Towards Optimal Scalability in Predictive Video Coding

Kenneth Rose and Shankar L. Regunathan
Department of Electrical and Computer Engineering
University of California
Santa Barbara, CA 93106 *

Abstract

We propose a method for efficient SNR scalability in predictive video coding which overcomes the known fundamental difficulties due to the prediction loop. The method is generally applicable to any scalable predictive coder but a special emphasis is given in this paper to standard DCT-based video compression. The compression efficiency of the enhancement-layer is substantially improved by casting the design of its prediction module within an estimation-theoretic framework. The resulting prediction of a transform coefficient in the enhancement-layer is optimal given both the corresponding reconstructed coefficient at the base-layer, and the enhancement-layer reconstructed coefficient in the previous frame. Simulation shows consistent, substantial improvement in PSNR of the reconstructed enhancement-layer frames. For two-layer systems, such optimal prediction can be closely approximated by a simple switched prediction scheme which is of negligible complexity. The proposed method may be easily combined with known temporal scalability methods to provide further improvement in compression performance over a wide range of bit rates.

1 Introduction

It is increasingly important for video coding algorithms to provide a scalable bitstream. Many applications (e.g., multi-party video conferencing, multicast over the internet, etc.) require the compressed information to be simultaneously transmitted to multiple receivers. The evolving global communication network is, in fact, a patchwork of a transmission media, which is highly nonuniform in its communication capabilities, and is characterized by vast variations in the channel bandwidth available for different links. Further, the feasible bit rate of a communication link is also constrained by the receiver and by the computational power and memory capacity of its decoder. A scalable[1] bitstream is one that allows decoding at a variety of bit rates (and corresponding levels of quality), and where the lower rate information streams are embedded within the higher rate bit-streams in a manner that minimizes redundancy.

The two main approaches to scalable video are: (i) three dimensional coding [1], and (ii) predictive video coding [2]. Predictive coders are more attractive for many applications because of their minimal requirements in terms of delay and memory and, further, because they allow straightforward incorporation of motion compensation. It is, therefore, of great interest to develop predictive video coders which can generate a scalable bitstream. However, there is a significant penalty in compression performance for a straightforward incorporation of scalability in predictive coding. The main difficulty is that of inefficient frame prediction at the enhancement-layer which seriously impacts the performance.

In this work, we develop an estimation-theoretic approach to enhancement-layer prediction in scalable coders. The frame prediction is optimal given the total information available at the enhancement-layer from the previous and current frames. The paper is organized as follows: In Section 2 we summarize standard approaches to the "prediction problem". Section 3 formulates the predictor design problem within an estimation-theoretic framework. Section 4 interprets the optimal estimate and its relation to standard predictors. It also contains discussion of the complexity-performance tradeoffs and derivation of a low complexity approximation. In section 5, we present simulation results on video sequences and demonstrate that optimal prediction at the enhancement layer can result in compression performance which is vastly superior to conventional approaches.

2 Background

In standard predictive video coding, temporal redundancy is removed by inter-frame prediction. The discrete cosine transform (DCT) is applied to the prediction error (residual), and the transform coefficients are then quantized. At the receiver, the residual is decoded and added to the prediction to form reconstruction of the current frame.

Figure 1 shows the prediction modules in a two-layer scalable coder. The prediction at the base-

*This work was supported in part by the National Science Foundation under grant no. NCR-9314335, the University of California MICRO program, ACT Networks, Inc, Advanced Computer Communications, Cisco Systems, Inc., DSP Group, Inc., DSP Software Engineering, Inc., Fujitsu Laboratories of America, Inc., General Electric Company, Hughes Electronics Corp., Intel Corp., Nokia Mobile Phones, Qualcomm, Inc., Rockwell International Corp., and Texas Instruments, Inc.

[1] We are chiefly concerned with SNR scalability in this work, though the work can be extended in a straightforward manner to spatial scalability. The term "scalable" should be understood as SNR scalable unless otherwise stated

0-8186-8821-1/98 $10.00 © 1998 IEEE

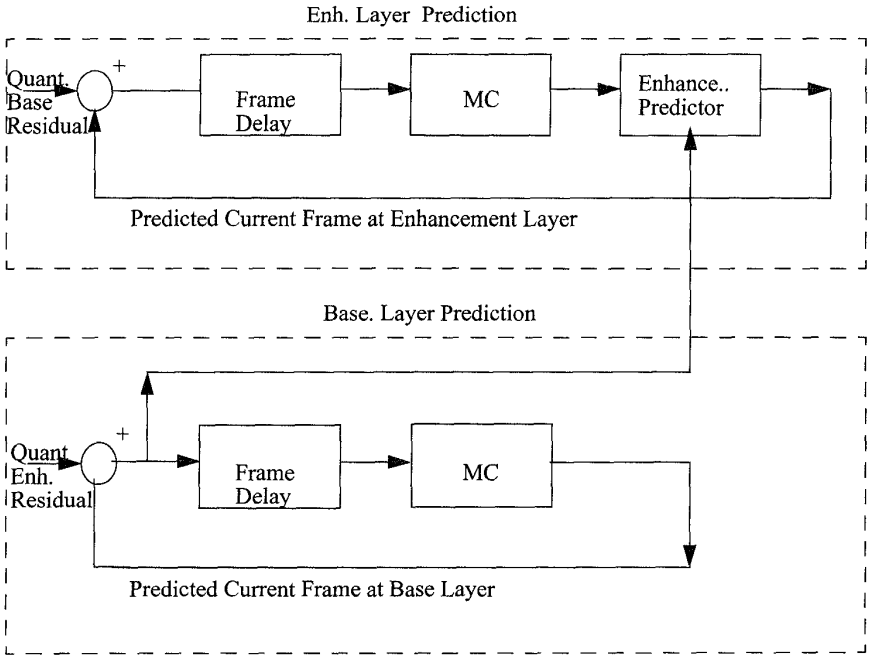

Figure 1: *A high-level sketch of frame prediction in a two-layer scalable coder. MC denotes motion compensation. The prediction is performed at both the encoder and decoder.*

layer is that of a standard (nonscalable) coder, and consists of motion-compensating the previous base-layer reconstruction. The main difficulty arises at the enhancement-layer. Two main sources of information are available for the enhancement-layer prediction: (i) enhancement-layer reconstruction of the previous frame, and (ii) base-layer reconstruction of the current frame. There are two main existing approaches to enhancement-layer prediction, which amount to the exclusive use of either one of the above sources of information:

P1: Use the base-layer reconstruction of the current frame as the prediction (e.g., [3]). While this method takes advantage of the base-layer compressed residual, it suffers from an obvious shortcoming: No advantage is taken of the superior quality motion compensated reconstruction of the previous frame which is available at the enhancement layer.

P2: Predict the current enhancement-layer frame from the motion compensated reconstruction of the previous enhancement-layer frame as in [4]. Note that in this case the enhancement-layer does not exploit the information in the current base layer residual. The two layers are, in fact, separately encoded except for possible savings on shared side-information such as motion vectors.

Coding schemes which can combine information from both sources to form the prediction have been suggested recently. The coder in H.263+[8] can choose to predict each macroblock by using either the corresponding block (after motion compensation) in the previous enhancement-layer reconstruction, or the block in the current base-layer reconstruction. The switching information is transmitted as side-information. To minimize the amount of side information, the predictor is fixed for each macroblock. This scheme can achieve significant gains over prediction schemes P1 and P2.

To achieve further efficiency in scalable video coding, we have developed a simple switching prediction algorithm in [6]. A similar algorithm was derived earlier in [5] for conditional replacement in the context of scalable video coding with pyramid and subband techniques. The choice of appropriate predictor is made for *each DCT coefficient* , while using *no side information.* The quantized DCT coefficient in the base-residual (freely available at both the encoder and decoder) is used for selecting the predictor of the corresponding DCT coefficient in the current frame. The switching algorithm can be summarized as follows:

FOR each DCT coefficient in the current frame: *IF* the quantized base residual is zero in this position, select the coefficient from previous motion-compensated enhancement-layer as predictor. *ELSE*, select the coefficient from current base-layer as predictor.

The switching rule was devised as a heuristic technique to improve prediction by using information available only to the enhancement-layer from the pre-

vious frame, without compromising the usefulness of the information provided by the current base-layer reconstruction. In simulations, scalable video with a switched predictor is shown to consistently outperform other existing approaches, at the cost of negligible complexity.

In this paper, we extend our approach to develop a predictor design algorithm within an estimation-theoretic framework. It turns out that the switching algorithm is a good approximation to the optimal predictor for a two-layer scalable coder. However, the optimal estimate can produce significant additional gains in the multi-layer scalable coding scenario.

3 Scalable Coder Design

The DCT domain is more convenient for the predictor design because the DCT coefficients of the residual are almost uncorrelated and, typically, scalar quantized. Thus, the predictor can be independently designed for each DCT coefficient with minimal loss of optimality. We assume that the evolution of a DCT coefficient in time ("from frame to frame") can be modeled by a first-order Markov process

$$x_n = \rho * x_{n-1} + z_n \tag{1}$$

where x_n is a DCT coefficient in the current frame and x_{n-1} is the corresponding (after motion compensation) DCT coefficient in the previous frame. Let $\hat{x}^b_{n-1}$ and $\hat{x}^e_{n-1}$ be the base and enhancement reconstructions of x_{n-1}, respectively. We assume that z_n is zero-mean, stationary, and independent of x_{n-1}.

Base-Layer: The optimal base layer predictor is given by

$$\tilde{x}^b_n = E[x_n|\hat{x}^b_{n-1}] \approx \hat{x}^b_{n-1}. \tag{2}$$

The approximation holds if $\rho \approx 1$ (a valid assumption for "low frequency" coefficients), and if successive quantization errors are nearly independent. The base encoder quantizes the residual, $r^b_n = x_n - \tilde{x}^b_n$, and transmits index i^b. Let (a, b) be the quantization interval associated with index i^b, hence, $r^b_n \in (a, b)$. Clearly, $x_n \in (\tilde{x}^b_n + a, \tilde{x}^b_n + b)$ captures all the information on x_n provided by the transmitted residual index. Therefore, the optimal base-layer reconstruction is given by,

$$\hat{x}^b_n = E[x_n|\hat{x}^b_{n-1}, x_n \in (\tilde{x}^b_n + a, \tilde{x}^b_n + b)]. \tag{3}$$

This estimate is computed (by definition) by calculating the expected value of x_n with respect to the density $p(x_n||\hat{x}^b_{n-1})$ over the interval $(\tilde{x}^b_n + a, \tilde{x}^b_n + b)$. We give more details on this operation when we discuss the case of the enhancement-layer.

Enhancement-Layer: In addition to the information provided by the base-layer, the enhancement-layer decoder has access to $\hat{x}^e_{n-1}$, the corresponding enhancement-layer reconstructed DCT coefficient of the previous frame. Thus, taking into account all the available information, the the optimal enhancement-layer predictor is

$$\tilde{x}^e_n = E[x_n|\hat{x}^e_{n-1}, \hat{x}^b_{n-1}, x_n \in (\tilde{x}^b_n + a, \tilde{x}^b_n + b)]. \tag{4}$$

It is reasonable to assume that $\hat{x}^b_{n-1}$ provides little information in addition to that contained in $\hat{x}^e_{n-1}$. Therefore, we can remove the conditioning on $\hat{x}^b_{n-1}$, and rewrite the optimal predictor as

$$\tilde{x}^e_n = E[x_n|\hat{x}^e_{n-1}, x_n \in (\tilde{x}^b_n + a, \tilde{x}^b_n + b)], \tag{5}$$

which can also be rewritten as

$$\tilde{x}^e_n = \hat{x}^e_{n-1} + E[z_n|z_n \in (\tilde{x}^b_n + a - \hat{x}^e_{n-1}, \tilde{x}^b_n + b - \hat{x}^e_{n-1})] \tag{6}$$

Thus, the optimal enhancement-layer predictor seeks the best estimate, based on the previous enhancement layer reconstruction (and the statistical relation between x_n and x_{n-1}), which is consistent with the quantization interval specified by the current base-layer reconstruction.

The enhancement-layer encoder quantizes the residual, $r^e_n = x_n - \tilde{x}^e_n$, and transmits index i^e. Let (c, d) be the quantization interval associated with index i^e. Therefore, $r^e_n \in (c, d)$ and $x_n \in (\tilde{x}^e_n + c, \tilde{x}^e_n + d)$. We now define $e = max[\tilde{x}^b_n + a, \tilde{x}^e_n + c]$ and $d = min[\tilde{x}^b_n + b, \tilde{x}^e_n + d]$. It is convenient to combine the information provided by the two quantization intervals in the statement

$$x_n \in (e, f). \tag{7}$$

The enhancement-layer reconstruction of the DCT coefficient is

$$\hat{x}^e_n = E[x_n|\hat{x}^e_{n-1}, x_n \in (e, f)] \tag{8}$$

To evaluate such expectations we employ an appropriate probabilistic model for z_n, the predictor error term. It is well known that the marginal density function of the DCT coefficient may be approximated by a Laplacian distribution. If x_n is a Markov-Laplace process, then the density of z_n is [7].

$$p_{z_n}(z) = \rho^2\delta(z) + (1 - \rho^2)\alpha e^{-|z|/\alpha} \tag{9}$$

The parameters ρ and α can be estimated from a training set. We found that $\rho \approx 1$ for "low and intermediate frequency" DCT coefficients. The optimal prediction consists of computing the centroid of the quantization interval (specified by the base layer) with respect to the density of (9) for each DCT coefficient, and is therefore of moderate complexity.

The predictor design can be extended in a straightforward manner to the multi-layer coding scenario. For the prediction at an enhancement layer, the interval, over which we evaluate the expectation, is determined by the quantization intervals of all the layers below it.

4 Interpretation and Implementation

We first show that the conventional prediction schemes, P1 and P2, are special cases of the optimal predictor, and are optimal under very specific circumstances:

Enhancement-Layer Rate $\approx$ Base-Layer Rate: In this case, the quality of the base-layer is

Rate (Kbps) of Enh. Layer	Conventional Pred.		H.263+ like Prediction	Switched Prediction	Optimal Prediction
	P1	P2			
32	30.16	29.85	30.64	31.00	31.01
40	30.62	30.78	31.52	31.84	31.86
48	31.07	31.66	32.25	32.58	32.61
64	31.84	33.07	33.51	33.75	33.78
128	34.21	36.54	36.81	36.98	36.98

Table 1: Performance of 2-layer scalable coders with different enhancement-layer prediction on the sequence *Carphone*. The entries provide the average PSNR of 89 reconstructed enhancement-layer frames (in dB) for different enhancement-layer rates. The base layer rate was fixed at 16 Kbps for all cases and the corresponding base layer PSNR was 29.30 dB (for all methods).

comparable to that of the enhancement layer and thus $\hat{x}^e_{n-1}$ in (5) may be replaced by $\hat{x}^b_{n-1}$.

$$\tilde{x}^e_n \approx E[x_n|\hat{x}^b_{n-1}, x_n \in (\tilde{x}^b_n + a, \tilde{x}^b_n + b)] = \hat{x}^b_n \quad (10)$$

Hence, the optimal predictor is approximated by P1 in this case.

Base-Layer Rate << Enhancement-Layer Rate The base quantizer is very coarse in comparison to the enhancement-layer quantizer. Thus the quantization interval specified by (a, b) is very large and captures almost all the probability of z_n.

$$\tilde{x}^e_n \approx \hat{x}^e_{n-1} + E[z_n|z_n \in (-\inf, \inf)] = \hat{x}^e_{n-1} \quad (11)$$

Hence, P2 approximates the optimal predictor.

We see that P1 and P2 provide close to optimal performance for two extreme target rates. At most rates of practical interest, however, neither P1 nor P2 approximate the optimal predictor well enough, and this is the main shortcoming of most scalable coders.

We next show that the switched predictor can give an excellent approximation to optimal prediction for all target rates, at negligible complexity. We consider two cases:

Case 1: Let $i^b = 0$ or $\hat{x}^e_{n-1} \in (\tilde{x}^b_n + a, \tilde{x}^b_n + b)$. We have

$$0 \in (\tilde{x}^b_n + a - \hat{x}^e_{n-1}, \tilde{x}^b_n + b - \hat{x}^e_{n-1}), \quad (12)$$

and

$$\tilde{x}^e_n = \hat{x}^e_{n-1} + \rho^2 * 0 + (1 - \rho^2) * [\ldots] \approx \hat{x}^e_{n-1}, \quad (13)$$

since $\rho \approx 1$.

Case 2: Let $i^b \neq 0$ or $\hat{x}^e_{n-1} \notin (\tilde{x}^b_n + a, \tilde{x}^b_n + b)$. In this case, we have to find the centroid of an exponential over an interval. The "memoryless" property of the exponential makes the centroid depend only on the quantization interval and not on the origin.

$$\tilde{x}^e_n = E[x_n|\hat{x}^b_{n-1}, x_n \in (\tilde{x}^b_n + a, \tilde{x}^b_n + b)] = \hat{x}^b_n \quad (14)$$

Thus, the optimal predictor is well approximated by the switching algorithm. Note that the approximation is valid for all base and enhancement layer rates.

5 Simulation Results

We developed a test bed for scalable coding by using the publicly available H.263 coder [9]. The H.263 algorithm was used for motion estimation, and compressing the prediction error residual of the base and enhancement layers. The advanced motion compensation and arithmetic encoding (of the prediction) options were turned off.

The following prediction modules for the enhancement-layer were implemented for the comparisons: (1) P1 (proposed in [3]), (2) P2 (proposed in [4]), (3) Switched Predictor [6], [5], and (4) Optimal Predictor. For additional comparison, we have implemented a coder similar to the H.263+ [8] coder, which chooses the predictor per macroblock and uses side information for signaling the predictor.

The z_n model parameters for each DCT coefficient were estimated from a training set extracted from the *Miss America* sequence. The experiments were performed on "qcif" sequences at bit rates ranging from 16 to 256 Kbps. The frameskip was 3, and the PSNR was averaged over 89 coded frames.

Table 1 shows the results of two layer scalable compression of the benchmark *Carphone* sequence. The base-layer rate was fixed at 16 kbps and results are presented for various ratios of base to enhancement rate. It is easy to see that optimal prediction (and switched prediction) substantially outperform all competing approaches. The gains in reconstructed PSNR of the enhancement layer is greater than $2.5dB$ in some cases. As expected (see section 3), P1 outperforms P2 at small ratios of enhancement to base rate, and underperforms P2 at the other extreme. The H.263+ predictor gains over P1 and P2 and performs very well at high enhancement rates. However, at lower rates for the enhancement-layer, the side information for transmitting the choice of predictor comprises a significant fraction of the total rate and leads to performance significantly lower than the optimal predictor. Further, we see that switched predictor can give near optimal performance for the two layer case, at negligible complexity.

Table 2 shows the performance for the multi-layer coding scenario. The rate for each layer and the corresponding performance is listed in the table. As earlier, the optimal and switched predictors substantially outperform other approaches. The gains in performance increase with the number of layers due to improved

Rate (Kbps) of Each Layer	Conventional Pred.		H.263+ like Prediction	Switched Prediction	Optimal Prediction
	P1	P2			
16	29.30	29.30	29.30	29.30	29.30
32	30.16	29.85	30.64	31.00	31.01
48	30.77	29.85	31.38	31.94	31.98
64	31.31	29.85	31.94	32.63	32.71
96	32.40	31.66	33.25	34.19	34.29
128	33.32	31.66	34.18	35.31	35.43
192	35.18	34.14	36.14	37.50	37.63
256	36.71	34.14	37.69	38.93	39.12

Table 2: Performance of multi-layer coder with different enhancement-layer prediction on the sequence *Carphone*. The entries provide the average PSNR of 89 reconstructed frames for the different layers.

prediction at every layer. Further, optimal prediction has a *significant advantage* over switched prediction, which grows with the number of layers.

6 Summary and Conclusions

This paper presents a new approach to optimal SNR scalability in predictive video coding. The predictor is designed within an estimation-theoretic framework. DCT coefficient prediction, at the current frame, is optimal given both the reconstructed DCT coefficient at the base-layer, and the previous enhancement layer reconstruction of the corresponding coefficient. Simulation results show that the proposed scalable video coding offers substantial gains in compression performance over conventional approaches over a wide range of bit rates. The optimal predictor may be well approximated with a low-complexity, switched predictor for two-layer coding. However, optimal prediction can offer additional gains in the multi-layer coding scenario. Although optimal prediction was applied here in conjunction with standard DCT-based coding systems, it is easily extendible to subband-based, and pixel-domain coders.

Acknowledgments

The authors thank Peng Wu for his contribution to the early simulations.

References

[1] D. Taubman and A. Zakhor, "Multirate 3-D subband coding of images, " *IEEE Trans. on Image Processing*, Sept. 1994, pp. 572-88.

[2] B. G. Haskell, A. Puri, and A. N. Netravali, *Digital video : an introduction to MPEG-2.* New York: Chapman and Hall, International Thomson Pub., 1997.

[3] D. Wilson and M. Ghanbari, "Transmission of SNR scalable two lay er MPEG-2 coded video through ATM networks," *Proc. 7th International Workshop on Packet Video*, pp. 185-189, March 1996.

[4] B. Girod, U. Horn, and B. Belzer, "Scalable video coding with multiscale motion compensation and unequal error protection," In Y. Wang, S. Panwar, S.-P. Kim, and H. L. Bertoni, editors, *Multimedia Communications and Video Coding*, pp. 475-482, New York: Plenum Press, 1196.

[5] T. K. Tan, K. K. Pang, and K. N. Ngan, "A frequency scalable coding scheme employing pyramid and subband techniques," *IEEE Transactions on Circuits and Systems for Video Technology*, April 1994,vol.4, pp. 203-7.

[6] K. Rose, P. Wu and S. L. Regunathan, "Efficient SNR-scalability in predictive video coding," Appeared in *ICASSP 98.*

[7] N. Farvardin and J. W. Modestino, "Rate-Distortion performance of DPCM schemes for autoregressive sources," *IEEE Transactions on Information Theory*, May 1985, vol. 31, pp. 402-18.

[8] Draft Text of H.263 Version 2(H.263+).

[9] TMN (H.263) coder version 2.0, Telenor R&D Norway, 1996.