Rate-Distortion Analysis and Streaming of SP and SI Frames

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

***Abstract*—The new SP and SI picture types, introduced in the latest video coding standard H.264, 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 SP and SI pictures and compare it to experimental results, obtained with our implementation of an SP/SI encoder, made publicly available and recently adopted by the Joint Video Team. The model predicts how the relative sizes of SP and SI slices 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 with 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*—Bitstream switching, H.264, SI frames, SP frames, video compression, video streaming.**

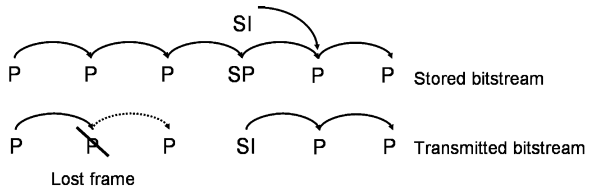

Fig. 1. SI frames share the instant refresh properties of I frames but are only sent after a frame is lost.

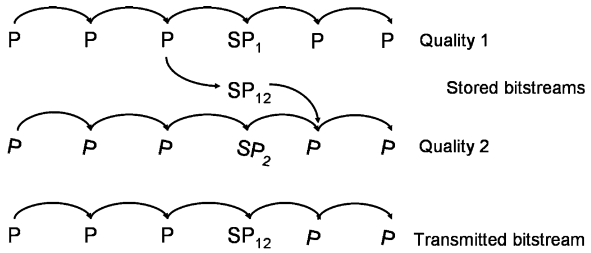

Fig. 2. Switching SP frames allow to switch streams using predictive frames only.

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 work by Färber and Girod [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 Figs. 1 and 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 paper 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 nonswitching SP slices and describe their encoding. In Section III, we model the rate-distortion performance of an idealized SP and SI encoder and compare it with experimental results obtained with our implementation of an SP/SI encoder,[1] recently adopted by the Joint Video Team. 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

Manuscript received June 23, 2005; revised November 7, 2005. This paper was recommended by Associate Editor F. Pereira.

The authors are with the Information Systems Laboratory, Stanford University, Stanford, CA 94305-9510 USA (e-mail: esetton@stanford.edu).

Digital Object Identifier 10.1109/TCSVT.2006.875208

[1]H.264 SP frame codec. [Online] Available: http://ivms.stanford.edu/~esetton/H264_2.htm

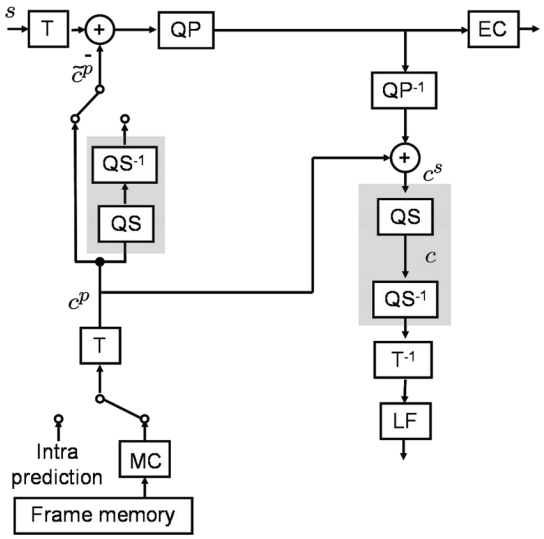

Fig. 3. H.264 primary SP encoder. Switches are indicated by small white circles. Quantizers are denoted by QP and QS, in-loop filtering by LF, transforms by T and motion-compensation by MC.

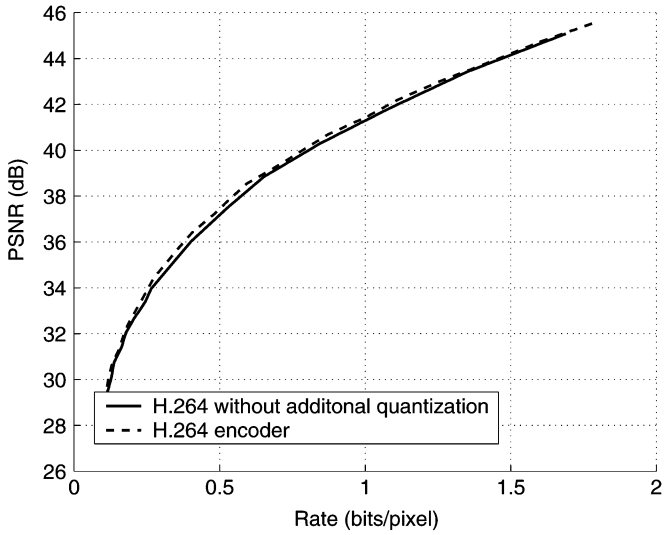

Fig. 4. H.264 SP picture encoder performance with and without the optional additional quantization of the signal c^p.

throughput-limited network to compare streaming with SP and SI frames with 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[2] can only be reconstructed without drift when their set of reference frames is decoded correctly. To alleviate this requirement, a nonswitching (also called primary) SP picture may be inserted in the bitstream as shown at the top of Figs. 1 and 2. Along with this nonswitching 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 Nonswitching 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 picture 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 of the transformed prediction signal c^p. In the current reference software implementation,[3] this step is performed at the encoder, on a block-by-block basis, only when it is beneficial to rate-distortion performance [7], [8]. The choice in performing this step or not is illustrated in the diagram of Fig. 3 by the presence of a switch that 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 by the rate-distortion performance curves of SP pictures, encoded with and without this enhancement. The results are for the CIF sequence *Foreman* and are shown in terms of peak-signal-to-noise-ratio (PSNR), measured in decibels, 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 Section III, we will assume c^p never undergoes this additional step. We will also neglect the effect of the final loop filter in the analysis.

B. Encoding of SI Slices and Switching SP Slices

The quantized coefficients, denoted by c in Fig. 3, 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.

For a given quality, the size of nonswitching SP slices and of SI slices may be traded off by varying the two parameters, QP and QS,[4] which control the quantizers shown in Fig. 3. At a given quality, making the quantizer QP finer (and the quantizer

[2]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 H.264 as a frame, encoded as one slice of this type.

[3]H.264/AVC Reference Software. [Online] Available: http://iphome.hhi.de/suehring/tml/download/

[4]These quantization control parameters are also denoted by QPSP and QPSP2 in the reference software implementation.

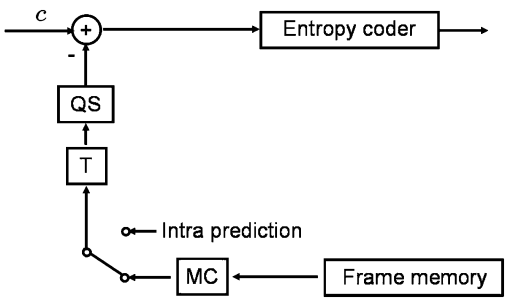

Fig. 5. H.264 SI and switching SP encoder.

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 modeled. Our analysis follows the model described in [9] 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 oversimplification, this model has been used in the literature to model the rate-distortion performance of image or video encoders (see, e.g., [9]–[11]). 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 defined as a *picture decoder*. The analysis presented in the following is based on the following result, obtained from [12]. The rate-distortion function of a stationary two-dimensional (2-D) zero-mean Gaussian signal a is expressed

$$R_a = \frac{1}{8\pi^2} \iint_{\Lambda} \max\left(0, \log_2\left(\frac{\Phi_{aa}(\Lambda)}{\theta}\right)\right) d\Lambda \text{ bit} \quad (1)$$

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

In (1) and (2), θ is a parameter that takes on all positive values to generate the rate-distortion curve. We denote by *ideal picture encoder* a picture encoder that 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 encoder shown in Fig. 3. This model is obtained by neglecting 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 rearranging the transforms, entropy coding,

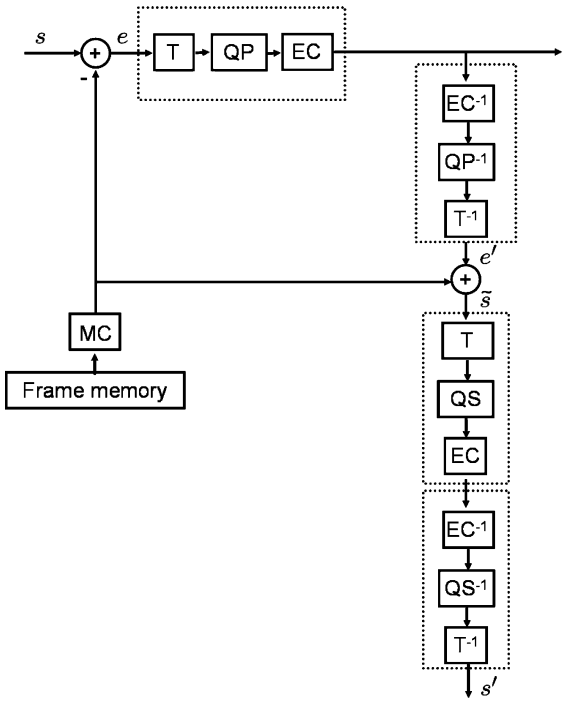

Fig. 6. Model of a primary SP encoder. Entropy coding is denoted by EC.

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 that the encoders are ideal.

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

$$R_{\mathrm{SP}_1} = \frac{1}{8\pi^2} \iint_{\Lambda} \max\left(0, \log_2\left(\frac{\Phi_{ee}(\Lambda)}{\theta_1}\right)\right) d\Lambda \text{ bit}. \quad (3)$$

This expression is identical to that given in [9] for P pictures, which is not surprising as the signal e, in the model of the SP encoder, is identical to the signal that would be obtained when encoding a P picture with an ideal 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, as follows:

$$D_{\mathrm{SP}_1} = D_1 + D_2 \quad (4)$$

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

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

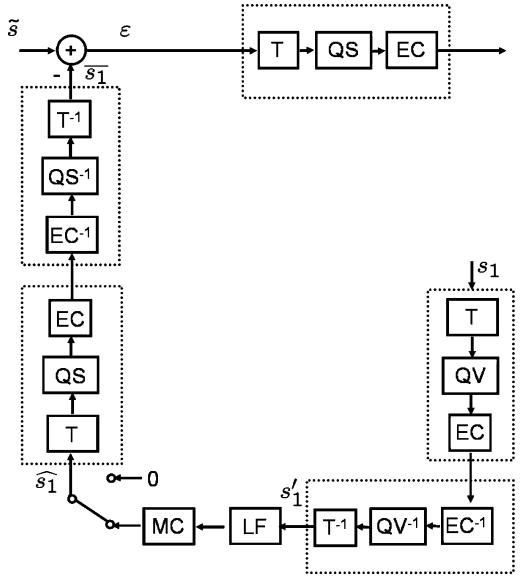

Fig. 7. Model of an SI and switching SP encoder. QV represents a quantizer.

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

B. RD Analysis of SI Pictures

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

$$D_{\mathrm{SI}} = D_{\mathrm{SP}_1}. \tag{7}$$

As stated in Section II, the signal c is compressed 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

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

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 the signal s_1 after this picture is encoded and decoded. 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 there is no prediction before the encoding of s_1.

The rest of the diagram in Fig. 7 can be obtained from the diagrams in Figs. 3 and 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. This requires assuming some quantizers are uniform (without deadzone). 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

$$D_{\mathrm{SP}_2} = D_{\mathrm{SI}} = D_{\mathrm{SP}_1}. \tag{9}$$

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

$$R_{\mathrm{SP}_2} = \frac{1}{8\pi^2}\iint_{\Lambda} \max\left(0, \log_2\left(\frac{\Phi_{\varepsilon\varepsilon}(\Lambda)}{\theta_2}\right)\right) d\Lambda \text{ bit.} \tag{10}$$

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

Fig. 8 shows the rate-distortion performance of SP and SI frames according to (3)–(10). The distortion is represented, in decibels, by its signal-to-noise ratio (SNR). As a reference, the rate-distortion curves of I and P frames, calculated according to [9], are also represented. At high rates, as expected, the slope of all the curves is equal to 6 dB/b, which represents 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 [9]. 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 tradeoff between the rate-distortion efficiency of nonswitching 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 Section 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-

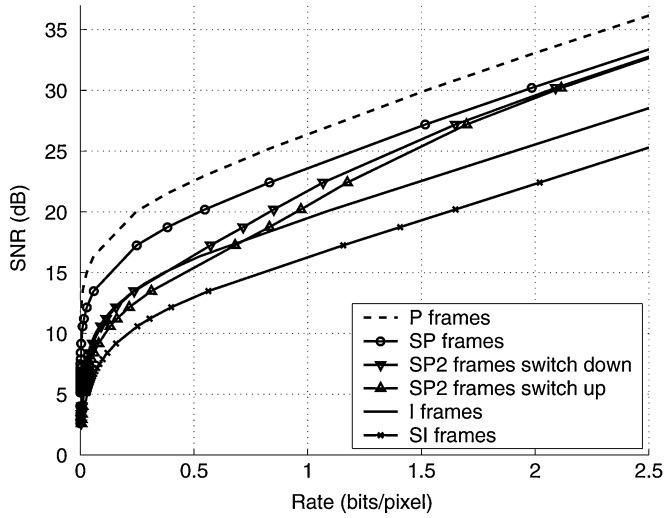

Fig. 8. Theoretical rate-distortion performance.

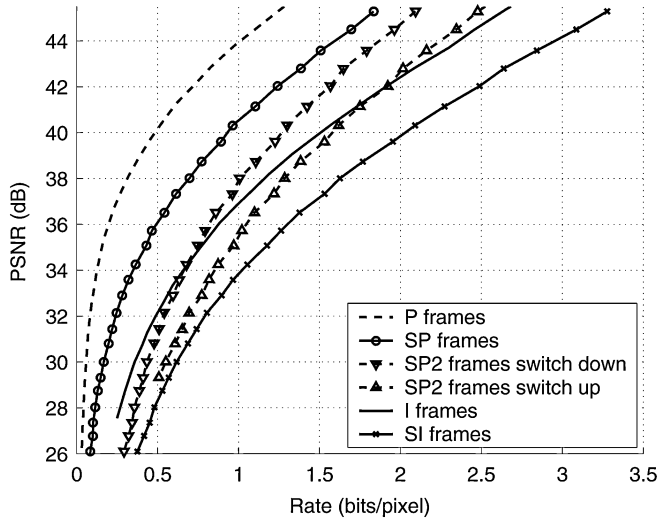

Fig. 9. Experimental rate-distortion performance.

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.

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.[5] 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 H.264 SP encoder. 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. We will analyze the setting of these parameters in the next section. For the results shown in Fig. 9, QP and QS are set to minimize the size of SI frames. As illustrated, SP frames are typically larger than P frames by approximately 90%, in this case. 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. Secondary SP frames rate-distortion performance is worse than that of I frames at low rates, and better at high rates.

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

IV. Optimal Setting for Streaming

In this section, the model is used to find how to set the SP encoder to minimize the expected bit rate when SP and SI frames are used for streaming with packet losses, as in the scenario depicted in Fig. 1. At each SP position, an SI picture can be sent instead of a primary SP picture to stop potential error propagation. One expects this to result in bit-rate savings compared with 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

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

In our model, R_{SI} and $R_{\mathrm{SP_1}}$ 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:

$$\text{Minimize} \quad \mathcal{R} \tag{12}$$

$$\text{such that} \quad D_{\mathrm{SP_1}} = D_{\mathrm{SI}} = D. \tag{13}$$

The equality constraint (13) sets the quality of SP and SI pictures equal to the quality of the rest of the encoded stream. In particular, D can be expressed as the distortion of P pictures and be written as a function of Φ_{ee} and of a positive parameter denoted by θ_{ref}, as follows:

$$D = \frac{1}{4\pi^2} \iint_{\Lambda} \min(\theta_{\mathrm{ref}}, \Phi_{ee}(\Lambda)) d\Lambda. \tag{14}$$

This expression is taken from [9]. At high rates, $\Phi_{ee}(\Lambda) \gg \theta_1$, $\Phi_{ee}(\Lambda) \gg \theta_{\mathrm{ref}}$, and $\Phi_{ss}(\Lambda) \gg \theta_2$, this simplifies the expression of $R_{\mathrm{SI}}, R_{\mathrm{SP_1}}$, and D and reduces (12) and (13) to

$$\text{Minimize} \quad (x-1)\log(\theta_1) - x\log(\theta_2) \tag{15}$$

$$\text{such that} \quad \theta_1 + \theta_2 = \theta_{\mathrm{ref}}. \tag{16}$$

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

$$\theta_2^* = \frac{x}{1-x}\theta_1^*. \tag{17}$$

In the following, we use (17) to derive the optimal setting of the two quantization control parameters QP and QS in the SP/SI

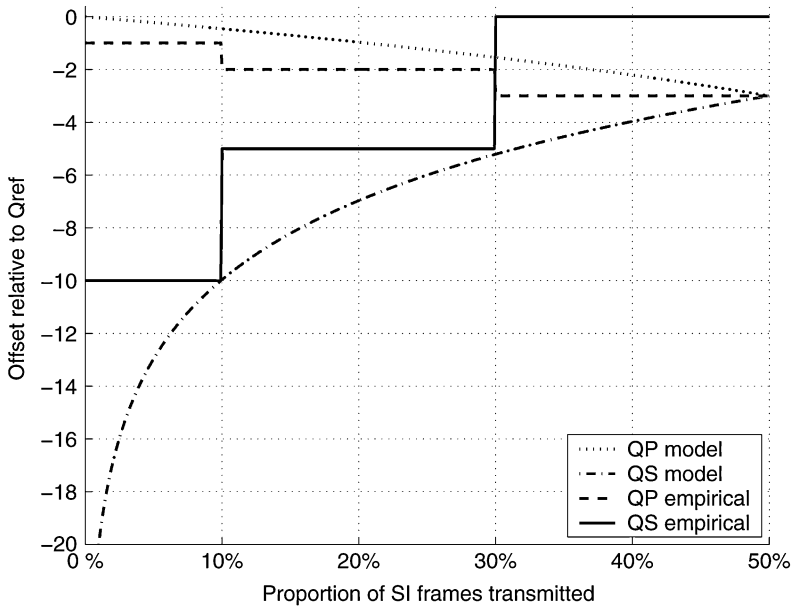

Fig. 10. Theoretical and empirical optimal settings of the two quantization parameters QP and QS.

encoder as a function of the parameter Q_{ref} used to control the quantization of P pictures.

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

$$\frac{dD}{dR} = -2\log(2)\Theta \tag{18}$$

where Θ is the parameter used to generate the rate-distortion curve [13]. Equation (18) can be used to express, as a function of θ_1, the slope of the distortion-rate function of the error signal e, represented in Fig. 6, denoted by $(dD_e)/(dR_e)$

$$\frac{dD_e}{dR_e} = -2\log(2)\theta_1. \tag{19}$$

Likewise, (18) can be used to express, as a function of θ_2, the slope of the distortion-rate function of $\tilde{s}$ denoted by $(dD_{\tilde{s}})/(dR_{\tilde{s}})$

$$\frac{dD_{\tilde{s}}}{dR_{\tilde{s}}} = -2\log(2)\theta_2. \tag{20}$$

In [14], the slope of distortion-rate performance of the H.264 encoder is expressed, empirically, as a function of the quantization parameter. Therefore, if we consider encoding e and $\tilde{s}$ with H.264, $(dD_e)/(dR_e)$ and $(dD_{\tilde{s}})/(dR_{\tilde{s}})$ are given by

$$\frac{dD_e}{dR_e}(H.264) = -0.85\,2^{\frac{\mathrm{QP}-12}{3}} \tag{21}$$

$$\frac{dD_{\tilde{s}}}{dR_{\tilde{s}}}(H.264) = -0.85\,2^{\frac{\mathrm{QS}-12}{3}}. \tag{22}$$

If we assume that H.264 approaches ideal rate-distortion performance, then the expressions in (19) and (21) are equal and so

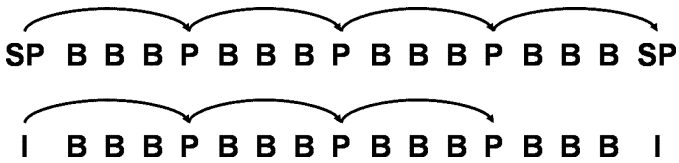

Fig. 11. Encoding structures used for streaming with SP and SI frames and for periodic I frame insertion.

are (20) and (22). The parameters θ_{ref} and Q_{ref}, which control the encoding of P pictures, can also be related using the same argument

$$2\log(2)\theta_{\mathrm{ref}} = 0.85\,2^{\frac{Q_{\mathrm{ref}}-12}{3}}. \tag{23}$$

The optimal setting QP^* and QS^* is represented in Fig. 10 and is obtained by combining (18)–(23), as follows:

$$\mathrm{QP} = Q_{\mathrm{ref}} + 3\log_2(1-x) \tag{24}$$

$$\mathrm{QS} = Q_{\mathrm{ref}} + 3\log_2(x). \tag{25}$$

As expected, QP^* and QS^* are finer than Q_{ref}, and QS^* is an increasing function of the proportion of SI pictures. Furthermore, the setting does not depend on the encoding rate. In Fig. 10, the settings derived from the model are compared with the optimal empirical settings. These were obtained by determining which settings minimized the expected bit rate for varying proportions of SI pictures. The results were consistent for six different sequences encoded at different bit rates. As illustrated in Fig. 10, the empirical settings follow the trend predicted by the model. Differences between the model and the experimental results are due to the simplifying assumptions that were made in the derivation and to the fact that H.264 restricts the three quantization parameters QP, QS, and Q_{ref} to integers.

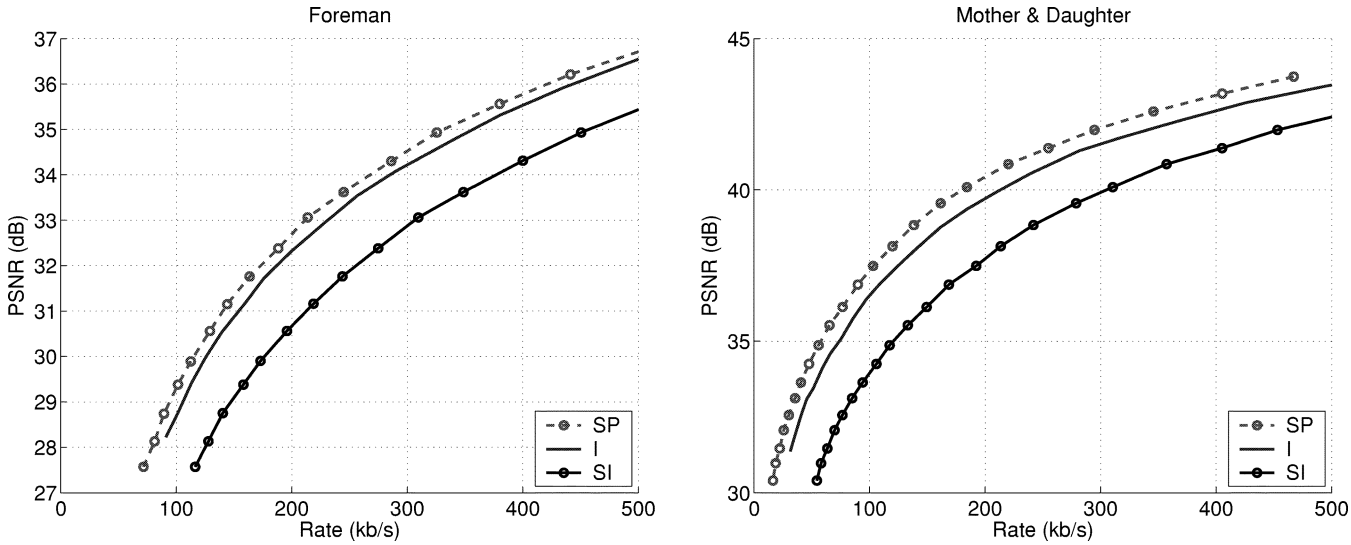

Fig. 12. Rate-distortion performance with periodic I frame, SP frame, or SI frame insertion for the sequences *Mother & Daughter* (left) and *Foreman* (right).

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. 11, 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 B frames to use as reference their two neighboring P frames or SP frames.[6] This ensures good error resilience properties and allows us to easily scale down the frame rate by 2 or even 4 if needed. For SP frames, we used the intermediate encoder settings shown in Fig. 10. The encoded video sequences used in the following experiments are made publicly available.[7]

Fig. 12 illustrates the rate-distortion characteristic of the CIF video sequences *Mother & Daughter* and *Foreman*, compressed using this coding structure. 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 larger amount of motion in the sequence *Foreman* reduces the coding efficiency of P frames and SP frames. As a consequence, the bit-rate savings expected by sending intra-coded frames on an as-needed basis will be higher for the sequence *Mother & Daughter*. For the first sequence, transmitting SP frames instead of I frames can lead to a performance gain of 1 dB at low rates and 0.6 dB at higher rates. For the second sequence, this gap is smaller and ranges from 1 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 2.5 dB at low rates and a little less than 1.5 dB at high rates.

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

[7]Encoded sequences with SP/SI frames. [Online] Available: http://ivms.stanford.edn/~eset-ton/'sequences.htm

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.

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 to 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 Maximum Transmission Unit size packet, whereas SP frames may necessitate two 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. 11. 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

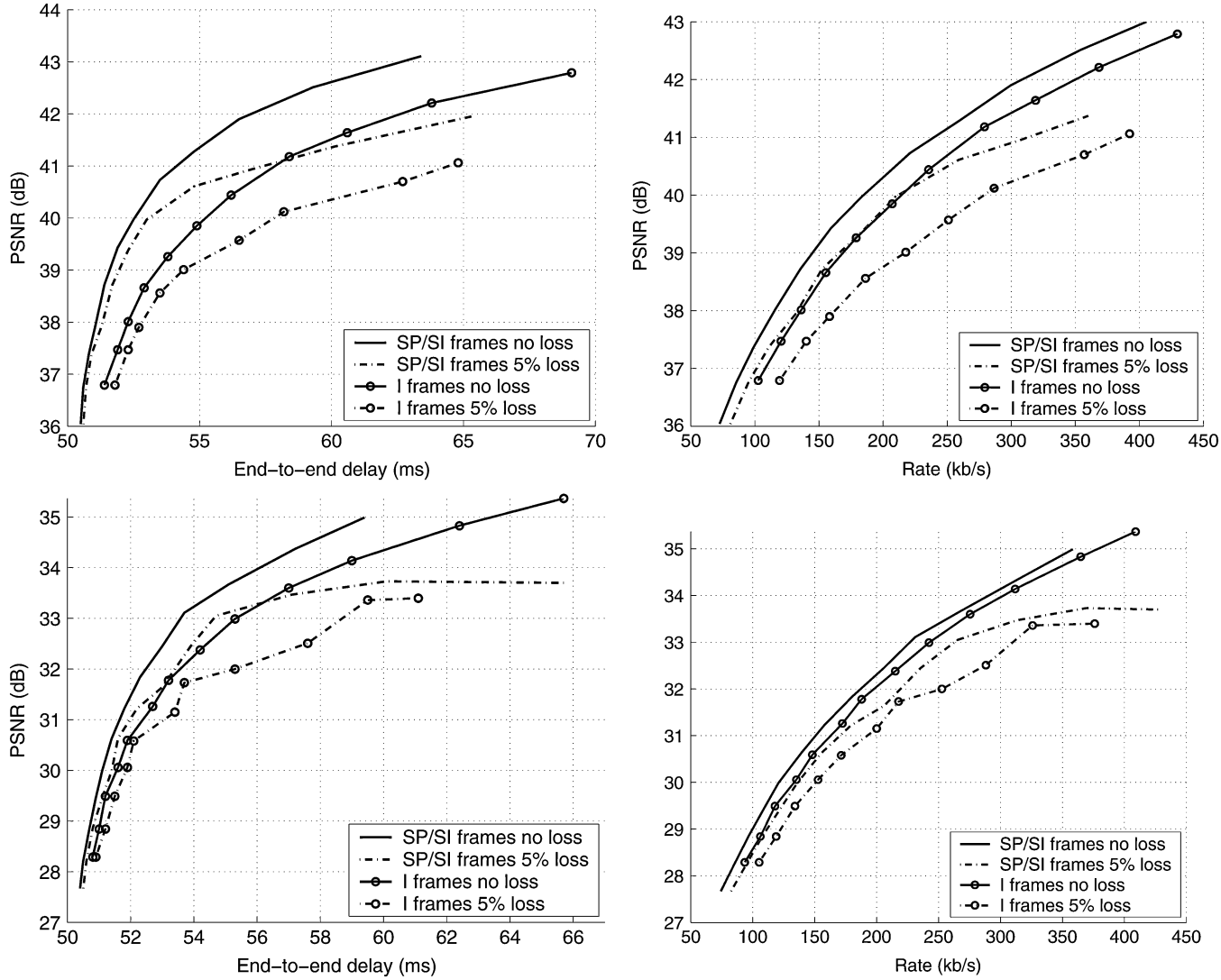

Fig. 13. Congestion-rate-distortion performance for *Mother & Daughter* (top) and *Foreman* (bottom) for varying loss rates.

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.

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. 13, 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 s, 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

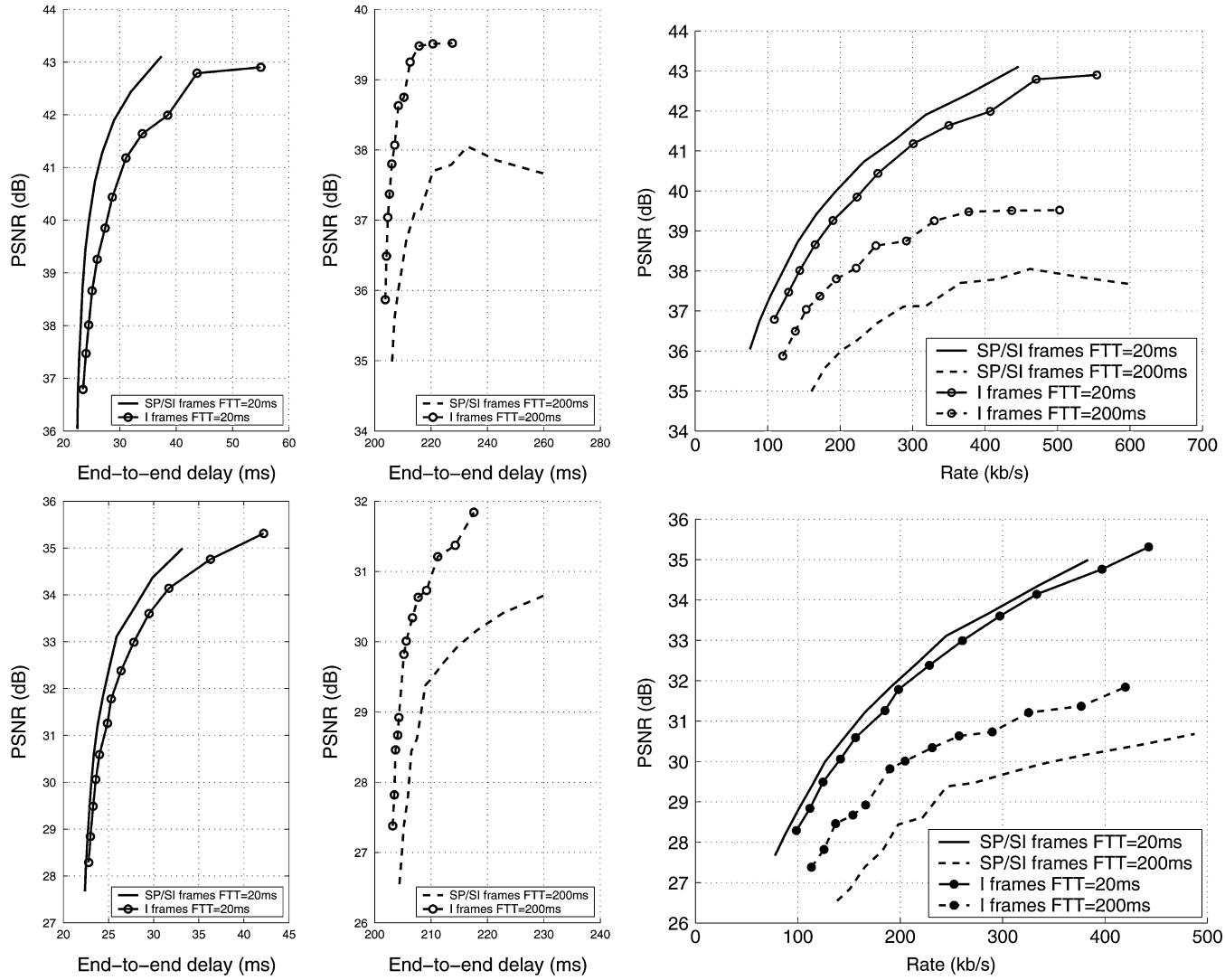

Fig. 14. Congestion-rate-distortion performance for *Mother & Daughter* (top) and *Foreman* (bottom) for varying propagation delays.

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 and 200 ms. This delay occurs on the high-bandwidth links and reflects the time needed for signal to propagate along links that 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. 14. For short propagation delays, the performance is only slightly worse than the performance in the absence of loss, discussed in Section 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 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 with 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 that predicts the rate-distortion performance of nonswitching SP frames, SI frames, and switching SP frames. Experimental results, obtained with our implementation of an SP encoder, based on H.264 and made publicly available, validate the theoretical results. The model predicts the relative performance of SP, SI, and switching SP pictures compared with 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 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

$$\varepsilon = \tilde{s} - \overline{s_1} \tag{26}$$

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 two-dimensional (2-D) bandlimited discrete space Fourier transform (FT) of a signal a, and $*$ denotes 2-D convolution. In this section, for simplicity, we consider continuous signals, which are obtained by interpolating the discrete signals they correspond to with a sinc function.

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

$$s_1(x,y) = s(x,y) * \delta(x - d_x, y - d_y) \tag{27}$$

where (d_x, d_y) is the displacement. The transformation between s_1 and s_1' is the optimum forward channel defined in [12] and used in [9], or more recently in [10] 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 nonwhite 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, as follows:

$$\Phi_{s_1's_1'}(\Lambda) = \Phi_{s_1s_1}(\Lambda)|G(\Lambda)|^2 + \Phi_{nn}(\Lambda). \tag{28}$$

The expression of G and Φ_{nn} are given in [12]. As s and s_1, as a consequence of (24), have the same PSD, G and Φ_{nn} are expressed

$$G(\Lambda) = \max\left(0, 1 - \frac{\theta_3}{\Phi_{ss}(\Lambda)}\right) \tag{29}$$

$$\Phi_{nn}(\Lambda) = \max\left(0, \theta_3\left(1 - \frac{\theta_3}{\Phi_{ss}(\Lambda)}\right)\right). \tag{30}$$

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

$$\widehat{s_1} = (s_1' * f) * \delta(x - \widehat{d_x}, y - \widehat{d_y}) \tag{31}$$

$$\Phi_{\widehat{s_1}\widehat{s_1}}(\Lambda) = \Phi_{s_1's_1'}(\Lambda)|F(\Lambda)|^2 \tag{32}$$

$$\Phi_{\widehat{s_1}\widehat{s_1}}(\Lambda) = \Phi_{ss}(\Lambda)|G(\Lambda)F(\Lambda)|^2 + \Phi_{nn}|F(\Lambda)|^2. \tag{33}$$

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 nonwhite noise $\widehat{n}$

$$\overline{s_1} = \widehat{s_1} * \widehat{g} + \widehat{n} \tag{34}$$

$$\Phi_{\overline{s_1}\overline{s_1}}(\Lambda) = \Phi_{\widehat{s_1}\widehat{s_1}}(\Lambda)|\widehat{G}(\Lambda)|^2 + \Phi_{\widehat{n}\widehat{n}}(\Lambda). \tag{35}$$

We use, once again, the expression of the optimum forward channel given in [12], to write the FT of the filter and the PSD of the noise

$$\widehat{G}(\Lambda) = \max\left(0, 1 - \frac{\theta_2}{\Phi_{\widehat{s_1}\widehat{s_1}}(\Lambda)}\right) \tag{36}$$

$$\Phi_{\widehat{n}\widehat{n}}(\Lambda) = \max\left(0, \theta_2\left(1 - \frac{\theta_2}{\Phi_{\widehat{s_1}\widehat{s_1}}(\Lambda)}\right)\right). \tag{37}$$

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:

$$\varepsilon = \tilde{s} - \overline{s_1} \tag{38}$$

$$\varepsilon = s - \overline{s_1} + n_s \tag{39}$$

$$\varepsilon = s - (((((s * \delta(x - d_x, y - d_y)) * g + n) * f) * \delta(x - \widehat{d_x}, y - \widehat{d_y})) * \widehat{g}) - \widehat{n} + n_s. \tag{40}$$

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 with 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

$$\Phi_{\varepsilon\varepsilon}(\Lambda) = \Phi_{ss}(\Lambda)E[|1 - FG\widehat{G}e^{(j(\omega_x\Delta d_x + \omega_y \Delta d_y))}|^2] + \Phi_{nn}(\Lambda)|F(\Lambda)\widehat{G}(\Lambda)|^2 + \Phi_{\widehat{n}\widehat{n}}(\Lambda) \quad (41)$$

where E is the expectation function taken with respect to the probability density function of the displacement error. Following the simplification derived in [9], (38) can be rewritten

$$\Phi_{\varepsilon\varepsilon}(\Lambda) = \Phi_{ss}(\Lambda)(1 + |F(\Lambda)G(\Lambda)\widehat{G}(\Lambda)|^2 - 2\mathrm{Re}\{F(\Lambda)G(\Lambda)\widehat{G}(\Lambda)P(\Lambda)\}) + \Phi_{nn}(\Lambda)|F(\Lambda)\widehat{G}(\Lambda)|^2 + \Phi_{\widehat{n}\widehat{n}}(\Lambda) \quad (42)$$

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

ACKNOWLEDGMENT

The authors would like to thank Dr. P. Ramanathan and Dr. M. Flierl for insightful discussions and the anonymous reviewers for their helpful comments to 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)* (in 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," in *Proc. Int. Conf. Image Processing (IC1P)*, Santa Barbara, CA, Oct. 1997, vol. 2, pp. 73–76.

[3] M. Karczewicz and R. Kurceren, "A proposal for SP-frames," in *Video Coding Experts Group Meeting, Doc. VCEG-L-27*, Eibsee, Germany, Jan. 2001.

[4] M. Karczewicz and R. Kurceren, "The SP- and SI-frames design for H.264/AVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 637–644, Jul. 2003.

[5] E. Setton, P. Ramanathan, and B. Girod, "Rate-distortion analysis of SP and SI frames," presented at the Visual Communications Image Processing (VCIP), San Jose, CA, Jan. 2006, unpublished.

[6] E. Setton and B. Girod, "Video streaming with SP and SI frames," presented at the Visual Communications and Image Processing (VCIP), Beijing, China, Jul. 2005, unpublished.

[7] X. Sun, S. Li, F. Wu, J. Shen, and W. Gao, "The improved SP frame coding technique for the JVT standard," in *Proc. Int. Conf. Image Processing*, Barcelona, Spain, Sep. 2003, vol. 3, pp. 297–300.

[8] X. Sun, F. Wu, S. Li, and R. Kurceren, "The improved JVT-B097 SP coding scheme," in *ITU-T SG16 Q6mJVT-C114*, Fairfax, VA, May 2002.

[9] B. Girod, "The efficiency of motion-compensating prediction for hybrid coding of video sequences," *IEEE J. Sel. Areas Commun.*, vol. 5, no. 7, pp. 1140–1154, Aug. 1987.

[10] G. Cook, J. Prades-Nebot, and E. Delp, "Rate-distortion bounds for motion-compensated rate scalable video coders," in *Proc. Int. Conf. Image Processing (ICIP)*, Oct. 2004, pp. 3121–3124.

[11] M. H. Flierl, "Video coding with superimposed motion-compensated signals," Ph.D. dissertation, Dept. of Elect. Eng., Univ. of Erlangen, Erlangen, Germany, 2003.

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

[13] R. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.

[14] T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, and G. Sullivan, "Rate-constrained coder control and comparison of video coding standards," *IEEE Trans. Circ. Syst. Video Technol.*, vol. 13, no. 7, pp. 688–703, Jul. 2003.

Eric Setton (S'03) received the B.S. degree from Ecole Polytechnique, Palaiseau, France, in 2001 and the M.S. degree in electrical engineering from Stanford University, Stanford, CA, in 2003. He is currently working toward the Ph.D. degree in the Department of Electrical Engineering of Stanford University and is part of the Image, Video and Multimedia Systems group.

He has spent time in industry in France at SAGEM and in the United States at HP laboratories and at Sony Electronics. He has six patents pending. His research interests range from multimedia communication over peer-to-peer networks, to streaming over wired and wireless networks and video compression.

Mr. Setton received the Carnot fellowship and the SAP Stanford Graduate fellowship in 2001. In 2003, he received the Sony SNRC fellowship.

Bernd Girod (S'80–M'80–SM'97–F'98) received the M. S. degree in electrical engineering from Georgia Institute of Technology (Georgia Tech), Atlanta, in 1980 and the Ph.D. degree (with highest honors) from the University of Hannover, Germany, in 1987.

Until 1987, he was a member of the research staff at the Institut für Theoretische Nachrichtentechnik und Informationsverarbeitung, University of Hannover. In 1988, he joined the Massachusetts Institute of Technology, Cambridge, first as a Visiting Scientist with the Research Laboratory of Electronics, then as an Assistant Professor of Media Technology at the Media Laboratory. From 1990 to 1993, he was Professor of Computer Graphics and Technical Director of the Academy of Media Arts, Cologne, Germany, jointly appointed with the Computer Science Section of Cologne University. He was a Visiting Adjunct Professor with the Digital Signal Processing Group at Georgia Tech in 1993. From 1993 until 1999, he was Chaired Professor of Electrical Engineering/Telecommunications at University of Erlangen-Nuremberg, Germany, and the Head of the Telecommunications Institute I, codirecting the Telecommunications Laboratory. He has served as the Chairman of the Electrical Engineering Department from 1995 to 1997 and as Director of the Center of Excellence "3-D Image Analysis and Synthesis" from 1995 to 1999. He has been a Visiting Professor with the Information Systems Laboratory of Stanford University, Stanford, CA, during the 1997–1998 academic year. Currently, he is Professor of Electrical Engineering in the Information Systems Laboratory of Stanford University. He also holds a courtesy appointment with Stanford University's Department of Computer Science. He serves as Director both of the Stanford Center for Image Systems Engineering (SCIEN) and the Max Planck Center for Visual Computing and Communication. Since 2004, he has also served as Chairman of the Steering Committee of the new Deutsche Telekom Laboratories at the Technical University of Berlin, Germany. As an entrepreneur, he has worked successfully with several start-up ventures as founder, investor, director, or advisor. Most notably, he has been a cofounder and Chief Scientist of Vivo Software, Inc., Waltham, MA from 1993 to 1998; after Vivo's aquisition, from 1998 to 2002, Chief Scientist of RealNetworks, Inc. He has served on the Board of Directors for 8 × 8, Inc., Santa Clara, CA, from 1996 to 2004 and for GeoVantage, Inc., Swampscott, MA, from 2000 to 2005. He is currently an advisor to start-up companies Mobilygen, Santa Clara, CA, and to NetEnrich, Inc., Santa Clara, CA. He has authored or coauthored one major textbook (printed in four languages), three monographs, and over 350 book chapters, journal articles, and conference papers in his field, and he holds over 20 U.S. patents. His current research interests include video coding and networked media systems.

Prof. Girod has served on numerous conference committees, e.g., as Tutorial Chair of the International Conference on Acoustics, Speech and Signal Processing (ICASSP) 1997 in Munich and the International Conference on Image Processing (ICIP) 2000, Vancouver, Canada; as General Chair of the 1998 IEEE Image and Multidimensional Signal Processing Workshop in Alpbach, Austria; as General Chair of the Visual Communication and Image Processing Conference (VCIP), San Jose, CA, in 2001; and General Chair of Vision, Modeling, and Visualization (VMV), Stanford, CA, in 2004. He has served on the Editorial Boards or as Associate Editor for several journals in his field, among them as Area Editor for Speech, Image, Video & Signal Processing of the IEEE TRANSACTIONS ON COMMUNICATIONS. He has been a member of the IEEE Image and Multidimensional Signal Processing Committee from 1989 to 1997 and was elected Fellow of the IEEE in 1998 "for his contributions to the theory and practice of video communications." He has been named Distinguished Lecturer for the year 2002 by the IEEE Signal Processing Society. He is recipient of the 2002 EURASIP Best Paper Award (with J. Eggers) and the 2004 EURASIP Technical Achievement Award.