the video coding standardization community has been the peak signal to noise ratio (PSNR). There are certainly some weaknesses to the PSNR-based BD-rate measure in terms of its correspondence with human perception of fidelity. Some other objective distortion metrics which have been asserted to have a better relationship with human perception, such as structural similarity (SSIM) index,[12] multi-scale SSIM (MS-SSIM),[13] and video multimethod assessment fusion (VMAF),[14] have also been considered with the BD-rate measurement process. However, the use of PSNR-based BD-rate measures is the most prevalent in the video coding standardization community. In this document, BD-rate is used to denote PSNR-based BD-rate unless a different distortion metric is explicitly mentioned.

This document is based on JVET-Q0826[15] and JVET-R2016[16].

6 The PSNR-based BD-rate concept

When developing a video coding standard, it is important to have a uniform way of reporting the compression results so that different contributions can be compared against each other.

The PSNR metric is based on the squared error of individual sample values and does not take into account how the human visual system works. A relevant question is therefore whether the PSNR metric is a good predictor of subjective quality. The answer depends at least partly on how different the encoding methods being compared are to each other. If the two methods differ greatly, their artefacts can be very different, and the perceived subjective quality will depend heavily on which type of artefact is psychovisually more disturbing. BD-rate measurements are most often used to compare between two versions of the same video encoder that only differ in that in one of them one tool has been turned on or has been modified versus the other. In this scenario it is much more likely that the BD-rate score between these two versions will correlate with a difference in subjective quality. A clear exception is when tools are considered that are only (or primarily) expected to affect subjective quality, such as deblocking filters. Here, decisions are almost always based on a subjective test or expert viewing, and BD-rate numbers are provided more as an assurance that the tool has not caused some unexpected problem.

An advantage with using PSNR is that it is mathematically simple and therefore straightforward to optimize for. As an example, if a tool depends on filter coefficients or other parameters, the reference encoder can search for the parameter value that minimizes mean square error (MSE) and thus optimizes PSNR, and this type of optimization is often straightforward to analyse and implement. The idea is that a real encoder can optimize for a different distortion metric that is psychovisually more relevant but where the parameter search can be a lot more complicated to implement. By choosing PSNR as the distortion metric in the BD-rate calculations, the work can concentrate on creating coding tools instead of spending time developing encoder optimizations for advanced distortion metrics. However, it is also possible to compute BD-rate measurements using other objective distortion metrics or subjective mean-opinion scores.

For high-dynamic range (HDR)/wide colour gamut (WCG) material and 360° video material, there are additional aspects that influence the usability of BD-rate calculations; these are addressed in Clauses 7 and 8, respectively. The JVET common test conditions also specify a separate category for screen content material (i.e. material that has not been captured by a camera). However, in the context of standardization development, a need for a special metric instead of PSNR-based BD-rate has not been identified for this category.

7 PSNR-based BD-rate calculation

7.1 General

There are several steps in the BD-rate calculation process, where the result in each step is calculated from the result obtained in the previous step:

a) Calculation of PSNR for individual frames.

b) Calculation of per-sequence PSNR and bit rate values for each quantization parameter (QP) value. The QP value influences the resulting bit rate. Hence, compressing the sequence several times with different QPs ensures that the final BD-rate measurement will reflect the performance at many different bit rates.

c) Calculation of per-sequence BD-rate values.

d) Calculation of an aggregate BD-rate value for all sequences.

These steps are further described in subclauses 7.2 through 7.6.

7.2 Calculation of PSNR for individual frames

For an individual frame, the mean square error is calculated between the luma channel $decY$ of the decoded output image and the luma channel $origY$ of the original image according to Formula (1).

$$MSE\_Y = \frac{1}{W*H}\sum_{y=0}^{H-1}\sum_{x=0}^{W-1}\left(decY(x,y) - origY(x,y)\right)^2 , \tag{1}$$

where

$decY(x,y)$ and $origY(x,y)$	are the luma sample values at position (x,y) of the decoded and original images at the same time instance, respectively;
W	is the width of the luma component;
H	is the height of the luma component.

A luma PSNR value for the frame can then be calculated using Formula (2).

$$PSNR\_Y = 10*\log10\left(\frac{\left(255 \ll (bitDepth-8)\right)^2}{MSE\_Y}\right), \tag{2}$$

where

$bitDepth = 10$ for 10-bit inputs;

$\ll$ denotes a bitwise left-shift operation.

If $MSE\_Y = 0$, i.e. if the decoded image exactly matches the original image, there is some adjustment applied to avoid a division by zero. Different implementations can use a different adjustment method. For example, the HEVC test model (HM) and VVC test model (VTM) software packages set the $PSNR\_Y$ value to 999.99. In this case, the AVC joint test model (JM) and HDRTools software packages impose a minimum MSE of 1 ÷ ($W * H$), and another approach could be to impose a minimum MSE of 1÷12, since that is the MSE that would theoretically result from rounding large numbers to the nearest multiple of 1.

The use of 255 ≪ ($bitDepth$ − 8) instead of $2^{bitDepth}$ − 1 in the numerator of the expression in Formula (2) is slightly unusual, but it provides a small adjustment so that if the same video content is coded using $bitDepth$ = 8 or is coded by shifting it up by two bits and using a 10-bit encoder, and when any error is also just scaled up accordingly, there will be no difference in the resulting fidelity measurement. The difference between the two types of measurement is just a constant offset of 0.0255 dB, so it is normally insignificant.

Three PSNR numbers are ordinarily calculated in this manner; one for luma (*PSNR_Y*), and two for chroma (*PSNR_U* and *PSNR_V*).

7.3 Calculation of sequence PSNR and bit rate numbers for each QP value

The aggregate PSNR for a test sequence is calculated as the average of the PSNR values for the individual frames according to Formula (3).

$$PSNR\_Y_{sequence} = \frac{1}{NumFrames} \sum_{k=0}^{NumFrames-1} PSNR\_Y_k , \tag{3}$$

where

$PSNR\_Y_k$	is the $PSNR\_Y$ value for frame k calculated according to subclause 7.2;
NumFrames	is the number of frames in the sequence.

An alternative to averaging PSNR would be to average the MSE value and then use Formula (2) to calculate the aggregate PSNR for the sequence. That would avoid the issue with dividing by a zero *MSE_Y* value in Formula (2) when a single decoded frame matches the original perfectly. More generally, it would avoid the case where a single frame with very high fidelity has a large influence on the average. However, that would also mean that a single frame with very poor fidelity could influence the final number considerably, although it can arguably be difficult to notice, especially at high frame rates. It has been the typical practice to average the PSNR scores instead. The bit rate for the sequence is calculated in kilobits per second and is calculated from the number of frames per second (fps), the number of frames in the sequence, and the size of the file in bytes according to Formula (4).

$$BitRate = \frac{8 * FilesizeInBytes * fps}{NumFrames * 1000} . \tag{4}$$

There is sometimes extra information in the bitstream, such as checksums, that is not necessary for decoding. This information is only used for bitstream validation and is not counted in FilesizeInBytes. Each test sequence is compressed using four different QP values (values 22, 27, 32 and 37 according to the JVET common test conditions). PSNR numbers and bit rate numbers are calculated for each QP.

For chroma, $PSNR\_U_{sequence}$ and $PSNR\_V_{sequence}$ are calculated in a similar fashion.

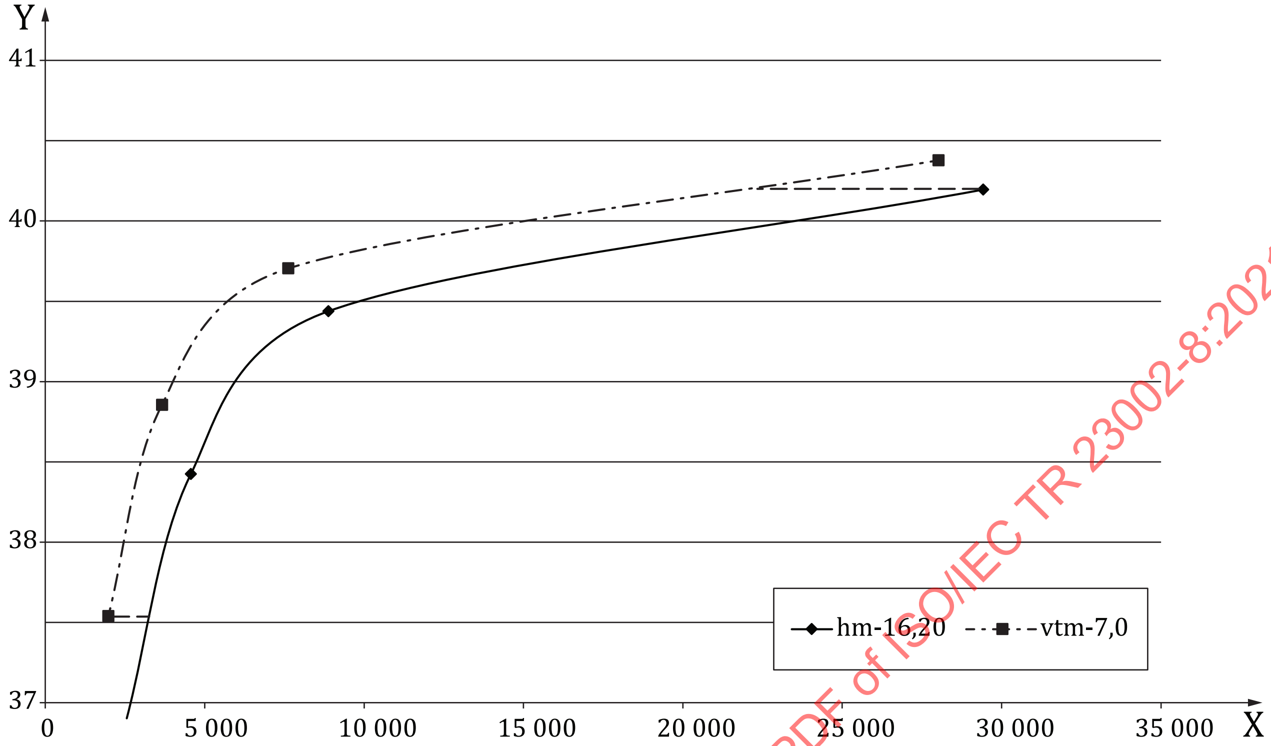
7.4 Calculation of sequence BD-rate number

Subclauses 7.2 and 7.3 have determined the PSNR values and bit rate values for each QP value, both for the anchor and for the tested method. The anchor here refers to the baseline that a tested method is compared against, such as the HEVC reference software HM-16.20, whereas the test is the tested method under investigation, for instance the VVC reference software VTM-7.0. Table 1 presents one example of how the values can differ between the anchor and test scenarios.

Table 1 — Example bit rates and PSNR values for anchor and test

QP value	Bit rate of anchor (kbps)	PSNR_Y anchor	Bit rate of test (kbps)	PSNR_Y test
22	29419.76	40.19	28020.45	40.38
27	8876.16	39.44	7622.83	39.70
32	4564.60	38.42	3661.62	38.86
37	2551.37	36.90	1979.02	37.54

The values in this table can be plotted as two curves as shown in Figure 1 and Figure 2.

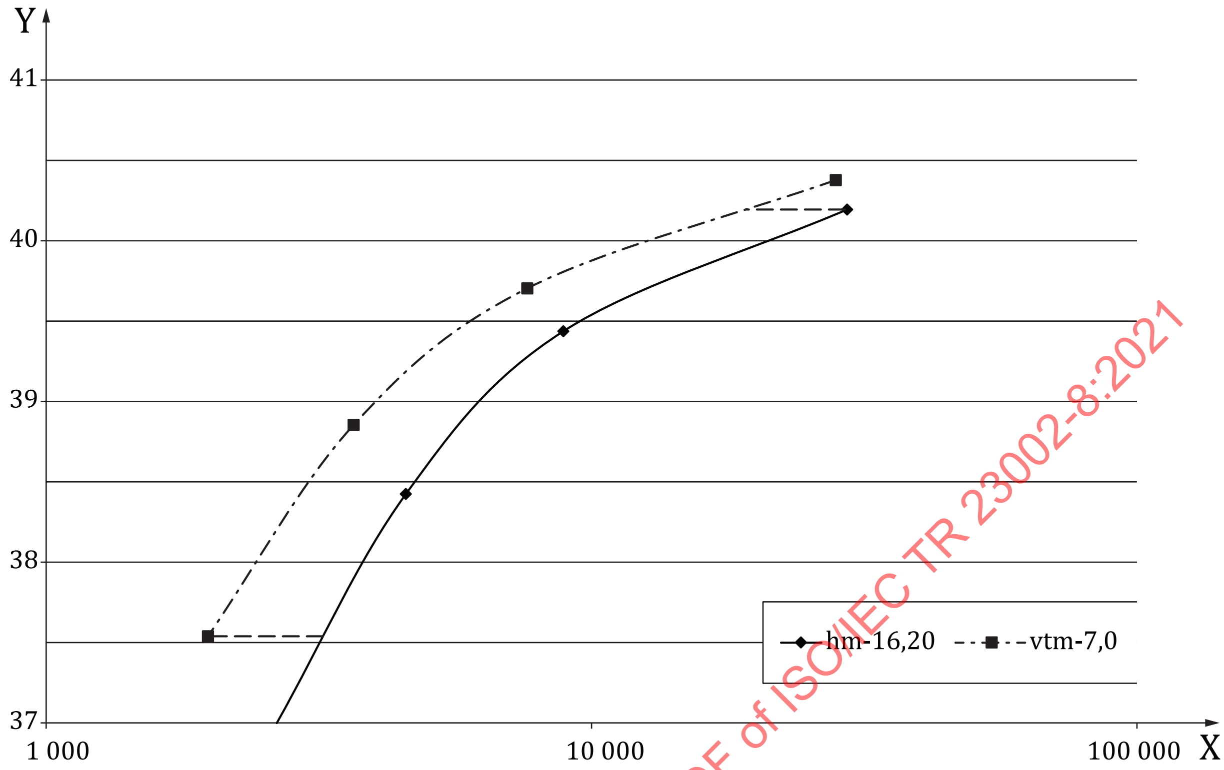


Key

X bit rate (kbps)

Y PSNR_Y (dB)

Figure 1 — Example of luma PSNR plotted as a function of bit rate for HM-16.20 (unbroken line) versus VTM-7.0 (dashed-dotted line)



Key

X bit rate (kbps)

Y PSNR_Y (dB)

Figure 2 — Example of luma PSNR plotted as a function of bit rate for HM-16.20 (unbroken line) versus VTM-7.0 (dashed-dotted line), where the bit rate axis is in log scale, as suggested in VCEG-AI11[3]

The area between these two curves is estimated to compute an average bit rate savings for equal measured quality, i.e. the Bjøntegaard delta bit rate (BD-rate).[1] This is done with the help of a piecewise cubic fitting1) of the *PSNR_Y* versus bit rate curves, where the bit rate is measured in the log domain. An integration is then performed to determine the area between the two curves. The area is further divided by the PSNR range to obtain an average rate difference. Further details are described in References [1],[2],[3],[4],[5],[8] and [9]. Currently a VBS script bdrate( ) from the Excel file available in JVET-O0003[9] is used. A VBS script bdrateOld( ) is also available and computes a BD-rate value using cubic fitting, which was used historically. The two BD-rate values obtained with piecewise cubic and cubic fitting can diverge significantly, indicating a numerical instability. When the two values are close to each other, the result is considered to be more reliable. There are also newer BD-rate functions in JVET-H0030[5] that can handle more than 4 points, can also calculate the BD-rate within a specified range, or can also consider extrapolation techniques to deal with cases with minimal overlap between the two curves.

The PSNR ranges of the curves generally do not overlap completely, and the current practice is to avoid measuring the part of the area between curves that are extrapolated2) rather than interpolated, which can give unpredictable results if the non-overlapping parts are large. Therefore, the area between the two curves is only measured in the region where there is an overlap, i.e. the area between the two dashed lines in Figure 1. In the above example, this would mean that BD-rate is calculated only within the area between minPSNR and maxPSNR, where

minPSNR = Max($\text{Min}_{i=0..N}(\text{PSNR\_anchor}_i)$, $\text{Min}_{i=0..N}(\text{PSNR\_test}_i)$) = 37.5394 in this example, and

1) A piecewise-cubic Hermite interpolating polynomial is used.

2) No extrapolation occurs when using piecewise cubic fitting. However, extrapolation occurs with cubic fitting.

maxPSNR = Min($\text{Max}_{i=0..N}$(PSNR_anchor$_i$), $\text{Max}_{i=0..N}$(PSNR_test$_i$)) = 40.1949 in this example.

This is done by a script provided in JVET-O0003[9].

Calculating the value only where there is overlap poses another challenge; if the overlap is only in a very small region, the BD-rate will be calculated using only a small (and possibly atypical) part of the available data. Therefore, it is important that the overlap is substantial for the BD-rate value to be meaningful.

Another issue to be careful of is if the shapes of the curves are very complicated or have unusual characteristics. Especially if they are crossing over each other multiple times, the BD-rate value can be unreliable.

A negative value indicates a gain, i.e. an improvement in coding efficiency. As an example, if the luma BD-rate value is −1.0 %, this means that it is possible to compress using the “test” method using 1 % fewer bits than using the “anchor” method while maintaining the same luma PSNR.

An alternative approach to calculating BD-rate is to calculate BD-PSNR. Instead of giving an answer like “at equal PSNR, the tested method has 1 % lower bit rate”, BD-PSNR gives an answer like “at equal bit rate, the tested method has 0.05 dB higher PSNR”. However, while 1 % lower bit rate at a given quality is a simple concept to understand, it is not immediately obvious what an increase in PSNR of 0.05 dB means. Hence the common practice is to use BD-rate.

7.5 Consideration of chroma fidelity

Whereas the PSNR values will be different for luma and chroma, the same bit rate is used in both cases, since the encoding represents the three components together and it is not very feasible to try to separate which bits to assign to which components. Since more of the bits are used to encode luma channels than are used for the chroma channels, this means that the chroma BD-rate values can become difficult to interpret if they deviate too much from the luma BD-rate values. If the BD-rate measures are very different for the luma and chroma components or have opposite signs, the results can be misleading. As an example, if the luma BD-rate value is +0.5 % (Y), while the chroma differences are −10.0 % (U) and −9.0 % (V), it can be difficult to judge which method (test or anchor) is actually better in terms of compression efficiency. A common way around this problem is to carry out a new test where bits are transferred from chroma to luma, for instance by increasing the step-size used for chroma quantization. If the new results are −0.5 % (Y), −0.03 % (U), −0.02 % (V), it is safer to say that the tested method is better. An alternative approach that provides a rough simplified measurement and does not require running a new simulation is to calculate a weighted per-sequence combined PSNR average, for example, using Formula (5).

$$PSNR\_YUV_{sequence} = \frac{1}{8}\left(6 * PSNR\_Y_{sequence} + PSNR\_U_{sequence} + PSNR\_V_{sequence}\right) \quad (5)$$

Here the stronger weighting of luma PSNR is to compensate for the fact that most of the bits are used to describe luma information. However, the values of the weights [6, 1, 1]/8 are not the result of careful study and are merely the ones that are most commonly used. In fact, using a larger relative weight for the luma channel might ordinarily be a more accurate reflection of the relative bit allocation effects.

The per-sequence $PSNR\_YUV_{sequence}$ values are then used together with the bit rate values to obtain a YUV-BD-rate value for each sequence. These YUV-BD-rate numbers can be helpful especially if there are many methods that are compared with each other. Another possibility, which is not recommended, is to simply create an average of the BD-rates. In such case a larger weight is typically used for the luma channel’s BD-rate. As an example, a BD-rate difference of +0.5 % (Y), −10.0 % (U), −9.0 % (V) would be averaged to (6 * 0.5 % − 10.0 % − 9.0 %) / 8 = −2.0 %. This other method can misrepresent gains by a substantial amount when the gain in the chroma channels is substantially larger than the gain in the luma channel.

7.6 Calculation of aggregate BD-rate value for all sequences

Once the BD-rate values for a set of test sequences have been determined, they are typically combined using an arithmetic average as in Formula (6).

$$\text{BD-rate aggregated over several sequences} = \frac{1}{N}\sum_{k=1}^{N}\text{BD-rate for test sequence } k \tag{6}$$

Typically, the test sequences are divided into classes that are categorized mainly by resolution or by other characteristics, such as whether they contain camera-captured content versus containing text and graphics with motion. One aggregate BD-rate value per class is reported. Hence one number is reported for all the HD sequences, one for WVGA resolution sequences, and so on. Finally, one aggregate BD-rate value can be calculated by averaging across sequences of different classes.

8 BD-rate calculation for HDR material

A PSNR-based BD-rate measurement is limited in its capability to predict subjective quality improvements. For high-dynamic range (HDR) material, there is a further complication in that there is a very non-linear mapping between the luma codewords that are used as an input to the encoder, and the luminance values that would be output by a display. This is especially true for content that is represented with the perceptual quantizer (PQ) transfer function.[10][11] This transfer function gives a much higher importance and allocates a large number of codewords to darker regions, as compared to transfer functions typically used for standard dynamic range content. This reduces the correlation between the PSNR-based BD-rate measurement and subjective quality.

For HDR content, calculating the PSNR on the luma and chroma codewords, as is done for the standard dynamic range (SDR) case, is still appropriate for HDR content that employs an hybrid log-gamma (HLG) transfer function.[11] However, it is necessary to complement the PSNR metric with a number of additional metrics for HDR content represented with the PQ transfer function.[7] These metrics include:

— PSNRL100: This metric is calculated using the luminance values rather than the luma codewords. It is based on the CIELab colour space representation of the input and output sample values of the codec.

— wPSNR: This is a PSNR-like metric calculated from the codewords, that attempts to compensate for the more significant distribution of luma codewords to the darker regions mentioned above by performing a weighting of the codewords before calculating the PSNR.

— DE100: This is a metric based on the CIELab colour space representation of the input and output sample values of the codec. It is specifically targeted at chrominance fidelity.

For more details on these metrics, see Reference [7].

The JVET and JCT-VC groups do not typically create combined metrics using the wPSNR information from the luma and chroma channels, as was described in Clause 7. Instead, the group relies on the PSNRL100 and DE100 metrics. If a combined metric is desired, then attention is given to the colour gamut used for the content. In particular, the weighting would need to be adjusted when using the ITU-R BT.2100 colour gamut[11] typically employed for HDR content. The JVET and JCT-VC groups have not codified any specific adjustment to date.

9 BD-rate calculation for 360° video

For 360° omnidirectional video applications, the video scene is theoretically positioned on the interior of a distant sphere surrounding the viewer, who is looking out from the centre of the sphere. Since video encoding ordinarily works by compressing rectangular arrays of video samples rather than spheres, a projection must be used to map the scene content on the sphere to the values of the samples of a rectangular array before compression can take place. This also means that measuring the PSNR value of the rectangular video array can be very misleading, since the proportion of area on the sphere that corresponds to the sample values at different positions can be very different. As an example, one possible projection is the equirectangular projection, which is similar to a simplistic mapping between the geography depicted on a globe and on a 2D map of the world. If the PSNR value were to be calculated on the rectangular video that was created using such a mapping, it would result in a strong over-emphasis of errors near the poles, since these regions are stretched out considerably. In order to get around this problem, a modified measurement known as the weighted to spherically uniform PSNR (WS-PSNR) has been used. This is a weighted form of PSNR that compensates for this stretching. For more information on WS-PSNR, see Reference [6].