Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG
(ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6)
17th Meeting: Nice, France 18-21 Oct., 2005

Document: JVT-Q025
Filename: JVT-Q025.pdf

Title:	Rate-Distortion analysis and streaming of SP and SI frames
Status:	Input Document to JVT
Purpose:	Implementation proposal / performance evaluation

Author(s) or Contact(s):
Eric Setton and Bernd Girod
Information Systems Laboratory
Department of Electrical Engineering
Stanford University
Stanford, CA94305-9510, USA

Tel: +1 (650) 723-3476
Fax: +1 (650) 723-8473
Email: esetton@stanford.edu

Source: Stanford University

Abstract – *Despite widespread interest, no reference implementation of an SI/SP slice encoder has been provided by JVT. We describe modifications to the reference software which enable the encoding of SI slices and of switching SP slices. We also describe a rate-distortion model which is useful in analyzing the performance of SP and SI frames. Through theoretical and empirical analysis we derive optimal settings for encoding SP and SI frames for streaming with packet losses. Our works identifies scenarios where SP and SI frames offer an attractive alternative to streaming with periodic I frames.*

1. Software implementation

All the results presented in the following were collected with our implementation of an SP/SI slice encoder. A patch which modifies the reference software to enable SI slice and switching SP slice coding is available on the following website: http://www.stanford.edu/~esetton/H264_2.htm . Although the results in the rate-distortion analysis were obtained for version JM7.6, the patch has been updated to modify the latest reference software, JM10.1.

1.1 SI slice coding

The encoder generates an SI slice in two passes:

- in the first pass, a primary or non-switching SP slice is created,
- in the second pass, the corresponding SI slice is encoded.

This is transparent to the user which only sees the result of the second pass, as illustrated in the following example, obtained for the first frames of the Foreman QCIF sequence:

SI stream (JM10.1)

Frame	Bits	QP	PSNRY	PSNRU	PSNRV
0000(IDR)	24976	28	36.948	39.744	41.996
0001(P)	2544	28	36.428	39.484	41.627
0002(P)	3528	28	36.215	39.498	41.431
0003(P)	3272	28	36.015	39.543	41.204
0004(SI)	**38904**	**26**	**35.664**	**39.245**	**41.088**
0005(P)	3376	28	35.572	39.193	40.913
0006(P)	4088	28	35.539	39.294	40.801
0007(P)	2768	28	35.456	39.164	40.972
0008(SI)	**39376**	**26**	**35.373**	**39.213**	**40.595**
0009(P)	3136	28	35.455	39.051	40.485

SP stream (to verify perfect reconstruction)

Frame	Bits	QP	PSNRY	PSNRU	PSNRV
0000(IDR)	24976	28	36.948	39.744	41.996
0001(P)	2544	28	36.428	39.484	41.627
0002(P)	3528	28	36.215	39.498	41.431
0003(P)	3272	28	36.015	39.543	41.204
0004(SP)	**5760**	**26**	**35.664**	**39.245**	**41.088**
0005(P)	3376	28	35.572	39.193	40.913
0006(P)	4088	28	35.539	39.294	40.801
0007(P)	2768	28	35.456	39.164	40.972
0008(SP)	**5760**	**26**	**35.373**	**39.213**	**40.595**
0009(P)	3136	28	35.455	39.051	40.485

The encoding of SI slices is controlled by a flag, which can be set in the encoder configuration file. The SI slice encoding is implemented for rate-distortion optimized encoding.

1.2 Switching SP slice coding

Switching SP pictures are used to switch between two bit-streams (e.g. a high quality stream and a low quality stream), without the need of intra-coded pictures. Encoding switching SP slices is done in three steps:

- a high quality bit-stream is encoded;
- a low quality bit-stream is encoded;
- a bit-stream alternating between the two qualities is encoded.

During each of the two first steps, an additional output file is created. This file contains the transformed and quantized coefficients (denoted by *c* in the standard) which are compressed losslessly, with motion-compensated prediction when switching pictures are encoded. Our implementation encodes SP slices which switch from the one bit-stream to the other. This allows, for example, alternating between a high and a low quality stream as in the following example, obtained for the first frames of the Foreman QCIF sequence:

Switched stream (JM10.1)

Frame	Bits	QP	PSNRY	PSNRU	PSNRV
0000(IDR)	9280	38	30.118	36.733	37.801
0001(P)	832	38	30.198	36.64	37.692

0002(P)	944	38	30.083	36.589	37.264
0003(P)	1064	38	30.006	36.587	37.13
0004(SP)	**35984**	**26**	**35.664**	**39.245**	**41.088**
0005(P)	3376	28	35.572	39.193	40.913
0006(P)	4088	28	35.539	39.294	40.801
0007(P)	2768	28	35.456	39.164	40.972
0008(SP)	**14184**	**36**	**28.59**	**35.486**	**35.605**
0009(P)	776	38	28.838	35.807	35.856

High quality stream (to verify perfect reconstruction)

Frame	Bits	QP	PSNRY	PSNRU	PSNRV
0000(IDR)	24976	28	36.948	39.744	41.996
0001(P)	2544	28	36.428	39.484	41.627
0002(P)	3528	28	36.215	39.498	41.431
0003(P)	3272	28	36.015	39.543	41.204
0004(SP)	**5760**	**26**	**35.664**	**39.245**	**41.088**
0005(P)	3376	28	35.572	39.193	40.913
0006(P)	4088	28	35.539	39.294	40.801
0007(P)	2768	28	35.456	39.164	40.972
0008(SP)	**5760**	**26**	**35.373**	**39.213**	**40.595**
0009(P)	3136	28	35.455	39.051	40.485

Low quality stream (to verify perfect reconstruction)

Frame	Bits	QP	PSNRY	PSNRU	PSNRV
0000(IDR)	9280	38	30.118	36.733	37.801
0001(P)	832	38	30.198	36.64	37.692
0002(P)	944	38	30.083	36.589	37.264
0003(P)	1064	38	30.006	36.587	37.13
0004(SP)	**896**	**36**	**28.868**	**35.925**	**36.377**
0005(P)	1152	38	29.11	36.231	36.682
0006(P)	1000	38	29.138	36.054	36.607
0007(P)	768	38	29.113	35.88	36.4
0008(SP)	**1136**	**36**	**28.59**	**35.486**	**35.605**
0009(P)	776	38	28.838	35.807	35.856

The encoding of switching SP slices and several other parameters which control the output and names of additional files can all be set in the encoder configuration file. Switching SP slice encoding is implemented for rate-distortion optimized encoding. Note that the number of reference frames of P and SP frames was limited to 1 to avoid mismatch.

2. Rate-Distortion analysis

Rate-Distortion Analysis and Streaming of SP and SI Frames

Eric Setton, *Student Member, IEEE,* and Bernd Girod, *Fellow, IEEE,*

***Abstract*— SP and SI frames are new picture types introduced in the latest video coding standard H.264. They allow drift-free bitstream switching and can also be used for error-resilience and random access. In this paper, we propose a model for the rate-distortion performance of SI and SP frames, and compare it to experimental results, obtained with our implementation of an SP/SI encoder, based on the H.264 reference software and made publicly available. The model predicts how the relative sizes of SP and SI frames can be traded off. We analyze, both theoretically and experimentally, how this can be used to minimize the transmitted bit-rate when SP frames are used for video streaming with packet losses, and derive optimal settings for our encoder. We investigate the benefits of SI and SP frames for error resilience as compared to periodic I frame insertion. Empirical rate-distortion curves predict rate-distortion gains may be obtained. Experiments carried out over a simulated throughput-limited network confirm this to be the case when the end-to-end delay is limited. We analyze the influence of loss rate and delay on the congestion-rate-distortion performance of streaming with SI and SP frames. Our results identify scenarios for which SI and SP frames provide an attractive alternative to streaming with I frames.**

***Index Terms*— Video compression, video streaming, H.264, bitstream switching, SP frames, SI frames.**

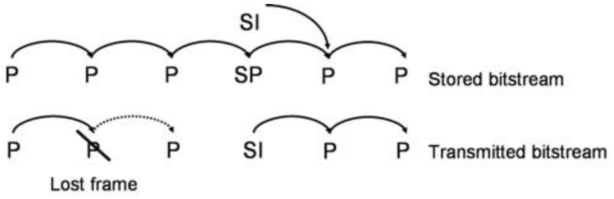

Fig. 1. SI frames share the instant refresh properties of I frames but are only sent after a frame is lost.

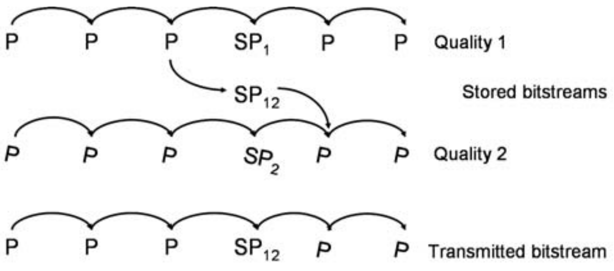

Fig. 2. Switching SP frames allow to switch streams using predictive frames only.

I. Introduction

THE video coding standard, H.264/AVC [1], accommodates the requirements of video streaming solutions which must adapt to varying network conditions. In addition to achieving superior coding efficiency, H.264 uses network-friendly syntax and incorporates several new encoding features which can be taken advantage of when designing flexible and adaptive streaming systems. The new picture types SP and SI are among these features.

SP/SI pictures are new types of predictively/intra coded pictures. Based on the seminal work by Färber et al. [2], they were proposed in 2001 by Karczewicz and Kurceren, as a solution for error resilience, bitstream switching and random access [3], [4]. They are now part of the Extended Profile of H.264. The main advantage of this new picture type is that it can be reconstructed without drift by using different sets of predictors or no predictor at all. This allows drift-free bitstream switching applications, e.g. to refresh a prediction chain or switch between different quality streams as depicted in Fig. 1 and Fig. 2.

Despite widespread interest in SP and SI frames, no work so far has addressed the following questions: how efficient are SP and SI frames? How can their relative sizes be traded off? How does streaming with SP and SI frames compare to traditional systems? This is, in part, due to the fact that no reference implementation of an SP encoder has been provided to the community. The purpose of this work is to address these questions. We propose a model for the rate-distortion functions of SP and SI frames and use it to analyze the properties of these pictures. From the model, we derive optimal encoder settings for typical streaming scenarios which strive to minimize the expected transmitted bit-rate. Finally, we investigate the benefits of SP and SI frames for streaming with packet losses and identify in which scenarios SP and SI frames offer an attractive alternative to streaming with I frames. This paper extends and completes work presented in [5] and in [6].

In the next section, we define switching and non-switching SP frames and describe their encoding. In Section III, we model the rate-distortion performance of SP frames and SI frames and compare it to experimental results obtained with our implementation of an SP frame encoder [7]. The model predicts the relative performance of P, SP, SI and I frames. It also indicates how the relative sizes of SP and SI frames can be traded off. We analyze, in Section IV, both theoretically and experimentally, how this can be used to minimize the transmitted bit-rate when SP frames are used for video streaming with packet losses. In Section VI, we present experimental results carried out over a simulated throughput-limited network to compare streaming with SP and SI frames to periodic I frame insertion. We analyze the influence of losses and delay on performance and conclude by identifying scenarios in which SP and SI frames offer an attractive alternative to streaming with I frames.

II. Encoding of SP and SI slices

Predictively encoded P pictures[1] can only be reconstructed without drift when their set of reference frames is decoded correctly. To alleviate this requirement, a non-switching (also called primary) SP picture may be inserted in the bitstream as shown at the top of Fig. 1 and 2. Along with this non-switching SP picture, a corresponding SI picture or a switching SP picture may be created. The SI picture can be decoded without any predictor and will correspond exactly to the initial primary SP picture. Likewise, the switching (also called secondary) SP picture, can be decoded from its own set of predictors. Its reconstruction corresponds exactly to the initial primary SP picture.

A. Encoding of non-switching SP slices

The diagram of an H.264 primary SP encoder is shown in Fig. 3. Notations for the signals and quantization control parameters follow the H.264 standard [1]. The differences between this encoder and a P slice encoder are highlighted in the diagram.

The first difference is an additional quantization followed by inverse quantization which operates on the signal c^s. It is this additional step that allows identical reconstruction from different predictors and provides the switching and restart functionalities of SP slices.

The second difference is an additional quantization step followed by inverse quantization which operates on the transformed prediction signal c^p. In the current reference software implementation [8], this step is performed at the encoder, on a block by block basis, only when it is beneficial to rate-distortion performance [9], [10]. The choice in performing this step or not is illustrated in Fig. 3 by the presence of a switch which controls how $\tilde{c}^p$ is obtained. However, this additional step has little influence on the rate-distortion performance of SP slices. This is illustrated in Fig. 4, for the CIF sequence *Foreman* where the rate-distortion performance of SP pictures, encoded with and without this enhancement are compared. Results are shown in terms of peak-signal-to-noise-ratio (PSNR), measured in dB, for 18 pictures, evenly spaced in the sequence. The quality improvement due to the conditional quantization never exceeds 0.4 dB. As SP slices, typically represent only a small fraction of an encoded bitstream, this loss in performance is negligible. Hence, in the rate-distortion analysis developed in Sec. III, we will assume c^p never undergoes this additional step. We will also neglect the effect of the final loop filter.

B. Encoding of SI slices and switching SP slices

The quantized coefficients, c, are subsequently losslessly compressed to produce SI or switching SP slices. For switching SP slices, only the residual of a motion-compensated prediction of c is entropy-coded, as depicted in Fig. 5. For SI slices, the prediction signal is obtained from other parts of the same slice. As these steps are lossless, the coefficients c may be obtained at the decoder whether an SP, SI or switching SP frame is transmitted. This ensures that the reconstructed image is identical in all cases.

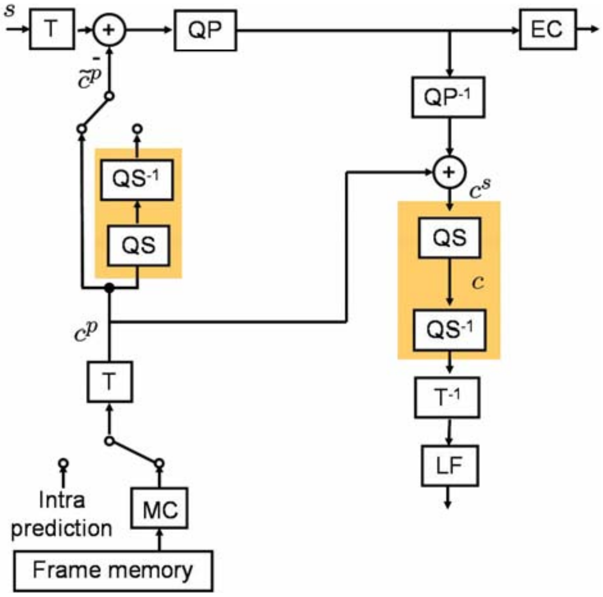

Fig. 3. H.264 primary SP encoder. Switches are indicated by small white circles. Quantizers are denoted by QP and QS, in-loop filtering by LF, transforms by T and motion-compensation by MC.

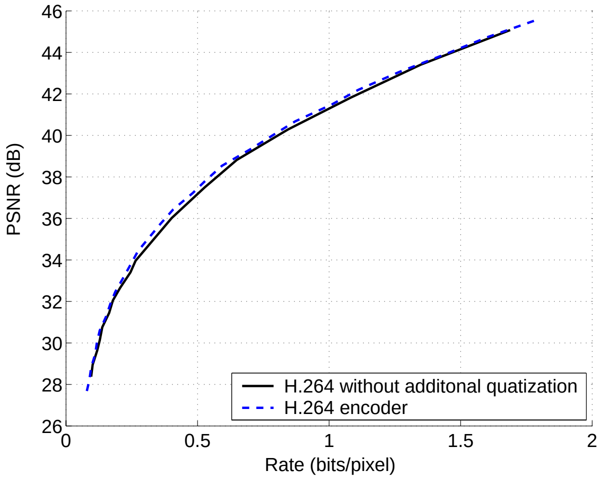

Fig. 4. H.264 SP picture encoder performance with and without the optional additional quantization of the signal c^p.

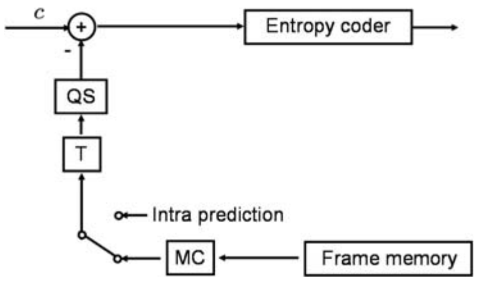

Fig. 5. H.264 SI and switching SP encoder.

[1] Throughout the paper, we employ the terms frame and picture interchangeably and associate them to picture types. These terms refer to what is defined in H264 as a frame, encoded as one slice of this type.

For a given quality, the size of non-switching SP slices and of SI slices may be traded off by varying the two quantization parameters, QP and QS[2], which control the quantizers shown in Fig. 3. At a given quality, making the quantizer QP finer (and the quantizer QS coarser) reduces the size of SI slices at the expense of larger SP slices. The optimal tradeoff depends on the application.

III. RD analysis of SP and SI frames

In this section we explain how the rate-distortion performance of primary and secondary SP frames can be modelled. Our analysis follows the model described in [11] for motion-compensated coding. The model is derived by assuming the different image signals, and the various error signals used throughout the paper are stationary and jointly Gaussian zero-mean signals. Although this is an over-simplification, this model has been used in the literature to model the rate-distortion performance of image or video encoders (see e.g. [11], [12], [13]). The rate-distortion functions we derive can be thought of as an upper bound to the rate-distortion function for a non-Gaussian signal with the same power spectral density (PSD).

In the rest of the paper, we denote the PSD of a signal a by $\Phi_{aa}(\Lambda)$, where $\Lambda = (\omega_x, \omega_y)$ is a vector representing spatial frequency. The independent variables will sometimes be omitted when there is no ambiguity. We define a *picture encoder* as a cascaded transform, quantizer and entropy coder. The inverse process is denoted defined as a *picture decoder*. The analysis presented in the following is based on the following result, obtained from [14]: the rate-distortion function of a stationary two-dimensional zero-mean Gaussian signal a is

$$R_a = \frac{1}{8\pi^2} \iint\limits_{\Lambda} \max(0, \log_2(\frac{\Phi_{aa}(\Lambda)}{\theta_1}))d\Lambda \text{ bit} \quad (1)$$

$$D_a = \frac{1}{4\pi^2} \iint\limits_{\Lambda} \min(\theta_2, \Phi_{aa}(\Lambda))d\Lambda. \quad (2)$$

In (1)-(2), θ is a parameter which takes on all positive values to generate the rate-distortion curve. We denote by *ideal picture encoder* a picture encoder which achieves this optimal rate-distortion performance.

A. RD analysis of primary SP pictures

The diagram in Fig. 6 is our model of the H.264 primary SP picture encoder shown in Fig. 3. This model is obtained by ignoring the effect of the loop filter and by assuming $\tilde{c}^p = c^p$, as stated in the previous section. We also assume that c^p is obtained simply by motion-compensation of the previous picture in the frame memory. The diagram in Fig. 6 can then be obtained from the diagram in Fig. 3 by re-arranging the transforms, entropy coding and entropy decoding blocks, in a way that does not change any of the signals output by the system. In this process we take advantage of the linearity of the transform. The resulting block diagram, is simply composed of two picture encoders and two picture decoders, delineated by dots in Fig. 6. In the following derivation, we assume the encoders are ideal.

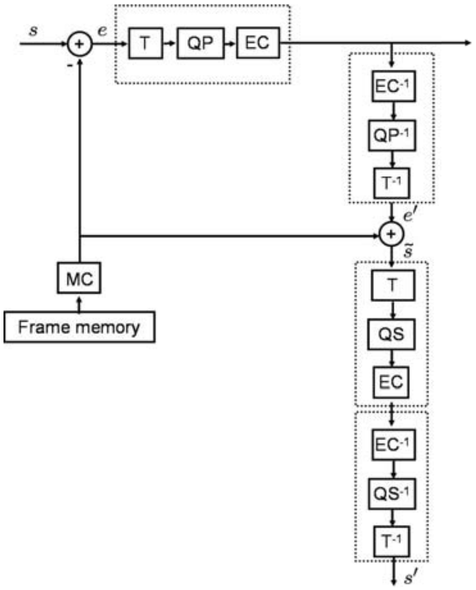

Fig. 6. Model of a primary SP encoder. Entropy coding is denoted by EC.

We consider the image signal s and the prediction error e shown in Fig. 6. As the signal e is assumed to be Gaussian, we obtain from (1) the expression for the rate of primary SP pictures:

$$R_{SP_1} = \frac{1}{8\pi^2} \iint\limits_{\Lambda} \max(0, \log_2(\frac{\Phi_{ee}(\Lambda)}{\theta_1}))d\Lambda \text{ bit.} \quad (3)$$

This expression is identical to that given in [11], for P pictures, which is not surprising as the signal e, in the model of the SP encoder, is identical to the signal which would be obtained when encoding a P picture with a conventional video encoder.

The second picture encoder depicted in Fig. 6 increases the distortion of the reconstructed signal $\tilde{s}$. At high rates, we can assume that the PSD of $\tilde{s}$ is close to that of the original signal s. We further assume that the distortion contributed by the second picture encoder is additive relative to the distortion introduced by the first encoder. Hence, we can express the mean square error distortion of the primary SP picture as a sum of two terms corresponding, respectively, to the distortion contribution of the first and the second encoders:

$$D_{SP_1} = D_1 + D_2 \quad (4)$$

$$D_1 = \frac{1}{4\pi^2} \iint\limits_{\Lambda} \min(\theta_1, \Phi_{ee}(\Lambda))d\Lambda \quad (5)$$

$$D_2 = \frac{1}{4\pi^2} \iint\limits_{\Lambda} \min(\theta_2, \Phi_{ss}(\Lambda))d\Lambda. \quad (6)$$

In (3)-(6), θ_1 and θ_2 are parameters which take on all positive values to generate the rate-distortion curves.

[2] These quantization control parameters are also denoted by $QPSP$ and $QPSP_2$ in the reference software implementation.

B. RD analysis of SI pictures

The diagram in Fig. 7 contains our model of the H.264 SI picture encoder. In this model, for SI pictures we assume the intra-prediction signal is 0, this corresponds to $\widehat{s_1} = 0$. Reconstructed SI slices correspond exactly to the primary SP slices they stem from. Therefore, the distortion of SI pictures, denoted D_{SI} is equal to D_{SP_1}:

$$D_{SI} = D_{SP_1}. \tag{7}$$

As stated in Section II, the signal c is entropy-coded to produce an SI slice. In the model, this corresponds to compressing $\tilde{s}$ (which in this case is equal to ε) by an ideal picture encoder. As the PSD of this signal, $\Phi_{\tilde{s}\tilde{s}}(\Lambda)$, is assumed to be Gaussian and equal to $\Phi_{ss}(\Lambda)$, at high rates, for ideal encoders, the encoding rate of SI pictures is:

$$R_{SI} = \frac{1}{8\pi^2} \iint_{\Lambda} \max(0, \log_2(\frac{\Phi_{ss}(\Lambda)}{\theta_2}))d\Lambda \text{ bit.} \tag{8}$$

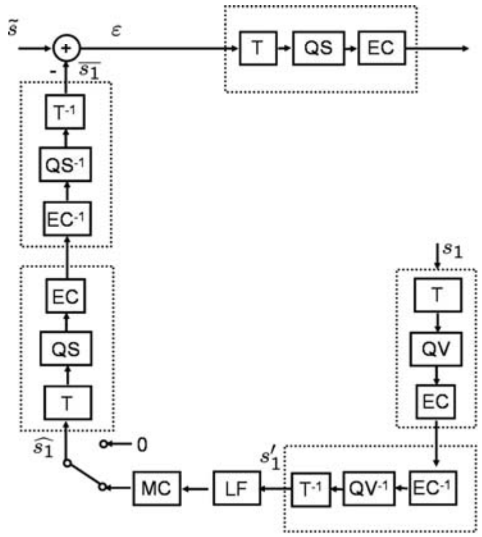

Fig. 7. Model of an SI and switching SP encoder. QV represents a quantizer.

C. RD analysis of secondary SP pictures

To encode a secondary SP picture, a different video stream is used for motion-compensated prediction. In our model, we assume only one picture from the frame memory is used to form the prediction and that there is no intra-prediction. This is illustrated in Fig. 7 where $\widehat{s_1}$ is obtained from a reconstructed version of the compressed picture s_1. If s_1 is compressed at a lower quality than the primary SP picture, the secondary SP picture will serve to switch from a low quality bitstream to a high quality bitstream and vice versa. The correlation between s and s_1 and the magnitude of the compression determine the efficiency of the prediction, whereas the nature of the compression (intra-coding, or motion-compensated predictive coding) has little influence. Hence, in our model we assume s_1 is encoded and decoded by picture encoders and do not consider any prediction when this signal is compressed.

The rest of the diagram in Fig. 7 can be obtained from the diagrams in Fig. 3 and in Fig. 5, by rearranging the transforms, quantization, entropy coding and entropy decoding blocks, in a way that does not change the signal output by the system. In this process, we assume the that some quantizers are uniform, as defined in [15]. Since our model assumes Gaussian signals the assumption is justified as, at high rates, the quantization used by ideal encoders is uniform. The distortion of secondary, or switching, SP pictures can easily be derived as these pictures are identical to the corresponding SI pictures and primary SP pictures:

$$D_{SP_2} = D_{SI} = D_{SP_1}. \tag{9}$$

The rate of a secondary SP picture is expressed as a function of $\Phi_{\varepsilon\varepsilon}$, the PSD of ε:

$$R_{SP_2} = \frac{1}{8\pi^2} \iint_{\Lambda} \max(0, \log_2(\frac{\Phi_{\varepsilon\varepsilon}(\Lambda)}{\theta_2}))d\Lambda \text{ bit.} \tag{10}$$

The expression of $\Phi_{\varepsilon\varepsilon}$ is derived in Appendix I. $\Phi_{\varepsilon\varepsilon}$ depends notably on θ_3, which indicates the level of compression of the picture s_1. This parameter reflects whether the secondary SP picture is used for switching from low quality to high quality (in which case $\theta_3 > \theta_1$), or from high quality to low quality.

D. Rate-distortion performance

Figure 8 shows the rate-distortion performance of SP and SI frames according to (3)-(10). The distortion is represented, in dB, by its SNR. As a reference, the rate-distortion curves of I and P frames, calculated according to [11], are also represented. At high rates, as expected, the slope of all the curves is equal to 6 dB/bit which represent the slope of a memoryless Gaussian process. All the curves are obtained by letting the parameter θ_1 take on all positive values. The expressions used for Φ_{ss} and Φ_{ee} are those suggested in [11]. The derivation of Φ_{ee} and $\Phi_{\varepsilon\varepsilon}$ is obtained by assuming the displacement error in the motion estimation is small and Gaussian with variance $\sigma^2_{\Delta d} = 0.04 \cdot f_{sx}^{-2}$, where f_{sx} is the sampling frequency.

One interesting design parameter is the parameter θ_2 which controls the trade-off between the rate-distortion efficiency of non-switching SP frames and SI frames. Decreasing θ_2 leads to smaller primary SP frames but to larger SI and secondary SP frames. The rate-distortion performance of primary SP frames never exceeds that of P frames (with equality when $\theta_2 = 0$). Likewise, the performance of SI frames is limited by that of I frames (with equality when $\theta_1 = 0$). To generate the rate-distortion performance curves of SP and SI frames shown in Fig. 8, we fix $\theta_2 = 0.9 \cdot \theta_1$. The form of this setting will be justified in Sec. IV.

In the example represented in Fig. 8, the rate-distortion curves for switching SP frames are obtained by setting $\theta_3 = 1.2 \cdot \theta_1$ when switching up, and $\theta_3 = 0.5 \cdot \theta_1$ when switching down. Note that the coding efficiency gap between these two different kind of switching SP frames vanishes at high rates. The rate-distortion curves of secondary SP frames crosses that of I frames at some intermediate rate which depends on the efficiency of the motion-compensated prediction.

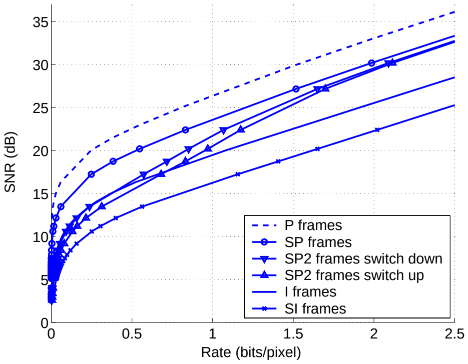

Fig. 8. Theoretical rate-distortion performance.

Although the model is derived for high rates, the theoretical curves correspond to the empirical performance of SP and SI frames shown in Fig. 9 even at low rates[3]. The results were obtained by encoding the QCIF sequence *Foreman* at 30 frames per second. Results are shown for 62 evenly spaced frames encoded with our implementation of an SP encoder [7], based on the H.264 codec. As illustrated, SP frames are typically larger than P frames by approximately 90% for these settings and this sequence. Similarly, SI frames are 35% larger than I frames. The main characteristics predicted by the model are verified experimentally. Namely, SP frame rate-distortion performance is between that of P frames and I frames, and SI frames are larger than I frames. The size of primary SP frames and SI frames can be traded off. In H.264, the counterparts to θ_1 and θ_2 are the two quantization control parameters QP and QS, which control the relative sizes of the frames. They are set according to the last column of Table I. Secondary SP frames rate-distortion performance is worse than that of I frames at low rates, and better at high rates.

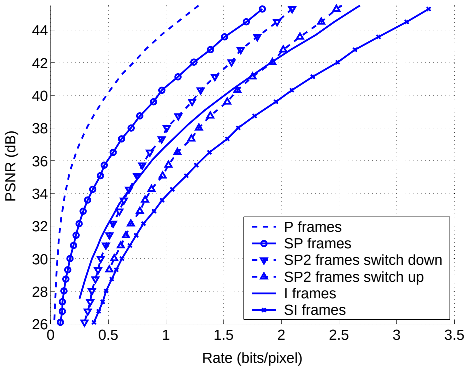

Fig. 9. Experimental rate-distortion performance.

IV. Optimal Setting for Streaming

In this section, the model is used to find how to set the SP frame encoder to minimize the expected bit-rate when SP and SI frames are used for streaming with packet losses.

We consider the scenario, depicted in Fig. 1, in which SP frame positions are spaced regularly in the transmitted video stream. At each of these positions, an SI frame can be sent instead of a primary SP frame to stop potential error propagation. One expects this to result in bit-rate savings compared to periodic I frame insertion which occurs regardless of the outcome of previous transmissions. To take full advantage of this effect, we seek an optimal tradeoff between the sizes of SP and SI frames. Depending on the packet error rate and on the spacing of SP frames, different relative proportions of SI and SP frames will be transmitted. We denote x the probability of transmitting an SI frame at an SP frame position. Minimizing the expected bit-rate, at a given quality, is equivalent to minimizing the expected size of a frame sent at an SP position:

$$\mathcal{R} \quad = \quad xR_{SI} + (1-x)R_{SP_1}. \tag{11}$$

In our model, R_{SI} and R_{SP_1} depend on the two encoding parameters θ_1 and θ_2. The optimal tradeoff corresponds an optimal setting of these parameters, θ_1^* and θ_2^*, derived by solving the following constrained optimization problem:

$$\text{Minimize} \qquad \mathcal{R} \tag{12}$$
$$\text{such that} \quad D_{SP_1} = D_{SI} = D \tag{13}$$

The equality constraint (13) sets the quality of SP and SI frames equal to the quality of the rest of the encoded stream. At high rates, $\Phi_{ee}(\Lambda) \gg \theta_1$ and $\Phi_{ss}(\Lambda) \gg \theta_2$, this simplifies the expression of R_{SI} and R_{SP_1}, and reduces (12)-(13) to:

$$\text{Minimize} \quad (x-1)\log(\theta_1) - x\log(\theta_2) \tag{14}$$
$$\text{such that} \qquad \theta_1 + \theta_2 = D \tag{15}$$

The solutions to the optimization, θ_1^* and θ_2^*, can easily be derived and are related linearly:

$$\theta_2^* = \frac{x}{1-x}\theta_1^*. \tag{16}$$

In the following, we use (16) to derive the optimal setting of the two quantization control parameters QP and QS in the SP frame encoder.

For any Gaussian signal with a continuous PSD, vanishing at high frequencies, the slope of the rate-distortion function is expressed:

$$\frac{dR}{dD} = \frac{-1}{2\log(2)\Theta} \tag{17}$$

where Θ is the parameter used to generate the rate-distortion curve [16]. A derivation of this result is included in Appendix II. Equation (17) can be used to express, as a function of θ_1, the slope of the rate-distortion function of the error signal e, represented in Fig. 6, denoted by $\frac{dR_e}{dD_e}$:

$$\frac{dR_e}{dD_e} = \frac{-1}{2\log(2)\theta_1}. \tag{18}$$

[3] Note that the experimental results, use PSNR in dB, while the theoretical results, that deal with a Gaussian random process, use SNR, also in dB. These two measures can be related with a vertical shift. The relative difference in dB for either SNR or PSNR, however, is equivalent.

Likewise, (15) can be used to express, as a function of θ_2, the slope of the rate-distortion function of $\tilde{s}$ denoted by $\frac{dR_{\tilde{s}}}{dD_{\tilde{s}}}$:

$$\frac{dR_{\tilde{s}}}{dD_{\tilde{s}}} = \frac{-1}{2\log(2)\theta_2}. \tag{19}$$

In H.264, the slope of distortion-rate of the encoder is expressed, empirically, as a function of the quantization parameter [17]. Let's consider encoding the signals e and $\tilde{s}$ with H.264. According to [17], $\frac{dD_e}{dR_e}$ and $\frac{dD_{\tilde{s}}}{dR_{\tilde{s}}}$ are given by:

$$\frac{dD_e}{dR_e}(H.264) \quad = \quad -0.85\ 2^{\frac{QP-12}{3}}, \tag{20}$$

$$\frac{dD_{\tilde{s}}}{dR_{\tilde{s}}}(H.264) \quad = \quad -0.85\ 2^{\frac{QS-12}{3}}. \tag{21}$$

If we assume the H.264 approaches ideal rate-distortion performance, then the expressions in (18) and (20) are equal and so are (19) and (21). The optimal setting QP^* and QS^* is given by combining (18)-(21), which results in a simple offset:

$$QS^* = QP^* + 3\log_2(\frac{x}{1-x}). \tag{22}$$

This offset is independent of the encoding rate. This allows us to formulate a simple setting for encoding SP and SI frames. As we require SP frames to have the same quality as the other frames of the video stream, both QP^* and QS^* should be finer than QP_{ref}, the quantization control parameter used while encoding P frames or I frames. In addition, H.264 restricts the values of the three quantization parameters QP, QS and QP_{ref} to integers. Based on these restrictions, we have determined experimentally that the only suitable settings are those indicated in Tab. I. According to (22), we derive for which values of x these settings should be used. The settings indicated by the model correspond to settings derived empirically, also reported in the table.

TABLE I

OPTIMAL SETTINGS FOR QP AND QS, FOR DIFFERENT PROBABILITIES OF TRANSMITTING AN SI FRAME

x model	≤ 0.2	≥ 0.2 and ≤ 0.5	≥ 0.5
x empirical	≤ 0.1	≥ 0.1 and ≤ 0.2	≥ 0.2
QP	$QP_{ref} - 1$	$QP_{ref} - 2$	$QP_{ref} - 3$
QS	$QP_{ref} - 10$	$QP_{ref} - 5$	QP_{ref}

V. PERFORMANCE ANALYSIS

In this section, we analyze the empirical rate-distortion efficiency of a stream encoded with SP and SI frames for streaming and derive bounds on bit rate savings when SP and SI frames are used instead of periodic I frame insertion.

The temporally layered encoding scheme, shown in Fig. 10, was chosen to encode the video. The first temporal layer is composed of SP frames (and their corresponding SI frames). The second temporal layer is composed of P frames. We restrict both P and SP frames to use as a reference the P or SP frame preceding them in display order, as illustrated in the figure. The last layer is composed of B frames. We restrict the frames to use as reference their two neighboring P frames or SP frames[4]. This ensures good error resilience properties and allows to easily scale down the frame rate by 2 or even 4 if needed. The encoded video sequences used in the following experiments are made publicly available [18].

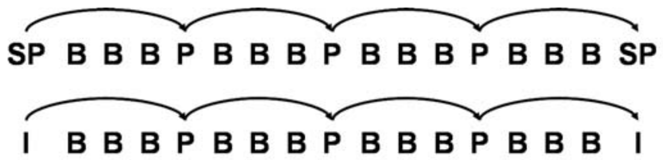
SP B B B P B B B P B B B P B B B SP
I B B B P B B B P B B B P B B B I

Fig. 10. Encoding structures used for streaming with SP and SI frames and for periodic I frame insertion.

Fig. 11 illustrates the rate-distortion characteristic of the video sequences *Mother and Daughter* and *Foreman*, in CIF format, compressed using the coding structure described in Fig. 10. As there are only one I, SP or SI frame in 16 frames, the difference between the curves is not as pronounced as in Fig. 9. The main difference between these two sequences is the relative size of I frames and P frames. For the sequence *Mother and Daughter*, I frames are approximately 3 times larger than P frames. For the sequence *Foreman*, I frames are approximately twice the size of P frames. This difference comes from larger motion in the sequence *Foreman* which increases the average size of P frames. As a consequence, the bit rates savings expected by sending intra-coded frames on an as-needed basis will be higher for the sequence *Mother and Daughter*. For the first sequence, transmitting SP frames instead of I frames can lead to a performance gain of 1.5 dB at low bit rates and 1 dB at higher rates. For the second sequence, this gap is smaller and ranges from 1 dB to 0.8 dB. These gaps represent a bound on the performance improvement, achieved when streaming takes place with no losses. If SI frames are used instead of I frames, the rate distortion performance is reduced by approximately 1.5 dB at low rates and a little less than 1 dB at high rates.

VI. SIMULATION RESULTS

To illustrate realistically the benefits of streaming with SP and SI frames we consider a low latency video streaming scenario, suitable for live streaming or for video-on-demand, where a sender transmits video frames sequentially to a receiver which sends acknowledgements (ACKs) back. We strive for end-to-end delays of no more than a few hundred milliseconds. When a packet arrives at the receiver after its playout deadline, it is discarded by the decoder as if it were lost. To avoid interruptions, the errors due to packet loss or to excessive delays are concealed by freezing the previous frame until the next decodable frame and the playout continues at the cost of higher distortion. The sender retransmits lost packets when ACKs are received out of sequence, and when there is still enough time to retransmit a packet before its playout deadline. When SP frames are used, if a P frame or an SP frame is lost and cannot be retransmitted, an SI frame is sent at the next SP frame position as depicted in Fig. 1.

[4] Please note that these restrictions are not dictated by the H.264 standard.

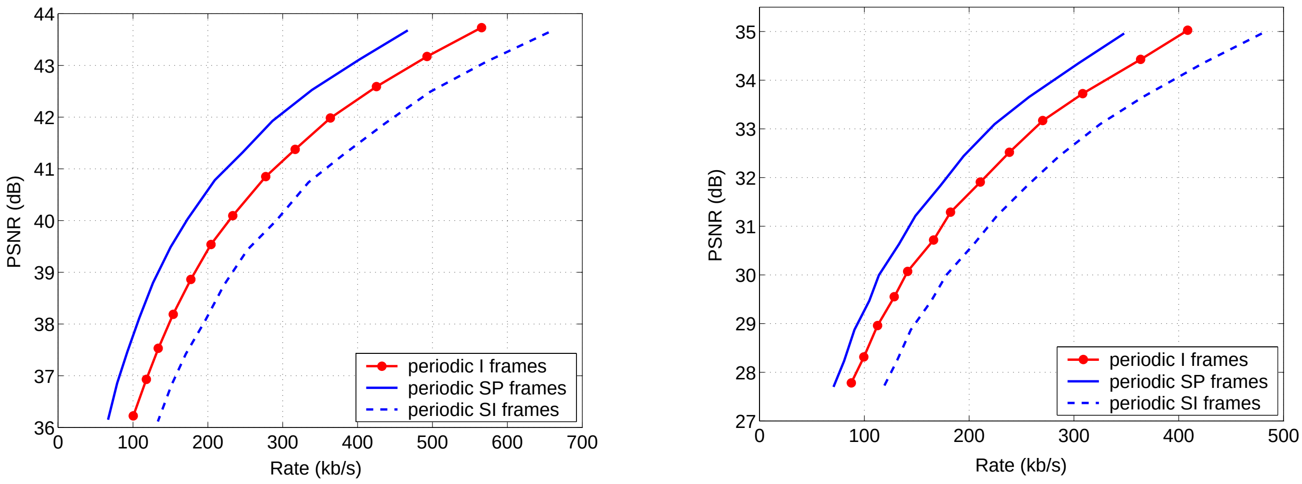

Fig. 11. Rate-distortion performance with periodic I frame, SP frame or SI frame insertion for the sequences *Mother & Daughter* (left) and *Foreman* (right).

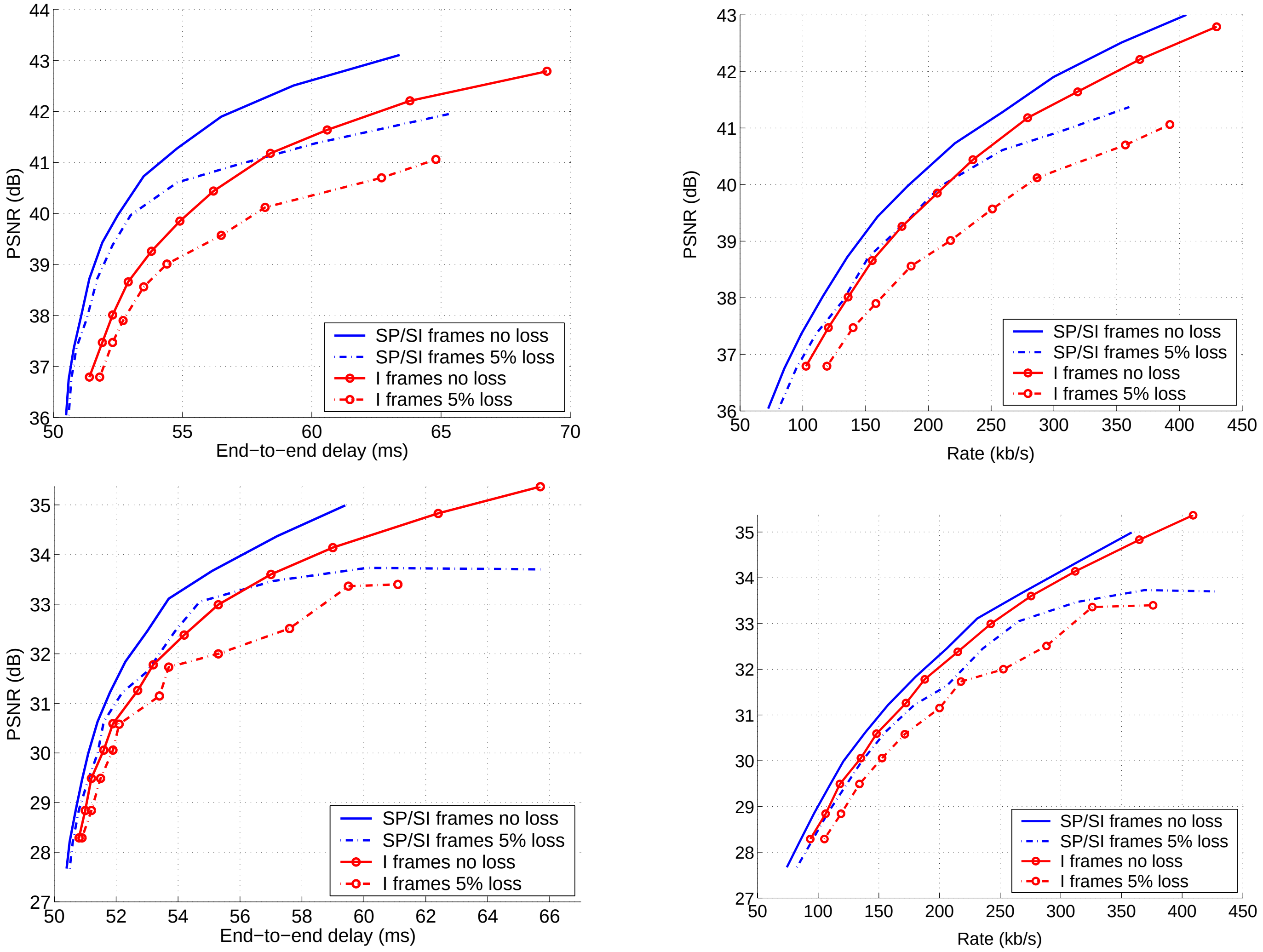

Fig. 12. Congestion-Rate-Distortion performance for *Mother & Daughter* (top) and *Foreman* (bottom) for varying loss rates.

We consider the route between sender and receiver as a succession of high bandwidth links ended by a bottleneck last hop which can support up 800 kb/s. Packet losses are simulated on this last hop in some of the following experiments. Packets containing an entire video frame are generated by our video encoder and are fragmented, if required by the transport layer. When a loss occurs, the entire frame is discarded, even though, in most cases, only one packet is lost. It is important to consider realistic packetization as different frame types have vastly varying sizes, as illustrated in Fig. 9. At low rates, for example, B and P frames may fit into one MTU size packet, whereas SP frames may necessitate 2 packets, I frames 3, and SI frames 6. Consequently, different frame types may experience different loss rates. The impact on the resulting PSNR may be significantly different from that induced by independent losses identically distributed among all the frames.

The sequences are encoded at 30 frames per second with the encoding structure shown in Fig. 10. The first 288 frames of the sequences are encoded, and the encoded sequence is looped 50 times when collecting results. Video quality is measured by taking the average of the PSNR over all the decoded frames. Performance is also evaluated in terms of the total transmitted rate, including retransmissions, and of the average end-to-end delay between the server and the client. This quantity reflects the congestion created by the stream on the network. The fact that this metric, unlike rate, depends on the capacity of the network path makes it well-suited to performance evaluation in a throughput-limited environment. It reflects the delay another stream would experience if it was sharing the link with the video stream. End-to-end delay is measured by taking the average end-to-end delay of header packets transmitted every 20 ms from the server to the client, as in [19].

A. *Influence of losses*

We first analyze the influence of packet losses. We consider a fixed 50 ms propagation delay and a 500 ms latency tolerance. In Fig. 12, both the congestion-distortion performance and the rate-distortion performance are shown for two sequences. In the absence of losses, the gains in terms of rate and distortion are close to those predicted in the previous section. The rate-distortion performance gap is a little smaller due in part to the fact that I frames are inserted every 10 seconds, each time the sequence is looped. For the sequence *Mother & Daughter* the performance gap is approximately 0.6 dB, and 0.4 dB for the sequence *Foreman* for different bit rates. The congestion-distortion performance gap is larger, it varies from 2 dB for low levels of congestion to 1 dB for higher levels of congestion for the sequence *Mother & Daughter*. The gap is smaller for the sequence *Foreman*. This illustrates the queueing delay spikes caused by I frames, which are not captured by the average rate of the sequence. When a 5% loss rate is introduced on the bottleneck link, the performance drops for all the curves. This drop is more significant at high bit rates as the packet loss rate translates into a higher frame loss rate. For higher rates than those shown, the average decoded video quality decreases. Surprisingly, the rate-distortion performance gap increases when losses are introduced. This is due to the fact that I frame retransmissions occur more frequently than SI frame insertions, and can be explained by the large size of I frames compared to SP frames. The congestion-distortion performance gap remains almost the same. These experiments show that streaming with SI and SP frames is beneficial in this experimental setup regardless of packet loss rate.

B. *Influence of delay*

In this section, we analyze the influence of the propagation delay. We consider a fixed 2% packet loss rate and a 500 ms latency tolerance. The propagation delay is varied between 20 ms and 200 ms. This delay occurs on the high bandwidth links and reflects the time needed for signal to propagate along links which can potentially be very long (e.g. transoceanic or transcontinental links). Please note that in addition to propagation delay, transmission delay is also taken into account. In the experiments, transmission delay is dominated by the delay over the 800 kb/s bottleneck link. Streaming performance for varying forward trip times (FTT) is shown in Fig. 13. For short propagation delays the performance is only slightly worse than the performance in the absence of loss, discussed in Sec. VI-A. The slight loss in performance is due to the 2% loss rate which induces retransmissions and an increase in rate. For long propagation delays, there is no time for retransmission and the performance drop for all the schemes is 2 to 4 dB. This causes significant quality impairments for streaming with SI and SP frames as well as for streaming with I frames. The performance gap in this case is reversed. Indeed, as SI frames need to be inserted almost constantly, the congestion-rate-distortion performance is worse than for periodic I frame insertion. The performance gap ranges from 1.5 to 2 dB for different rates and congestion levels for *Mother & Daughter*. Likewise, for *Foreman*, it ranges from 1 dB to 2 dB. For high propagation delays, in the absence of retransmission, streaming with periodic I frames is more efficient and the performance gap is significant.

As a summary to this analysis, SP and SI frames provide an attractive alternative to streaming with I frames when feedback is available and propagation delay is small compared to the maximum tolerable latency. In these cases, the performance is superior both in terms of congestion-distortion and in terms of rate-distortion. The performance gap is larger for low motion sequences as this causes larger differences between I and P frames. It is also more pronounced at lower bit rates and can reach up to 1.5 dB. For the case when retransmissions are not possible, periodic I frame insertion remains the best alternative and the performance gap is over 1 dB.

VII. Conclusion

In this paper we analyze and discuss the encoding and benefits for streaming of the new H.264 picture types SP and SI. We propose a theoretical model which predicts the rate-distortion performance of non-switching SP frames, SI frames and switching SP frames. Experimental results, obtained with our implementation of an SP encoder, based on

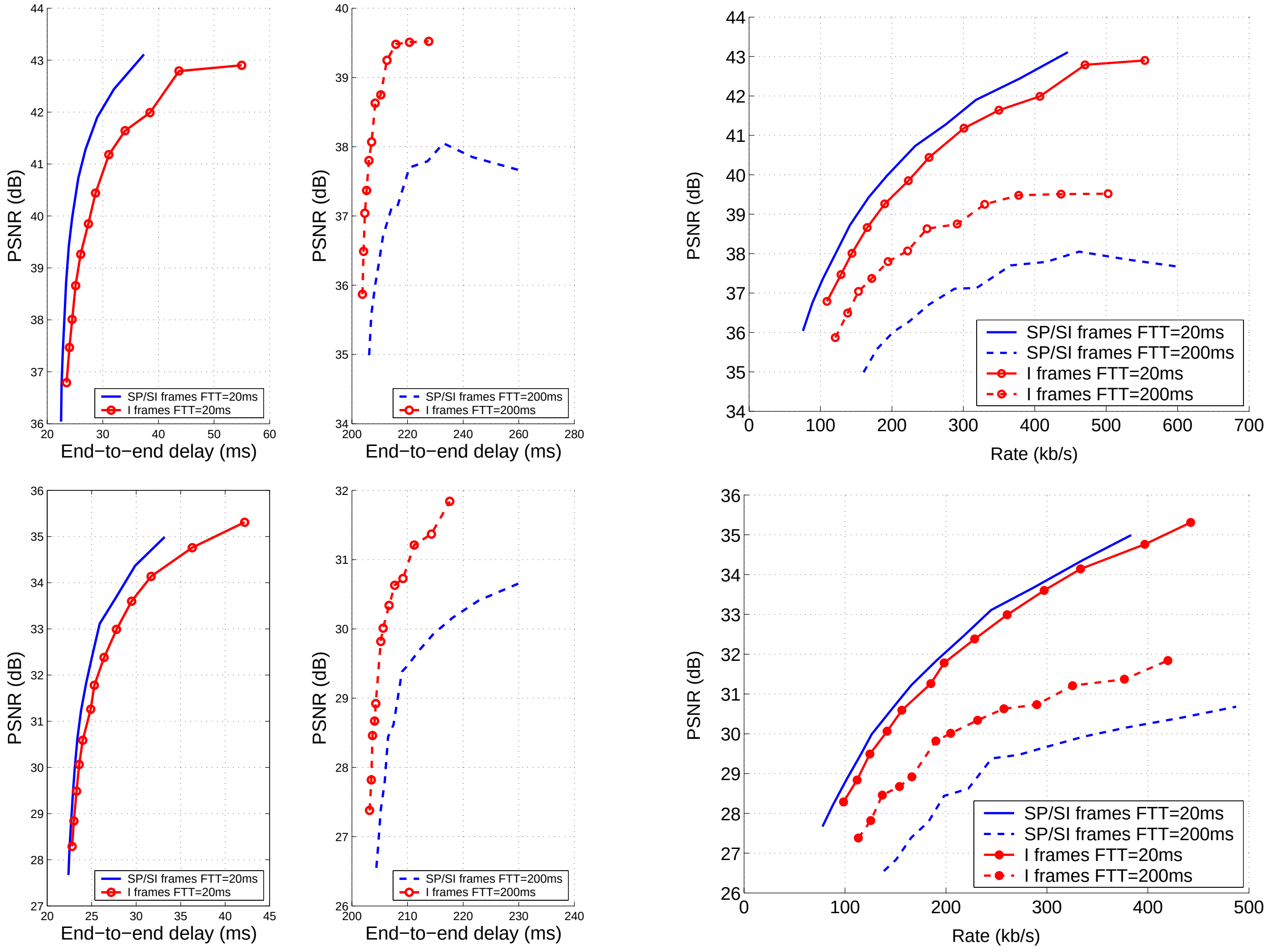

Fig. 13. Congestion-Rate-Distortion performance for *Mother & Daughter* (top) and *Foreman* (bottom) for varying propagation delays

H.264 and made publicly available, validate the theoretical results. The model predicts the relative performance of SP, SI and switching SP pictures compared to other picture types and the tradeoff in the relative efficiency of SP and SI frames. We apply the model to determine the optimal tradeoff and the corresponding practical encoder settings, which minimize the expected bit-rate when SP and SI frames are used for streaming with packets losses. Performance analysis reveals that distortions gains of up to 1.5 dB can be obtained for video rates between 100 kb/s and 600 kb/s when SP and SI frames are used instead of periodic I frame insertion. Experimental results obtained on a simulated bandwidth limited network, for low latency streaming, with varying loss rates and propagation delays confirm these expectations. The experiments also illustrate that streaming with SP and SI frames reduces the congestion created by the stream on the network. The use of SP and SI frames is beneficial for scenarios where feedback is available from the receiver and the propagation delay is low enough to allow ACK-based retransmissions.

Appendix I
Expression of the Power Spectral Density $\Phi_{\varepsilon\varepsilon}$

As illustrated in Fig. 7, the signal ε, is defined as:

$$\varepsilon = \tilde{s} - \overline{s_1}, \qquad (23)$$

where, $\overline{s_1}$ is the reconstructed version of the compressed signal $\widehat{s_1}$. In the following, we derive the expression of the PSD of $\overline{s_1}$ and use it to derive a closed form expression of $\Phi_{\varepsilon\varepsilon}$.

We denote by A, the 2-D band-limited discrete space Fourier transform (F.T.) of a signal a and $*$ denotes generalized 2-D convolution. Details on both operators are given in [20].

If we assume motion between pictures s and s_1 to be a constant translation, s_1 can be expressed as a convolution with a discrete impulse:

$$s_1(x, y) = s(x, y) * \delta(x - d_x, y - d_y), \qquad (24)$$

where (d_x, d_y) is the displacement. The transformation between s_1 and s_1' is the optimum forward channel defined in [14] and used in [11] or more recently in [12] to derive rate-distortion functions of video encoders. In this model, s_1' is the result of filtering s_1 by g and subsequently adding non-white noise n, assumed to be uncorrelated with s_1. The PSD of s_1' can be expressed as a function of G, and of N:

$$\Phi_{s_1's_1'}(\Lambda) = \Phi_{s_1s_1}(\Lambda)|G(\Lambda)|^2 + \Phi_{nn}(\Lambda). \tag{25}$$

The expression of G and Φ_{nn} are given in [14]. As s and s_1, as a consequence of (24), have the same PSD, G and Φ_{nn} are expressed:

$$\begin{aligned} G(\Lambda) &= \max(0, 1 - \frac{\theta_3}{\Phi_{ss}(\Lambda)}), & (26) \\ \Phi_{nn}(\Lambda) &= \max(0, \theta_3(1 - \frac{\theta_3}{\Phi_{ss}(\Lambda)})). & (27) \end{aligned}$$

In (26) and (27), θ_3 determines the level of compression of the signal s_1. Following the encoding process depicted in Fig. 7, the signal s_1' is loop-filtered and motion-compensated to produce $\widehat{s_1}$. We will assume motion-compensation is a spatially constant translation $(\widehat{d_x}, \widehat{d_y})$ resulting in a random displacement error $(\Delta d_x, \Delta d_y)$, and will denote by f the loop-filter. Consequently, $\widehat{s_1}$ and its PSD are:

$$\begin{aligned} \widehat{s_1} &= (s_1' * f) * \delta(x - \widehat{d_x}, y - \widehat{d_y}), & (28) \\ \Phi_{\widehat{s_1}\widehat{s_1}}(\Lambda) &= \Phi_{s_1's_1'}(\Lambda)|F(\Lambda)|^2, & (29) \\ \Phi_{\widehat{s_1}\widehat{s_1}}(\Lambda) &= \Phi_{ss}(\Lambda)|G(\Lambda)F(\Lambda)|^2 + \Phi_{nn}|F(\Lambda)|^2. & (30) \end{aligned}$$

The transformation between $\widehat{s_1}$ and $\overline{s_1}$ is again an optimum forward channel. The signal $\overline{s_1}$ is the result of filtering $\widehat{s_1}$ by $\widehat{g}$ and subsequently adding non-white noise $\widehat{n}$:

$$\begin{aligned} \overline{s_1} &= \widehat{s_1} * \widehat{g} + \widehat{n}, & (31) \\ \Phi_{\overline{s_1s_1}}(\Lambda) &= \Phi_{\widehat{s_1}\widehat{s_1}}(\Lambda)|\widehat{G}(\Lambda)|^2 + \Phi_{\widehat{n}\widehat{n}}(\Lambda). & (32) \end{aligned}$$

We use, once again, the expression of the optimum forward channel given in [14], to write the F.T. of the filter and the PSD of the noise:

$$\begin{aligned} \widehat{G}(\Lambda) &= \max(0, 1 - \frac{\theta_2}{\Phi_{\widehat{s_1}\widehat{s_1}}(\Lambda)}), & (33) \\ \Phi_{\widehat{n}\widehat{n}}(\Lambda) &= \max(0, \theta_2(1 - \frac{\theta_2}{\Phi_{\widehat{s_1}\widehat{s_1}}(\Lambda)})). & (34) \end{aligned}$$

By combining (30) and (32)-(34), the PSD of $\overline{s_1}$ can be expressed as a function of s.

We define $n_s = \tilde{s} - s$. The derivation of $\Phi_{\varepsilon\varepsilon}$ follows:

$$\begin{aligned} \varepsilon = & \quad \tilde{s} - \overline{s_1}, & (35) \\ \varepsilon = & \quad s - \overline{s_1} + n_s, & (36) \\ \varepsilon = & \quad s - (((((s * \delta(x - d_x, y - d_y)) * g + n) \\ & \quad *f) * \delta(x - \widehat{d_x}, y - \widehat{d_y})) * \widehat{g}) - \widehat{n} + n_s. & (37) \end{aligned}$$

We make the assumption that s, $\widehat{n}$, n and n_s are statistically independent. We also assume that, at high rates, the PSD of n_s can be neglected compared to the other noise terms. The displacement error $(\Delta d_x, \Delta d_y) = (d_x, d_y) + (\widehat{d_x}, \widehat{d_y})$, is spatially constant but is random. Hence, the PSD of ε is:

$$\begin{aligned} \Phi_{\varepsilon\varepsilon}(\Lambda) = & \quad \Phi_{ss}(\Lambda)E[|1 - FG\widehat{G}e^{(j(\omega_x\Delta d_x + \omega_y\Delta d_y))}|^2] \\ + & \quad \Phi_{nn}(\Lambda)|F(\Lambda)\widehat{G}(\Lambda)|^2 + \Phi_{\widehat{n}\widehat{n}}(\Lambda). & (38) \end{aligned}$$

where E is the expectation function taken with respect to the probability density function of the displacement error. Following the simplification derived in [11], (38) can be rewritten:

$$\begin{aligned} \Phi_{\varepsilon\varepsilon}(\Lambda) = \Phi_{ss}(\Lambda)(1 + |F(\Lambda)G(\Lambda)\widehat{G}(\Lambda)|^2 \\ -2\text{Re}\{F(\Lambda)G(\Lambda)\widehat{G}(\Lambda)P(\Lambda)\}) \\ +\Phi_{nn}(\Lambda)|F(\Lambda)\widehat{G}(\Lambda)|^2 + \Phi_{\widehat{n}\widehat{n}}(\Lambda), \end{aligned} \tag{39}$$

where $P(\Lambda)$ is the continuous Fourier transform of the displacement error probability density function.

Appendix II
Derivation of the slope of the rate-distortion function

In this section we express the slope of the rate-distortion function of a Gaussian signal. We assume the PSD of this signal, denoted by $\mathcal{S}$, is continuous and vanishes at high frequencies. The proof is given for a 1-D signal and is easily extendable to a 2-D signal.

Any point of the rate-distortion function can be written as a function of a positive parameter Θ:

$$\begin{aligned} R &= \frac{1}{8\pi^2} \int\limits_{\Lambda} \max(0, \log_2(\frac{\mathcal{S}(\Lambda)}{\Theta}))d\Lambda \text{ bit}. & (40) \\ D &= \frac{1}{4\pi^2} \int\limits_{\Lambda} \min(\Theta, \mathcal{S}(\Lambda))d\Lambda & (41) \end{aligned}$$

We denote $[a_i(\Theta)\ b_i(\Theta)]$ the intervals over which $\mathcal{S} > \Theta$. As $\mathcal{S}$ vanishes at high frequencies, the bounds of these intervals are different from the bounds of the integration domain. We note that by continuity:

$$\mathcal{S}(a_i(\Theta)) = \mathcal{S}(b_i(\Theta)) = \Theta. \tag{42}$$

By differentiating R and D with respect to Θ, we obtain:

$$\begin{aligned} \frac{dR}{d\Theta} = & \quad \tfrac{1}{8\pi^2} \textstyle\sum_i (\log_2(\mathcal{S}(b_i(\Theta)))b_i'(\Theta) - \log_2(\mathcal{S}(a_i(\Theta)))a_i'(\Theta) \\ & \quad -\tfrac{1}{\log(2)\Theta}(b_i(\Theta) - a_i(\Theta)) - \log_2(\Theta)(b_i'(\Theta) - a_i'(\Theta))) & (43) \\ = & \quad \tfrac{-1}{8\pi^2\log(2)\Theta} \textstyle\sum_i (b_i(\Theta) - a_i(\Theta)) \\ \frac{dD}{d\Theta} = & \quad \tfrac{1}{4\pi^2} \textstyle\sum_i (\Theta(b_i'(\Theta) - a_i'(\Theta)) + (b_i(\Theta) - a_i(\Theta)) \\ & \quad -(\mathcal{S}(b_i(\Theta))b_i'(\Theta) - \mathcal{S}(a_i(\Theta))a_i'(\Theta))) & (44) \\ = & \quad \tfrac{1}{4\pi^2} \textstyle\sum_i (b_i(\Theta) - a_i(\Theta)) \end{aligned}$$

By dividing both expressions, we get:

$$\frac{dR}{dD} = \frac{-1}{2\log(2)\Theta}. \tag{45}$$

Acknowledgment

The authors would like to thank Dr. P. Ramanathan and Dr. M. Flierl for insightful discussions and the anonymous reviewers who helped improve the presentation of this work.

References

[1] *Advanced Video Coding for Generic Audiovisual services, ITU-T Recommendation H.264 - ISO/IEC 14496-10(AVC)*, ITU-T and ISO/IEC JTC 1, 2003.

[2] N. Färber and B. Girod, “Robust H.263 Compatible Video Transmission for Mobile Access to Video Servers,” *Proc. ICIP-97, Santa Barbara, CA, USA*, vol. 2, pp. 73–76, Oct. 1997.

[3] M. Karczewicz and R. Kurceren, “A Proposal for SP-Frames,” *Video Coding Experts Group Meeting, , Doc. VCEG-L-27, Eibsee, Germany*, Jan. 2001.

[4] ——, “The SP- and SI-frames design for H.264/AVC,” *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 637–644, July 2003.

[5] E. Setton, P. Ramanathan, and B. Girod, “Rate-distortion analysis of SP and SI frames,” *submitted to MMSP 2005*.

[6] E. Setton and B. Girod, “Video Streaming with SP and SI frames,” *Proceedings VCIP, Beijing, to appear*, July 2005.

[7] “H.264 SP frame codec,” *http://www.stanford.edu/˜ esetton/H264_2.htm*.

[8] “H.264/AVC Reference Software,” *http://iphome.hhi.de/suehring/tml/download/, seen on Aug. 28 2005*.

[9] X. Sun, S. Li, F. Wu, J. Shen, and W. Gao, “The improved SP frame coding technique for the JVT standard,” *International Conference on Image Processing, Barcelona, Spain*, vol. 3, pp. 297–300, Sept. 2003.

[10] X. Sun, F. Wu, S. Li, and R. Kurceren, “The improved JVT-B097 SP coding scheme,” *ISO/IEC JTC1/SC29/ WG11 and ITU-T SG16 Q.6, JVT-C114, Fairfax, Virginia, USA*, May 2002.

[11] B. Girod, “The Efficiency of Motion-Compensating Prediction for Hybrid Coding of Video Sequences,” *IEEE Journal on Selected Areas in Communications*, vol. 5, no. 7, pp. 1140–1154, Aug. 1987.

[12] G. Cook, J. Prades-Nebot, and E. Delp, “Rate-distortion bounds for motion-compensated rate scalable video coders,” *International Conference on Image Processing (ICIP)*, pp. 3121–3124, Oct. 2004.

[13] M. H. Flierl, *Video coding with superimposed motion-compensated signals, Ph.D. dissertation, University of Erlangen*, 2003.

[14] T. Berger, *Rate Distortion Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1971.

[15] A. Gersho and R. M. Gray, *Vector quantization and signal compression*. Norwell, MA, USA: Kluwer Academic Publishers, 1991.

[16] R. Gallager, *Information Theory and Reliable Communication*. NY: Wiley, 1968.

[17] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, and G. Sullivan, “Rate-constrained coder control and comparison of video coding standards,” *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 688–703, July 2003.

[18] “Encoded sequences with SP/SI frames,” *http://ivms.stanford.edu/˜ esetton/sequences.htm*.

[19] E. Setton and B. Girod, “Congestion-Distortion Optimized Scheduling of Video,” *Multimedia Signal Processing Workshop (MMSP), Siena, Italy*, pp. 99–102, Oct. 2004.

[20] B. Girod, “Efficiency Analysis of Multi-Hypothesis Motion-Compensated Prediction for Video Coding,” *IEEE Trans. Image Processing*, vol. 9, no. 2, pp. 173–183, Feb. 2000.