EFFICIENT SNR-SCALABILITY IN PREDICTIVE VIDEO CODING

Kenneth Rose, Peng Wu and Shankar L. Regunathan

Department of Electrical and Computer Engineering
University of California, Santa Barbara, CA 93106.

ABSTRACT

A new method is proposed for efficient SNR scalability in predictive video coding. It is of low complexity, and it is applicable to standard DCT-based video compression with motion compensation. Information that is only available to the enhancement layers is exploited to improve the quality of their frame prediction without compromising the usefulness of the compressed data provided by the base layer(s). More specifically, the next frame prediction for use by an enhancement-layer decoder is obtained by combining, or switching between transform coefficients from: i) the reconstructed base-layer frame; and ii) the predicted enhancement-layer frame. The combining rule depends on the compressed residual of the base layer, and on the parameters used for this compression. The method is applied to standard DCT-based predictive video coding, and preliminary simulation shows consistent, substantial improvement in the performance of enhancement layers. The proposed method may be easily combined with known temporal scalability methods to provide further improvement of the performance of enhancement layers over a wide range of bit rates.

1. INTRODUCTION

Scalable signal compression algorithms are a major requirement of the rapidly evolving global network which involves a variety of channels with widely differing capacities, and even more so due to the recent trend toward incremental capacity and bandwidth reservation channels. Many applications require data to be simultaneously decodable at a variety of rates. Examples include applications such as multicast in a heterogenous network, where the channels dictate the feasible bit rates for each user. Similarly it is motivated by the co-existence of receivers of differing complexity (and cost). A compression technique is scalable[1] if it offers a variety of decoding rates using the same basic algorithm, and where the lower rate information streams are embedded within the higher rate bit-streams in a manner that minimizes redundancy.

There are two main approaches to scalable video compression: (i) three dimensional coding, and (ii) predictive coding. Three dimensional coding schemes [1] buffer up a set of consecutive video frames and apply a 3-D transform (or subband decomposition) to decorrelate the pixels. The resulting coefficients are encoded by a hierarchical quantization strategy which provides a scalable bit-stream. However, three dimensional schemes suffer from two major drawbacks: (i) buffering multiple frames causes a substantial delay and requires much memory, and (ii) motion compensation is not incorporated in this framework and the compression efficiency is compromised.

Predictive coders have low delay and small memory requirements and allow straightforward incorporation of motion compensation. Thus, standards such as H.26x and MPEG use DCT-based predictive compression. But there is a well-known, important performance penalty for incorporating scalability in a predictive coding framework. The main difficulty is how to improve the enhancement layer prediction of the current frame by using the additional previous frame information which is only available to the enhancement layer, without undermining the usefulness of the current frame information sent for the base layer. We propose a novel method of generating the enhancement layer current frame prediction by efficiently combining transform domain coefficients from: (i) motion compensated, previous enhancement layer reconstruction , and (ii) current base-layer reconstruction.

The organization of this paper is as follows: We summarize conventional approaches to scalable predictive compression of video in section 2. The new switched prediction approach is described in section 3. Section 4 presents preliminary simulation results on video sequences which demonstrate the feasibility and potential of our approach.

[1] In this work we are concerned with SNR scalability. The term "scalable" should be understood as "SNR scalable" unless otherwise stated.

0-7803-4428-6/98 $10.00 © 1998 IEEE

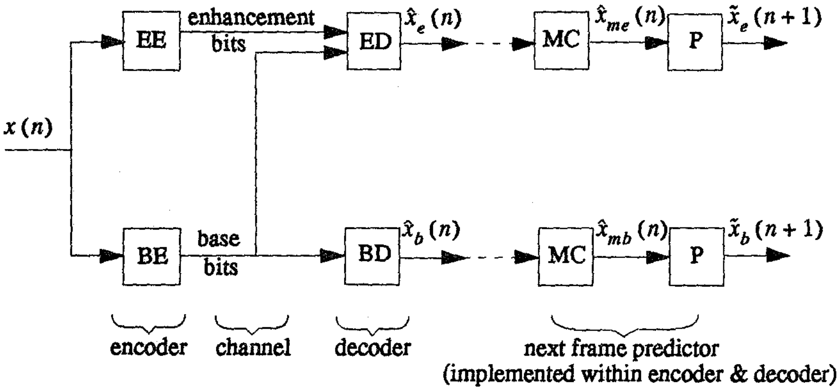

Figure 1: *A simplified sketch of a two-layer scalable predictive coding system. The predictors are included to show the relation of the next frame prediction to the reconstructed current frame. In fact, they are implemented within the encoders and the decoders.*

2. CONVENTIONAL SCALABLE PREDICTIVE VIDEO CODING

To appreciate the fundamental difficulties in conventional predictive coding, let us consider Figure 1 which depicts a two-layer scalable coder. The input frame $x(n)$ is compressed by the base encoder (BE) which produces the base bit-stream. The enhancement encoder (EE) has access to the input frame and to any information produced by, or available to BE. It uses all this to generate the enhancement bit-stream. A base decoder (BD) receives the base bit-stream and produces a reconstruction $\hat{x}_b(n)$, while the enhancement decoder (ED) has access to both bit-streams and produces an enhanced reconstruction $\hat{x}_e(n)$. The base and enhancement reconstructed frames are motion compensated to form $\hat{x}_{mb}(n)$ and $\hat{x}_{me}(n)$, respectively. The motion compensated frames are used to generate a prediction for the next frame: $\tilde{x}_b(n+1)$ or $\tilde{x}_e(n+1)$. Note that while BE, EE and ED can compute both predictions, BD can only produce $\tilde{x}_b(n+1)$.

The prediction loop poses severe difficulties on the design of scalable coding. There are several well known approaches to scalable predictive coding, depending on how this prediction is handled at the base and enhancement layers.

2.1. Approaches with "drift"

This class of approaches (A) allows a decoder to use a prediction different from the one used by the corresponding encoder. Each decoder employs the best available prediction. Hence, BD uses the prediction $\tilde{x}_b(n+1)$, while ED which has access to the enhancement bit stream uses the prediction $\tilde{x}_e(n+1)$. On the other hand, both the encoders, BE and EE, use the same prediction. In this case, "drift" is unavoidable. The term "drift" refers to a form of mismatch where the decoder uses a different prediction than the one assumed by the encoder. This mismatch tends to grow as the "corrections" provided by the encoder are misguiding, and hence, the decoder "drifts away" (similar to the use of open-loop prediction). At the encoder a choice must be made: If the encoder uses the base layer prediction, $\tilde{x}_b(n+1)$, then there is drift of the enhancement decoder. Use of the enhancement layer prediction, $\tilde{x}_e(n+1)$, at the encoder results in drift of the base decoder. See [2] for more about this approach and its shortcomings.

2.2. Approaches without "drift"

A second class of approaches (B) constrain each encoder/decoder pair to use the same prediction. Therefore BE and BD must use the base prediction $\tilde{x}_b(n+1)$, while EE and ED may use the enhanced prediction $\tilde{x}_e(n+1)$. Thus the encoder and decoder are always in step and the "drift" is eliminated.

The base-layer prediction $\tilde{x}_b(n+1)$ is easily obtained as the motion compensated previous base-layer reconstruction $\hat{x}_{mb}(n)$. BE compresses the residual $r_b(n) = x(n) - \tilde{x}_b(n)$ and produces $\hat{r}_b(n)$. BD produces the reconstruction $\hat{x}_b(n) = \tilde{x}_b(n) + \hat{r}_b(n)$.

However, an efficient way to generate the enhancement layer prediction $\tilde{x}_e(n+1)$ is a more difficult problem. The two main methods used in practice to produce

$\tilde{x}_e(n+1)$ are:

B1: Use the base-layer reconstruction of the next frame as the prediction, $\tilde{x}_e(n+1) = \hat{x}_b(n+1)$. Thus, EE, in effect, compresses the base-layer's reconstruction error of the next frame $x(n+1) - \hat{x}_b(n+1) = x(n+1) - \tilde{x}_b(n+1) - \hat{r}_b(n+1)$. See, e.g., [3] for further details. 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 current frame $\hat{x}_{me}(n)$ which is available to ED.

B2: Generate enhancement-layer prediction from the enhancement-layer reconstruction of previous frame after motion compensation, $\tilde{x}_e(n+1) = \hat{x}_{me}(n)$. Note that here EE does not exploit the knowledge of $\hat{r}_b(n+1)$ which is available to the enhancement layer. The reason is that due to the use of different prediction in the layers, the residuals are not necessarily correlated, and hence, the usefulness of the compressed base-layer residual to the enhancement-layer is largely compromised. The two layers are, in fact, separately encoded except for savings on unrepeated overhead information (such as motion vectors) [4].

In this paper we propose ways to exploit information available only to the enhancement layer, while taking full advantage of the base-layer reconstructed residual.

3. THE SWITCHED PREDICTION APPROACH

The objective is to upgrade the next frame prediction of the enhancement layer as much as possible by judiciously using information from $\hat{x}_{me}(n)$, with minimal conflict with the base-layer reconstructed frame $\hat{x}_b(n+1)$, whose information may be largely exploited by the enhancement-layer. EE computes a new predicted frame by combining transform coefficients from $\hat{x}_b(n+1)$ and from $\hat{x}_{me}(n)$, as depicted in Figure 2. The combining rule may depend on the reconstructed residual, $\hat{r}_b(n+1)$, and compression parameters of the base-layer (such as quantization step and threshold). The exact definition of the combination rule depends on the level of complexity allowed for the module. In this section we describe a low complexity option which consists of switching between transform coefficients from the two sources.

A major feature of DCT-based video coding is that the residual is quantized in the transform domain, where a large number of coefficients are typically quantized to zero (thresholded). The information on the positions of thresholded coefficients in the compressed base-layer residual (where $\hat{r}_b(n+1) = 0$), is available to the enhancement-layer. At these positions, the transform coefficient of the base-layer reconstruction, $\hat{x}_b(n+1) = \hat{x}_{mb}(n)$. However, we know that $\hat{x}_{me}(n)$ is typically a better estimate than $\hat{x}_{mb}(n)$. We can, therefore, improve the prediction at these positions by substituting $\hat{x}_b(n+1)$ with $\hat{x}_{me}(n)$.The switching rule is summarized as follows: *If the reconstructed base residual $\hat{r}_b(n+1)$ is zero in this position, select the coefficient from $\hat{x}_{me}(n)$. Else, select the coefficient from $\hat{x}_b(n+1)$.*

This switching rule provides an improved prediction to the enhancement layer without sacrificing any information available from the base-layer residual. No conflict with the base-layer reconstructed residual is possible because base-layer prediction is used wherever the reconstructed residual does not vanish. On the other hand, wherever the base-layer residual is quantized to zero, we can only gain by using the enhancement-layer coefficient.

This simple basic idea provides substantial gains over the known approaches to SNR scalability. Rather than discard the additional information available to the enhancement decoder, we judiciously exploit it wherever it does not interfere with the data provided through the compressed residual of the base-layer. It is easy to see that the proposed approach captures the advantages of both methods B1 and B2. Preliminary simulation results in the section 4 demonstrate the potential for gains from this approach.

The approach can be easily extended to more than two layers. The prediction at any layer is obtained by switching between the transform coefficients of the (motion compensated) previous frame reconstruction of that layer and the current reconstructed frame of the layer immediately below it. The gains in performance increase with the number of layers due to improved prediction at every layer.

4. SIMULATION RESULTS

We applied the proposed technique to scalable compression of a benchmark video sequence at various bit rates. The results are compared to those obtained by existing methods (B1) [3] and (B2) which was extracted from [4] (without the elements irrelevant to SNR scalability). The PSNR results are given in Table 1. These preliminary results are for two layer scalable coding and for various enhancement/base rate ratios. One should expect B1 to outperform B2 at low rate ratios, and the

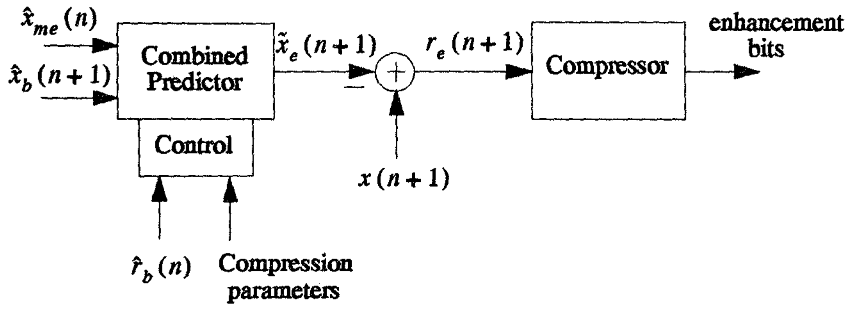

Figure 2: *The improved enhancement-layer prediction is obtained by combining information from both layers. The combining rule depends on the reconstructed base residual and the compression parameters.*

Rate (in Kbps)	Conventional B1	Conventional B2	Switched Prediction
64	30.66	30.47	31.44
128	32.46	33.99	34.61
256	35.22	37.84	38.27
512	39.17	42.09	42.49

Table 1: *Performance of different predictive scalable coding techniques on the sequence Carphone. The entries provide the average PSNR of reconstructed enhancement layer frames (in dB) for different enhancement layer rates. The base layer rate was fixed at 32 Kbps for all cases and the corresponding base layer PSNR was 29.62 dB (for all methods).*

opposite at high ratios (where the enhancement layer had much higher rate than the base layer). This is because at high ratios, the loss incurred by B2 for neglecting the base layer information becomes negligible.

However, the main observation is that the new scalable method outperforms both competitors at all ratios, and the gains vary from 0.4 dB to 3.4 dB in average PSNR of the reconstructed frames. The gains over the conventional approaches are expected to grow when scalability over a sequence of layers is implemented. Unfortunately, these results are not yet ready at the time of submission.

5. EXTENSIONS AND FUTURE WORK

Returning to Figure 2, we re-emphasize that it is possible to use more sophisticated combination rules than the simple switching rule we described above. In particular, a linear prediction of the current enhancement layer frame can be obtained from the current base layer reconstruction and the previous enhancement layer motion compensated reconstruction, given the compressed residual and the compression parameters. Such an extension increases the complexity, though it seems easily manageable, and allows us to soften the previous rule. We no longer restrict ourselves not to compromise the usefulness of the reconstructed residual, but instead we trade gains from enhancement layer prediction for gains from the base layer compressed residual, in an optimal way. This direction is currently under investigation, and we expect to publish performance results in a future paper.

6. 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 layer 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, 1996.