Analysis of a Large Library of MPEG–4 FGS Rate–Distortion Traces for Streaming Video
Institut Eurecom Technical Report RR-02-068

Philippe de Cuetos[*], Martin Reisslein[†], Keith W. Ross

Abstract

Recently, the MPEG–4 standard has been enriched with a new encoding tool designed to be used for video streaming, namely, the Fine–Grained Scalability (FGS). In this report, we present a publicly available library of frame size and quality traces of long MPEG–4 FGS encoded videos. A first analysis of these traces give useful insights for video streaming algorithms. In particular, we find that the base layer encoding process and the internal structure of the FGS–EL bitstream play an important role in the statistical characteristics of FGS–encoded videos.

1 Introduction

Fine Grained Scalability (FGS) has been recently added to the MPEG-4 standard [1] in order to increase the flexibility of streaming of encoded videos over networks. Similar to the other types of video scalability (SNR, spatial, or temporal), FGS encodes the video into a Base Layer (BL) and one or several Enhancement Layers (EL). However, with FGS, any number of bits of the enhancement layer can be suppressed at the server before transmission, and the decoder can use all of the truncated bitstream to maximize the video quality at the client [9]. When streaming FGS–encoded videos, the server typically chooses the number of bits to stream for each FGS–EL image (or any sequence of images) given the network conditions, as in [7].

[*] P. de Cuetos and K. W. Ross are with Institut Eurecom, 2229 Route des Cretes, 06904 Sophia–Antipolis, France (email: {decuetos,ross}@eurecom.fr)

[†] M. Reisslein is with the Telecommunications Research Center, Dept. of Electrical Engineering, Arizona State University, Goldwater Center, Tempe AZ 85287–7206 (email: reisslein@asu.edu)

In order to optimize the streaming of a given video, streaming algorithms should take the specific video characteristics into account [11]. In particular, by exploiting the rate–distortion characteristics of the individual images of the video, the video application may maximize the overall quality, while meeting the rate constraints imposed by the underlying network [3]. Maximizing the overall perceived quality of the video can be achieved by maximizing the quality of all individual images and minimizing the variations in quality between successive images [16]. This type of problem is in general approached by complex optimization algorithms that take the rate–distortion functions of all individual images for all video layers into account [3, 16]. In order to assess the performance of such algorithms for various video types and encoding methods, it is essential to have a database of rate–distortion functions of individual images from a range of complete videos. Generally, rate (i.e., frame sizes) and distortion depend strongly on the semantic video content. Therefore, in order to obtain results with meaningful statistical confidence levels, we need to use a large representative library of long videos (of several minutes). In this report, we present and analyze a publicly available library of frame size and quality traces of long MPEG–4 FGS encoded videos.

This report is organized as follows. We first review the main properties of MPEG–4 FGS encoding. Then, we present a framework for evaluation of streaming mechanisms that aim to maximize the overall video quality. Our framework considers the streaming at different aggregation levels: images, GoPs and scenes. Next, we explain how we obtained the rate-distortion traces for a representative library of long videos. Based on our framework, we analyze the statistical properties of these traces. We find that the base layer coding and the internal structure of the FGS–EL bitstream (in particular the bit–plane structure) play an important role in the statistical characteristics of FGS–encoded videos. All traces and statistics for long videos are made available on our web site (`http://peach.eas.asu.edu/index.html`).

2 Properties of the MPEG–4 FGS encoding

The FGS layered–encoding technique, as defined by MPEG–4 [1], encodes the quantization error between the original image and the corresponding BL–encoded image [9]. Figure 1 shows the architecture of the MPEG–4 FGS decoder as defined by the standard. Before transmission by the server, any number of bits can be truncated from the FGS–EL and all the remaining bits can be used by the decoder to enhance the quality of the BL at the client. The rendered

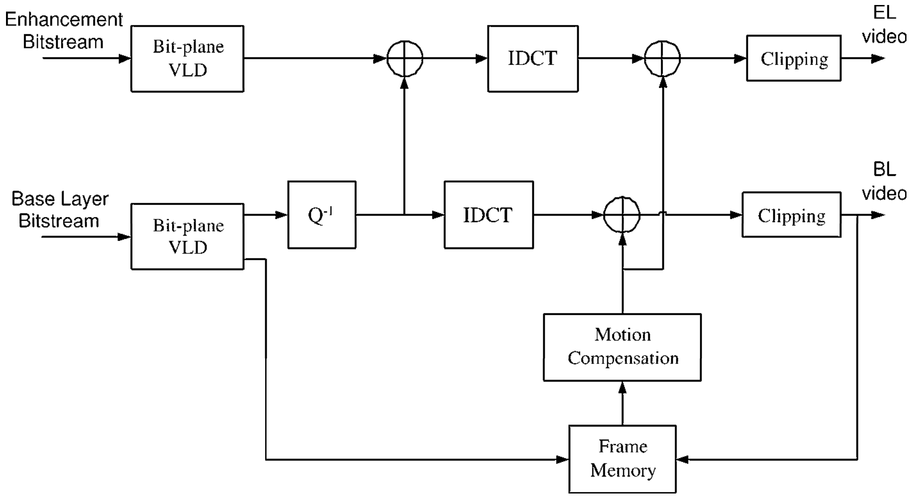

Figure 1: MPEG–4 FGS decoder structure

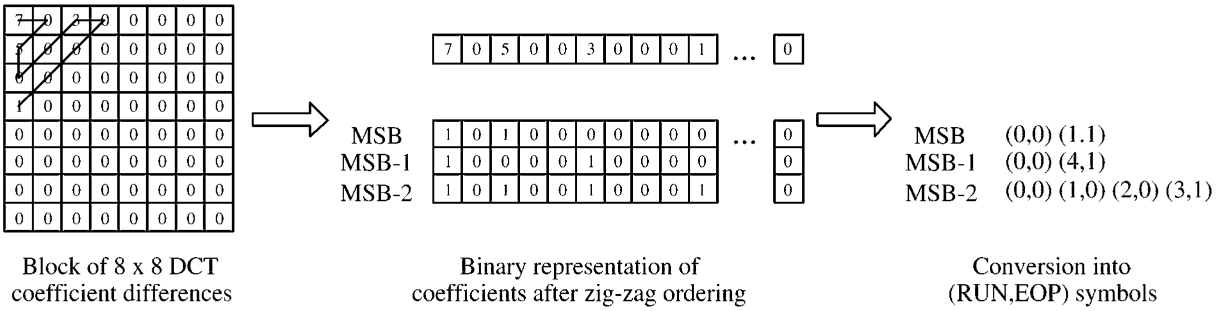
7 0 3 0 0 0 0 0
5 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0
1 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0
0 0 0 0 0 0 0 0
Block of 8 x 8 DCT coefficient differences
7 0 5 0 0 3 0 0 0 1 ... 0
MSB 1 0 1 0 0 0 0 0 0 0 ... 0
MSB-1 1 0 0 0 0 1 0 0 0 0 0
MSB-2 1 0 1 0 0 1 0 0 0 1 0
Binary representation of coefficients after zig-zag ordering
MSB (0,0) (1.1)
MSB-1 (0,0) (4,1)
MSB-2 (0,0) (1,0) (2,0) (3,1)
Conversion into (RUN,EOP) symbols

Figure 2: Example of bit–plane coding

image quality is directly proportional to the number of bits decoded.

The fine granularity property of the FGS–EL comes from the bit–plane encoding of the DCT coefficients (of the difference between the original image and the BL image after reconstruction). As shown in the example from Figure 2, the bit–plane coding method considers each DCT coefficient (for each 8×8 block) as a binary number of several bits [9]. After zig–zag ordering, the bits are ordered in bit–planes (from MSB to LSB) and each bit–plane is converted into (RUN,EOP) symbols. The number of bit–planes to code for each block depends on the block visual complexity, and should not exceed 8 (the MSB plane of each block is the first bit–plane that has a non zero bit). Less significant bit–planes usually have to be compressed using more bits because they have higher entropy.

With conventional non–layered encoders which feature different types of images (e.g. I, P and B in MPEG–1,2,4), the rate–distortion characteristics depend strongly on the type of the image which is considered (e.g., in general, fewer bits are needed to code B pictures compared to I or P pictures). This is thus also true for the BL of any scalable video. Moreover, since the FGS coding method, such as in MPEG–4, does not include motion compensation, one would expect that the amount of quality improvement brought by the EL frames depends only on the complexity of the individual images, and on the quality of the corresponding BL frame (which in turn depends on its encoding type). This is reflected in our results as we demonstrate later in the report.

3 Framework for Evaluating Streaming Mechanism

One of the particularities of videos is that successive images have strong temporal correlations. We define a *sequence* of images or frames as a temporal alignment of consecutive frames. Consecutive images of the same encoding type I, P, or B are likely to have roughly the same image quality for the same encoding rate. Therefore, the problem of maximizing the overall quality of the video can be approached by maximizing the overall quality of a sequence of images, and the problem of minimizing the variations in image quality can be approached by minimizing the differences in quality between the consecutive sequences. In this case, the server chooses the bit rate of the FGS–EL to stream to the client for every sequence of images, instead of for every single image. Because there is no motion compensation in the MPEG–4 FGS–EL and the decoding of the VBR–BL is supposed to achieve a roughly constant image quality, the server can choose to stream the same number of EL bits for each image within the sequence. This approach has the potential to reduce the computational load on video servers. In this study, we propose to analyze the traces image by image, but also GoP by GoP, and scene by scene. An example of streaming algorithm for stored FGS–encoded videos that operates scene by scene is given in [6].

Typically, long videos feature many different scenes composed of successive images with similar visual characteristics. Following Saw [13], we define a *video scene* as a sequence of images between two scene changes, where a scene change is defined as any distinctive difference between two adjacent images (this includes changes in motion as well as changes in the visual content).

3.1 Notations

In this paper, we use the terms "images" and "video frames" interchangeably. The frame period is constant and denoted by T seconds. The video is composed of N frames, which are indexed by n, $n = 1, \ldots, N$. Frame n is supposed to be decoded at discrete instant $t = n \cdot T$. The BL and FGS–EL of the video are VBR–encoded, with constant bitrates $r_b(t)$ and $r_e(t)$ during frame period n from $t = (n-1) \cdot T$ to $t = n \cdot T$, $n = 1, \ldots, N$. According to the FGS property, the EL can be truncated anywhere before decoding. We refer to any part of the EL which is to be added to the BL as an *EL substream*. We say that an EL substream is encoded at rate $C(t) \in [0, r_e(t)]$, when the last $T \cdot [r_e(t) - C(t)]$ bits of the frame which is supposed to be decoded at time t, $t = T, \ldots, N \cdot T$, have been removed from the original EL bitstream. We allow $C(t)$ to change only at discrete time instants $n \cdot T$. The BL Group of Pictures (GoP) is composed of 12 images throughout our study, and its pattern is fixed to IBBPBBPBBPBB.

We suppose that the video is partitioned into consecutive scenes. Let S denote the total number of scenes in a given video. Let s, $s = 1, \ldots, S$, denote the scene index and N_s the length (in number of images) of scene number s. (Note that $\sum_{s=1}^{S} N_s = N$.) For simplicity of presentation, we extend all notations that relate to scenes to GoPs, by considering that grouping the successive images within a same GoP is a special case of video segmentation. In the rest of the report, we specify explicitly each time the notations relate to GoPs instead of visual scenes.

3.2 Image-based metrics

Let $Q_n(C)$, $n = 1, \ldots, N$, denote the quality of the n^{th} decoded image, when the EL is encoded with rate C. Let $Q_n^b = Q_n(0)$, denote the quality of the same image, when only the BL is decoded. We define $Q_n^e(C) = Q_n(C) - Q_n^b$ as the improvement (increase) in quality which is achieved when decoding the EL encoded with rate C, as well as the BL of frame n.

The mean and sample variance of the individual images qualities, are estimated as:

$$\bar{Q}(C) = \frac{1}{N} \sum_{n=1}^{N} Q_n(C), \tag{1}$$

$$\sigma_Q^2(C) = \frac{1}{N-1} \sum_{n=1}^{N} [Q_n(C) - \bar{Q}(C)]^2 = \frac{1}{N-1} \left\{ \sum_{n=1}^{N} [Q_n(C)]^2 - [\bar{Q}(C)]^2 \right\}. \tag{2}$$

The coefficient of quality variation is given by:

$$CoV_Q = \frac{\sigma_Q(C)}{\bar{Q}(C)}. \tag{3}$$

The autocorrelation coefficient of the image qualities $\rho_Q(C,k)$ for lag k, $k = 1,\ldots,N$, is estimated as:

$$\rho_Q(C,k) = \frac{1}{N-k}\sum_{n=1}^{N-k}\frac{[Q_n(C)-\bar{Q}(C)][Q_{n+k}(C)-\bar{Q}(C)]}{\sigma_Q^2(C)}. \tag{4}$$

We denote the total size of image n by $X_n(C) = X_n^b + X_n^e(C)$, when the EL is encoded with rate C. Let $X_n^{e_i}$, $i = 1,\ldots,8$, $n = 1,\ldots,N$, denote the size of EL bit–plane i of image n and $Y_n^{e_i}$, $i = 1,\ldots,8$, $n = 1,\ldots,N$ denote the aggregate size of bit–planes $1,\ldots,i$, that is, $Y_n^{e_i} = \sum_{j=1}^{i} X_n^{e_j}$. In our notations, bit–plane number 1 denotes the Most Significant Bit–plane (MSB).

Let $Q_{s,n}(C)$, $s = 1,\ldots,S$, $n = 1,\ldots,N_s$, denote the quality of the n^{th} decoded image of scene s, when the EL is encoded with rate C, i.e., $Q_{s,n}(C) = Q_{N_1+\ldots+N_{s-1}+n}(C)$. Similar to $Q_n(C)$, we denote the quality of image n within scene s, when only the BL is decoded by $Q_{s,n}^b = Q_{s,n}(0)$, and the improvement in quality achieved when decoding the EL by $Q_{s,n}^e(C) = Q_{s,n}(C) - Q_{s,n}^b$. The Rate–Distortion (RD) characteristics of each image n within scene s are obtained by plotting the curves $Q_{s,n}(C)$.

The mean and sample variance of the qualities of the images within scene s, $s = 1,\ldots,S$, are denoted by $\bar{Q}_s(C)$ and $\sigma_{Q_s}^2(C)$. They are estimated in the same way as the mean and sample variance of individual image quality over the entire video.

3.3 Scene-based metrics

We define a general framework for studying the quality of long videos scene by scene. The mean in quality of all individual images of a scene, $\bar{Q}_s(C)$, may not be appropriate to account for the overall quality of the scene. First, the quality of individual images does not measure temporal artifacts, such as mosquito noise (moving artifacts around edges) or drifts (moving propagation of prediction errors after transmission). Secondly, high variations in quality between successive images within the same scene may decrease the perceptual global quality of the scene. For instance, a scene with alternating high and low quality images may have the same mean image quality as if the scene was played with medium but constant image quality, but the quality perceived by the user is likely to be much lower.

Let $\Theta_s(C)$ be the *overall quality* of video scene number s, $s = 1, \ldots, S$, when the EL has been coded at rate C for all images of the scene. Similar to the measure of quality of the individual images, we define $\Theta_s(C) = \Theta_s^b + \Theta_s^e(C)$, where $\Theta_s^b = \Theta_s(0)$ denotes the overall quality of scene s when only the BL is decoded, and $\Theta_s^e(C)$ the improvement in quality achieved by the EL coded at rate C. We analyze the mean, sample variance, and autocorrelation coefficients of the scene qualities, denoted by $\bar{\Theta}(C)$, $\sigma_\Theta^2(C)$ and $\rho_\Theta(C, k)$. They are estimated using the same way as for the image–based metrics.

For each scene s, the rate–distortion characteristics are obtained by plotting the curves $\Theta_s(C)$. The mean and variance of the scenes' qualities give an overall indication of the perceived quality of the entire video. However, the variance of scene quality does not capture the differences in quality between successive video scenes, which degrade the perceived overall quality of the video. To this end, we introduce a new metric, called variability, which is defined as:

$$V(C) = \frac{1}{S-1} \sum_{s=2}^{S} |\Theta_s(C) - \Theta_{s-1}(C)|. \tag{5}$$

Note that our measure for overall scene quality does not account for differences in the length of the successive scenes. However, our analysis with a simple weighted measure of quality gave very similar results, so they are not presented here. Moreover, the perception of the overall quality of a scene may not be linearly proportional to the length of the scene, but may also depend on other factors, such as the content of the scene.

Also, we define the maximum quality variation between two consecutive scenes as:

$$V_{max} = \max_{2 \leq s \leq S} |\Theta_s(C) - \Theta_{s-1}(C)|. \tag{6}$$

Recalling that $\bar{X}_s(C)$ denotes the mean size of the frames in scene s, the correlation coefficient between the mean frame size $\bar{X}_s(C)$ of a scene and the overall quality $\Theta_s(C)$ of a scene is estimated as

$$\rho_{\bar{X},\Theta}(C) = \frac{1}{S-1} \sum_{s=1}^{S} \frac{(\bar{X}_s(C) - \bar{X}(C))(\Theta_s(C) - \bar{\Theta}_s(C))}{\sigma_{\bar{X}}(C) \cdot \sigma_\Theta(C)}, \tag{7}$$

where $\bar{X}(C)$ denotes the mean of the successive mean frame sizes of all scenes composing the video ($\bar{X}(C) = \sum_{s=1}^{S} \bar{X}_s(C)/S$). We denote the correlation coefficient between the BL quality and the total (BL+EL) quality of a scene by $\rho_{\Theta^b,\Theta}(C)$. It is estimated the same way as $\rho_{\bar{X},\Theta}(C)$.

Finally, we monitor the length (in video frames) of the successive scenes N_s, $s = 1, \ldots, S$. We denote the mean and sample variance of N_s as $\bar{N} = N/S$ and σ_N^2.

3.4 MSE and PSNR measures

Although our analysis method is intended to be independent of the particular quality measure which is considered, we give results in terms of the MSE (Mean Squared Error) and the PSNR (Peak Signal to Noise Ratio). Recently, the Video Quality Expert Group (VQEG) released its first report, in which are described a series of experiments to account for the performance of some objective quality measures [12]. Subjective results showed that no quality measure performed better than PSNR. As it is recalled in [15], for video pictures of size $X \times Y$ pixels, the PSNR of the video sequence between images n_1 to n_2 is defined by:

$$\mathrm{PSNR}(n_1, n_2) = 10 \log \frac{M^2}{\mathrm{MSE}(n_1, n_2)}, \tag{8}$$

where M is the maximum value of a pixel (255 for 8–bit grayscale images), and $\mathrm{MSE}(n_1, n_2)$ is defined as:

$$\mathrm{MSE}(n_1, n_2) = \frac{1}{XY(n_2 - n_1 + 1)} \sum_{n=n_1}^{n_2} \sum_{y=1}^{Y} \sum_{x=1}^{X} [I(x, y, n) - \tilde{I}(x, y, n)]^2, \tag{9}$$

where $I(x, y, n)$ and $\tilde{I}(x, y, n)$ are the gray-level pixel values of the original and decoded frames number n, respectively. The MSE and PSNR measures are not always reliable but they give a good estimate of the perceived quality and can be easily computed. According to [15], the PSNR and MSE are only well-defined for luminance values, not for color. Moreover, as noted in [12], the HVS (Human Visual System) is much more sensitive to the sharpness of the luminance component than that of the chrominance component. Therefore, we only consider luminance PSNR.

Henceforth, assuming that the EL is encoded with constant bitrate C, we set:

$$\begin{aligned} Q_n(C) &= \mathrm{PSNR}(n, n), & (10)\\ Q_{s,n}(C) &= \mathrm{PSNR}(T_s + n - 1, T_s + n - 1), & (11)\\ \Theta_s(C) &= \mathrm{PSNR}(T_s, T_{s+1} - 1). & (12) \end{aligned}$$

where $T_s \in \{1, \ldots, N\}$ is the absolute frame number of the first frame of scene s, i.e. $T_s = 1 + \sum_{j=1}^{s-1} N_j$.

4 Generation of Traces

In our experiments, we used the Microsoft encoder/decoder [4] with FGS functionality. We generated our traces according to the following methodology:

1. First, we encode a sequence `video.yuv` using 2 different sets of quantization parameters for the BL. This gives compressed BL bitstreams of high quality (with quantization parameters $(4, 4, 4)$ for (I, P, B) frames) and low quality (with quantization parameters $(10, 14, 16)$), as well as the associated EL bitstreams.

2. We segment the video into S successive scenes. This can be done based on the compressed BL bitstream or the raw yuv video, according to the segmentation tool which is used. This yields a file containing the image numbers delimiting the scenes (scene-nb(s), last-image-nb($T_s + N_s - 1$)).

3. For each BL quality, we cut the corresponding FGS–EL bitstream into m EL CBR–encoded substreams of increasing and equally spaced bitrates $C = c,\ 2 \cdot c, \ldots, m \cdot c \leq r_e$. c is the step size at which we increase the FGS cutting rate. In all our experiments we used a step size of $c = 200$ Kbps.

4. For each tuple of compressed bitstreams (BL quality, EL substream encoded at rate C), we compute the PSNR for each image after decoding, and then the PSNR for each scene.

Finally, for each BL quality, we obtain the following traces:

- a file containing the BL statistics for each image number (image-nb(n), decoding-timestamp($n \cdot T$), image-type, frame-size(X_n^b), PSNR-Y(Q_n^b), PSNR-U, PSNR-V),

- a file containing the size of each EL bit–plane (up to 8 bit–planes) for each image number (image-nb(n), size-of-BP1($X_n^{e_1}$), ..., size-of-BP8($X_n^{e_8}$)),

- a file, for each EL encoding rate C, containing the image quality (in PSNR) obtained after decoding the BL and the truncated EL for all frames (image-nb(n), PSNR-Y($Q_n(C)$), PSNR-U, PSNR-V).

4.1 Limitations

Due to an encoder limitation, we had to encode separately two 30 minute sequences of our 1 hour videos and then concatenate the traces. For the video *News*, a few bidirectionally predicted frames at the end of the sequence are skipped at the encoder, so we repeated the last encoded frame until the original end of the sequence (this is visible on the BL traces when the frame–type stays constant for some frames at the end of a 54000 image sequence). Since this only concerns 4 frames of the videos, we do not expect it to change the statistical results.

Because of a software bug (probably loss of synchronization between the BL and the EL due to the cut of a system header during the partitioning of the EL into substreams), some PSNR results (particularly at low EL bit-rates) are not coherent. It has a minor impact for short videos, because the trend of the RD curves for all individual images and video scenes is clear enough, so that we can estimate the quality that will be reached without this bug at these particular low bit–rates. However, for long videos, only high quality BL encoding gives valid results for most EL bit–rates, so we can only study the case of high BL quality for long videos.

Because the automatic extraction of scene boundaries is still a subject of ongoing research ([5], [8], [10]), we restricted the segmentation of the video to scene shot changes only. However, our approach remains valid for any finer segmentation. Many commercial applications can now detect shot cuts with good efficiency, and we used the MyFlix software [14]. MyFlix is a MPEG–1 editing software which can find cuts directly on MPEG–1 videos. Therefore we had to encode each video into MPEG–1 once in order to detect shot cuts. We have observed during our experiments with long videos, that MyFlix was giving overall good performance.

4.2 Organization of the Web Site

All our traces for long videos, together with some statistics, can be found on our public web site. The site is organized as follows. For each movie encoded at high BL quality, we have the following directories:

- **stats/**, which contains the traces of the bit–plane sizes, the boundaries of the scenes and the total (BL+EL) coding rate by scene and GoP. It also features some overall statistics, such as statistics for scene length (S, $\bar{N}$ and σ_N) and the graphs of scene and GoP quality statistics as a function of the FGS rate ($\bar{\Theta}(C)$, $\sigma_\Theta(C)$, $V(C)$, $\rho_{\bar{X},\Theta}(C)$, $\rho_{\Theta^b,\Theta}(C)$) for $C = 0, 800, 1000, \cdots, 2000$ Kbps. Note that for the graphs in this directory, we did not plot the statistics corresponding to the FGS cutting rates $C = 200, 400, 600$ Kbps because of the phenomenon explained in section 4.1.

- **srd/**, which contains the Rate–Distortion trace files for each scene ($\Theta_s(C)$).

- **q0/** ... **q2000/**, which contain, for each FGS cutting rate $C = 0, \cdots, 2000$ Kbps, the trace of individual image quality ($n, Q_n(C)$), the graphs of the autocorrelation in scene or

GoP quality ($\rho_\Theta(C,k)$), the graph of the scene quality as a function of the scene number and the graph of the GoP quality as a function of the GoP number ($\Theta_s(C)$).

5 Analysis of traces - Short Clip

In this section, we present the analysis of a short video clip of 828 frames encoded in CIF format, obtained by concatening the sequences “coastguard”, “foreman” and “table” in this order. We segmented (by hand) the resulting clip into 4 scenes ($T_1 = 1$, $T_2 = 301$, $T_3 = 601$, $T_4 = 732$) corresponding to the 4 shots of the video (the table sequence is composed of 2 shots). The (I, P, B) quantization parameters used to encode the BL were fixed to $(4, 4, 4)$ and $(10, 14, 16)$ for the high and low quality versions of the BL, respectively.

Figures 3 and 4 show the quality of the successive images Q_n for the 4 scenes, when only the BL is decoded and when a substream of the FGS–EL (at rate $C = 3$ Mbps) is added to the BL before decoding. We make the following observations for both low and high BL qualities. (*i*) First, the average image quality changes from one scene to the other for the BL–only stream and also when a constant rate EL is added. This is confirmed for all EL rates in Figures 13 and 15 which show the average image quality achieved for the images within the same scene. This trend in the image quality time series suggests to analyze the quality statistics of each scene separately [2]. (*ii*) For a given scene, we see that for the BL there are significant differences in the quality achieved for successive images. Most of these differences are introduced by the different types of BL images (I, P, B) — the frames with the highest quality correspond to I–frames. When adding a part of the EL (at rate $C = 3$ Mbps in the figures), we see that these differences are still present, even if they have changed in magnitude. Therefore, this suggests to distinguish between the different types of images in order to study the RD characteristics of the FGS–EL. (*iii*) We notice that scenes 2 and 3 feature high variations of the quality for a given frame type within the same scene. Scene 2 corresponds to the ’foreman sequence’ in which the camera pans from the foreman’s face to the building. A finer scene segmentation tool would have segmented scene 2 into two different scenes, since the foreman’s face and the building have different complexities.

Figures 5 and 6 show the aggregate size of the EL bit–planes $Y_n^{e_i}$ and Figures 7 and 8 the size of the BL frames X_n^b. We observe that, in general, I pictures have fewer bit–planes than P or B pictures and the total number of bits for the EL images is larger for P and B pictures than for I pictures. This is because I images have higher BL quality. Therefore, fewer bit–planes and fewer bits are required to code the EL of I images. For the same reason, when comparing high and low BL qualities, we see that the EL corresponding to the high BL quality needs, for most images, fewer bit-planes than the EL corresponding to the low BL quality. Also, when comparing the average size of EL frames for all scenes with the average size of the corresponding BL frames, we see that the larger the average BL frame size the larger the average EL frame size. This can be explained by the different complexities of the scenes. For example, we see that it requires fewer bits to code I images in the first part of scene 2 than to code I images in scene 1, meaning that the complexity of scene 1 images is larger than those of scene 2. Therefore, the average number of bits required to code the EL of scene 1 images is larger than for the first part of scene 2. Note that, for both BL qualities, the bit–planes number 5, 6, 7 and 8 are empty for all frames, which indicates that the EL contains up to 4 bit–planes.

We plot in Figures 9, 10, 11, and 12, the RD functions $Q^e_{1,13}(C)$, $Q^e_{1,14}(C)$, $Q^e_{2,213}(C)$, and $Q^e_{2,214}(C)$ (improvement in quality brought by the EL as a function of the FGS encoding rate) for different types of images within the same GOP. Note that some RD functions feature a few outliers (at low FGS bitrate) which correspond to the phenomenon that we explained in Section 4.1. The plots confirm that the RD functions of the EL do depend on the type of image of the BL and the particular scene. We first see that RD functions are different for each bit–plane, indicating that bit–planes have different characteristics (in general, the RD functions of lower bit–planes are closer to a linear function than the functions of higher bit–planes). Also, the maximum gain in quality for the same amount of EL data added to the BL, i.e. $Q^e((k+1)\cdot c) - Q^e(k \cdot c)$, for $c = 200$ Kbps and $k = 1, \dots, m-1$, is always achieved when we get closer to the end of a bit–plane. This may be due to the bit–plane headers. Indeed, the more bits there are in a given bit–plane after truncation, the smaller the share of the bit–plane header in the total data for this bit–plane.

Figures 14 and 16 give the standard deviation of the image quality for the different scenes of the video for both low and high BL qualities. Scene 2 is the scene with the largest variance, because of the variations in average image quality from the beginning to the end of the scene. We see that, for a given scene, the variance in quality can change considerably with the FGS rate. These variations include I, P, B variations, as well as longer term variations. Figures 17 and 18 give the standard deviation of both image quality σ_Q and GoP quality σ_Θ for the entire video. As we see, the standard deviation of GoP quality is negligible compared to the standard deviation of image quality. This means that most of the variations in quality are due to variations in image quality between the images of different type within a given GoP — which are in turn due to the BL encoding (single-layer encoder and quantization parameters used).

Finally, Figures 19 and 20 give the autocorrelation function of the image quality for the BL and the FGS rates $C = 1$, 2 and 3 Mbps. We observe periodic spikes which correspond to the GoP pattern. We verify that, at small lags, there are high correlations in quality for the different types of pictures at all FGS rates. Moreover, we see that autocorrelation at small lags slightly increases with the FGS rate. This means that the FGS layer smoothes the difference in quality between near images. Indeed, for the same number of FGS–EL bits added to the BL, the gain in quality is different for consecutive I, P and B frames. In general, the gain in quality for I frames increases slower than the gain in quality for P or B frames: as pointed out earlier, the BL achieves a higher quality for I frames, so for I frames the EL bits which are added to the BL correspond to higher (less visible) frequencies than for P and B frames.

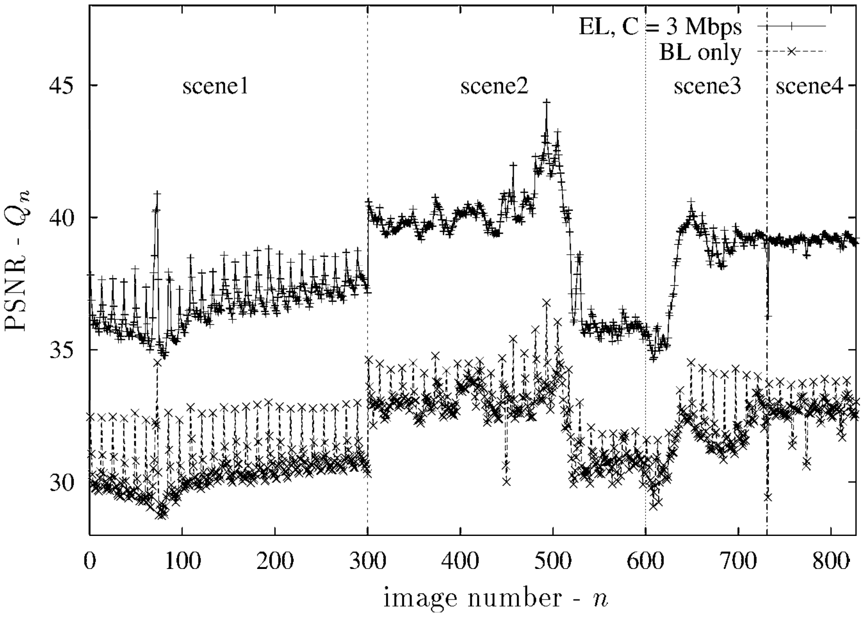

Figure 3: Image PSNR Q_n as a function of image number n for "Clip" encoded with low quality BL

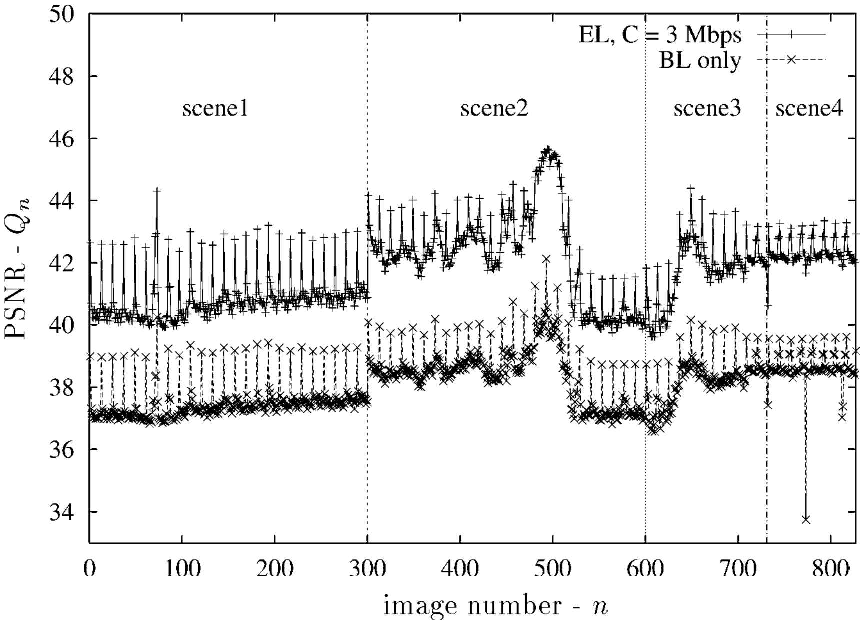

Figure 4: Image PSNR Q_n as a function of image number n for "Clip" encoded with high quality BL

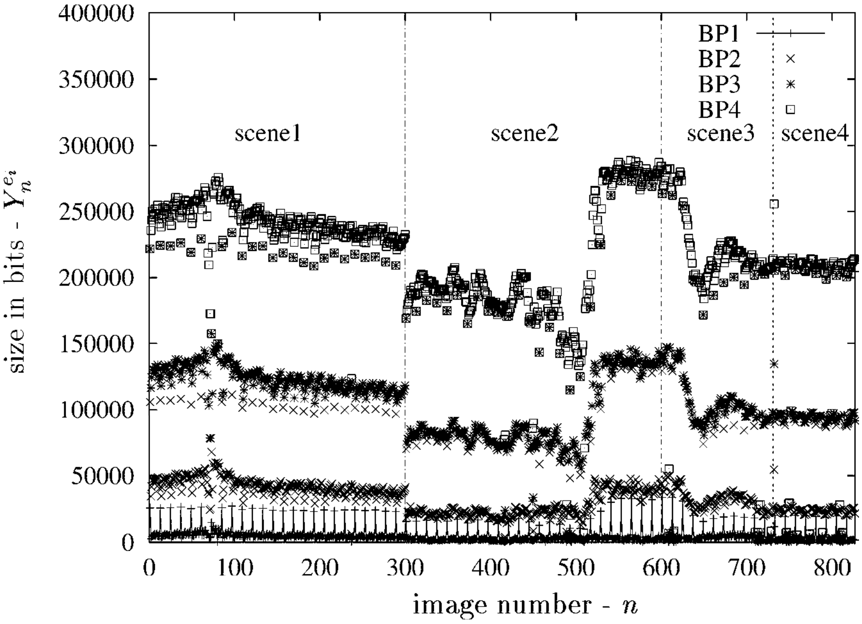

Figure 5: Aggregate size of the EL bit–planes $Y_n^{e_i}$ as a function of image number n for "Clip" encoded with low quality BL

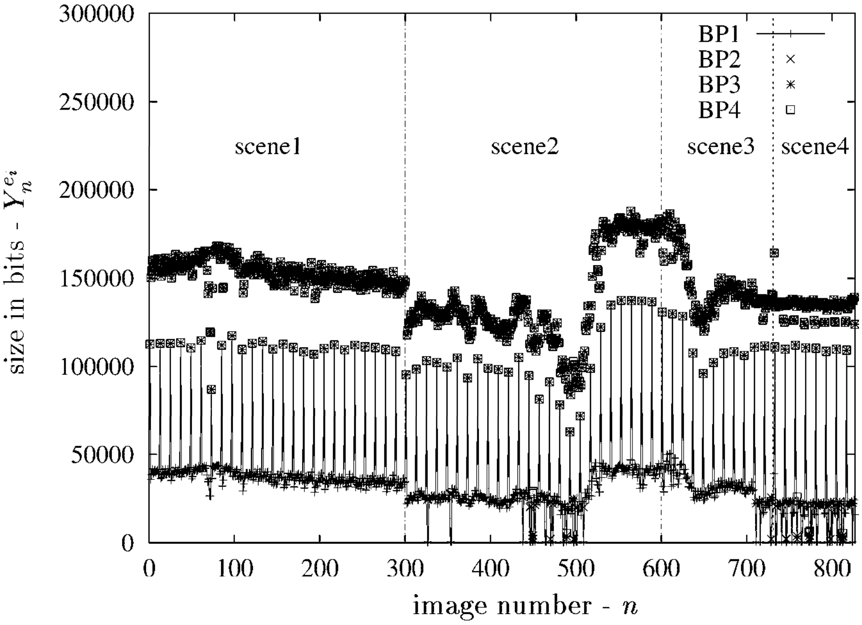

Figure 6: Aggregate size of the EL bit–planes $Y_n^{e_i}$ as a function of image number n for "Clip" encoded with high quality BL

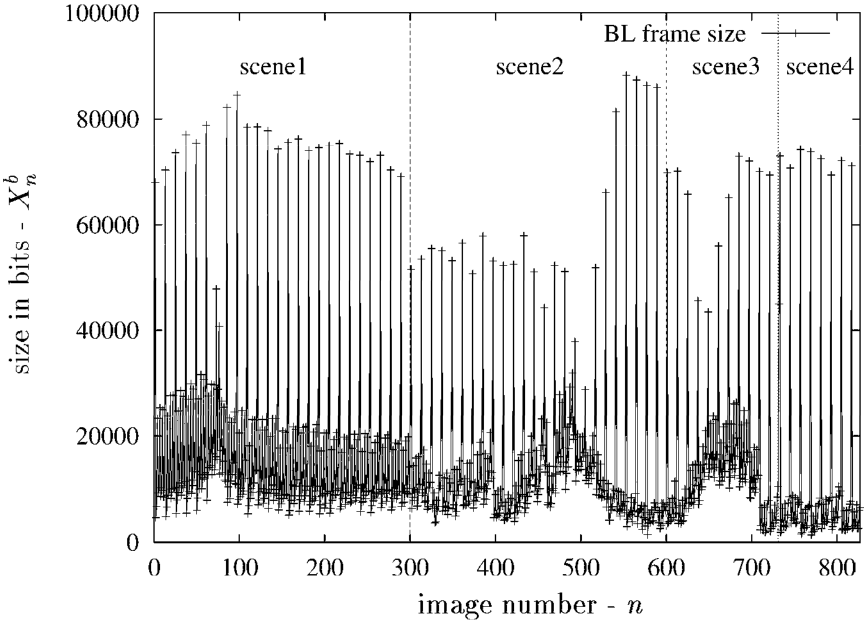

Figure 7: Size of BL images X_n^b as a function of image number n for “Clip” encoded with low quality BL

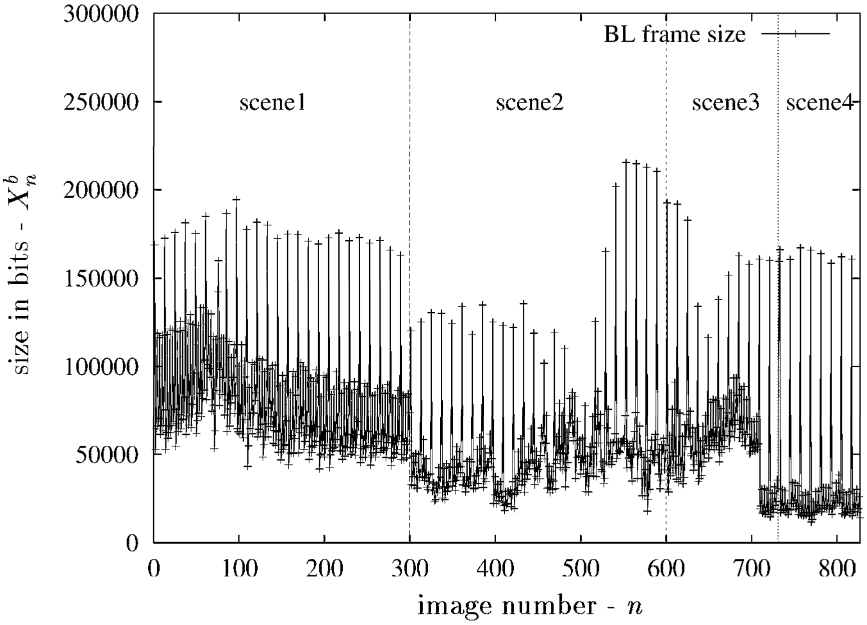

Figure 8: Size of BL images X_n^b as a function of image number n for “Clip” encoded with high quality BL

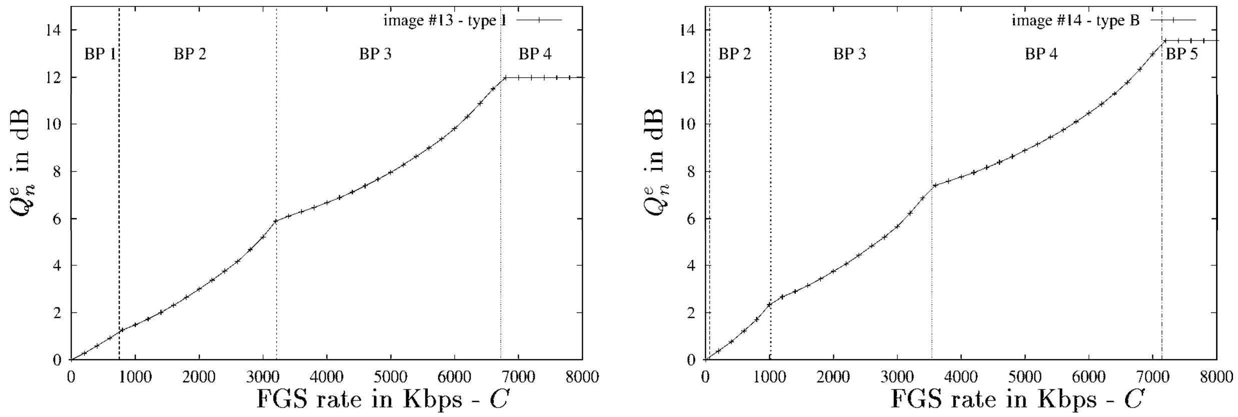

Figure 9: Improvement in PSNR $Q^e_{1,13}$ and $Q^e_{1,14}$ as function of the FGS bitrate C for successive I and B images in scene 1 of "Clip" encoded with low quality BL

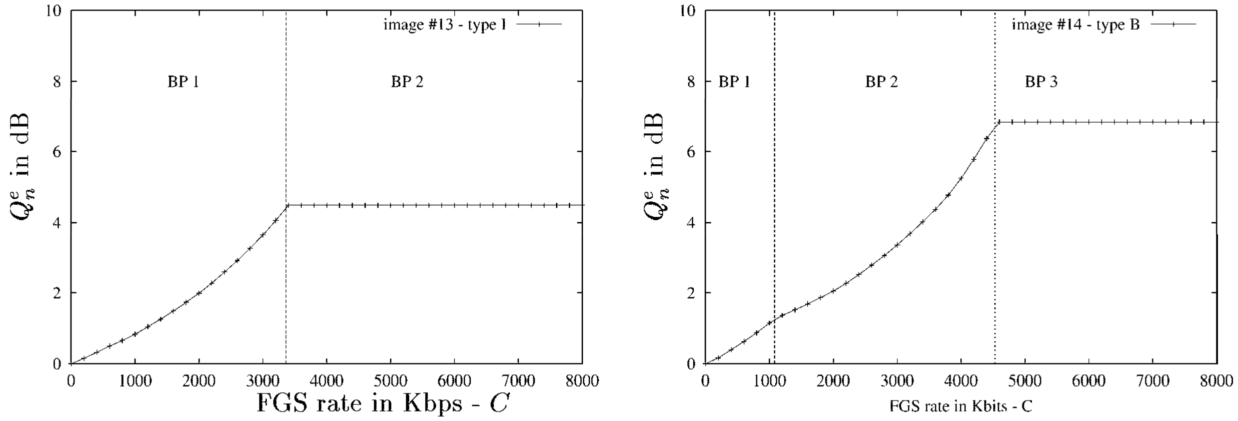

Figure 10: Improvement in PSNR $Q^e_{1,13}$ and $Q^e_{1,14}$ as function of the FGS bitrate C for successive I and B images in scene 1 of "Clip" encoded with high quality BL

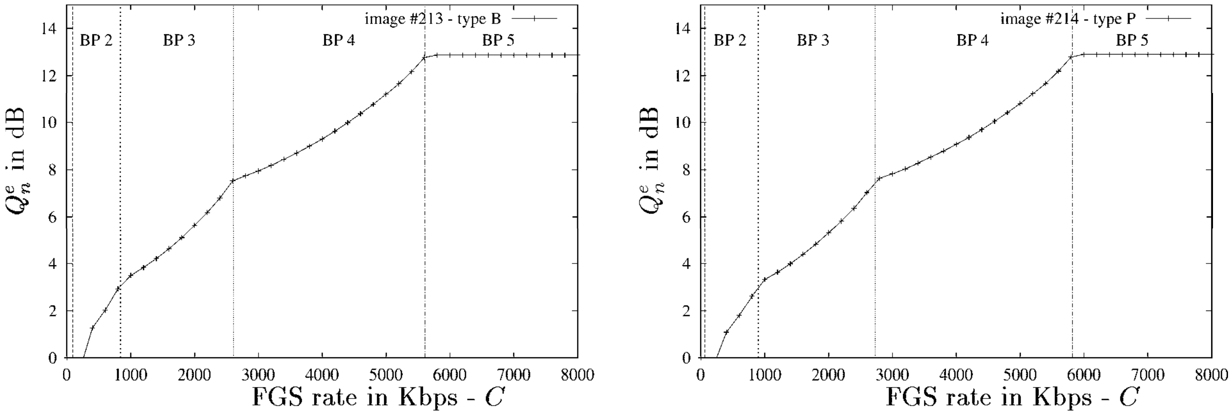

Figure 11: Improvement in PSNR $Q^e_{2,213}$ and $Q^e_{2,214}$ as function of the FGS bitrate C for successive B and P images in scene 2 of "Clip" encoded with low quality BL

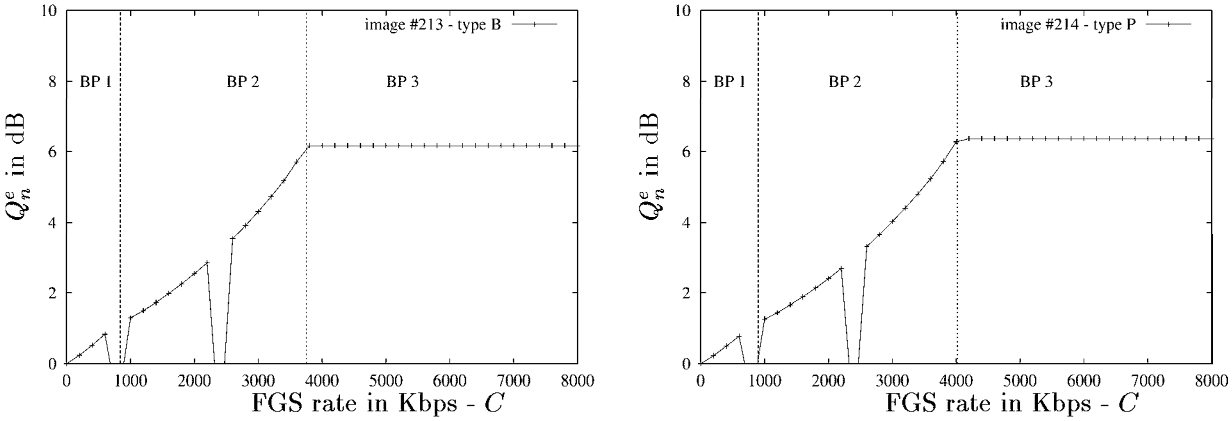

Figure 12: Improvement in PSNR $Q^e_{2,213}$ and $Q^e_{2,214}$ as function of the FGS bitrate C for successive B and P images in scene 2 of "Clip" encoded with high quality BL

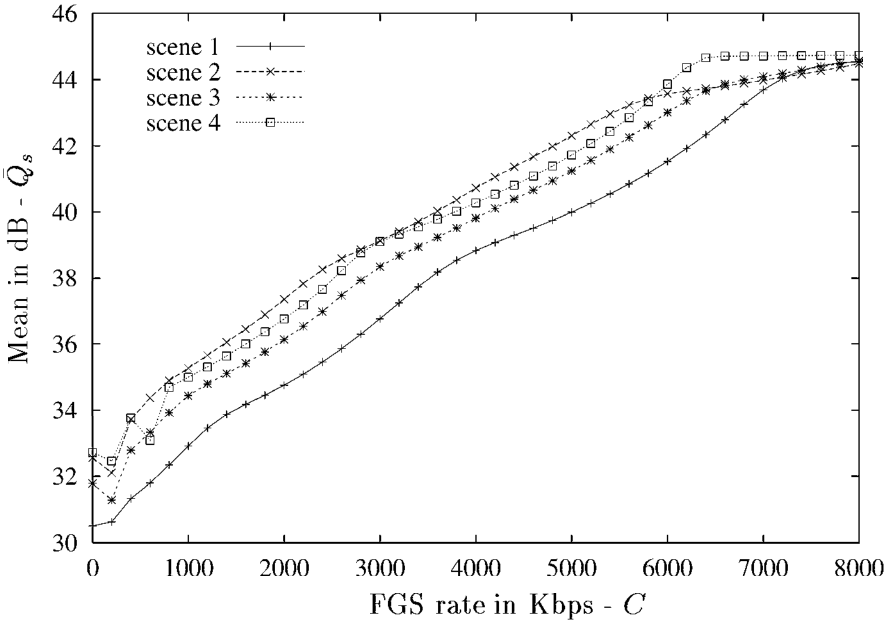

Figure 13: Average image quality by scene $\bar{Q}_s$ as a function of the FGS bitrate C for all scenes of "Clip" encoded with low quality BL

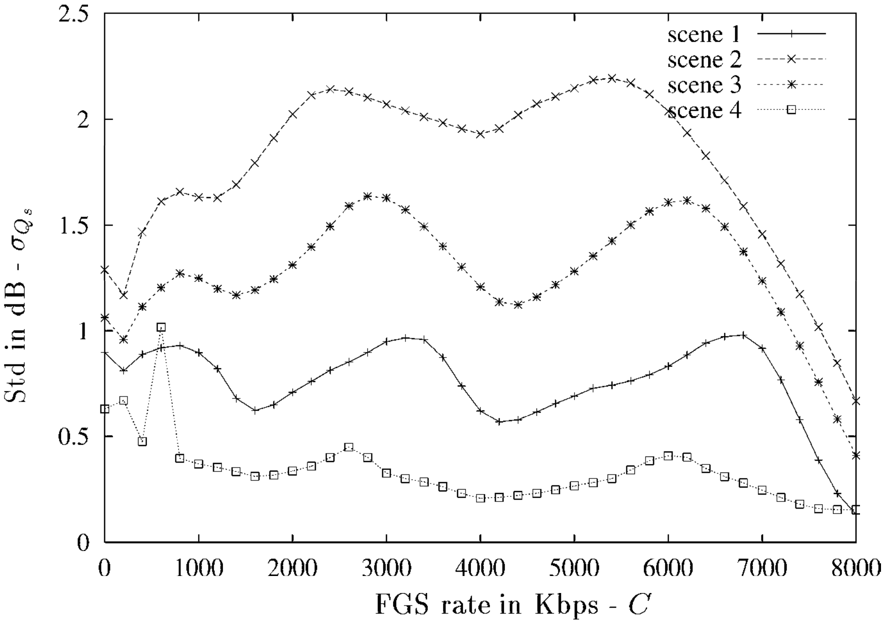

Figure 14: Standard deviation of image quality by scene σ_{Q_s} as a function of the FGS bitrate C for all scenes of "Clip" encoded with low quality BL

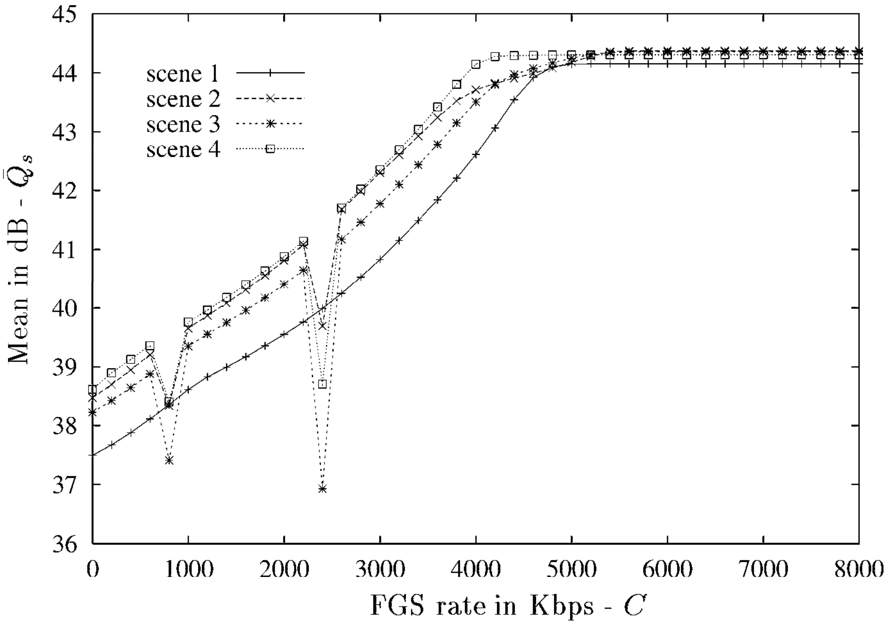

Figure 15: Average image quality by scene $\bar{Q}_s$ as a function of the FGS bitrate C for all scenes of "Clip" encoded with high quality BL

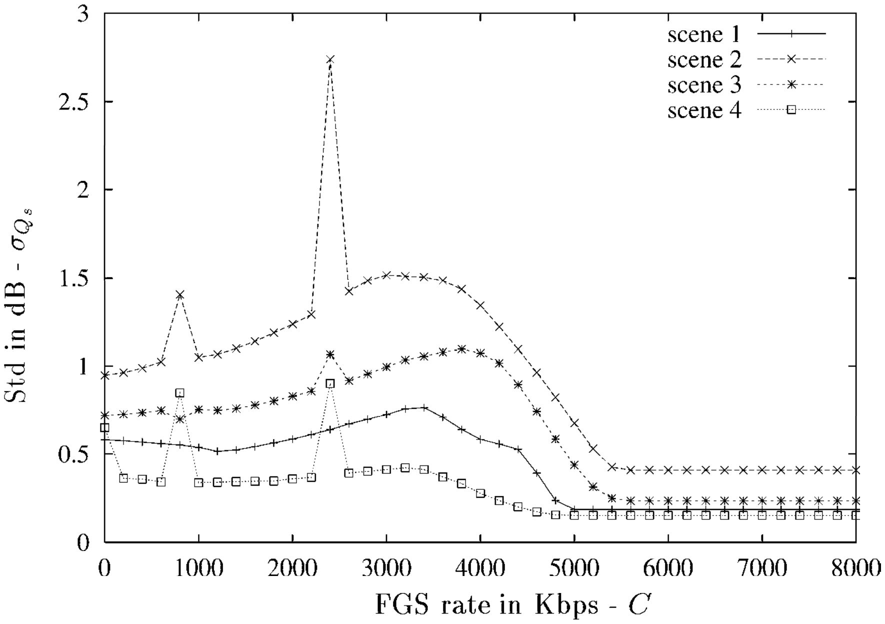

Figure 16: Standard deviation of image quality by scene σ_{Q_s} as a function of the FGS bitrate C for all scenes of "Clip" encoded with high quality BL

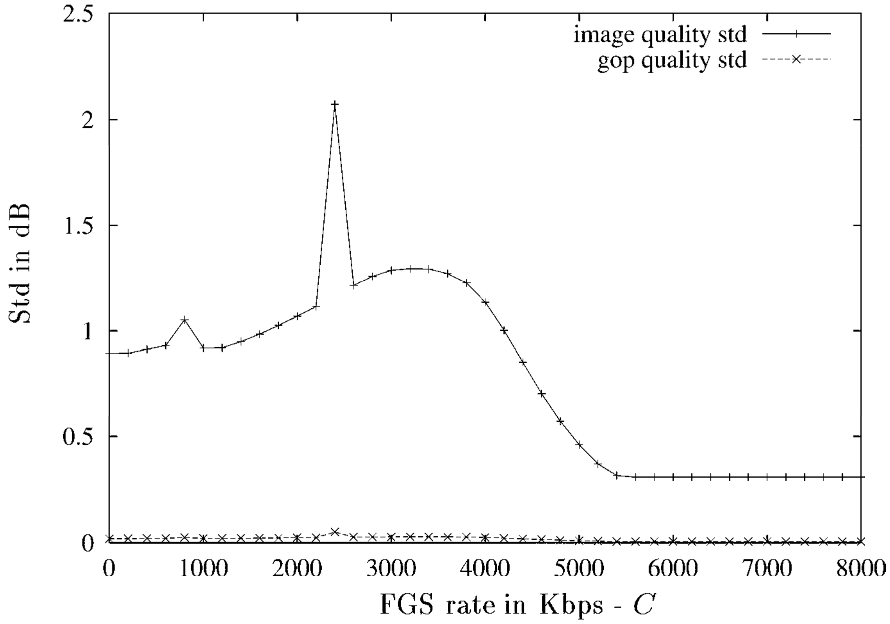

Figure 17: Standard deviation of image quality σ_Q and GoP quality σ_Θ for "Clip" encoded with high quality BL

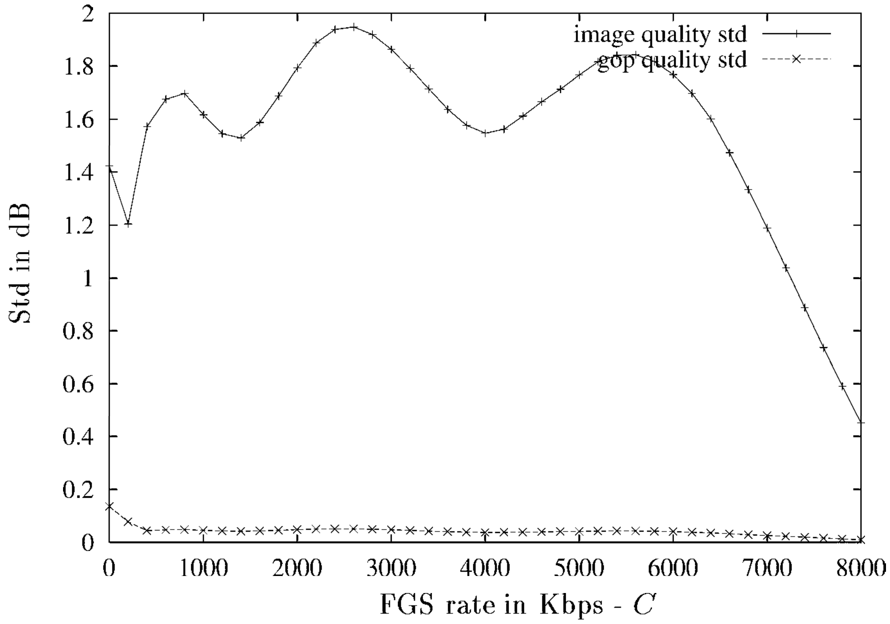

Figure 18: Standard deviation of image quality σ_Q and GoP quality σ_Θ for "Clip" encoded with low quality BL

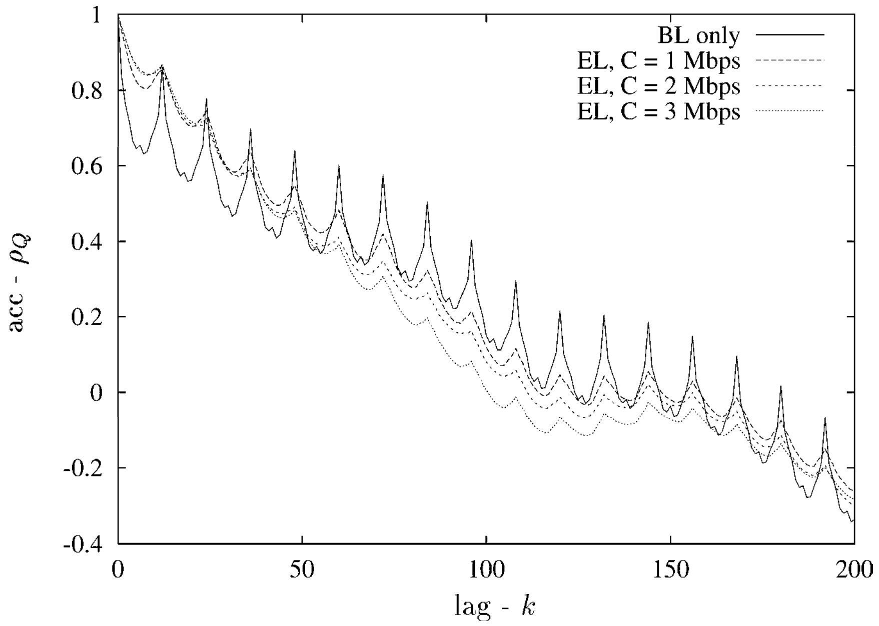

Figure 19: Autocorrelation coefficient of image quality ρ_Q for "Clip" encoded with low quality BL

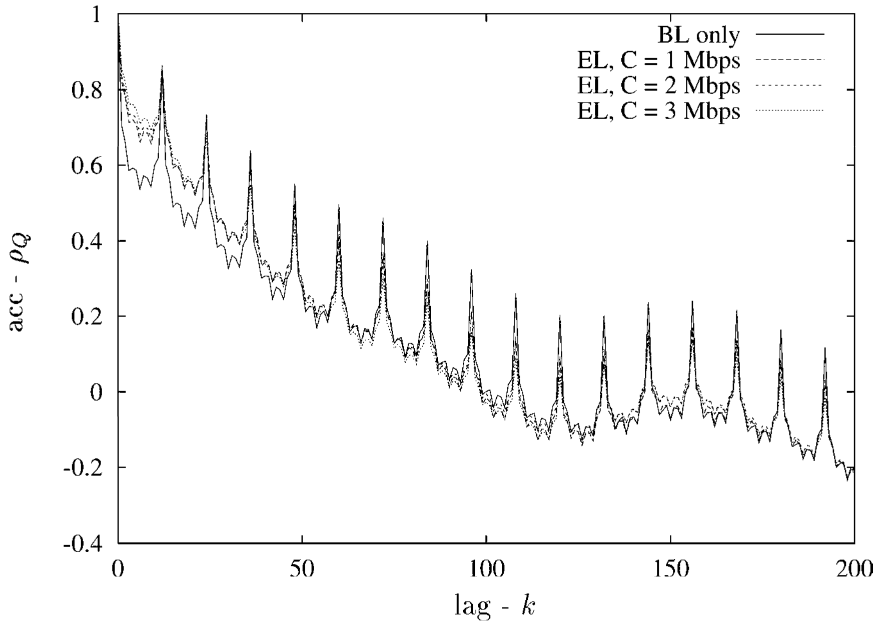

Figure 20: Autocorrelation coefficient of image quality ρ_Q for "Clip" encoded with high quality BL

	run time	S	$\bar{N}$	CoV_N	$N_{max}/\bar{N}$	$\bar{r}_b$ (Mbps)	$\bar{X}^b$	CoV_{X^b}	$X^b_{max}/\bar{X}^b$
The Firm	1h	890	121	0.94	9.36	0.65	21765	0.65	6.52
Oprah+com	1h	621	173	2.46	39.70	2.73	91129	0.14	1.94
Oprah	38mn	320	215	1.83	23.86	1.69	56200	0.19	2.33
News	1h	399	270	1.67	9.72	0.74	24645	0.54	5.30
Star Wars	1h	984	109	1.53	19.28	0.49	16363	0.65	6.97
Silence CIF	30mn	184	292	0.96	6.89	1.74	57989	0.72	7.85
Toy Story	1h	1225	88	0.95	10.74	1.08	36141	0.49	5.72
Football	1h	876	123	2.34	31.47	0.97	32374	0.53	3.90
Lecture	49mn	16	5457	1.62	6.18	1.54	51504	0.29	2.72

Table 1: Scene and Base Layer traffic characteristics of long videos

6 Analysis of long videos

We now focus on the statistics for long videos encoded with high BL quality. All videos have been captured and encoded in QCIF format(176 × 144 pixels), except for the movie *Silence of the Lamb* which has been captured and encoded the CIF format(352 × 288 pixels). We show in Tables 1, 2, 3 and 4 a summary of some statistics for nine videos of different types.

In Table 1, we first see that our nine long videos have various scene characteristics and BL statistics. The total number of scenes S, mean length $\bar{N}$, and standard deviation σ_N of the length of the scenes of a video can all have an impact on the resources required by streaming algorithms that perform optimization scene by scene, in particular on the amount of client buffering needed. Of particular interest are the videos *Oprah with commercials* and *Oprah* (same video but where we removed the commercials). Both videos have very high average BL bit–rate $\bar{r}_b$, small coefficient of variation in BL frame size CoV_{X^b} as well as small peak–to–mean ratio of frame sizes $X^b_{max}/\bar{X}^b$ compared to the other movies. We observe that, according to the common intuition, commercials have reduced the average length of scenes $\bar{N}$ and increased the coefficient of variation in scene length CoV_N as well as the peak–to–mean ratio of scene length $N_{max}/\bar{N}$. The *Toy Story* video has the longest number of scenes, whereas *Lecture*, which is a recording of a class by Professor Martin Reisslein at ASU, has the smallest number of scenes.

Table 2 indicates that the mean scene quality is very different from one video to the other. In particular, while *Oprah with commercials* and *Oprah* have the highest BL encoding rates (see Table 1), the average overall quality achieved for the BL of both videos is low compared

	BL only			$C = 1$ Mbps			$C = 2$ Mbps		
	$\bar{\Theta}$	CoV_{Θ}	$\Theta_{min}/\bar{\Theta}$	$\bar{\Theta}$	CoV_{Θ}	$\Theta_{min}/\bar{\Theta}$	$\bar{\Theta}$	CoV_{Θ}	$\Theta_{min}/\bar{\Theta}$
The Firm	36.76	0.013	0.97	40.10	0.017	0.88	43.70	0.003	1.99
Oprah+com	35.71	0.015	0.99	38.24	0.013	0.99	42.30	0.010	0.99
Oprah	35.38	0.003	1.00	38.18	0.003	1.00	42.84	0.007	0.99
News	36.66	0.018	0.97	39.65	0.027	0.96	43.76	0.021	0.98
Star Wars	37.48	0.025	0.95	41.14	0.031	0.94	43.83	0.013	0.99
Silence CIF	37.88	0.015	0.96	NA	NA	NA	39.70	0.020	0.96
Toy Story	36.54	0.021	0.97	39.57	0.029	0.97	43.95	0.013	0.97
Football	37.42	0.034	0.95	40.69	0.041	0.94	43.97	0.018	0.99
Lecture	35.54	0.000	1.00	38.48	0.000	1.00	43.64	0.001	0.99

Table 2: Scene quality statistics of long videos for the BL and FGS bit–rates $C = 1$ and 2 Mbps

	BL only		$C = 1$ Mbps		$C = 2$ Mbps	
	V	V_{max}	V	V_{max}	V	V_{max}
The Firm	0.06	1.83	0.16	2.37	0.00	1.26
Oprah+com	0.04	12.15	0.05	11.31	0.03	7.32
Oprah	0.00	0.36	0.00	0.42	0.00	1.13
News	0.11	3.15	0.29	3.17	0.12	2.36
Star Wars	0.29	8.25	0.57	8.83	0.13	6.28
Silence CIF	0.05	1.42	NA	NA	0.25	4.45
Toy Story	0.19	9.77	0.40	11.25	0.12	6.23
Football	0.51	9.79	0.72	10.12	0.19	6.36
Lecture	0.00	0.14	0.00	0.20	0.00	0.64

Table 3: V and V_{max} statistics of long videos for the BL and FGS bit–rates $C = 1$ and 2 Mbps

	BL only		$C = 1$ Mbps		$C = 2$ Mbps	
	$\rho_{X,\Theta}$	$\rho_{\Theta^b,\Theta}$	$\rho_{X,\Theta}$	$\rho_{\Theta^b,\Theta}$	$\rho_{X,\Theta}$	$\rho_{\Theta^b,\Theta}$
The Firm	-0.71	1.00	0.00	0.92	0.52	-0.18
Oprah+com	-0.20	1.00	0.01	0.99	0.07	0.83
Oprah	0.42	1.00	0.00	084	-0.04	0.48
News	-0.66	1.00	0.00	0.83	0.25	0.25
Star Wars	-0.47	1.00	-0.02	0.97	0.51	0.67
Silence CIF	-0.80	1.00	NA	NA	0.00	0.53
Toy Story	-0.39	1.00	-0.01	0.98	0.16	0.90
Football	-0.54	1.00	-0.02	0.97	0.33	0.81
Lecture	-0.14	1.00	0.00	0.52	-0.11	-0.17

Table 4: Scene-based correlation statistics of long videos for the BL and FGS bit–rates $C = 1$ and 2 Mbps

to the average quality achieved by the other videos. The normalized minimum scene quality $\Theta_{min}/\bar{\Theta}$ is very closed to 1 for all videos, more particularly at $C = 2$ Mbps. Note that some scene PSNRs can be artificially high, for example when the scene is very simple (such as a black screen). Therefore, in order to make our statistics more significant, we limited the PSNRs to the maximum value of 50 for all videos.

In Table 3, we first observe that *Oprah* has the smallest quality variability V at all FGS rates, whereas the *Football* video has the largest quality variability for the BL and $C = 1$ Mbps. For most videos, V and V_{max} are both minimum at $C = 2$ Mbps. We see, from V_{max}, that the difference in quality between successive scenes can be as high as 12 dB. However, in general, such a high difference is due to low complexity scenes which can have an artificially high PSNR as noted earlier. We expect that a better quality measure than PSNR would not give such high variations. Still, for most videos, the value of V_{max} is higher than 2 dB at all FGS rates, indicating that there are significant variations in quality between some successive video scenes.

In the following, we provide plots for the QCIF videos *The Firm*, *Oprah with commercials*, *Oprah* (without commercials) and *News*. We first show in Figures 21, 22, 23, 24, 25, 26, 27 and 28, the scene quality achieved as a function of the scene number ($\Theta_s(C)$) for the BL and FGS cutting rates $C = 1$ and 2 Mbps, as well as the BL and BL+EL average encoding bitrates. As we see, for a given video, the statistics of the BL dictate the statistics of the aggregate

BL+EL stream. In particular, significant differences between the quality or the coding rate of successive scenes in the BL remain also in the EL (for instance, we see that the smooth quality achieved for *Oprah with commercials* is conserved). At a high FGS rate ($C = 2$ Mbps), the quality achieved by all videos is almost constant for all scenes. In this situation, we may have reached the maximum encoding rate for all scenes (all scenes are encoded at maximum achievable quality).

Figure 29 shows the average scene quality achieved by the four videos as a function of the FGS rate ($\bar{\Theta}(C)$). We notice that there is a significant difference between the average quality achieved by the BL for *The Firm* or *News* and the average quality achieved by the BL for *Oprah* or *Oprah with commercials* (around 1 dB). This difference roughly remains at all FGS rates, indicating that the average quality achieved by a video at all FGS rates strongly depends on the average quality of the BL. This is confirmed in Figure 32 which shows the coefficient of correlation between the BL and the BL+EL quality of the scenes as a function of the FGS rate ($\rho_{\Theta^b,\Theta}(C)$). The correlation decreases slightly with the FGS rate but stays high at all rates. This is confirmed for most videos in Table 4.

Figure 30 shows the variability as a function of the FGS rate ($V(C)$). As we see, the difference in quality between the successive scenes first increases with the FGS rate for *The Firm* and *News*. At high FGS rates the variability starts to decrease because some scenes have reached their maximum quality (this is confirmed in Table 3 for most videos). For *Oprah* and *Oprah with commercials* the variability stays very low at all FGS rates. Indeed, for these videos, the VBR–BL encoder has been able to smooth the differences in scene quality almost perfectly, as we have seen earlier.

Figure 31 shows the coefficient of correlation between the average size and the quality of the scenes ($\rho_{\bar{X},\Theta}(C)$). Except for *Oprah*, the coefficient of correlation is negative for the BL because the BL encoder allocates more bits to the more complex scenes. Then, the coefficient of correlation globally increases with the FGS rate and become positive for most videos as shown in Table 4. Indeed, most of the complexity of the diverse scenes has been absorbed by the VBR–BL and then adding the EL increases the quality of the scenes more evenly. The *Oprah* video is a special case: the coefficient of correlation is already positive for the BL and then decreases with the FGS rate. In this case, the diversity of scene complexity has been almost totally absorbed by the BL, which explains the very low scene quality variability V achieved for *Oprah*.

Finally, Figure 33 shows, for each video, the autocorrelation in scene quality ρ_Θ for the BL and FGS rates $C = 0$, 1 and 2 Mbps. As we can see, for the four videos, the autocorrelation function for the BL+EL quality follows closely the autocorrelation function for the BL only, except for *Oprah with commercials* at $C = 2$ Mbps. The difference in autocorrelation at low lags between *Oprah* and *Oprah with commercials* can be explained by the higher diversity of successive scene types when adding commercials.

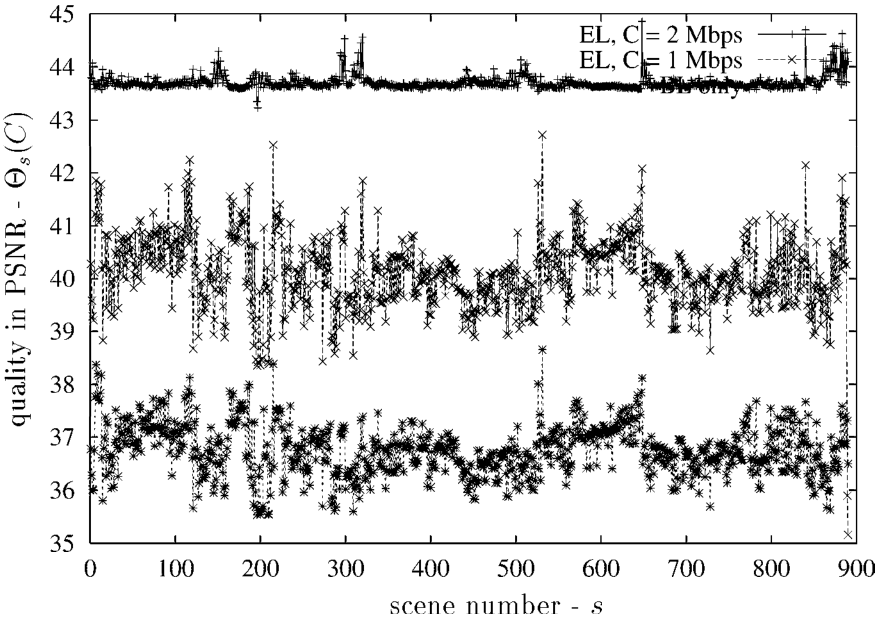

Figure 21: Scene PSNR $\Theta_s(C)$ as a function of scene number s for *The Firm* encoded with high quality BL

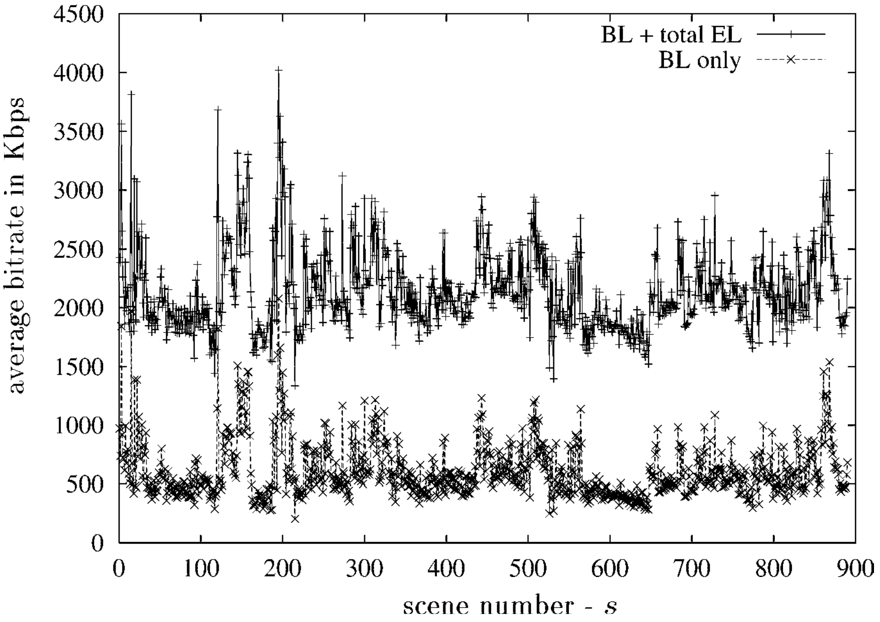

Figure 22: Average encoding bitrate $\bar{X}_s/T$ as a function of scene number s for *The Firm* encoded with high quality BL

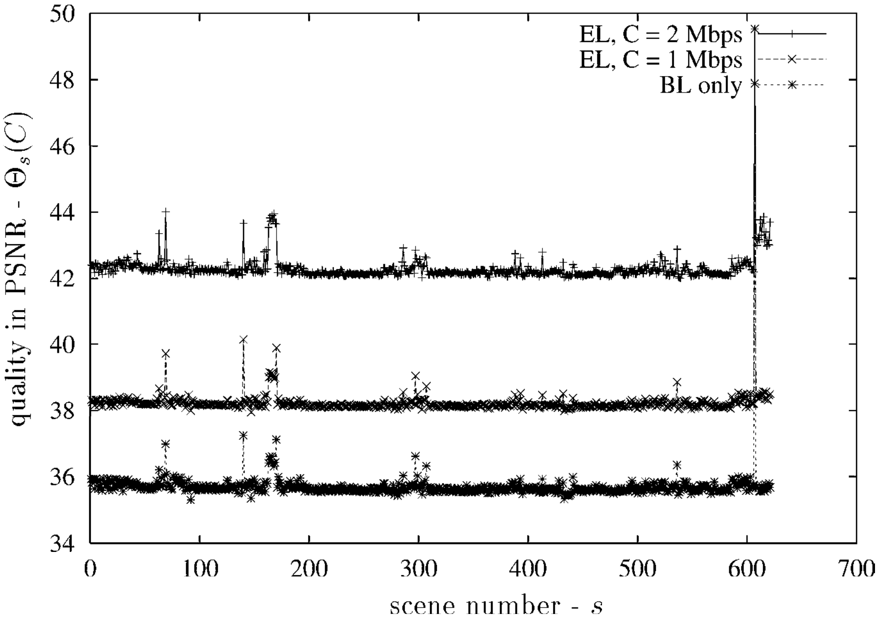

Figure 23: Scene PSNR Θ as a function of scene number s for *Oprah with commercials* encoded with high quality BL

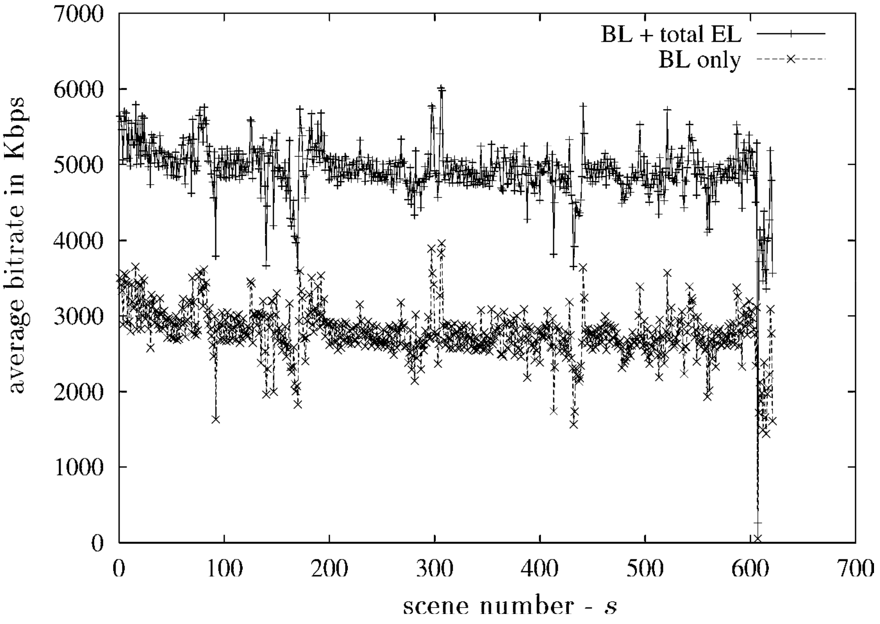

Figure 24: Average encoding bitrate as a function of scene number s for *Oprah with commercials* encoded with high quality BL

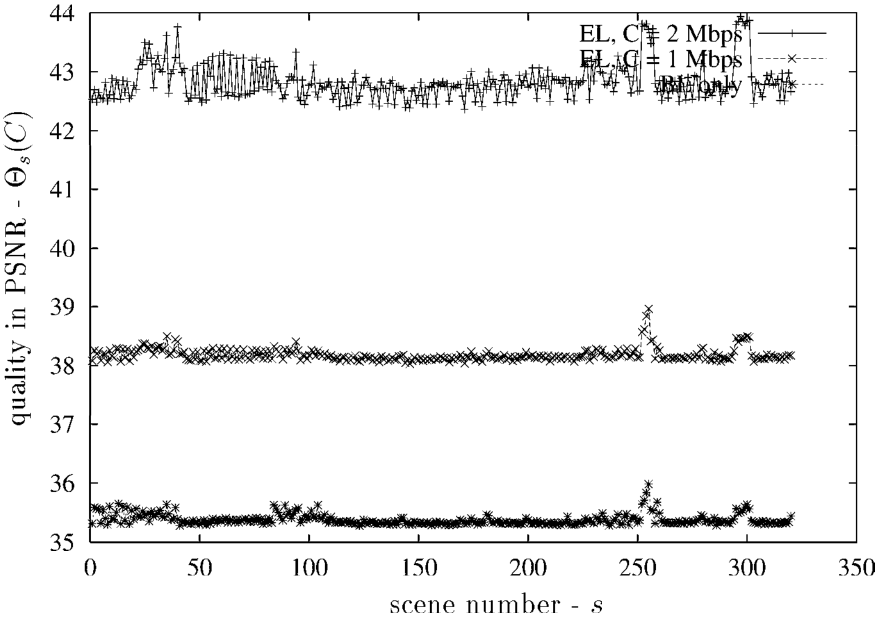

Figure 25: Scene PSNR Θ as a function of scene number s for *Oprah without commercials* encoded with high quality BL

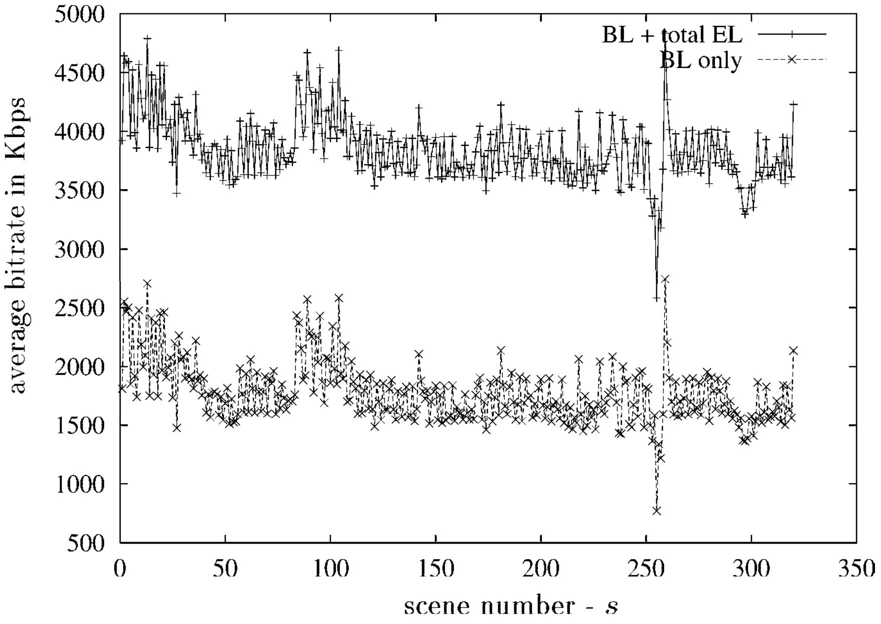

Figure 26: Average encoding bitrate as a function of scene number s for *Oprah without commercials* encoded with high quality BL

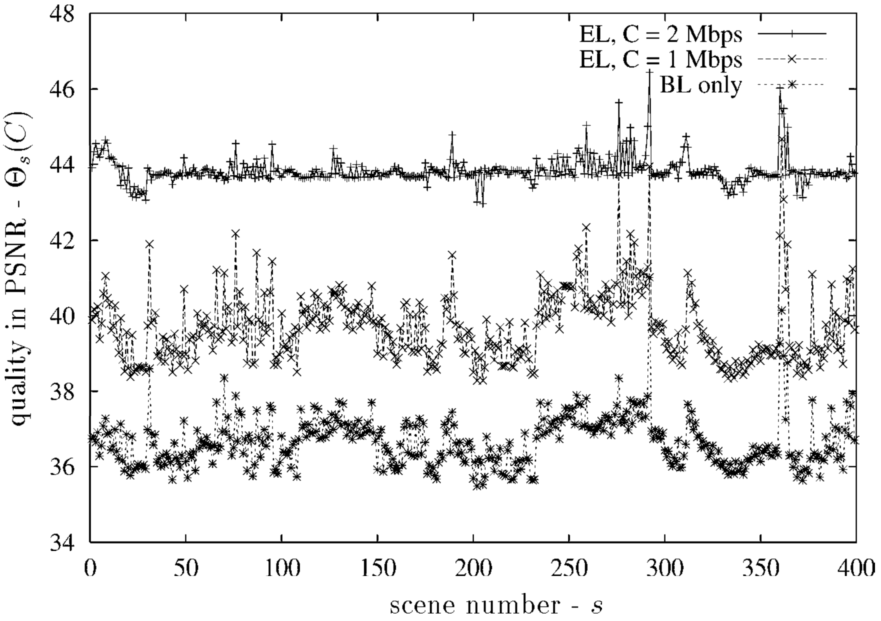

Figure 27: Scene PSNR Θ as a function of scene number s for *News* encoded with high quality BL

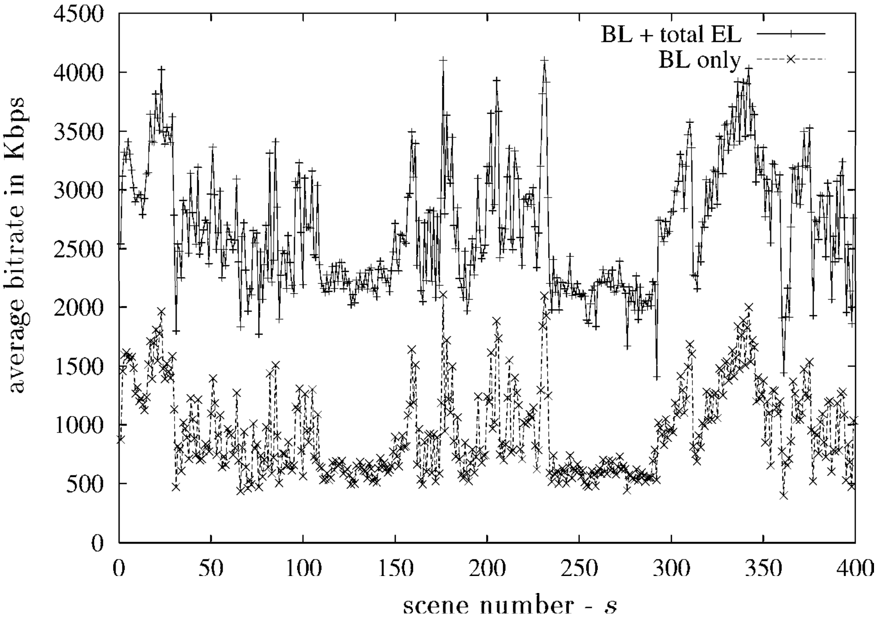

Figure 28: Average encoding bitrate as a function of scene number s for *News* encoded with high quality BL

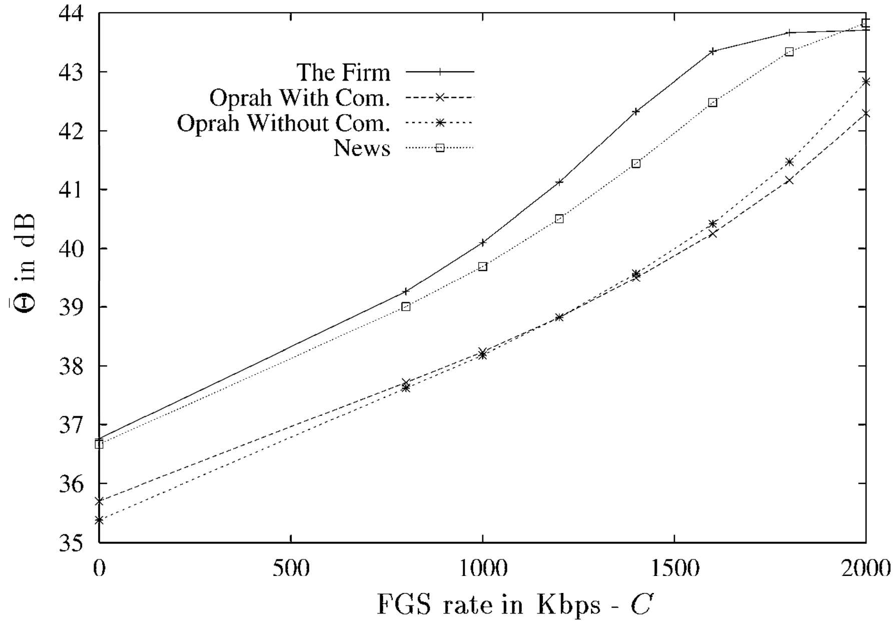

Figure 29: Average scene quality $\bar{\Theta}$ as a function of the FGS bitrate C for videos encoded with high quality BL

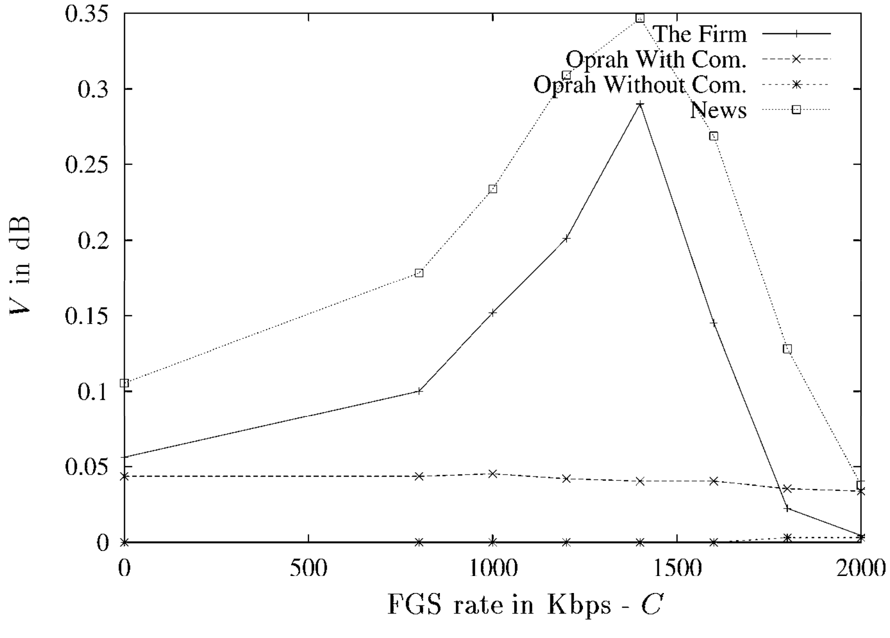

Figure 30: Variability in scene quality V as a function of the FGS bitrate C for videos encoded with high quality BL

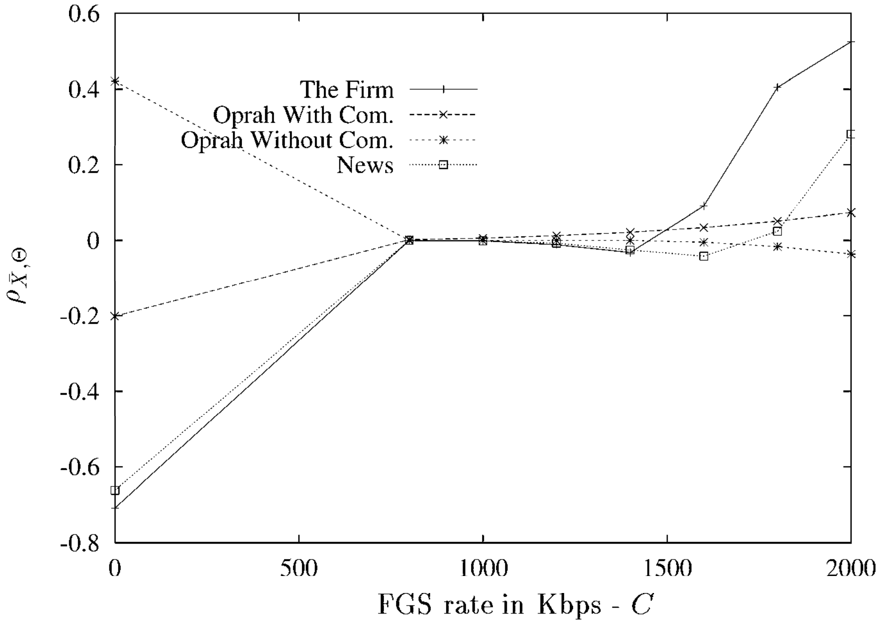

Figure 31: Coefficient of correlation between average size and quality of scenes $\rho_{\bar{X},\Theta}$ as a function of the FGS bitrate C for videos encoded with high quality BL

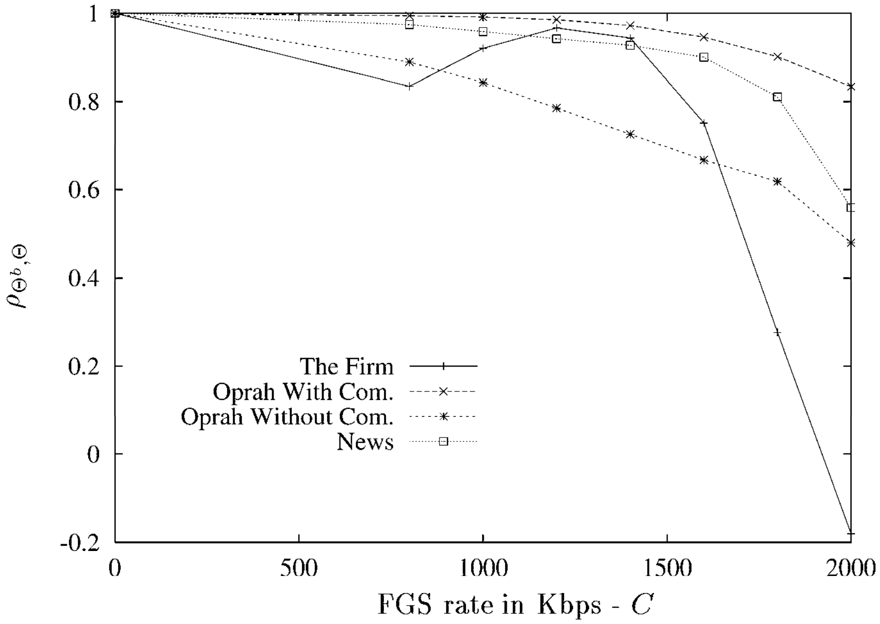

Figure 32: Coefficient of correlation between scene BL quality and scene overall quality $\rho_{\Theta^b,\Theta}$ as a function of the FGS bitrate C for videos encoded with high quality BL

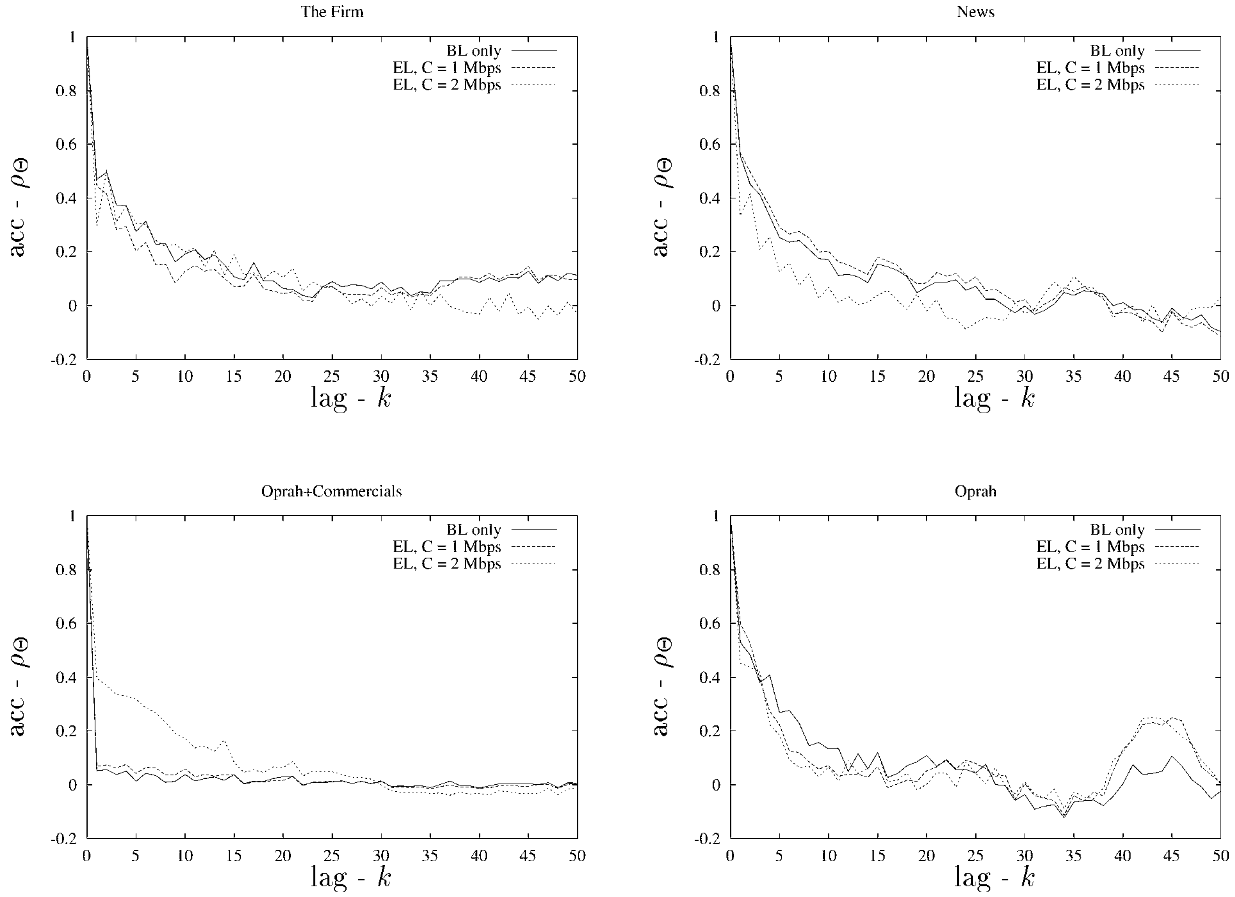

Figure 33: Autocorrelation in scene quality ρ_Θ for videos encoded with high quality BL

7 Conclusion

In this report, we have presented a publicly available library of traces from long MPEG–4 FGS encoded videos. It is intended to be used by researchers world–wide, in particular in order to test streaming algorithms with FGS–encoded videos.

A first analysis of the quality of individual images has shown that in order to maximize the gain in quality brought by the FGS–EL, streaming algorithms should take into account the internal structure of the MPEG–4 FGS bitstream. In particular, by prioritizing cutting the bitstream close to the end of an EL bit–plane, streaming algorithms could maximize the quality gain achieved by the EL for each image. This may suggests that the streaming algorithm only cuts the EL bitstream at bit–plane boundaries. However, such an approach does not exploit the main advantage of FGS–encoded videos over standard layered–encoding methods, namely the fine–granularity property which provides more flexibility in adapting the video encoding rate to varying network conditions [7].

Another main result of our analysis is that the EL statistics of individual images and video scenes depend strongly on the BL statistics — it is then essential to optimize the encoding of the BL. We have observed that the VBR–BL typically features significant variations in quality between the different types of successive images and the different scenes. Alternatively, streaming algorithms could tend to smooth these variations by selecting the FGS coding rate to stream for each image or each scene accordingly. In a sense, the FGS–EL could also be used to compensate the flaws of the BL encoding process.

Finally, we have presented a framework for analyzing quality statistics of videos scene by scene, rather than just considering the quality of individual images separately. We believe that optimizing the streaming for successive complete visual scenes has the potential to decrease the complexity of quality optimization procedures when streaming video over best–effort networks, without significantly decreasing the overall perceived quality. This is what we intend to show in future work, using the traces that we have generated in this study.

References

[1] *ISO/IEC JTC1/SC29/WG11 Information Technology - Generic Coding of Audio-Visual Objects : Visual ISO/IEC 14496-2 / Amd X*, December 1999.

[2] C. Chatfield. *The Analysis of Time Series.* Chapman and Hall, 1989.

[3] P. A. Chou and Z. Miao. Rate-Distortion Optimized Streaming of Packetized Media. *submitted to IEEE Transactions on Multimedia*, February 2001.

[4] Microsoft Corp. ISO/IEC 14496 Video Reference Software. Microsoft-FDAM1-2.3-001213.

[5] A. M. Dawood and M. Ghanbari. Scene Content Classification From MPEG Coded Bit Streams. In *Proc. of the 3rd Workshop on Multimedia Signal Processing*, pages 253–258, 1999.

[6] P. de Cuetos, P. Guillotel, K. W. Ross, and D. Thoreau. Implementation of Adaptive Streaming of Stored MPEG–4 FGS Video. *to appear in Proc. of IEEE International Conference on Multimedia and Expo*, August 2002.

[7] P. de Cuetos and K. W. Ross. Adaptive Rate Control for Streaming Stored Fine-Grained Scalable Video. In *Proc. of NOSSDAV*, pages 3–12, Miami, Florida, May 2002.

[8] C.-L. Huang and B.-Y. Liao. A Robust Scene-Change Detection Method for Video Segmentation. *IEEE Trans. on Circuits and Systems for Video Technology*, 11(12):1281–1288, December 2001.

[9] W. Li. Overview of Fine Granularity Scalability in MPEG-4 Video Standard. *IEEE Trans. on Circuits and Systems for Video Technology*, 11(3):301–317, March 2001.

[10] G. Lupatini, C. Saraceno, and R. Leonardi. Scene break detection: a comparison. In *Proc. of the Int. Workshop on Continuous–Media Databases and Applications*, pages 34–41, 1998.

[11] R. Rejaie and A. Reibman. Design Issues for Layered Quality—Adaptive Internet Video Playback. In *Proc. of the Workshop on Digital Communications*, pages 433–451, Taormina, Italy, September 2001.

[12] A. M. Rohaly and al. Video Quality Experts Group: Current Results and Future Directions. In *Proc. SPIE Visual Communications and Image Processing*, volume 4067, pages 742–753, Perth, Australia, June 2000.

[13] Y.-S. Saw. *Rate Quality Optimized Video Coding.* Kluwer Academic Publishers, 1999.

[14] MediaWare Solutions. MyFlix 3.0. http://www.mediaware.com.au/MyFlix.html.

[15] S. Winkler. *Vision Models and Quality Metrics for Image Processing Applications.* PhD thesis, EPFL, Switzerland, 2000.

[16] Q. Zhang, W. Zhu, and Y-Q. Zhang. Resource Allocation for Multimedia Streaming over the Internet. *IEEE Transactions on Multimedia*, 3(3):339–355, September 2001.