PARAMETER ESTIMATION FOR AN H.264-BASED DISTRIBUTED VIDEO CODER

B. Macchiavello, R. L. de Queiroz[*]

Departamento de Engenharia Eletrica
Universidade de Brasilia
Brasilia, DF, Brazil
Email: bruno,queiroz@image.unb.br

D. Mukherjee

Hewlett Packard Labs
Palo Alto, CA, USA
debargha.mukherjee@hpl.hp.com

ABSTRACT

In this paper we present a statistical model used to select coding parameters for a mixed resolution Wyner-Ziv framework implemented using the H.264/AVC standard. This paper extends the results of a previous work for the H.263+ case to the H.264/AVC coder, since the parameters need to be recalculated for the H.264 case. The proposed correlation estimation mechanism guides the parameter choice process, and also yields the statistical model used for decoding. This mechanism is proposed based on extracting edge information and residual error rate in co-located blocks from the low resolution base layer that is available at both ends.

***Index Terms*—** distributed video coding, Wyner-Ziv, parameter estimation

1. INTRODUCTION

Distributed source coding (DSC), which has its roots in the theory of coding correlated sources developed by Slepian and Wolf [1], for the lossless case, and Wyner and Ziv [2], for the lossy case, has been recently applied to video coding to enable a reversed complexity coding mode [3]–[8]. In reversed mode, the encoder complexity is reduced by eliminating the motion estimation task or obviating the need for full motion search. The performance loss is partially recovered by a more complex decoding process exploiting source statistics.

In previous works [9],[10] we proposed a mixed resolution framework that can be implemented as an optional coding mode in any existing video codec standard. In this framework, the reference frames are coded exactly as in a regular codec as I-, P- or reference B-frames, at full resolution. For the non-reference P- or B- frames the encoding complexity is reduced by low resolution (LR) encoding. At the decoder, a high quality version of the non-reference frames are generated by a multi-frame motion-based mixed super-resolution mechanism [9]–[12]. The interpolated LR reconstruction is subtracted form this frame to obtain the side-information (SI), which is a Laplacian residual frame. Thereafter, the Wyner-Ziv (WZ) layer is channel decoded to obtain the final reconstruction.

In realistic usage scenarios for video communication using power-constrained devices, it is not necessary for a video decoder to reproduce the signal immediately after reception. Therefore, a feedback channel may not always be available. In the mixed resolution approach, the LR layer can be immediately decoded for real-time communications. More important, since the framework does not use a feedback channel for rate-estimation, it enables the enhancement layer to be decoded offline. However, the elimination of the feedback channel requires a sophisticated mechanism for estimating the correlation statistics at the encoder, followed by mapping the estimated statistics to actual encoding parameters. A previous work [13] presented a such estimation model using H.263+ as the regular codec. We present in this paper, as a continuation [13], a statistical model as well as a mechanism to estimate the model parameters for a memoryless coset code using H.264/AVC.

2. WYNER-ZIV CODING MODE ON H.264/AVC

The basic architecture for the WZ coding mode can be found elsewhere [9], [10]. Summarizing, at the encoder (shown in Fig.1), the non-reference frames are decimated and coded using decimated versions of the reconstructed reference frames in the frame store. Then the Laplacian residual, obtained by taking the difference between the original frame and an interpolated version of the LR layer reconstruction, is WZ coded to form the enhancement layer. Related work [14] has also explored spatial reduction. Nevertheless, our mixed resolution approach, while less aggressive in complexity reduction, may achieve better compression efficiency.

At the decoder, the LR image is decoded and interpolated. The optional process of enhancement begins with the generation of the SI. The interpolated decoded frame and the reference frames are used to create a semi super-resolution version of the current frame [11]. Then, it is subtracted from the interpolated LR decoded frame. The resulting residual frame is

[*] Work supported by HP Brasil.

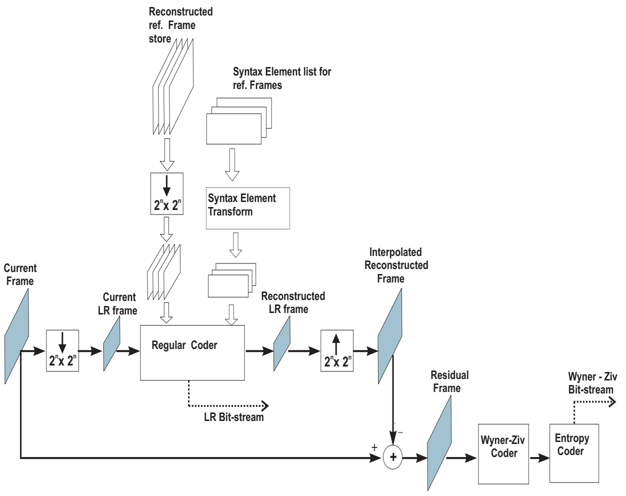

Fig. 1. Architecture for the DVC encoding mode.

the actual SI frame to be used for channel decoding.

2.1. Enhancement Layer

Let the random variable X denote the transform coefficients of the residual error frame. Then, the quantization of X yields $Q : Q = \phi(X, QP), QP$ being the quantization step-size. Next, the cosets $C : C = \psi(Q, M) = \psi(\phi(X, QP), M), M$ being the coset modulus, are computed:

$$\psi(q, M) = \begin{cases} (Q) - M\lfloor Q/M \rfloor, & (Q) - M\lfloor q/M \rfloor < M/2 \\ (Q) - M\lfloor q/M \rfloor - M, & (Q) - M\lfloor Q/M \rfloor \geq M/2 \end{cases} \tag{1}$$

If quantization bin q corresponds to interval $[x_l(q), x_h(q)]$, then the probability of the bin $q \in \Psi_q$, and $c \in \Psi_c$ are given by:

$$p(q) = \int_{xh(q)}^{xl(q)} f_X(x)dx \tag{2}$$

$$p(c) = \sum_{q \in \Psi_q, \psi(q,M)=c} p(q) = \sum_{q \in \Psi_q, \psi(q,M)=c} \int_{xh(q)}^{xl(q)} f_X(x)dx, \tag{3}$$

The entropy coder that already exists in the regular coder can be reused for C, but a different entropy coder conditioned on M should yield better compression. For decoding, the minimum MSE reconstruction function based on unquantized side information y and received coset index c, is given by:

$$\hat{X}_{YC}(y, c) = \frac{\sum\limits_{q \in \Psi_q, \psi(q,M)=c} \int_{xh(q)}^{xl(q)} x f_{X|Y}(x, y)dx}{\sum\limits_{q \in \Psi_q, \psi(q,M)=c} \int_{xh(q)}^{xl(q)} f_{X|Y}(x, y)dx}. \tag{4}$$

The regularly coded reference frames and the LR layer frames are assumed to be coded with quantization step-size QP_t. Therefore, the enhancement layer frames should be ideally coded such that the distortion is at about the same level as that obtained by regular coding with QP_t. A rate-distortion analysis to find the optimal encoding parameters QP, M based on our statistical model, can be found elsewhere [10], [15].

3. CORRELATION STATISTICS ESTIMATION

We assume a general enough statistical model: $Y = \rho X + Z$, where X is a Laplacian distributed transform coefficient, Z is additive Gaussian noise uncorrelated with X and $0 < \rho \leq 1$ is an attenuation factor expected to decay at higher frequencies. It is necessary to have an accurate estimation of σ_X and σ_Z for the encoder parameter choice and for minimum MSE reconstruction at the decoder. Note that this is a generalization of the simpler model: $Y = X + Z$ [10], [11], [15]. However, we can rewrite it as $Y/\rho = X + Z/\rho$. Then, the same procedure described in [9]–[11] can be applied by simply replacing σ_Z^2 with $(\sigma_Z/\rho)^2$ and replacing Y with Y/ρ during decoding.

We specialized the model parameters for each frequency band (FB) within a block, where the FB is defined as diagonals in a transform block. Also note that the correlation is obviously dependent on the quantization step-size QP_t for the reference frame and the LR layer. Besides, other vital information can be extracted from the LR layer to direct the estimation process. Note that since any data from the LR layer is available at both decoder and encoder, no overhead bits need to be transmitted to convey this information. An alternative approach may explicitly transmit some statistical information. In this work, we adopt a no-overhead approach. To generate the estimation models we use a training-based approach where X (transform coefficients of Laplacian residual of original frame) and Y (transform coefficients of residual after multi-frame processing) data for each FB is collected for a set of training video sequences for varying values of QP_t along with the corresponding values of additional information extracted from the LR layer. Then, we need to estimate σ_X, σ_Z, ρ that are used to select QP, M, for coset creation and for MSE reconstruction.

3.1. Estimation of σ_X^2- variance of Laplacian residual coefficients

The variance of a Laplacian residual coefficient (σ_X^2) is not the same at every block of a coded frame. It does not only depend on QP_t and FB, but also on the high frequency content of the block. If the original frame has a high edge content it is likely that the error between the decimated-interpolated version and the original one would be larger. Even though the exact high frequency content in an original frame is not available at the decoder, we can use an edge activity measure of the reconstructed LR block as a parameter to estimate σ_X^2. It is intuitive to think that the edge activity in the LR block will be correlated with the energy of the high frequency coefficients of the Laplacian residual, while the energy at the lower frequencies in the Laplacian residual will be more related to QP_t. The edge activity, denoted E, is computed as the accumulated sum of the difference between neighbor pixels along the lines and columns of a macroblock in the reconstructed version of the interpolated LR frame. Then, σ_X^2 is modeled as a function of QP_t, FB and E. That is:

$$\sigma_X^2 = f_1(QP_t, FB, E). \qquad (5)$$

We next assume σ_X^2 to be proportional to QP_t^2. Further, after processing the training data we find that it is enough to linearly model the remaining part for each FB, so that:

$$\sigma_X^2 = (k_{1,FB}E + k_{2,FB})QP_t^2 \qquad (6)$$

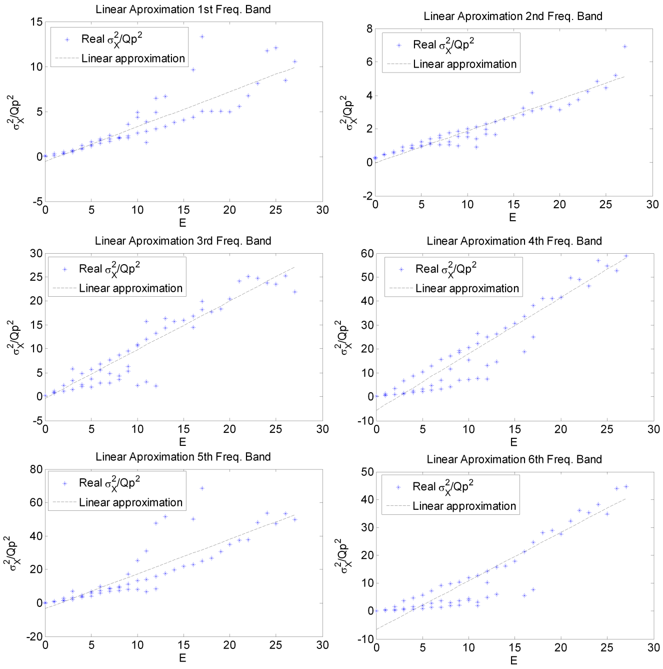

Fig. 2. Statistics estimation. Real σ_x^2/QP_t^2 vs. E and linear approximation.

where $k_{i,FB}$ are constants that vary for each frequency band. In Figure 2, we show the linear approximations used for σ_x^2/QP_t^2 vs. E, compared to the real training data for the first 6 frequency bands.

3.2. Estimation of the correlation parameter

To estimate the correlation parameter, we use a simplified model assuming that it only depends on QP_t and FB:

$$\rho_X = f_2(QP_t, FB). \qquad (7)$$

Note that with higher QP_t the variables X and Y should be less correlated. The values of ρ obtained from the training data set can be stored as pre-calculated tables at both encoder and decoder.

3.3. Estimation of the variance of the Gaussian noise

To estimate σ_Z^2 from the training data set, we first calculate $Z = Y - \rho X$. Further, we conjecture that σ_Z^2, for a macroblock in the enhancement layer depends on the residual error rate R used to code a colocated 8×8 block in the LR base layer along with QP_t, FB and E. A higher rate in the LR base layer indicates greater inaccuracy of motion estimation at reduced resolution. Therefore, the multi-frame super-resolution process is also likely to yield more inaccurate estimate of the high-resolution frame at the decoder, leading an increase in σ_Z^2. However, since R also depends on QP_t we use normalized rate $R_n = R \times QP_t^2$ in order to remove the effect of QP_t. Thus, we can model σ_Z^2 as:

$$\sigma_Z^2 = f_3(QP_t, FB, E, R_n). \qquad (8)$$

We next assume σ_Z^2 to be proportional to σ_X^2 for a given FB and R_n, and the effect of QP_t and E to be within σ_X^2. Furthermore, the remaining part is modeled linearly for each FB, such that

$$\sigma_Z^2 = f_3(k_{3,FB}R_n + k_{4,FB})\sigma_X^2. \qquad (9)$$

4. RESULTS AND CONCLUSION

The parameter choice mechanism [10] and the proposed model for estimating the correlation statistics were applied to our mixed resolution framework using H.264/AVC as the regular codec. In Figs.3 and 4 we compare the performances of: (i) a regular H.264 codec working in $IBPB$... mode; (ii) the LR base where the non-reference B-frames are encode at quarter resolution and interpolated at the decoder; and (iii) the results from decoding both layers using the proposed statistics model. In Fig.5 a similar result is presented, where the regular H.264 codec is working in $IP_eP_rP_e$... mode. P_r is a reference P-frame, and P_e is a disposable non-reference P-frame, here the P_e frames are at quarter resolution. For low-motion and low-high-frequency content sequences (like “mother and daughter”) our architecture may even outperform regular coding. However, for most sequences the reversed coding mode will perform below regular coding, but it will be competitive. In Table 1, a coding time comparison is made.

The reader may note that this statistic estimation method can be used in other frameworks as well[16].

Table 1. Coding Time comparison. (FPS: frame per second)

Foreman CIF Sequence (299 frames)			
Mode IPBPB (B-Frames are non-reference frames)			
Conventional H.264/AVC			
Total time (sec)	303.12	FPS	0.98
Wyner-Ziv mode on H.264/AVC (B-frames are WZ coded)			
Total time (sec)	173.65	FPS	1.72
Reduced Total time	42.71%	FPS Gain	75.51%

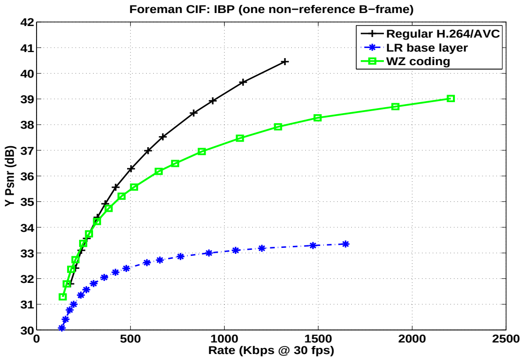

Fig. 3. PSNR (dB) plots for Foreman CIF sequence encoded in IBP mode.

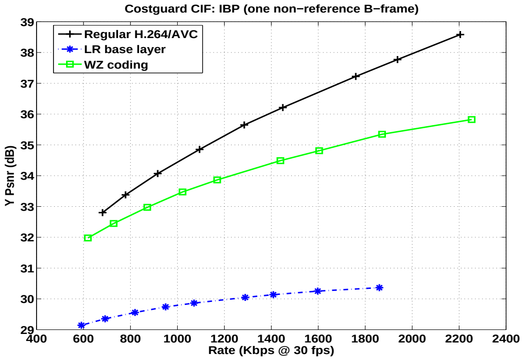

Fig. 4. PSNR (dB) plots for Coastguard CIF sequence encoded in IBP mode.

5. REFERENCES

[1] J. Slepian and J. Wolf, "Noiseless coding of correlated information sources," *IEEE Trans on Inf. Theory*, vol. 19, no. 4, pp. 471–480, Jul 1973.

[2] A. Wyner and J. Ziv, "The rate-distortion function for source coding with side information at the decoder," *IEEE Trans on Inf. Theory*, vol. 2, no. 1, pp. 1–10, Jan 1976.

[3] A. Aaron, R. Zhang, and B. Girod, "Transform-domain Wyner-Ziv codec for video," *In Proc. SPIE Visual Com. and Img. Proc.*, vol. 5308, pp. 520–528, January 2004.

[4] Q. Xu and Z. Xiong, "Layered WynerZiv video coding," *IEEE Trans on Img. Proc.*, vol. 15, no. 12, pp. 3791–3809, Dec 2006.

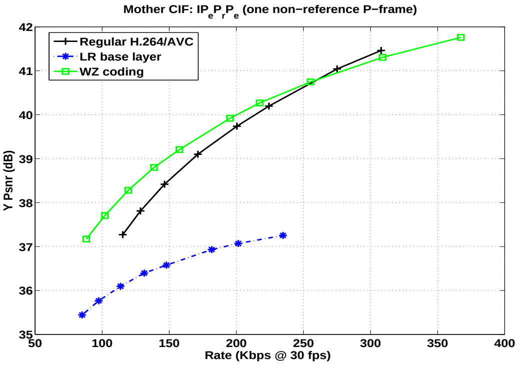

Fig. 5. PSNR (dB) plots for Mother and Daughter CIF sequence encoded in IP_eP_r mode.

[5] H. Wang, N. M. Cheung, and A. Ortega, "A framework for adaptive scalable video coding using Wyner-Ziv techniques," *EURASIP Journal on Applied Signal Proc.*, pp. 1–18, 2006.

[6] M. Tagliasacchi, A. Majumdar, and K. Ramchandram, "A distributed-source-coding based robust spatio-temporal scalable vedo codec," *Picture Coding Symposium*, December 2004.

[7] X. Wang and M. T. Orchard, "Desing of trellis codes for source coding with side infotmation at the decoder," *In Proc. of IEEE Data Compression Conf.*, pp. 361–370, 2001.

[8] B. Girod, A.M. Aaron, S. Rane, and D. Rebollo-Monedero, "Distributed video coding," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 71–83, Jan 2005.

[9] D. Mukherjee, "A robust reversed complexity Wyner-Ziv video codec introducing sign-modulated codes," *HP Labs Tech. Report, HPL-2006-80*, May 2006.

[10] D. Mukherjee and B. Macchiavello and R. L. de Queiroz, "A simple reversed-complexity Wyner-Ziv video coding mode based on a spatial reduction framework,"*Proc. of SPIE Visual Com. and Img. Proc.*, vol 6508, pp. 1Y1-1Y12, Jan 2007.

[11] B. Macchiavello, R.L de Queiroz and D. Mukherjee, "Motion-based side-inforamtion generation for a scalable Wyner-Ziv video Coding," *Proc. of the IEEE Int. Conf. on Img. Proc.*, pp. VI-413–VI-416, San Antonio, 2007.

[12] L. W. Kang and C. S. Lu, "Wyner-Ziv video coding with coding mode-aided motion compensation," *In Proc. IEEE Int. Conf. on Img. Proc.*, pp. 237–240, 2006.

[13] B. Macchiavello, D. Mukherjee and R.L de Queiroz, "A statistical model for a mixed resolution Wyner-Ziv framework," *Picture Coding Symposium*, Lisboa, November 2007.

[14] M. Wu, G. Hua, and C. W. Chen, "Syndrome-based lightweight video coding for mobile wireless application," *In Proc. Int. Conf. on Multimedia and Expo*, pp. 2013–2016, 2006.

[15] D. Mukherjee, "Optimal parameter choice for Wyner-Ziv coding of Laplacian sources with decoder side-information," *HP Labs Technical Report*, HPL-2007-34.

[16] R. Puri, A. Majumdar and K. Ramchandran, "PRISM: A Video Coding Paradigm With Motion Estimation at the Decoder," *IEEE Transactions on Image Processing*, vol 16, no 10, pp. 2436-2448, 2007.