ANALYSIS OF THE EFFICIENCY OF SNR-SCALABLE STRATEGIES FOR MOTION COMPENSATED VIDEO CODERS

Josep Prades-Nebot †, *Gregory W. Cook* ‡ *and Edward J. Delp*‡

† Departamento de Comunicaciones
Universidad Politécnica de Valencia
Valencia 46071, SPAIN
jprades@dcom.upv.es

‡ Video and Image Processing Laboratory (VIPER)
Purdue University
West Lafayette, IN 47907-1285, USA
cook@ieee.org, ace@purdue.ecn.edu

ABSTRACT

In this paper, an analysis of the efficiency of three signal-to-noise ratio (SNR) scalable strategies for motion compensated video coders and their non-scalable counterpart is presented. After assuming some models and hypotheses with respect to the signals and systems involved, we have obtained the SNR of each coding strategy as a function of the decoding rate. To validate our analysis, we have compared our theoretical results with data from encodings of real video sequences. Results show that our analysis describes qualitatively the performance of each scalable strategy, and therefore, it can be useful to understand main features of each scalable technique and what factors influence their efficiency.

1. INTRODUCTION

Scalable video can be decoded at two or more different bit-rates each corresponding to a different level of quality. Although scalability is a desirable property when video has to be transmitted in channels with errors and bandwidth fluctuations, scalable video coders are not commonly being used in practice. One of the reasons is that all scalable coders are lower in efficiency than their non-scalable (NS) counterparts [1, 2, 3, 4, 5]. Consequently, it is important to know main features of each scalable technique and what factors influence their efficiency. In this paper, we present a theoretical study of the efficiency of three signal-to-noise (SNR) scalable strategies used in video coders with single-loop motion compensated prediction (MCP).

Figure 1 shows the scheme of a SNR-scalable MCP-based video coder. At the transmitter, the predicted error frames (PEF) represented by signal e are encoded at a rate R_e to generate the bit-stream, and decoded at the *loop rate* R_l to provide signal e' to the MCP loop. At the decoder, the bit-stream is decoded at R_l' (for the MCP loop) and at the *decoding rate* R. Depending on the values of these four rates (R_e, R_l, R_l', R) we have different coding strategies. If $R_e = R_l = R_l' = R$, then we have a NS coder, which sets the maximum performance for scalable coders. In all the SNR-scalable strategies: $R_e = R_{\max}$ and the decoding rate can vary between the minimum and the maximum rate of the service ($R_{\min} \leq R \leq R_{\max}$). In *Scalable encodings Below the Loop Rate* (SBLR), $R_l = R_{\max}$ and $R = R_l'$. This is the encoding strategy proposed in the SNR-scalable MPEG-2 standard [1]. As the transmitter and the receiver have different reference frames, prediction drift is introduced (unless $R = R_{\max}$) which reduces the efficiency. In *Scalable Encodings Above the Loop Rate* (SALR), prediction drift is avoided by setting $R_l = R_l' = R_{\min}$. This is the scalable strategy used in the fine granular scalability (FGS) profile of the MPEG-4 standard [2]. In a SALR coder, the reference frames s' are decoded at $R_{\min}$ which limits the quality of the prediction, and therefore, the efficiency of the coder.

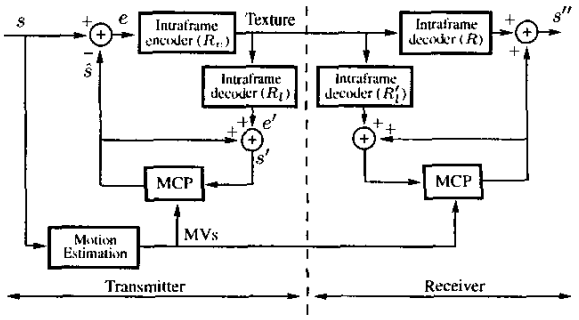

Fig. 1. Scheme of a SNR-scalable MCP-based video coder.

To improve their efficiency, some coders set R_l between $R_{\min}$ and $R_{\max}$ and allow decoding both above and below R_l [3, 4, 5]. In the following, we call this type *Scalable encoding Above and Below the Loop Rate* (SABLR). In [6], these three scalable schemes were studied considering one dimensional signals and linear prediction. In this paper, we have extended the study in [6] to video signals and motion compensated coders.

In our theoretical analysis we make some assumptions about the signals and systems involved. With respect to the intra-frame encoding, we assume that embedded quantization is used and the quantization noise q is modeled as an additive white noise with variance

$$\sigma_q^2 = \sigma_e^2 \, 2^{-\beta R}, \tag{1}$$

where σ_e^2 is the power of the PEF, β is a parameter that measures the efficiency of the of the intra-frame coding, and R is the intra-frame encoding rate [7]. We also assume that q and e are uncorrelated.

The rest of hypotheses are similar to the ones assumed in [8, 9]. With respect to the input video signal s, we assume that its frames constitutes a stationary random field. We also assume that

This work has been supported by a grant for the Secretaría de Estado de Educación y Universidades of the Spanish Government, by the program CICYT TIC-2002-02469, and by an Indiana Twenty-First Century Research and Technology Fund grant.

0-7803-8554-3/04/$20.00 ©2004 IEEE.

the only difference between consecutive frames is a constant-in-time and uniform-in-space displacement (d_x,d_y). Although these hypothesis are not accurate in real encodings (MVs change in time and space, motion can be non-translatory, at low rates q is not white and is correlated with e), our analysis can still be useful to study the relative performance of every scalable strategy.

In our analysis, we ignore the bits necessary to encode motion vectors (MVs). In practice, this does not introduce significant differences in analyzing the relative performance of each scalable strategy, if the number of bits aimed to encode MVs are approximately the same at all rates and is low compared to the number of bits used to encode PEF texture.

In the following, x and y are the spatial variables, and t is the temporal variable of the video sequence. Their corresponding frequency variables are ω_x, ω_y and ω_t respectively, although for simplicity, $\Lambda = (\omega_x, \omega_y)$ and $\Omega = (\omega_x, \omega_y, \omega_t)$ are used sometimes. The predictor is modeled as a random linear time-invariant system whose frequency response is

$$H(\omega_x, \omega_y, \omega_t) = F(\omega_x, \omega_y)\, e^{-j(\omega_x \hat{d}_x + \omega_y \hat{d}_y + \omega_t)} \qquad (2)$$

where $F(\omega_x, \omega_y)$ is the frequency response of the spatial filtering performed in the MCP loop and $(\hat{d}_x,\hat{d}_x)$ is the estimated (random) displacement vector. In general, there is a displacement error vector $\Delta d = (\Delta d_x, \Delta d_y)$

$$(\Delta d_x, \Delta d_y) = (d_x, d_y) - (\hat{d}_x, \hat{d}_y). \qquad (3)$$

2. ANALYSIS OF THE NON-SCALABLE CODER

The block diagram of a non-scalable MCP-based video coder is shown in Figure 2. Notice that the reconstruction error $r = s'' - s$ is equal to the quantization noise q, and thus $\sigma_q^2 = \sigma_r^2$.

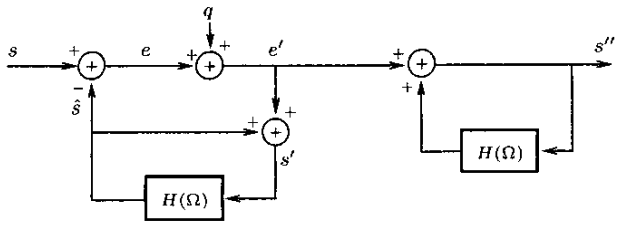

Fig. 2. Block diagram of the non-scalable coder.

The power spectral density (PSD) of the error frames is [9]:

$$\begin{aligned} S_{ee}(\Lambda) &= S_{ss}(\Lambda)\,\left[1 - 2\,\mathrm{Re}\left\{F^*(\Lambda)\,P(\Lambda)\right\} + |F(\Lambda)|^2\right] \\ &+ |F(\Lambda)|^2\, S_{qq}(\Lambda) \end{aligned} \qquad (4)$$

where $S_{ss}(\Lambda)$ and $S_{qq}(\Lambda)$ are the PSD of the input frames and the quantization noise respectively, $\mathrm{Re}\{\cdot\}$ denotes "real part", and $P(\Lambda)$ is the 2-D Fourier Transform of the probability density function $p_{\Delta d}(\Delta d)$. Then, the power of e is

$$\sigma_e^2 = E_s + \sigma_q^2\, E_f \qquad (5)$$

where E_s is

$$E_s = \frac{1}{4\pi^2} \iint_D S_{ss}(\Lambda)\,\left[1 - 2\,\mathrm{Re}\left\{F^*(\Lambda)\,P(\Lambda)\right\} + |F(\Lambda)|^2\right] d\Lambda,$$

where $D = \{\Lambda : |\omega_x| < \pi, |\omega_y| < \pi\}$, and E_f is

$$E_f = \frac{1}{4\pi^2} \iint_D |F(\Lambda)|^2\, d\Lambda. \qquad (6)$$

Finally, from (1) and (5), the SNR of the NS coder as a function of the decoding rate is

$$\mathrm{SNR}_{\mathrm{NS}}(R) = \frac{\sigma_s^2}{\sigma_r^2} = \frac{\sigma_s^2}{E_s}\left(2^{\beta R} - E_f\right). \qquad (7)$$

If R is large enough so that $2^{\beta R} \gg E_f$, then the SNR (in dB) of the NS coder is an affine function of R with slope 3β.

3. ANALYSIS OF THE SALR SCHEME

Figure 3 shows the block diagram of a SALR coder. The quantization noise q_b is generated by the encoding e at $R_{\max}$ and its further decoding at R_l. With respect to the quantization noise source q, is generated by encoding e at $R_{\max}$ and decoding it at R.

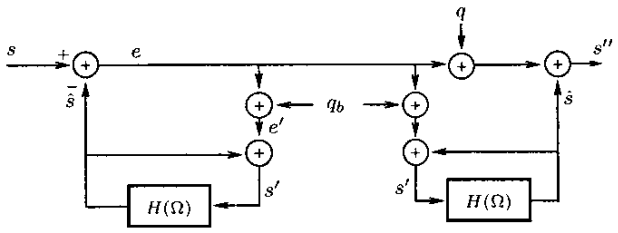

Fig. 3. Block diagram to compute the SNR of the SALR coder.

Similarly to the NS coder, $\sigma_r^2 = \sigma_q^2$, but now

$$\sigma_e^2 = E_s + \sigma_{q_b}^2\, E_f \qquad (8)$$

and the variance of q_b is

$$\sigma_{q_b}^2 = \sigma_e^2\, 2^{-\beta R_{\min}}. \qquad (9)$$

From (1), (8) and (9), the SNR of the SALR coder is

$$\mathrm{SNR}_{\mathrm{SALR}}(R) = \mathrm{SNR}_{\mathrm{NS}}(R_{\min})\, 2^{\beta(R - R_{\min})}. \qquad (10)$$

Notice there is no loss with respect to the NS coder at $R_{\min}$. Above this rate, the SNR (in dB) is an affine funtion of R with slope 3β.

4. ANALYSIS OF THE SBLR CODER

In a SBLR coder, two quantization noise sources must be taken into account (Figure 4). The first one (q_m) is placed in the transmitter and is the result of encoding and decoding the predicted error frames at $R_{\max}$. The second one (q) is placed in the receiver and is the result of decoding the compressed PEF at R.

In this case, the reconstruction error r is

$$r = q_m + \Delta q * h_d \qquad (11)$$

where $\Delta q = q - q_m$, h_d represents the end-to-end decoder transfer function, and $*$ is the convolution operator. We assume that $E\{q_m\, \Delta_q\} = 0$ and that Δq is white noise, which provides

$$\sigma_r^2 = \sigma_{q_m}^2 + \sigma_{\Delta_q}^2\, E_d \qquad (12)$$

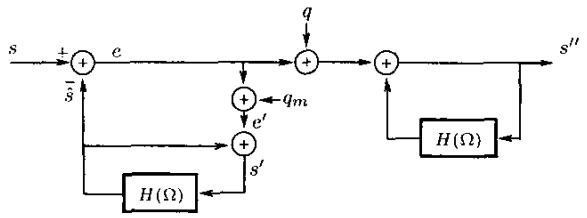

Fig. 4. Block diagram to compute the SNR of the SBLR coder.

where $\sigma^2_{q_m}$ and $\sigma^2_{\Delta_q}$ are the variances of q_m and Δ_q respectively, E_d is

$$E_d = \frac{1}{8\pi^3} E\left\{ \iiint_{D'} |1 - H(\Omega)|^{-2} \, d\Omega \right\} \qquad (13)$$

where $E\{\cdot\}$ is the expectation operator, and $D' = \{\Omega : |\omega_x| < \pi, |\omega_y| < \pi, |\omega_t| < \pi\}$. As $\sigma^2_{\Delta_q} = \sigma^2_q - \sigma^2_{q_m}$, Expression (12) transforms into

$$\begin{aligned} \sigma^2_r &= \sigma^2_{q_m} + \left(\sigma^2_q - \sigma^2_{q_m}\right) E_d \\ &= \sigma^2_{q_m} \left[1 + \left(2^{\beta(R_{\max}-R)} - 1\right) E_d\right]. \end{aligned} \qquad (14)$$

Finally, from (14) and $\sigma^2_{q_m} = E_s/(2^{\beta R_{\max}} - E_f)$, we obtain

$$\mathrm{SNR}_{\mathrm{SBLR}}(R) = \frac{\mathrm{SNR}_{\mathrm{NS}}(R_{\max})}{\left[1 + \left(2^{\beta(R_{\max}-R)} - 1\right) E_d\right]} \qquad (15)$$

The SBLR coder has no loss with respect the NS coder at $R_{\max}$. Below this rate, prediction drift is introduced. Note that if R is far below $R_{\max}$ so that $E_d \, 2^{\beta(R_{\max}-R)} \gg 1$, the SNR of the SBLR coder (in dB) is an affine function of R with slope 3β.

5. ANALYSIS OF SABLR CODER

In SABLR coders, according to the decoding rate R, we can distinguish two operating intervals:

- the *SBLR interval* ($R_{\min} \le R \le R_l$) where prediction drift is introduced. In this interval, the SABLR coder has a higher SNR than the SBLR coder.
- the *SALR interval* ($R_l \le R \le R_{\max}$) where there is a loss of performance with respect to the NS coder because the prediction is based on previous frames decoded at $R_{\min}$ instead of R. In this interval, the SABLR coder has a higher SNR than the SALR coder.

From Sections 3 and 4, the SNR for the SABLR coder is:

$$\mathrm{SNR}_{\mathrm{SABLR}}(R, R_l) = \begin{cases} \dfrac{\mathrm{SNR}_{\mathrm{NS}}(R_l)}{\left[1 + \left(2^{\beta(R_l - R)} - 1\right) E_d\right]}, & R \le R_l \\ \mathrm{SNR}_{\mathrm{NS}}(R_l)\, 2^{\beta(R - R_l)}, & R \ge R_l \end{cases} \qquad (16)$$

Notice that the SABLR coder has no loss with respect to its NS counterpart at R_l.

6. EXPERIMENTAL RESULTS

In this section, we compare our theoretical analysis with data from encodings of real video sequences using the MCP-based SNR-scalable *SAMCoW* video coder. As *SAMCoW* uses embedded quantization to encode the PEF [10], it can operate in any of the four coding modes (NS, SALR, SBLR and SABLR).

To obtain specific numerical simulation results, some parameters have to be set. With respect to the video signals, we assume s has an isotropic PSD

$$S_{ss}(\omega_x, \omega_y) = \frac{2\pi\sigma_s^2}{\omega_0^2} \left(1 + \frac{\omega_x^2 + \omega_y^2}{\omega_0^2}\right)^{-3/2} \qquad (17)$$

where σ_s^2 is the signal power and ω_0 has been set to provide an adjacent step correlation coefficient equal to 0.93 [9]. It is assumed that Δd follows a zero mean isotropic Gaussian distribution with $\sigma^2_{\Delta d} = 0.2\,T^2$ where T is the spatial sampling period. With respect to the coder, parameter β has been set to 3 and, although spatial filtering is not considered, we introduce a leaky factor equal to 0.95, and then $F(\Lambda) = 0.95$. The use of a leaky factor limits the effect of prediction drift in SBLR and SABLR coders. Practical coders usually introduce some implicit or explicit spatial filtering in the MCP loop which can be considered as a frequency-dependent leaky factor. The rate interval chosen is $R_{\min} = 0.066$ bits/pixel and $R_{\max} = 0.33$ bits/pixel which for CIF sequences at 30 frames/s is equivalent to $R_{\min} = 200$ kbits/s and $R_{\max} = 1000$ kbits/s.

Figure 5 shows the SNR(R) function of the NS, SALR, SBLR and SABLR coder for the set of parameters previously described. In the case of the SABLR coder three curves, corresponding to $R_l = 0.131$, 0.197 and 0.263 bits/pixel, have been plotted. These three rates correspond to 400, 600 and 800 kbits/s respectively, if CIF video sequences at 30 frames/s are used. In the SABLR curves, the R_l value is the rate at which the SABLR and the NS curve intersect. The portions of the three SABLR curves where $R > R_l$ are equivalent to the curves of a SALR coder using $R_{\min} = R_l$. Equivalently, the portions of the SABLR curves where $R < R_l$ can be considered SBLR curves with $R_{\max} = R_l$.

In the SALR intervals of the curves in Figure 5, notice that the larger $R_{\min}$ is, the lower the loss is with respect to the NS coder, but the interval of rates where decoding is possible is also lowered. In fact, if $R_{\min}$ is large enough so that $2^{\beta R_{\min}} \gg E_f$, the loss is insignificant. With respect to the SBLR intervals of the curves, the contrary effect in the SALR ones is noted: the loss decreases with a decrease in $R_{\max}$ (again, at the expense of reducing the interval of decoding rates). SABLR coders allow a balancing of both effects and by setting R_l properly, the mean SNR (MSNR) can be improved with respect to the SALR and the SBLR coders. For the encoding parameters of Figure 5, a maximum MSNR of 10.15 dB is achieved at $R_l = 0.162$ bits/pixel (or, equivalently, at 550.3 kbits/s with CIF sequences at 30 frames/s). With respect to the SALR and the SBLR coders, the MSNR are 8.86 dB and 8.33 dB respectively.

To test the efficiency of the strategies *in practice*, we have encoded several test CIF sequences (352 × 288 pixels/frame) at 30 frames/s with *SAMCoW*. The quality of each encoding is measured by computing the mean PSNR (in dB) of the luminance component of 100 decoded frames. As our theoretical analysis only accounts for the steady-state performance of coders, in every encoding an initial portion of each decoded sequence containing frames with transient response was not considered. Motion estimation is performed at integer-pixel accuracy with no loop filter and, as in theory, a leaky factor $c = 0.95$ is introduced. Figure 6 shows the SNR(R) function obtained by encoding *Foreman* with *SAMCoW* running in the four strategies. By comparing Figures 5 and 6, we

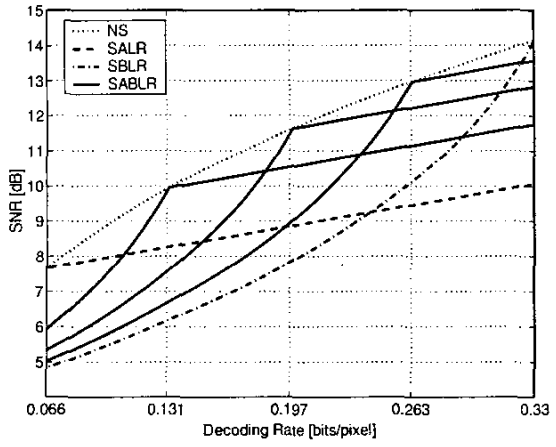

Fig. 5. Numerical simulation of the theoretical SNR(R) of the four video strategies using the assumptions outlined in Section 6.

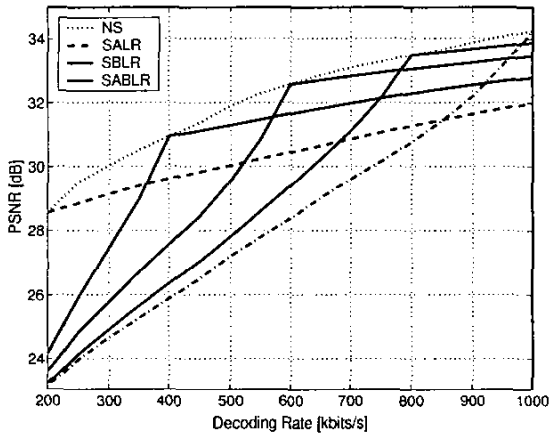

Fig. 6. PSNR(R) of the four video strategies using *SAMCoW*

can study the differences between theory and practice. No attempt of using similar parameters values (β, ω_0) in theory and practice has been made, and therefore, our comparison is qualitative.

With respect to the SALR intervals of the scalable strategies, while in theory all the SALR curves have the same slope, in practice the slope decreases when R_l increases. The reason is that, in practice β is not constant but depends on R_l: starting in $R_l = 0$, β decreases rapidly with increase in R_l, but tends to a constant value at high R_l. The consequence of this is that, in practice, the gain obtained by increasing the value of $R_{\min}$ is lower than the one obtained in theory.

With respect to the SBLR intervals of the scalable strategies, although theory and practice tend to be similar at high decoding rates, there is a great divergence at low decoding rates where the loss in practice is higher than the theoretical one. The reasons of this divergence is that, at low rates, some of our hypothesis do not hold (β changes largely with R and, Δq and q_m are correlated). We have checked that when rate intervals with higher $R_{\min}$ values are used, theory and practice are much closer. Differences between theory and practice in both the SALR and SBLR intervals, have two main consequences for the SABLR coder. First, R_l cannot be increased much above $R_{\min}$ because the improvement in the SALR interval could not compensate the loss introduced in the SBLR interval. Second, in practice, gains with respect to the SALR are lower than in theory. In fact, the optimum R_l value is 300 kbits/s which provides a mean PSNR of 30.72 dB, compared to the 30.41 dB and 28.44 dB of the SALR and SBLR coders respectively.

7. CONCLUSIONS AND FUTURE WORK

In this paper, we have theoretically analyzed the performance of three sorts of MCP-based SNR-scalable video coders and have compared them to their non-scalable counterpart. Results show that main trends in the efficiency described by the theory match practical results obtained from the encoding of real video sequences. Consequently, our analysis is useful to understand the main features of each scalable strategy and what factors influence their efficiency.

Although the present work only takes into account the steady-state response of SALR and SABLR coders, we are currently extending our analysis by considering also their transitory response. This will allow us to analyze the efficiency of these strategies in coders using periodic intra-frames. We are also studying the optimum values of parameters c and R_l when different degrees of motion estimation accuracy exist.

8. REFERENCES

[1] B. G. Haskell, A. Puri, and A. N. Netravali, *Digital Video: An introduction to MPEG-2*, Chapman and Hall, 1997.

[2] W. Li, "Overview of fine granularity scalability in MPEG-4 video standard," *IEEE Trans. on CSVT*, vol. CSVT-11, pp. 301–317, 2001.

[3] C. Buchner, T. Stockhammer, D. Marp, G. Blatterman, and G. Heising, "Efficient fine granular scalable video coding," in *Proceedings of the ICIP*, Thessaloniki, Greece, October 7–10 2001, pp. 997–1000.

[4] J. Prades-Nebot, G. Cook, and E. J. Delp, "Rate control for FFGS video coders," in *Proceedings of the SPIE VCIP*, San Jose, California, 2002, vol. 4310, pp. 828–839.

[5] M. van der Schaar and H. Radha, "Adaptive motion-compensation fine-granular-scalability (AMC-FGS) for wireless video," *IEEE Transactions on CSVT*, vol. 12, no. 6, pp. 360–370, June 2002.

[6] J. Prades-Nebot and G. W. Cook, "Analysis of the performance of predictive SNR scalable coders," in *Proceedings of the ICIP*, Barcelona, Spain, Sept. 2003, vol. 3, pp. 861–864.

[7] P.-Y. Cheng, J. Li, and C.-C. J. Kuo, "Rate control for an embedded wavelet video coder," *IEEE Trans. on CSVT*, vol. 7, pp. 696–701, 1997.

[8] B. Girod, "The efficiency of motion-compensating prediction for hybrid coding of video sequences," *IEEE Journal on SAC*, vol. SAC-5, no. 7, pp. 1140–1154, 1987.

[9] B. Girod, "Motion-compensating prediction with fractional-pel accuracy," *IEEE Trans. on Communications*, vol. 41, pp. 604–611, 1993.

[10] K. Shen and E. J. Delp, "Wavelet based rate scalable video compression," *IEEE Trans. on CSVT*, vol. 9, pp. 109–122, 1999.