Calculation of average coding efficiency based on subjective quality scores

Philippe Hanhart*, Touradj Ebrahimi

Multimedia Signal Processing Group, EPFL, Lausanne, Switzerland

Abstract

The Bjøntegaard model is widely used to calculate the coding efficiency between different codecs. However, this model might not be an accurate predictor of the true coding efficiency as it relies on PSNR measurements. Therefore, in this paper, we propose a model to calculate the average coding efficiency based on subjective quality scores, i.e., mean opinion scores (MOS). We call this approach Subjective Comparison of ENcoders based on fItted Curves (SCENIC). To consider the intrinsic nature of bounded rating scales, a logistic function is used to fit the rate-distortion (R-D) values. The average MOS and bit rate differences are computed between the fitted R-D curves. The statistical property of subjective scores is considered to estimate corresponding confidence intervals on the calculated average MOS and bit rate differences. The proposed model is expected to report more realistic coding efficiency as PSNR is not always correlated with perceived visual quality.

Keywords:
Coding efficiency, Rate-distortion curves, Subjective quality assessment, Mean opinion scores, Objective quality assessment, PSNR, Bjntegaard delta, BD-Rate, BD-PSNR

1. Introduction

"If you cannot measure it, you cannot improve it" (Lord Kelvin). This statement is especially true in the case of image and video compression. To design efficient compression algorithms, it is necessary to benchmark the performance of new algorithms against well-established and state-of-the-art algorithms on a dataset containing different contents. The quality of the compressed images and video sequences can be assessed by means of objective and subjective evaluations. Objective quality assessment relies on the use of objective metrics, which have been designed to predict the perceived quality of media content. The Peak Signal-to-Noise Ratio (PSNR) metric is commonly accepted and used by coding experts to objectively measure the performance of coding algorithms. However, it is known that PSNR does not accurately reflect human perception of visual quality [1]. In the case of subjective quality assessment, the quality of the decoded data is evaluated by a pool of human subjects (typically more than 15 people), following a common methodology [2]. Subjective tests are time consuming, expensive, and not always feasible. However, if the subjective evaluations are properly designed and conducted according to guideline recommendations [2], the subjective results are more accurate as they reflect the quality perceived by end-users.

The coding efficiency of different compression algorithms can be adequately compared only by means of subjective tests, carried out according to common evaluation methodologies defined by experts. During the development phase of their compression standards, Joint Photographic Experts Group (JPEG), Moving Picture Experts Group (MPEG), and Video Coding Experts Group (VCEG) have relied during past years on both objective and subjective evaluations to select and evaluate potential coding technologies, as well as for verification purposes. For example, subjective evaluations were conducted during the development of JPEG XR [3], MPEG-4 [4], H.264/MPEG-4 AVC [5, 6, 7], Scalable Video Coding (SVC) extension of H.264/MPEG-4 AVC [8, 7], and H.265/HEVC [9, 7]. Independent researchers have also conducted subjective evaluations, both during and after the development phase of compression standards, as a validation process or to evaluate the codecs in different scenarios.

Objective evaluations based on PSNR measurements are widely used by most researchers as they are simple and can be performed automatically. To calculate the coding efficiency between different codecs based on PSNR measurements, a model was proposed by Gisle Bjøntegaard during the development of H.264/MPEG-4 AVC [10]. The Bjøntegaard model is used to calculate the average PSNR and bit rate differences between two rate-distortion (R-D) curves obtained from the PSNR measurement when encoding a content at different bit rates. The model reports two values:

a) the so-called Bjøntegaard delta PSNR (BD-PSNR), which corresponds to the average PSNR difference in dB for the same bit rate,
b) the so-called Bjøntegaard delta rate (BD-Rate), which corresponds to the average bit rate difference in percent for the same PSNR.

*Corresponding author: Philippe Hanhart, EPFL/STI/IEL/GR-EB, Station 11, CH-1015 Lausanne, Switzerland; Phone: +41-21-693-46-20

Email addresses: philippe.hanhart@epfl.ch (Philippe Hanhart), touradj.ebrahimi@epfl.ch (Touradj Ebrahimi)

The Bjøntegaard model is used by various experts to calculate the coding efficiency of compression standards. For example, this model was used during the development of H.264/MPEG-4 AVC [11], Multiview Video Coding (MVC) extension of H.264/MPEG-4 AVC [12], H.265/HEVC [13], and multiview extensions of H.265/HEVC [14]. The Bjøntegaard model is also widely used by researchers working on image and video compression, to benchmark the performance of their algorithms against well-established and state-of-the-art compression algorithms. However, the Bjøntegaard model might not be an accurate predictor of the true coding efficiency as this model relies on PSNR measurements.

To estimate a more realistic coding efficiency, subjective quality scores should be considered instead of PSNR measurements. Therefore, in this paper, we propose a model to calculate the average coding efficiency based on mean opinion scores (MOS) gathered during subjective evaluations instead of PSNR measurements. We call this approach Subjective Comparison of ENcoders based on fItted Curves (SCENIC). To consider the intrinsic nature of bounded rating scales, as well as nonlinearities and saturation effects of the human visual system, a logistic function is used to fit the R-D values. The average MOS and bit rate differences are computed between the fitted R-D curves similarly to the Bjøntegaard model. To consider the statistical property of subjective scores, the 95% confidence intervals associated with the MOS are considered to estimate corresponding confidence intervals on the calculated average MOS and bit rate differences. To provide meaningful measures, the R-D curves should ideally cover the full range of the rating scale. This recommendation is considered in the proposed model to estimate a confidence index on the calculated average MOS and bit rate differences. The contributions of this paper are:

- a model to estimate more realistic coding efficiency based on subjective quality scores,
- an estimated confidence interval on the calculated average coding efficiency to consider the statistical property of subjective scores,
- an estimated confidence index on the calculated average coding efficiency to consider the range of conditions in the subjective evaluation.

The remainder of this paper is organized as follows. A brief overview of the Bjøntegaard model is given in Sec. 2. Each component of the proposed model is described in details in Sec. 3. Three case studies where the Bjøntegaard and proposed models were used to calculate average coding efficiency are presented in Sec. 4. Finally, concluding remarks and discussion on future work are given in Sec. 5.

2. Bjøntegaard model

In [10], Gisle Bjøntegaard has proposed a model to measure the coding efficiency between two different compression algorithms. To approximate a rate-distortion (R-D) curve given by a set of N bit rate values $(R_1, ..., R_N)$ with corresponding PSNR measurements $(D_1, ..., D_N)$, a third order logarithmic polynomial fitting has been proposed in the Bjøntegaard model, based on experimental observations

$$\hat{D}(R) = a\log^3 R + b\log^2 R + c\log R + d \tag{1}$$

where $\hat{D}$ is the fitted distortion in PSNR, R is the bit rate, and a, b, c, and d are the parameters of the fitting function.

To simplify notation, in the rest of the paper, we use lower case r when referring to the logarithm of the bit rate, i.e., $r = \log R$. Therefore, Eq. (1) is rewritten as

$$\hat{D}(r) = ar^3 + br^2 + cr + d \tag{2}$$

At least four R-D values are required to determine the fitting parameters of Eq. (2). If more than four values are used, then the R-D values are fitted in a least square sense.

The average PSNR difference between two R-D curves is approximated by the difference between the integrals of the fitted R-D curves divided by the integration interval [10]

$$\Delta D = \mathrm{E}\left[D_2 - D_1\right] \approx \frac{1}{r_H - r_L}\int_{r_L}^{r_H}\left[\hat{D}_2(r) - \hat{D}_1(r)\right]\mathrm{d}r \tag{3}$$

where ΔD is the so-called Bjøntegaard delta PSNR (BD-PSNR) computed between the two fitted R-D curves $\hat{D}_1(r)$ and $\hat{D}_2(r)$, respectively, and the integration bounds, r_L and r_H, are

$$\begin{aligned} r_L &= \max\left\{\min(r_{1,1}, ..., r_{1,N_1}), \min(r_{2,1}, ..., r_{2,N_2})\right\} \\ r_H &= \min\left\{\max(r_{1,1}, ..., r_{1,N_1}), \max(r_{2,1}, ..., r_{2,N_2})\right\} \end{aligned} \tag{4}$$

To express the (logarithm of the) rate as a function of the distortion, a third order polynomial fitting has been proposed in the Bjøntegaard model to fit the R-D values

$$\hat{r}(D) = aD^3 + bD^2 + cD + d \tag{5}$$

Note that a second fitting process is required to fit the bit rate values and that $\hat{r}(D)$ (see Eq. (5)) is not the inverse function of $\hat{D}(r)$ (see Eq. (2)).

The average bit rate difference between two R-D curves is approximated as [10]

$$\begin{aligned} \Delta R &= \mathrm{E}\left[\frac{R_2 - R_1}{R_1}\right] = \mathrm{E}\left[\frac{R_2}{R_1}\right] - 1 = \mathrm{E}\left[10^{r_2 - r_1}\right] - 1 \\ &\approx 10^{\mathrm{E}[r_2 - r_1]} - 1 \approx 10^{\frac{1}{D_H - D_L}\int_{D_L}^{D_H}[\hat{r}_2(D) - \hat{r}_1(D)]\mathrm{d}D} - 1 \end{aligned} \tag{6}$$

where ΔR is the so-called Bjøntegaard delta rate (BD-Rate) computed between the two fitted R-D curves $\hat{r}_1(r)$ and $\hat{r}_2(r)$, respectively, and the integration bounds, D_L and D_H, are

$$\begin{aligned} D_L &= \max\left\{\min(D_{1,1}, ..., D_{1,N_1}), \min(D_{2,1}, ..., D_{2,N_2})\right\} \\ D_H &= \min\left\{\max(D_{1,1}, ..., D_{1,N_1}), \max(D_{2,1}, ..., D_{2,N_2})\right\} \end{aligned} \tag{7}$$

Thanks to the logarithmic bit rate scale, the estimation of the average bit rate reduction is also simplified.

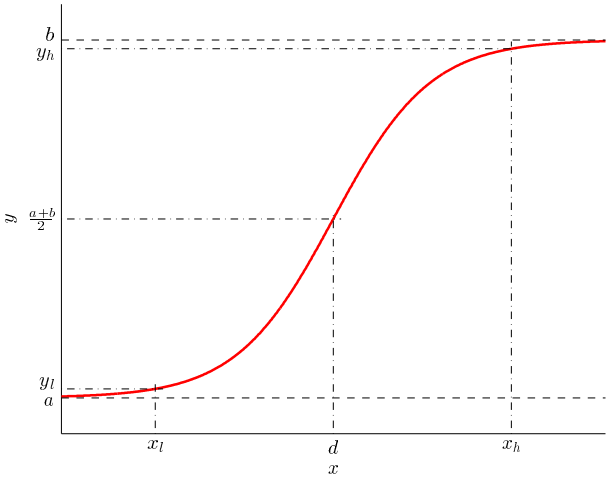


Figure 1 Logistic function $y = f(x) = a + \frac{b-a}{1+\exp[-c(x-d)]}$.

3. Proposed model

In this section, we propose a method for Subjective Comparison of ENcoders based on fItted Curves (SCENIC). First, the function used to fit the R-D values is described. Then, the calculation of average MOS and bit rate differences between two fitted R-D curves is presented. Finally, the confidence intervals and reliability index on the calculated average MOS and bit rate differences are presented. A MATLAB implementation of the proposed model can be downloaded from: `http://mmspg.epfl.ch/scenic`

3.1. Fitting function

According to recommendation ITU-R BT.500-13 [2], the relationship between MOS and the objective measure of picture distortion tends to have a sigmoid shape, provided that the natural limits of picture distortion extend far enough from the region in which the MOS varies rapidly. If the distortion parameter is measured in a physical unit, e.g., a time delay [ms], then a non-symmetrical function should be used to approximate this relationship [2]. If the picture distortion is measured in a related unit, e.g., PSNR [dB], then a logistic function is commonly used [2, 15, 16]. The logistic function (see Fig. 1) is

$$y = f(x) = a + \frac{b-a}{1+\exp[-c(x-d)]} \tag{8}$$

where a, b, c, and d are the parameters of the fitting function.

As bit rate is not a direct measure of picture distortion, a non-symmetrical function should be used to map bit rate values to MOS, according to recommendation ITU-R BT.500-13. However, Gisle Bjøntegaard has observed that R-D values expressed in $(\log(\text{bit rate}), \text{PSNR})$ do not deviate much from straight lines [17], meaning that there is a somewhat linear relationship between log(bit rate) and PSNR. Therefore, based on this observation, and following the common practice to map PSNR values to MOS, we propose to use a logistic function to fit the R-D values expressed in $(\log(\text{bit rate}), \text{MOS})$.

Fitting a logistic function to a set of observed values is a nonlinear curve-fitting problem and can be expressed in least-squares sense. Several solutions have been proposed to solve this class of problem. However, the initial conditions may be critical to converge towards the optimal solution. Nevertheless, in most cases, constraints can be applied on the different parameters based on *a priori* knowledge to restrict the parameter search.

Most rating scales defined in recommendation ITU-R BT.500-13 are divided into five categories with associated labels, such as (*Bad*; *Poor*; *Fair*; *Good*; and *Excellent*) or (*Very annoying*; *Annoying*; *Slightly annoying*; *Perceptible, but not annoying*; and *Imperceptible*). The asymptotes of the relationship between MOS and bit rate, which are caused by the use of bounded rating scales and the saturation effects of the human visual system, are typically associated with the two extreme categories of the rating scale. Moreover, the subjective scores should increase from the lower to the upper categories as the bit rate increases. Therefore, constraints are imposed on the logistic function such that the lower and upper asymptotes are associated with the lower and upper categories, respectively, and that the function is strictly increasing

$$\begin{gathered} u_{\min} \le a \le u_{\min} + \tfrac{1}{5}\Delta u \\ u_{\max} - \tfrac{1}{5}\Delta u \le b \le u_{\max} \\ c > 0 \end{gathered} \tag{9}$$

where $\Delta u = u_{\max} - u_{\min}$, $u_{\min}$ and $u_{\max}$ are the boundaries of the rating scale, and $\frac{1}{5}\Delta u$ corresponds to the "length" of one category in a five categories scale.

3.2. Integration bounds

Whereas the R-D curve based on PSNR measurements is unbounded, the R-D curve based on MOS is bounded due to the use of a bounded rating scale, the fact that many evaluation methods consist in comparing the quality of a test stimulus against the quality of a reference stimulus, and the saturation effect of the human visual system. Therefore, we think that it is not meaningful to compute average MOS or bit rate differences when both R-D curves have reached the saturation phase.

In statistics, it is common to consider only the values lying within the 95% confidence interval. In the proposed model, we consider a similar approach by discarding the lower and upper parts of the fitted R-D curve and keeping only the values between y_l and y_h (see Fig. 1), which covers 95% of the range spanned by the fitted R-D curve

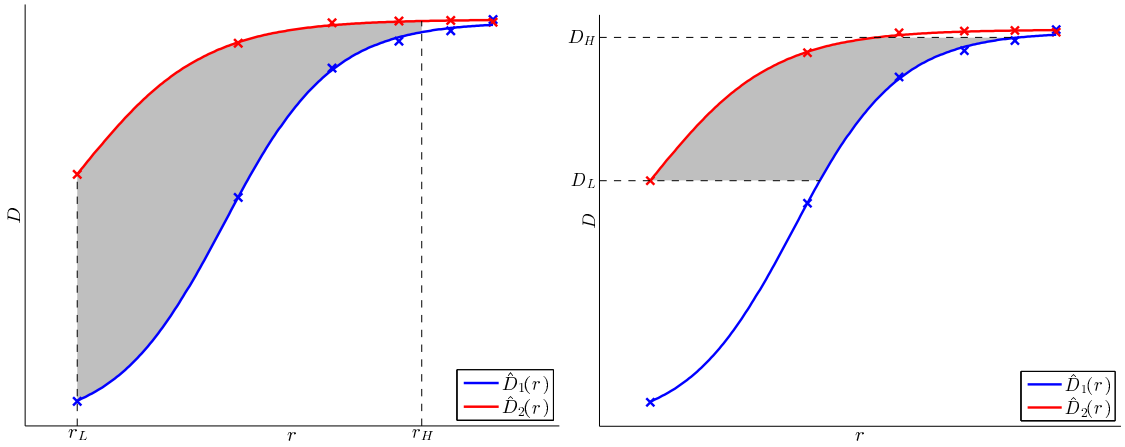
$$y_l = a + 0.025(b-a) \quad y_h = a + 0.975(b-a) \tag{10}$$

The x values corresponding to y_l and y_h are determined as

$$x_l = f^{-1}(y_l) \quad x_h = f^{-1}(y_h) \tag{11}$$

where f^{-1} is the inverse function of the logistic function

$$x = f^{-1}(y) = g(y) = -\frac{1}{c}\ln\frac{b-y}{y-a} + d \tag{12}$$



(a) Computation of delta MOS. (b) Computation of delta rate.

Figure 2 Integration bounds: the shaded area represents the integral of the difference of the two curves, evaluated between the lower and upper bounds.

3.3. Average MOS difference

To approximate the R-D curve given by a set of N bit rate values $(R_1, ..., R_N)$ with corresponding MOS $(D_1, ..., D_N)$, the R-D values are fitted in a least square sense using a logistic function with the constraints specified in Eq. (9)

$$\hat{D}(r) = a + \frac{b-a}{1+\exp\left[-c(r-d)\right]} \tag{13}$$

where $\hat{D}$ is the fitted distortion in MOS, r is the logarithm of the bit rate, and a, b, c, and d are the parameters of the fitting function.

Similarly to the Bjøntegaard model [10], the average MOS difference between two R-D curves is approximated by the difference between the integrals of the fitted R-D curves divided by the integration interval

$$\Delta D = \mathrm{E}\left[D_2 - D_1\right] \approx \frac{1}{r_H - r_L} \int_{r_L}^{r_H} \left[\hat{D}_2(r) - \hat{D}_1(r)\right] \mathrm{d}r \tag{14}$$

where ΔD is the delta MOS computed between the two fitted R-D curves $\hat{D}_1(r)$ and $\hat{D}_2(r)$, respectively, and the integration bounds, r_L and r_H, are

$$\begin{aligned} r_L &= \max\left\{\min\left(r_{1,1}, ..., r_{1,N_1}\right), \min\left(r_{2,1}, ..., r_{2,N_2}\right), \min\left(r_{1,l}, r_{2,l}\right)\right\} \\ r_H &= \min\left\{\max\left(r_{1,1}, ..., r_{1,N_1}\right), \max\left(r_{2,1}, ..., r_{2,N_2}\right), \max\left(r_{1,h}, r_{2,h}\right)\right\} \end{aligned} \tag{15}$$

where $r_{1,l}$ and $r_{1,h}$ ($r_{2,l}$ and $r_{2,h}$) are lower and upper rate bounds on $\hat{D}_1(r)$ ($\hat{D}_2(r)$) determined according to Eq. (11).

To compute ΔD, the analytical expression of the integral of the logistic function is used

$$F(x) = \int f(x)\mathrm{d}x = \frac{b-a}{c} \ln\left\{1 + \exp\left[-c(x-d)\right]\right\} + bx + (a-b)d + C \tag{16}$$

where C is an arbitrary constant.

Figure 2(a) illustrates the computation of the average MOS difference between two fitted R-D curves.

3.4. Average bit rate difference

Instead of applying another fitting to express the (logarithm of the) bit rate as a function of the distortion, as in the Bjøntegaard model [10], the inverse function of Eq. (13) is used

$$\hat{r}(D) = -\frac{1}{c} \ln \frac{b-D}{D-a} + d \tag{17}$$

where $\hat{r}$ is the fitted bit rate, D is the distortion in MOS, and a, b, c, and d are the parameters determined for Eq. (13). Therefore, the logistic fitting is applied only once for a given set of R-D values.

Similarly to the Bjøntegaard model, the average bit rate difference between two R-D curves is approximated as

$$\Delta R = \mathrm{E}\left[\frac{R_2 - R_1}{R_1}\right] \approx 10^{\frac{1}{D_H - D_L} \int_{D_L}^{D_H} \left[\hat{r}_2(D) - \hat{r}_1(D)\right] \mathrm{d}D} - 1 \tag{18}$$

where ΔR is the delta rate computed between the two fitted R-D curves $\hat{r}_1(r)$ and $\hat{r}_2(r)$, respectively, and the integration bounds, D_L and D_H, are

$$\begin{aligned} D_L &= \max\left\{\min(\hat{D}_{1,1}, ..., \hat{D}_{1,N_1}), \min(\hat{D}_{2,1}, ..., \hat{D}_{2,N_2}), \min(\hat{D}_{1,l}, \hat{D}_{2,l})\right\} \\ D_H &= \min\left\{\max(\hat{D}_{1,1}, ..., \hat{D}_{1,N_1}), \max(\hat{D}_{2,1}, ..., \hat{D}_{2,N_2}), \max(\hat{D}_{1,h}, \hat{D}_{2,h})\right\} \end{aligned} \tag{19}$$

where $\hat{D}_{1,l}$ and $\hat{D}_{1,h}$ ($\hat{D}_{2,l}$ and $\hat{D}_{2,h}$) are the lower and upper distortion bounds on $\hat{D}_1(r)$ ($\hat{D}_2(r)$) determined according to Eq. (10).

To compute ΔR, the analytical expression of the integral of the inverse logistic function is used

$$G(y) = \int g(y)\mathrm{d}y = \frac{b-y}{c}\left[\ln(b-y) - 1\right] + \frac{y-a}{c}\left[\ln(y-a) - 1\right] + dy + C \tag{20}$$

where C is an arbitrary constant.

Figure 2(b) illustrates the computation of the average bit rate difference between two fitted R-D curves.

3.5. Confidence interval

The mean opinion score ($\overline{u}_i$) is a statistical measure

$$\overline{u}_i = \frac{1}{M}\sum_{j=1}^{M} u_{ij} \tag{21}$$

where M is the number of valid subjects and u_{ij} is the score by subject j for the test condition i (specific combination of content, codec, and bit rate).

The relationship between the estimated mean values based on a sample of the population (i.e., the subjects who took part in the experiment) and the true mean values of the entire population is given by the confidence interval of the estimated mean. In our experiments, the $100 \times (1-\alpha)\%$ confidence intervals for mean opinion scores were computed using the Student's t-distribution

$$[\overline{u}_i - \delta_i, \overline{u}_i + \delta_i] \tag{22}$$

with

$$\delta_i = t(1-\alpha/2, M)\frac{s_i}{\sqrt{M}} \tag{23}$$

$$s_i = \sqrt{\frac{1}{M-1}\sum_{j=1}^{M}\left(u_{ij} - \overline{u}_i\right)^2} \tag{24}$$

where M is the number of valid subjects, s_i is the sample standard deviation of a single test condition i across the subjects j, and $t(1-\alpha/2, M)$ is the t-value corresponding to a two-tailed Student's t-distribution with $M-1$ degrees of freedom and a desired significance level α (equal to 1-degree of confidence). The confidence intervals are computed for an α equal to 0.05, which corresponds to a degree of confidence of 95%.

To consider the statistical property of mean opinion scores, the corresponding confidence intervals should be considered in the proposed model when computing the average MOS and bit rate differences. In recommendation ITU-R BT.500-13 [2], it is proposed to consider three series of grades, constructed from the mean opinion scores for each test condition and associated 95% confidence intervals

a) minimum grade series $(\overline{u}_1 - \delta_1, ..., \overline{u}_N - \delta_N)$,
b) mean grade series $(\overline{u}_1, ..., \overline{u}_N)$,
c) maximum grade series $(\overline{u}_1 + \delta_1, ..., \overline{u}_N + \delta_N)$.

According to this recommendation, the three grade series should be fitted independently.

Figure 3 depicts an example of mean opinion scores and associated 95% confidence interval. The fitting functions $\hat{D}^-(r)$, $\hat{D}(r)$, and $\hat{D}^+(r)$ (see Table 1) for the minimum, mean, and maximum grade series, respectively, are drawn on the same graph to provide an estimate of the 95% continuous confidence region, which can be used to determine a tolerance range. The space between $\hat{D}^+(r)$ and $\hat{D}^-(r)$ is not an exact 95% confidence interval, but a mean estimate thereof [2].

Table 1 Fitting functions for the different grade series.

Fitting functions	Fitting of	Values
$\hat{D}^-(r), \hat{r}^-(D)$	minimum grade series	$(\overline{u}_1 - \delta_1, ..., \overline{u}_N - \delta_N)$
$\hat{D}(r), \hat{r}(D)$	mean grade series	$(\overline{u}_1, ..., \overline{u}_N)$
$\hat{D}^+(r), \hat{r}^+(D)$	maximum grade series	$(\overline{u}_1 + \delta_1, ..., \overline{u}_N + \delta_N)$

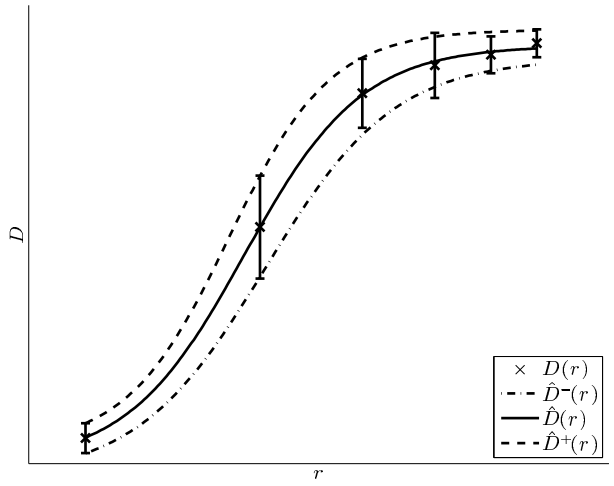


Figure 3 Different grade series: $\hat{D}^-(r)$, $\hat{D}(r)$, and $\hat{D}^+(r)$ are the fitting functions for the minimum, mean, and maximum grade series, respectively, constructed from the mean opinion scores for each test condition and associated 95% confidence intervals.

The parameters a, b, and c of the logistic function are constrained as the subjective scores should increase from the lower to the upper categories as the bit rate increases (see Sec. 3.1). These constraints should be modified when fitting the minimum and maximum grade series to consider the confidence intervals. If we consider a typical R-D curve and rating scale divided into five categories, at the extreme parts of the curve, the confidence intervals generally tends to become smaller, due to the intrinsic nature of bounded rating scales, but they may slightly span outside of the extreme categories. Therefore, for the fitting of the minimum (maximum) grade series, we decrease (increase) the lower (upper) bound on parameters a and b by half of the "length" of one category (see Table 2).

The average MOS and bit rate differences are computed from the mean grades series according to Secs. 3.3 and 3.4, respectively. The corresponding 95% confidence interval is estimated using the minimum and maximum grade series to consider the confidence intervals associated with the mean opinion scores.

The average MOS difference, ΔD, and its corresponding estimated 95% confidence interval $[\Delta D_{\min}, \Delta D_{\max}]$, are

$$\begin{aligned}
\Delta D &= \phi\left(\hat{D}_1(r), \hat{D}_2(r), r_L, r_H\right) \\
\Delta D_{\min} &= \min\left\{\phi\left(\hat{D}_1^-(r), \hat{D}_2^+(r), r_L, r_H\right), \phi\left(\hat{D}_1^+(r), \hat{D}_2^-(r), r_L, r_H\right)\right\} \\
\Delta D_{\max} &= \max\left\{\phi\left(\hat{D}_1^-(r), \hat{D}_2^+(r), r_L, r_H\right), \phi\left(\hat{D}_1^+(r), \hat{D}_2^-(r), r_L, r_H\right)\right\}
\end{aligned} \tag{25}$$

where r_L and r_H are the integration bounds computed from $(r_{1,1}, ...r_{1,N_1})$, $(r_{2,1}, ...r_{2,N_2})$, $\hat{r}_1(D)$, and $\hat{r}_2(D)$ according to Eq. (15), and ϕ is a generic function to compute the average MOS difference between two fitted R-D curves $\hat{D}_1(r)$ and $\hat{D}_2(r)$, between r_L and r_H

$$\phi\left(\hat{D}_1(r), \hat{D}_2(r), r_L, r_H\right) = \frac{1}{r_H - r_L}\int_{r_L}^{r_H}\left[\hat{D}_2(r) - \hat{D}_1(r)\right]\mathrm{d}r \tag{26}$$

Table 2 Constraints for the different fitting functions.

Fitting functions	Constraints on parameter		
	a	b	c
$\hat{D}^-(r), \hat{r}^-(D)$	$u_{\min} - \frac{1}{10}\Delta u \le a \le u_{\min} + \frac{1}{5}\Delta u$	$u_{\max} - \frac{3}{10}\Delta u \le b \le u_{\max}$	$c > 0$
$\hat{D}(r), \hat{r}(D)$	$u_{\min} \le a \le u_{\min} + \frac{1}{5}\Delta u$	$u_{\max} - \frac{1}{5}\Delta u \le b \le u_{\max}$	$c > 0$
$\hat{D}^+(r), \hat{r}^+(D)$	$u_{\min} \le a \le u_{\min} + \frac{3}{10}\Delta u$	$u_{\max} - \frac{1}{5}\Delta u \le b \le u_{\max} + \frac{1}{10}\Delta u$	$c > 0$

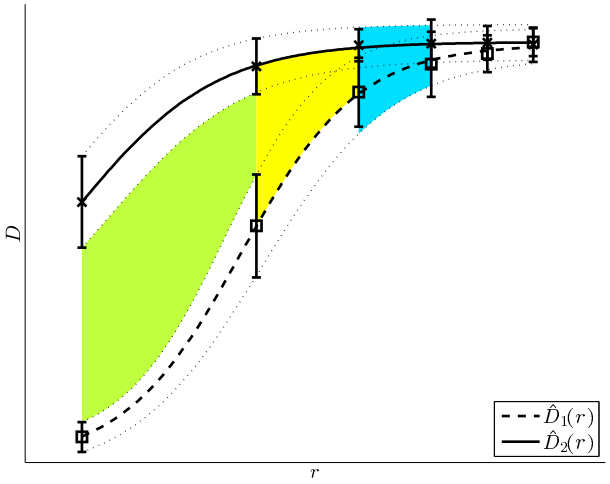


Figure 4 Confidence interval: the green, yellow, and blue areas illustrate the calculation of $\Delta D_{\min}$, ΔD, and $\Delta D_{\max}$, respectively. For illustration purpose, only part of the total area used in the calculation of each value is represented. The same principle applies for the calculation of $\Delta R_{\min}$, ΔR, and $\Delta R_{\max}$.

Figure 4 illustrates the calculation of $\Delta D_{\min}$, ΔD, and $\Delta D_{\max}$.

The average bit rate difference, ΔR, and its corresponding estimated 95% confidence interval $[\Delta R_{\min}, \Delta R_{\max}]$, are

$$\begin{aligned} \Delta R &= \psi\left(\hat{r}_1(D), \hat{r}_2(D), D_L, D_H\right) \\ \Delta R_{\min} &= \min\left\{\psi\left(\hat{r}_1^-(D), \hat{r}_2^+(D), D_L, D_H\right), \psi\left(\hat{r}_1^+(D), \hat{r}_2^-(D), D_L, D_H\right)\right\} \\ \Delta R_{\max} &= \max\left\{\psi\left(\hat{r}_1^-(D), \hat{r}_2^+(D), D_L, D_H\right), \psi\left(\hat{r}_1^+(D), \hat{r}_2^-(D), D_L, D_H\right)\right\} \end{aligned} \tag{27}$$

where D_L and D_H are the integration bounds computed from $(D_{1,1}, ...D_{1,N_1})$, $(D_{2,1}, ...D_{2,N_2})$, $\hat{D}_1(r)$, and $\hat{D}_2(r)$ according to Eq. (15), and ψ is a generic function to compute the average bit rate difference between two fitted R-D curves $\hat{r}_1(D)$ and $\hat{r}_2(D)$, between D_L and D_H

$$\psi\left(\hat{r}_1(D), \hat{r}_2(D), D_L, D_H\right) = 10^{\frac{1}{D_H - D_L}\int_{D_L}^{D_H}[\hat{r}_2(D) - \hat{r}_1(D)]\mathrm{d}D} - 1 \tag{28}$$

3.6. Confidence index

To provide confident measures, the R-D curves should ideally cover the full range of the rating scale. In most quality evaluations, both objective and subjective, a predefined set of targeted bit rates is usually considered. In well-designed subjective tests, the lower bit rate is chosen such that at least one test stimulus (specific combination of content, codec, and bit rate) would have a quality corresponding to the lower category. However, care should be taken to avoid too low quality test stimuli. Therefore, it is possible that at the lower bit rate, one codec produces bad quality, whereas another codec produces fair or good quality, if there is a significant difference in terms of compression efficiency between the two codecs.

These considerations are incorporated in the proposed model to produce a confidence index on the calculated average MOS and bit rate differences. As it is impossible in most practical situations to cover the full range of the rating scale with both R-D curves for the above-mentioned reason, we assume that at least one of the two R-D curves should cover 80% of the rating scale to have a valid measure of the average MOS and bit rate differences. The range of the rating scale, Δu_1 and Δu_2, covered by the two R-D curves is

$$\begin{aligned} \Delta u_1 &= \max(D_{1,1}, ..., D_{1,N_1}) - \min(D_{1,1}, ..., D_{1,N_1}) \\ \Delta u_2 &= \max(D_{2,1}, ..., D_{2,N_2}) - \min(D_{2,1}, ..., D_{2,N_2}) \end{aligned} \tag{29}$$

We also consider the goodness of the fitting functions, measured in terms of the Pearson correlation coefficient

$$\begin{aligned} \rho_1 &= r\left((D_{1,1}, ..., D_{1,N_1}), (\hat{D}_{1,1}, ..., \hat{D}_{1,N_1})\right) \\ \rho_2 &= r\left((D_{2,1}, ..., D_{2,N_2}), (\hat{D}_{2,1}, ..., \hat{D}_{2,N_2})\right) \end{aligned} \tag{30}$$

where $r(\cdot)$ is the Pearson correlation coefficient.

The confidence index is computed as

$$\text{Confidence index} = \min\left\{1, \frac{\max(\Delta u_1, \Delta u_2)}{0.8(u_{\max} - u_{\min})}\rho_1\rho_2\right\} \tag{31}$$

where $u_{\min}$ and $u_{\max}$ are the boundaries of the rating scale.

4. Applications and discussions

In this section, three case studies are presented where the Bjøntegaard and proposed models were used to calculate average coding efficiency. The aim of these examples is twofold. The first objective is to show that the Bjøntegaard model does not always provide an accurate measure of coding efficiency, whereas the proposed model should report more realistic coding efficiency. However, as there is no ground truth for the coding efficiency, it is impossible to quantify the performance of the two models, but rather to discuss when the two models do not agree. The second objective is to illustrate the usefulness of the confidence intervals and confidence index provided by the proposed model.

4.1. Quality of high resolution images

In this case study, we used the results of an evaluation of four compression algorithms on a dataset of high resolution images. First, the dataset is described. Then, the results of the Bjøntegaard and proposed models are given and analyzed.

Table 4 Proposed model.

(a) Delta rate.

Encoding	Average bit rate difference relative to			
	JPEG	JPEG 2000 4:2:0	JPEG 2000 4:4:4	HEVC
JPEG	-	−9% [−25%,+6%] (100%)	+34% [+13%,+68%] (100%)	+5% [−14%,+26%] (100%)
J2000 4:2:0	+10% [−6%,+33%] (100%)	-	+61% [+24%,+109%] (77%)	+19% [−9%,+53%] (85%)
J2000 4:4:4	−25% [−40%,−12%] (100%)	−38% [−52%,−19%] (77%)	-	−26% [−45%,−3%] (86%)
HEVC	−5% [−21%,+17%] (100%)	−16% [−35%,+10%] (85%)	+36% [+3%,+81%] (86%)	-

A negative (positive) value indicates a decrease (increase) of bit rate for the same MOS. Reading: ΔR [$\Delta R_{\min}$, $\Delta R_{\max}$] (Confidence index)

(b) Delta MOS.

Encoding	Average MOS difference relative to			
	JPEG	JPEG 2000 4:2:0	JPEG 2000 4:4:4	HEVC
JPEG	-	−0.5 [−11.4,+11.2] (100%)	−18.7 [−30.0,−7.0] (100%)	−7.2 [−18.9,+3.8] (100%)
J2000 4:2:0	+0.5 [−11.2,+11.4] (100%)	-	−18.2 [−29.0,−5.5] (77%)	−6.6 [−17.8,+5.3] (85%)
J2000 4:4:4	+18.7 [+7.0,+30.0] (100%)	+18.2 [+5.5,+29.0] (77%)	-	+11.5 [−1.1,+23.0] (86%)
HEVC	+7.2 [−3.8,+18.9] (100%)	+6.6 [−5.3,+17.8] (85%)	−11.5 [−23.0,+1.1] (86%)	-

A negative (positive) value indicates a decrease (increase) of MOS for the same bit rate. Reading: ΔD [$\Delta D_{\min}$, $\Delta D_{\max}$] (Confidence index)

Table 3 Bjøntegaard model.

(a) Delta rate.

Encoding	Average bit rate difference relative to			
	JPEG	J2000 4:2:0	J2000 4:4:4	HEVC
JPEG	-	+78%	+26%	+112%
J2000 4:2:0	−44%	-	−31%	+18%
J2000 4:4:4	−21%	+44%	-	+73%
HEVC	−53%	−15%	−42%	-

A negative (positive) value indicates a decrease (increase) of bit rate for the same PSNR.

(b) Delta PSNR.

Encoding	Average PSNR difference relative to			
	JPEG	J2000 4:2:0	J2000 4:4:4	HEVC
JPEG	-	−3.4dB	−1.3dB	−4.4dB
J2000 4:2:0	+3.4dB	-	+2.1dB	−1.0dB
J2000 4:4:4	+1.3dB	−2.1dB	-	−3.1dB
HEVC	+4.4dB	+1.0dB	+3.1dB	-

A negative (positive) value indicates a decrease (increase) of PSNR for the same bit rate.

4.1.1. Dataset

The dataset was composed of ten high resolution image contents, four for the training and six for the test. All the images had a resolution of 1280×1600 pixels and were available in RGB 4:4:4 uncompressed format. The images were compressed using JPEG, JPEG 2000 (both YCbCr 4:2:0 and RGB 4:4:4 configurations), and HEVC intra profile at 0.25, 0.5, 0.75, 1, 1.25, and 1.5bpp.

An adaptation of the double-stimulus continuous quality scale (DSCQS) method [2] was used to evaluate the quality of the test stimuli. The selected methodology implies that two images were displayed simultaneously by splitting the screen horizontally into two parts. One of the two images was always the reference (unimpaired) image. The other was the test image, which in this study was a compressed version of the reference. Instead of judging the quality of both images, the subject was asked to detect the impaired image in the pair and rate its quality, using a continuous quality scale ranging from 0 to 100, associated with five distinct quality levels (*Bad*, *Poor*, *Fair*, *Good*, and *Excellent*).

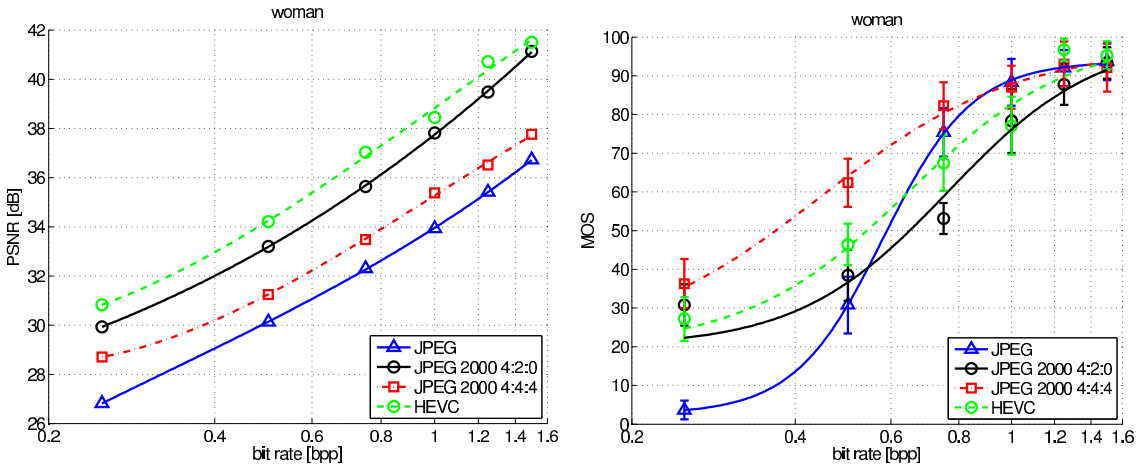
The subjective results were processed by first detecting and removing subjects whose scores appeared to deviate strongly from others in each test session. Then, the mean opinion score was computed for each test stimulus as the mean across the rates of the valid subjects. Readers can refer to [18] for more details about the dataset and subjective evaluation.

4.1.2. Results and discussion

Tables 3 and 4 report the coding efficiency calculated for content *woman* using the Bjøntegaard and proposed models, respectively. Figure 5 shows the fitted R-D curves for content *woman*.

Table 3(a) reports an average bit rate difference for JPEG 2000 4:2:0 over JPEG 2000 4:4:4 of −31% based on the Bjøntegaard model. However, Table 4(a) reports an average bit rate difference of +61% [+24%,+109%] based on the proposed model. Note that the 95% confidence interval resulting from the proposed model does not contain the value calculated by the Bjøntegaard model. These results show that JPEG 2000 4:2:0 has better coding efficiency than JPEG 2000 4:4:4 according to the Bjøntegaard model, whereas the proposed model dictates the opposite. To understand why the two models lead to different conclusions, it is necessary to analyze the objective and subjective scores. According to PSNR measurements, JPEG 2000 4:2:0 performed always better than JPEG 2000 4:4:4 (see Fig. 5(a)), whereas the subjective results dictate the opposite (see Fig. 5(b)).

Visual weighting was disabled for JPEG 2000 4:2:0, whereas



(a) Bjøntegaard model.

(b) Proposed model.

Figure 5 Rate-distortion curves for content *woman*.

it was enabled for JPEG 2000 4:4:4. The lack of visual weighting creates distortions, particularly at lower bit rates, as reported during the development of JPEG 2000. This example shows that when PSNR fails to capture a specific distortion, the comparison of coding efficiency using the Bjøntegaard model may lead to wrong conclusion. In this case, the proposed model, which relies on subjective scores, should report more realistic estimation of coding efficiency.

Table 3(a) reports an average bit rate difference over JPEG of −44% and −53% for JPEG 2000 4:2:0 and HEVC, respectively, based on the Bjøntegaard model. However, Table 4(a) reports an average bit rate difference over JPEG of +10% [−6%,+33%] and −5% [−21%,+17%] for JPEG 2000 4:2:0 and HEVC, respectively, based on the proposed model. Note that the 95% confidence intervals resulting from the proposed model do not contain the values calculated by the Bjøntegaard model. As it can be observed from Fig. 5, HEVC outperformed JPEG by at least 3dB on all bit rates, whereas JPEG was evaluated better than or equal to HEVC at 0.75bpp and above based on the subjective results. This example shows that the coding efficiency reported by the Bjøntegaard model may be over-estimated in some cases.

It is known that PSNR does not accurately reflect human perception of visual quality [1]. As the Bjøntegaard model relies on PSNR measurements, it is not surprising that the coding efficiency calculated with this model may not accurately reflect the true coding efficiency in some cases. Using a different model relying on a perceptual metric that better correlates with perceived quality, e.g., Structural Similarity Index (SSIM), would probably result in more accurate estimation of coding efficiency.

4.2. Quality of ultra-high definition video sequences

In this case study, we used the results of an evaluation of H.264/AVC and H.265/HEVC on a dataset of ultra-high definition video sequences. First, the dataset is described. Then, the results of the Bjøntegaard and proposed models are given and analyzed.

4.2.1. Dataset

The dataset was composed of four ultra-high definition video contents, one for the training (*Sintel39*) and three for the test (*PeopleOnStreet*, *Traffic*, and *Sintel2*), with different visual characteristics, resolutions, and frame rates. All test sequences were stored as raw video files, progressively scanned, with YCbCr 4:2:0 color sampling, and 8 bits per sample. The video sequences were compressed with H.264/AVC and H.265/HEVC using the *Random Access* configuration. For each content and codec, five different bit rates were selected.

The double stimulus impairment scale (DSIS) method [2], Variant II, with a continuous impairment scale ranging from 0 to 100, associated with five distinct impairment levels (*Very annoying*, *Annoying*, *Slightly annoying*, *Perceptible but not annoying*, and *Imperceptible*), was chosen to perform the subjective quality assessment experiments.

The subjective results were processed by first detecting and removing subjects whose scores appeared to deviate strongly from others in each test session. Then, the mean opinion score was computed for each test stimulus as the mean across the rates of the valid subjects. Readers can refer to [19] for more details about the dataset and subjective evaluation.

4.2.2. Results and discussion

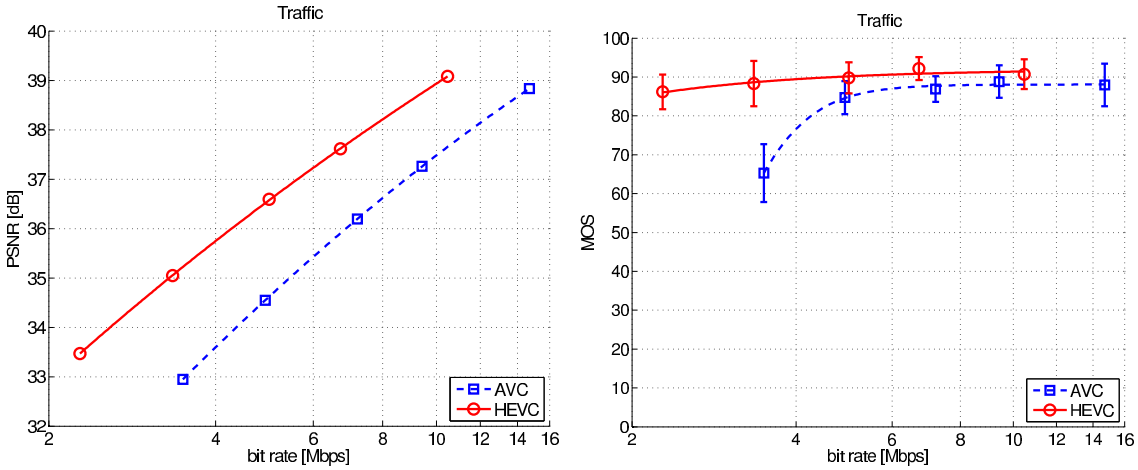
Table 5 reports the coding efficiency for H.265/HEVC over H.264/AVC calculated on each test content using the Bjøntegaard and proposed models. Figure 6 shows the fitted R-D curves for content *Traffic*.

For content *Traffic*, subjects evaluated nine out of ten video sequences as *Imperceptible* (see Fig. 6(b)). These results show that, at the selected bit rates, the R-D curves are mostly in the upper saturation phase. However, it is impossible to predict this behavior from the PSNR measurements as the two curves are

Table 5 Average coding efficiency for H.265/HEVC over H.264/AVC.

Model	Bjøntegaard		Proposed		
Content	Delta rate	Delta PSNR	Delta rate ΔR [$\Delta R_{\min}, \Delta R_{\max}$]	Delta MOS ΔD [$\Delta D_{\min}, \Delta D_{\max}$]	Confidence index
PeopleOnStreet	−27%	+1.6dB	−53% [−69%,−27%]	+25.8 [+13.0,+38.4]	79%
Traffic	−38%	+1.8dB	−59% [-,−5%]	+10.8 [−2.2,+20.3]	28%
Sintel2	−68%	+4.4dB	−73% [-,−60%]	+40.7 [+28.9,+52.4]	62%
Overall	−44%	+2.6dB	−62% [-,−31%]	+25.8 [+13.2,+37.1]	56%

Bjøntegaard model: A negative (positive) delta rate indicates a decrease (increase) of bit rate for the same PSNR. A negative (positive) delta PSNR indicates a decrease (increase) of PSNR for the same bit rate. Proposed model: A negative (positive) delta rate indicates a decrease (increase) of bit rate for the same MOS. A negative (positive) delta MOS indicates a decrease (increase) of MOS for the same bit rate.



(a) Bjøntegaard model.

(b) Proposed model.

Figure 6 Rate-distortion curves for content *Traffic*.

continuously increasing and the PSNR values are below 40dB, which is often considered as excellent quality.

For this particular content, the R-D values were mostly measured in the upper saturation phase, and not across the entire rating scale, as recommended. Therefore, the average PSNR/MOS and bit rate differences calculated using the two models are not representative of the true coding efficiency for this content. Nevertheless, for the proposed model, this problem is reflected in the low confidence index (28%) and wide confidence interval reported in Table 5. Note that the value for $\Delta R_{\min}$ could not be determined as there was no overlap between the two R-D curves. However, the the Bjøntegaard model does not consider the saturation effect of the human visual system and does not provide such indication regarding the confidence of the calculated coding efficiency.

4.3. Quality of 3D video sequences

In this case study, we used the results of the evaluations of the MPEG Call for Proposals (CfP) on 3D Video Coding Technology [20]. First, the dataset is described. Then, the results of the Bjøntegaard and proposed models are given and analyzed.

4.3.1. Dataset

The test material used in the MPEG CfP is composed of eight different contents encoded at four target bit rates. The contents were divided in two classes: Class A, with a spatial resolution of 1920 × 1088 pixels and a temporal resolution of 25 frames per seconds, and Class C, with 1024 × 768 pixels at 30 frames per second. All contents are ten seconds long. All test sequences were stored as raw video files, progressively scanned, with YCbCr 4:2:0 color sampling, and 8 bits per sample. Twenty-two coding algorithms, submitted by the proponents, and two anchors were evaluated in the tests. In this paper, only the results for the 3-view configuration, fixed stereo pair, of two of the best HEVC-compatible proposals (*P18* and *P25*) and the two anchors (*P09* and *P10*) were used. In the evaluations of the 3-view configuration, the displayed stereo pair was formed from two synthesized views. In this paper, the PSNR was computed as the average PSNR of the left and right views of the displayed stereo pair.

The double stimulus impairment scale (DSIS) method [2], Variant II, with an 11-grade numerical categorical scale ranging from 0 (lowest quality) to 10 (highest quality), was chosen to perform the subjective quality assessment experiments.

In this paper, mean opinion score and corresponding 95% confidence intervals that were computed by the MPEG test coordinator on a total of 36 naïve viewers from three different laboratories [20] have been used. Outlier detection was performed by the MPEG test coordinator according to the procedure adopted by the ITU Video Quality Experts Group (VQEG)

Table 7 Proposed model.

(a) Delta rate.

Encoding	Average bit rate difference relative to			
	P09	*P10*	*P18*	*P25*
P09	-	+10% [+2%,+22%] (71%)	+146% [+131%,+164%] (72%)	+155% [+134%,+178%] (71%)
P10	−9% [−18%,−2%] (71%)	-	+131% [+111%,+149%] (71%)	+142% [+115%,+180%] (70%)
P18	−59% [−62%,−57%] (72%)	−57% [−60%,−53%] (71%)	-	−8% [−14%,+2%] (44%)
P25	−61% [−64%,−57%] (71%)	−59% [−64%,−53%] (70%)	+8% [−2%,+16%] (44%)	-

A negative (positive) value indicates a decrease (increase) of bit rate for the same MOS. Reading: ΔR [$\Delta R_{\min}$, $\Delta R_{\max}$] (Confidence index)

(b) Delta MOS.

Encoding	Average MOS difference relative to			
	P09	*P10*	*P18*	*P25*
P09	-	−0.5 [−1.1,−0.0] (71%)	−4.6 [−5.1,−4.1] (72%)	−4.3 [−4.8,−3.8] (71%)
P10	+0.5 [+0.0,+1.1] (71%)	-	−4.1 [−4.5,−3.6] (71%)	−3.8 [−4.2,−3.3] (70%)
P18	+4.6 [+4.1,+5.1] (72%)	+4.1 [+3.6,+4.5] (71%)	-	+0.3 [−0.1,+0.8] (44%)
P25	+4.3 [+3.8,+4.8] (71%)	+3.8 [+3.3,+4.2] (70%)	−0.3 [−0.8,+0.1] (44%)	-

A negative (positive) value indicates a decrease (increase) of MOS for the same bit rate. Reading: ΔD [$\Delta D_{\min}$, $\Delta D_{\max}$] (Confidence index)

Table 6 Bjøntegaard model.

(a) Delta rate.

Encoding	Average bit rate difference relative to			
	P09	*P10*	*P18*	*P25*
P09	-	+16%	+159%	−11%
P10	−14%	-	+123%	−22%
P18	−61%	−55%	-	-
P25	+12%	+28%	-	-

A negative (positive) value indicates a decrease (increase) of bit rate for the same PSNR.

(b) Delta PSNR.

Encoding	Average PSNR difference relative to			
	P09	*P10*	*P18*	*P25*
P09	-	−1.1dB	−5.9dB	+1.8dB
P10	+1.1dB	-	−4.9dB	+2.9dB
P18	+5.9dB	+4.9dB	-	+7.8dB
P25	−1.8dB	−2.9dB	−7.8dB	-

A negative (positive) value indicates a decrease (increase) of PSNR for the same bit rate.

for its Multimedia Project. Then, the mean opinion score was computed for each test stimulus as the mean across the rates of the valid subjects. Readers can refer to [21] for more details about the dataset and subjective evaluation.

4.3.2. Results and discussion

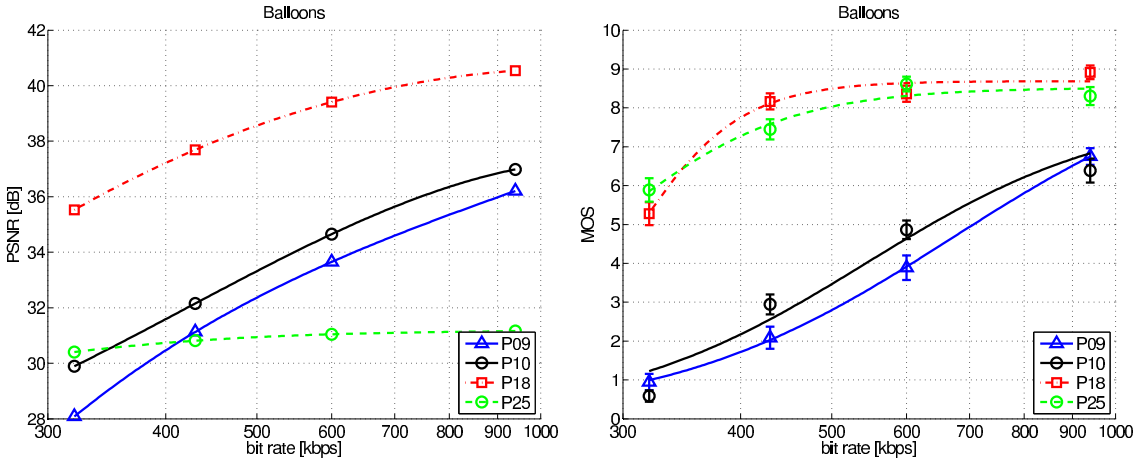
Tables 6 and 7 report the coding efficiency calculated for content *Balloons* using the Bjøntegaard and proposed models, respectively. Figure 7 shows the fitted R-D curves for content *Balloons*.

The average bit rate reduction values calculated using the Bjøntegaard model (see Table 6(a)) are in general similar to those calculated using the proposed model (see Table 7(a)), except for the values related to proponent *P25*. To understand why the two models differ for this particular proponent, it is necessary to analyze the objective and subjective scores. As it can be observed from Fig. 7, proponent *P25* obtained constant low PSNR values, whereas it obtained high subjective scores.

It is known that one proposal submitted in response to the CfP used a different view synthesis algorithm. As the data submitted by the proponents is anonymous, we cannot be certain that proponent *P25* used a different view synthesis algorithm. However, these results show that coding efficiency calculated based on PSNR measurements might not accurately reflect the true coding efficiency in the case of stereoscopic content formed from synthesized views, as PSNR is not accurate to assess perceived quality of synthesized views [22]. Using a different model relying on a perceptual metric that better correlates with perceived quality of stereoscopic content, e.g., Visual Information Fidelity (VIF) or Video Quality Metric (VQM) [22], would probably result in more accurate estimation of coding efficiency.

5. Conclusion and future work

In this paper, we proposed a model to calculate the average coding efficiency based on subjective quality scores. To consider the intrinsic nature of bounded rating scales, as well as nonlinearities and saturation effects of the human visual system, a logistic function was used to fit the R-D values. The average MOS and bit rate differences were computed between the fitted R-D curves. To consider the statistical property of subjective scores, the 95% confidence intervals associated with the MOS were considered to estimate corresponding confidence intervals on the calculated average MOS and bit rate differences. We presented three case studies where the Bjøntegaard and proposed models were used to calculate average coding efficiency. Results showed that the Bjøntegaard model does not always report



(a) Bjøntegaard model.

(b) Proposed model.

Figure 7 Rate-distortion curves for content *Balloons*.

an accurate measure of the true coding efficiency as it relies on PSNR measurements, which does not accurately reflect human perception of visual quality. Using a different model relying on a perceptual metric that better correlates with perceived quality would probably result in more accurate estimation of coding efficiency. However, the proposed model, which relies on subjective scores, is expected to report more realistic estimation of coding efficiency.

In future investigations, we plan to determine the influence of subjective evaluation methods and rating scales on the coding efficiency reported by the proposed model. Even though quality can be adequately assessed only by means of subjective tests, objective quality assessment is still preferred by most researchers due to its simplicity. Therefore, some of the concepts used in the proposed model, such as the computation of the integration bounds to consider the saturation effect of the human visual system or the reliability index, could be incorporated in the Bjøntegaard model to improve its prediction accuracy. A similar model for other objective metrics than PSNR could be used to compare codecs performance.

6. Acknowledgments

This work was conducted in the framework of the Swiss National Foundation for Scientific Research (FN 200021-143696-1) and COST IC1003 European Network on Quality of Experience in Multimedia Systems and Services QUALINET.

References

[1] H. Sheikh, M. Sabir, A. Bovik, A Statistical Evaluation of Recent Full Reference Image Quality Assessment Algorithms, IEEE Transactions on Image Processing 15 (2006) 3440–3451.

[2] ITU-R BT.500-13, Methodology for the subjective assessment of the quality of television pictures, International Telecommunication Union, 2012.

[3] F. De Simone, L. Goldmann, V. Baroncini, T. Ebrahimi, Subjective evaluation of JPEG XR image compression, in: Proceedings of SPIE 7443, Applications of Digital Image Processing XXXII, 2009.

[4] T. Alpert, V. Baroncini, D. Choi, L. Contin, R. Koenen, F. Pereira, H. Peterson, Subjective evaluation of MPEG-4 video codec proposals: Methodological approach and test procedures, Signal Processing: Image Communication 9 (1997) 305–325.

[5] T. Oelbaum, V. Baroncini, T. K. Tan, C. Fenimore, Subjective quality assessment of the emerging AVC/H. 264 video coding standard, in: International Broadcasting Conference (IBC), 2004.

[6] C. Fenimore, V. Baroncini, T. Oelbaum, T. K. Tan, Subjective testing methodology in MPEG video verification, in: Proceedings of SPIE 5558, Applications of Digital Image Processing XXVII, 2004, pp. 503–511.

[7] V. Baroncini, S. Quackenbush, MPEG Video/Audio Quality Evaluation, in: L. Chiariglione (Ed.), The MPEG Representation of Digital Media, Springer New York, 2012, pp. 249–261.

[8] T. Oelbaum, H. Schwarz, M. Wien, T. Wiegand, Subjective performance evaluation of the SVC extension of H.264/AVC, in: 15th IEEE International Conference on Image Processing (ICIP), 2008, pp. 2772–2775.

[9] F. D. Simone, L. Goldmann, J.-S. Lee, T. Ebrahimi, Towards high efficiency video coding: Subjective evaluation of potential coding technologies, Journal of Visual Communication and Image Representation 22 (2011) 734–748.

[10] G. Bjøntegaard, Calculation of average PSNR differences between RD-curves, Technical Report VCEG-M33, ITU-T SG16/Q6, Austin, Texas, USA, 2001.

[11] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, G. Sullivan, Rate-constrained coder control and comparison of video coding standards, IEEE Transactions on Circuits and Systems for Video Technology 13 (2003) 688–703.

[12] P. Merkle, A. Smolic, K. Muller, T. Wiegand, Efficient Prediction Structures for Multiview Video Coding, IEEE Transactions on Circuits and Systems for Video Technology 17 (2007) 1461–1473.

[13] J. Ohm, G. Sullivan, H. Schwarz, T. K. Tan, T. Wiegand, Comparison of the Coding Efficiency of Video Coding Standards - Including High Efficiency Video Coding (HEVC), IEEE Transactions on Circuits and Systems for Video Technology 22 (2012) 1669–1684.

[14] A. Vetro, D. Tian, Analysis of 3D and multiview extensions of the emerging HEVC standard, in: Proceedings of SPIE 8499, Applications of Digital Image Processing XXXV, 2012.

[15] ITU-T J.149, Method for specifying accuracy and cross-calibration of Video Quality Metrics (VQM), International Telecommunication Union, 2004.

[16] J. Korhonen, N. Burini, J. You, E. Nadernejad, How to evaluate objective video quality metrics reliably, in: 4th International Workshop on Quality of Multimedia Experience (QoMEX), 2012, pp. 57–62.

[17] G. Bjøntegaard, Improvements of the BD-PSNR model, Technical Report VCEG-AI11, ITU-T SG16/Q6, Berlin, Germany, 2008.

[18] P. Hanhart, M. Rerabek, P. Korsunov, T. Ebrahimi, Subjective evaluation of HEVC intra coding for still image compression, in: Proceedings of the

7th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM), 2013.

[19] P. Hanhart, M. Rerabek, F. D. Simone, T. Ebrahimi, Subjective quality evaluation of the upcoming HEVC video compression standard, in: Proceedings of SPIE 8499, Applications of Digital Image Processing XXXV, 2012.

[20] ISO/IEC JTC1/SC29/WG11, Report of Subjective Test Results from the Call for Proposals on 3D Video Coding Technology, Doc. N12347, Geneva, CH, 2011.

[21] P. Hanhart, F. De Simone, T. Ebrahimi, Quality Assessment of Asymmetric Stereo Pair Formed From Decoded and Synthesized Views, in: 4th International Workshop on Quality of Multimedia Experience (QoMEX), 2012.

[22] P. Hanhart, T. Ebrahimi, Quality Assessment of a Stereo Pair Formed From Two Synthesized Views Using Objective Metrics, in: 7th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM), 2013.