

**ISO/IEC JTC 1/SC 29/WG 11 N 19168**

**ISO/IEC JTC 1/SC 29/WG 11**

**Coding of moving pictures and audio**

**Convenorship: UNI (Italy)**

|  |  |
| --- | --- |
| **Document type:** | **Approved WG 11 document** |
|  |  |
| **Title:** | **Summary information on BD-rate experiment evaluation practices** |
|  |  |
| **Status:** | **Approved** |
|  |  |
| **Date of document:** | **2019-01-17** |
|  |  |
| **Source:** | **WG 11 (via JVET)** |
|  |  |
| **Expected action:** | **Study and discussion** |
|  |  |
| **No. of pages:** | **9** |
|  |  |
| **Email of convenor:** | **leonardo@chiariglione.org** |
|  |  |
| **Committee URL:** | **https://isotc.iso.org/livelink/livelink/open/jtc1sc29wg11** |

|  |  |
| --- | --- |
| **Joint Video Experts Team (JVET)**  **of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11**  17th Meeting: Brussels, BE, 7–17 January 2020 | Document: JVET-Q2016-v3 |

|  |  |  |  |
| --- | --- | --- | --- |
| *Title:* | **Summary information on BD-rate experiment evaluation practices** | | |
| *Status:* | Approved output document of JVET | | |
| *Purpose:* | Information | | |
| *Author(s) or Contact(s):* | **Jacob Ström Kenneth Andersson Rickard Sjöberg** (Ericsson Research)  **Andrew Segall Frank Bossen** (Sharp Labs of America)  **Gary Sullivan**  (Microsoft)  **Jens-Rainer Ohm** (RWTH Aachen) | Email: | {jacob.strom, kenneth.r.andersson, rickard.sjoberg} @ericsson.com  asegall@sharplabs.com fbossen@sharpsec.com  [garysull@microsoft.com](mailto:garysull@microsoft.com)  [ohm@ient.rwth-aachen.de](mailto:ohm@ient.rwth-aachen.de) |
| *Source:* | JVET | | |

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Abstract

This document provides general summary information about typical quality measurement practices for video coding experiments in recent work in MPEG and VCEG and their JVET and JCT-VC joint teams. In particular, it describes the use of Bjøntegaard Delta rate (BD-rate) measurements. It aims at providing a concept-level overview of recent practices and to provide some references to technical papers that describe further details. It provides comments on why some of the choices were made, and points out situations where caution must be taken when interpreting the results.

# Introduction

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 may 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.

In our community, we have typically used formal subjective testing at the Call for Proposals and Verification Testing stages of projects (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 metrics, and what has commonly been used in our community for this is the technique known as BD-rate (Bjøntegaard Delta bit rate) comparison [1].

We ordinarily perform encoding in the *Y′CBCR* domain (nicknamed YUV herein), and 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, we typically measure and report the fidelity of all three components and review the balance between luma and chroma fidelity when interpreting the results.

There are certainly some weaknesses to the BD-rate metric, in terms of its correspondence with human perception of fidelity. Some other objective metrics have been developed, although BD-rate measures remain the most common practice in the video coding standardization community, for several reasons which we will not try to expound upon here in the interest of brevity and since this document is intended merely to describe the common practice.

# The BD-rate concept

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 a lot, their artefacts will be very different, and the perceived subjective quality will depend heavily on which type of artefact is psychovisually more disturbing. BD-rate figures are most often used to compare between two versions of the same code that only differs in that one tool has been turned on or been modified. In this scenario it is much more likely that a better BD-rate score will correlate with improved subjective quality. A clear exception is when tools are considered that are only there to improve subjective quality, such as deblocking filter tools. Here, decisions are almost always based on a subjective test or expert viewing, and BD-rate figures are provided more as an assurance that the tool has not broken things.

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 MSE and thus optimizes PSNR, and this type of optimization is often easy to implement. The idea is that a real encoder can optimize for a different metric that is psychovisually more relevant but where the parameter search may be a lot more complicated to implement. By choosing PSNR, the work can concentrate on creating coding tools instead of spending time on difficult encoder optimizations for advanced metrics.

For high-dynamic range (HDR) material and 360° video material, there are additional aspects that influence the usability of BD-rate calculations, they are addressed in Section 4 and Section 5 respectively.

# BD-rate calculation

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:

1. Calculation of PSNR for individual frames
2. Calculation of sequence PSNR and bit rate figures for each QP value
3. Calculation of sequence BD-rate figures
4. Calculation of an aggregate BD-rate figure for all sequences.

We will go through these steps one by one.

## Calculation of PSNR for individual frames

For an individual frame, the mean square error is calculated between the luma channel of the decoded output image and the luma channel of the original image according to

where and are the luma sample values at position of the decoded and original image. and are the width and height of the luma component respectively. The luma PSNR value for the frame is then calculated as

where bitDepth = 10 for 10-bit input and where denotes leftwards shift. If , i.e., if the decoded image exactly matches the original image, the value is instead set to 999.99 to avoid a division by zero. The use of 255 ≪ (bitDepth − 8) instead of 2bitdepth − 1 is a small adjustment so that if the same video content is coded using bitDepth = 8 or is coded by shifting it up two bits and using a 10-bit encoder, and 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).

## Calculation of sequence PSNR and bit rate figures for each QP value

The aggregate PSNR for a sequence is calculated as the average of the PSNR values for the individual frames:

where is the value for frame calculated according to the previous section, and is the number of frames in the sequence. An alternative to averaging PSNR would be to average the MSE value and then use Equation 2 to calculate the aggregate PSNR for the sequence. That would avoid the issue with dividing by a zero MSE\_Y value in Equation 2 where 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 may 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 (), the number of frames in the sequence and the size of the file in bytes according to

It should be noted that there is sometimes extra information in the bit stream such as checksums that are not necessary for decoding, and these are not counted in . Each sequence is compressed using four different QP values (values 22, 27, 32 and 37 for JVET common test conditions). PSNR numbers and rate numbers are calculated for each QP.

For chroma, and are calculated in a similar fashion.

## Calculation of sequence BD-rate figure

The previous sections have given us PSNR values and bit rate figures for each QP value, both for the anchor and for the test. The anchor here refers to the baseline which is compared against, such as the HEVC reference software HM-16.20, whereas the test is what we want to investigate, for instance the VVC reference software VTM-7.0. As an example, we may have the values shown in Table 1:

*Table 1: Example bit rates and PSNR values for anchor and test*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **QP value** | **Bit rate anchor (kbps)** | **PSNR-Y anchor** | **Bit rate 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 |

This table can be plotted as two curves as shown in Figures 1 and 2.

*Figure 1: The luma PSNR plotted as a function of bit rate of HM-16.20 (black) vs VTM-7.0 (red).*

*Figure 2: The luma PSNR plotted as a function of the bit rate of HM-16.20 (black) vs VTM-7.0 (red).*

The idea is now to estimate the area between these two curves to compute an average bit rate savings for equal measured quality, i.e., the Bjøntegaard-Delta rate (BD-rate) [1]. This is done with the help of a piecewise cubic fitting[[1]](#footnote-1) of the PSNR-Y/rate curves, where the 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. The details of how this is done are described in references [1], [2] [3], [4] [5] [8] and [9]. Currently a VBS script bdrate( ) from the Excel file available in [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 may diverge significantly, indicating a numerical instability. When the two values are close to each other, the result is considered to be more reliable.

The PSNR ranges of the curves often do not overlap completely, and it is best to avoid trying to measure the area between curves that are extrapolated[[2]](#footnote-2) rather than interpolated, which can give unpredictable results. Therefore, the area between the two curves is only measured in the region where there is overlap, i.e., the area between the two green lines in Figure 1. In the above example, this would mean to only calculate the area between minPSNR and maxPSNR, where

minPSNR = max( min(PSNR\_anchor), min(PSNR\_test) ) = 37.5394 in this example, and

maxPSNR = min( max(PSNR\_anchor), max(PSNR\_test) ) = 40.1949 in this example.

This is done by the script in [9]. Calculating the value only where there is overlap poses another challenge; if the overlap is very small, the BD-rate will be calculated using only a small part of the 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 overlapping several 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 figure is −1.0%, this means that it is possible to compress using the “test” method using 1% less bits than using the “anchor” method while maintaining the same luma PSNR.

## 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 most 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 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 may 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 method that provides a rough simplified measurement and does not require running a new simulation is to calculate a weighted per-frame combined PSNR average, for example:

Here the stronger weighting of luma PSNR is to compensate for the fact that most of the bits are used to describe luma information. The per-frame values thus obtained are then averaged together to get a value using Equation 3 and then BD-rate is calculated using these values. These YUV-BD-rate numbers can be helpful especially if there are many methods that should be 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. As an example, a BD-rate difference of +0.5% (Y), −10.0% (U), −9.0% (V) would be averaged to (8 \* 0.5% − 10.0% − 9%) / 10 = −1.6%. This other method may misrepresent gains by a substantial amount when the gain in the chroma channels is substantially larger than the gain in the chroma channel.

## Calculation of aggregate BD-rate figure for all sequences

Once the BD-rate figures for all the test sequences have been determined, they are combined using an arithmetic average:

Typically, the test sequences are divided into classes sorted 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 figure is reported for all the HD sequences, one for WVCGA resolution sequences, and so on. Finally, one aggregate BD-rate value may be calculated by averaging across sequences of different classes.

# BD-rate calculations for HDR material

As discussed in the introduction, 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 code values that are used as an input to the encoder, and the luminance values that a screen would output during display. This is especially true for content that is represented in a SMPTE ST 2084 container, also known as the Perceptual Quantizer (PQ) transfer function [10]. Here, a change from a code value of 100 to 101 may only change the luminance from 0.039 cd/m2 to 0.041 cd/m2, a difference of only 0.002 cd/m2, whereas a change from 900 to 901 may change the luminance from 6487.17 cd/m2 to 6557.22 cd/m2, a much bigger difference of 70.05 cd/m2.

For HDR content, it is the experience of the JVET and JCT-VC groups that calculating the PSNR on the code values, as is done for the standard dynamic range (SDR) case, is appropriate for HDR content that employs an Hybrid Log-Gamma (HLG) transfer function. However, it does not work well as a predictor of subjective quality for SMPTE ST 2084 coded images. Due to this fact, when reporting results on HDR content, it has been necessary to complement the PSNR metric with a number of additional metrics for this PQ content [7], including:

* PSNRL100: This metric is calculated on the luminance values rather than on the code values. It is based on the CIELab colour space.
* wPSNR: This is a PSNR metric calculated from the code values, but it attempts to compensate for the difference in luminance mentioned above by performing a weighting of the code values before taking calculating the PSNR.
* DE100: This is a metric based on the CIELab colour space specifically targeted at chrominance fidelity.

For more details on the metrics, see [7].

It is also noted that 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 the section above. Instead, the group relies on the PSNRL100 and DE100 metrics. It is noted that if a combined metric is desired, then attention should be given to the colour gamut used for the content. And, specifically, the weighting would need to be adjusted when using the ITU-R BT.2020 colour gamut typically employed for HDR content. The JVET and JCT-VC groups have not codified any specific adjustment to date.

# BD-rate calculations for 360° video

For 360° omnidirectional video applications, the video is projected on a sphere around the viewer. Since video encoding ordinarily works by compressing rectangular blocks of video rather than spheres, a projection must be used to map the samples of the sphere to samples on a rectangular block before compression can take place. This also means that taking the PSNR value of the rectangular video can be very misleading. As an example, one possible projection is the equirectangular projection, which is similar to the mapping between the globe and 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 amplification of errors near the poles, since these regions are stretched out considerably. In order to get around this problem, a modified measurement known as WSPSNR has been used, which is a weighted form of PSNR that compensates for this stretching. For more information on this WSPSNR measure, see [6].

# References

1. G. Bjøntegaard, “Calculation of average PSNR differences between RD-curves,” in Proc. VCEG-M33, 13th VCEG Meeting, Austin, TX, USA, Apr. 2001.
2. S. Pateux and J. Jung, “An Excel add-in for computing Bjøntegaard metric and its evolution,” ITU-T SG16 Q.6 Document, VCEG-AE07, Marrakech, Jan. 2007.
3. G. Bjøntegaard, “Improvements of the BD-PSNR model,” ITU-T SG16 Q.6 Document, VCEG-AI11, Berlin, Germany, July 2008.
4. S. Pateux and J. Jung, “Improvements of Excel macro for BD-gain computation,” ITU-T SG16 Document, C358, Geneva, CH, October 2009.
5. A. M. Tourapis, D. Singer, Y. Su, K. Mammou, “BD-Rate/BD-PSNR Excel extensions,” Joint Video Exploration Team (JVET) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11, JVET-H0030, 8th Meeting: Macao, CN, 18–25 Oct. 2017.
6. Y. Ye, J. Boyce, “Algorithm descriptions of projection format conversion and video quality metrics in 360Lib (Version 9),” Joint Collaborative Team on Video Coding (JCT-VC) of ISO/IEC MPEG and ITU-T VCEG, JCTVC-M1004, Feb. 2019.
7. A. Segall, E. François, W. Husak, S. Iwamura, D. Rusanovskyy, “JVET common test conditions and evaluation procedures for HDR/WCG video,” Joint Collaborative Team on Video Coding (JCT-VC) of ISO/IEC MPEG and ITU-T VCEG, JCTVC-P2011, Oct. 2019.
8. F. Bossen, “Common HM test conditions and software reference configurations,” Joint Collaborative Team on Video Coding (JCT-VC) of ISO/IEC MPEG and ITU-T VCEG, JCTVC-G1200, Jan. 2011.
9. F. Bossen, X. Li, A. Norkin, K. Sühring, “JVET AHG report: Test model software development (AHG3),” Joint Collaborative Team on Video Coding (JCT-VC) of ISO/IEC MPEG and ITU-T VCEG, JVET-O0003, July 2019.
10. “High Dynamic Range Electro-Optical Transfer Function of Mastering Reference Displays,” SMPTE ST 2084:2014, 2014.

1. A Piecewise Cubic Hermite Interpolating Polynomial is used [↑](#footnote-ref-1)
2. It should be noted that no extrapolation occurs when using piecewise cubic fitting. However extrapolation occurs with cubic fitting. [↑](#footnote-ref-2)