2012 IEEE International Conference on Multimedia and Expo

A UNIFIED ESTIMATION-THEORETIC FRAMEWORK FOR ERROR-RESILIENT SCALABLE VIDEO CODING

Jingning Han, Vinay Melkote, and Kenneth Rose*

Department of Electrical and Computer Engineering, University of California Santa Barbara, CA 93106
{jingning,melkote, rose}@ece.ucsb.edu

ABSTRACT

A novel scalable video coding (SVC) scheme is proposed for video transmission over lossy networks, which builds on an estimation-theoretic (ET) framework for optimal prediction and error concealment, given all available information from both the current base layer and prior enhancement layer frames. It incorporates a recursive end-to-end distortion estimation technique, namely, the *spectral coefficient-wise optimal recursive estimate* (SCORE), which accounts for all ET operations and tracks the first and second moments of decoder reconstructed transform coefficients. The overall framework enables optimization of ET-SVC systems for transmission over lossy networks, while accounting for all relevant conditions including the effects of quantization, channel loss, concealment, and error propagation. It thus resolves longstanding difficulties in combining truly optimal prediction and concealment with optimal end-to-end distortion and error-resilient SVC coding decisions. Experiments demonstrate that the proposed scheme offers substantial performance gains over existing error-resilient SVC systems, under a wide range of packet loss and bit rates.

Index Terms— Scalable video coding, error resilience, end-to-end distortion, error concealment

1. INTRODUCTION

Scalable video coding (SVC) is an attractive approach for applications that cater to receivers with varying reception bandwidth or for transmission over networks with diverse communication links [1, 2]. Typically, the SVC base layer consists of information about the video sequence that can be decoded independently to obtain a reconstruction of coarse quality, whereas the enhancement layers' information allows a decoder to successively refine the reconstruction. Enhancement layer packets may be dropped on-the-fly at intermediate network nodes to adjust the transmission rate, while retaining a baseline decoding quality. The base layer encoding is essentially the same as single layer video coding, where macroblocks are encoded after either motion-compensated prediction, or intra-frame prediction. Prediction at the enhancement layer, however, has access to information from both the base and enhancement layers. Standard approaches [1] perform the enhancement layer prediction in the pixel domain by selecting amongst the available prediction modes the one that minimizes the rate-distortion cost, and are inherently suboptimal: they cannot fully exploit information from both base and enhancement layers (see Sec. 2 for more details). As an alternative, an optimal enhancement layer prediction approach was proposed in [3], where the enhancement layer motion-compensated reference is optimally combined with base layer quantization information, in a suitably derived estimation-theoretic (ET) framework, directly in the transform domain. This approach, henceforth referred to as ET-SVC, substantially outperforms existing pixel domain prediction methods, in the context of lossless channel condition.

Practical deployment of video codecs often requires careful consideration of the impact of the potential channel distortion during transmission over packet-based networks, as well as the interaction with subsequent error concealment. Errors due to packet losses propagate via the prediction loop, and can significantly affect the reconstruction quality. A variant of the ET approach was proposed in [4] to efficiently conceal the lost enhancement layer packets of pre-compressed sequences, assuming a loss free guarantee for the base layer. However, due to the inability of end-to-end distortion estimation tools to accurately capture, at the encoder, the effects of such operations, it has never been included in the prediction loop for joint optimization of *encoding* decisions. In this work, the ET concealment method is further generalized to encompass lossy base layer settings, and is fully accounted for at the encoder, in conjunction with ET prediction and other error-resilient modes, within a rate-distortion optimization framework. It builds on and significantly expands the preliminary work in [5], where the derivation required substantial simplifying assumptions including the guarantee of lossless base layer, and simple (non-ET) concealment at the decoder.

In general, the base layer is assumed to be transmitted through a relatively reliable but capacity limited channel, unlike enhancement layers. This setting can be implemented via e.g., error correction codes, priority packetization, etc. A major strategy to achieve error resilience is thus to judiciously select the prediction mode (i.e., intra/inter-mode at the base layer, or inter-frame/inter-layer at the enhancement layer), and other encoding parameters, so that the end-to-end distortion (EED) versus rate tradeoff is optimized. EED measures the distortion in the decoder reconstruction, and accounts for the various components of the video compression and networking system, including quantization, packet loss, concealment, and error propagation. Estimating EED at the encoder is central to optimize its decisions. The recursive optimal per-pixel estimate (ROPE) [6] is an optimal EED estimation method that recursively calculates the first and second moments of reconstructed *pixels* via update equations that explicitly account for the prediction modes, concealment methods, channel uncertainties, etc. ROPE and its variants have been successfully incorporated in standard SVC coders, e.g., [7]-[10], and

This work is supported in part by Qualcomm Inc., and by the NSF under grant CCF-0917230.

*Vinay Melkote is now with Dolby Laboratories Inc., 100 Potrero Avenue, San Francisco, CA 94103.

978-0-7695-4711-4/12 $26.00 © 2012 IEEE
DOI 10.1109/ICME.2012.76

significantly improved the coding performance.

We note that the applicability of ROPE is inherently limited to account for operations that are recursive in the pixel domain. While this is sufficient for standard (suboptimal) SVC coders, the ET approaches achieve optimality by performing both prediction and concealment directly in the transform domain (see discussion in Sec. 3). Combining optimal prediction and concealment with optimal encoding and error-resilience decisions has long been an open challenge. This difficulty has been preliminarily addressed in [5] (under simplifying assumptions) and is fully resolved in this work by an error-resilient variant of ET-SVC for transmission over lossy networks, which employs a ROPE-like EED estimate technique that performs its update recursions entirely in the transform domain, and is naturally suitable to capture such ET operations. The EED estimation leverages the *spectral coefficient-wise optimal recursive estimate* (SCORE). SCORE was initially proposed in [11] for single layer video coding, and is readily applicable to the base layer in SVC. It recursively computes up to second moments of decoder reconstructed *transform coefficients* in rough analogy to what ROPE does per-pixel. We extend the scope of SCORE to encompass the challenging setting of ET prediction and concealment. In particular, the non-linear recursive transform domain operations of ET prediction and concealment are incorporated into the SCORE update equations via a quadratic approximation, conditioned on the statistical knowledge of the current base layer reconstruction and prior enhancement layer reference. The coding parameters in this ET-SVC scheme are then optimized by exploiting the EED estimates provided by SCORE. It is experimentally demonstrated that the proposed overall ET-SVC-SCORE coder substantially outperforms standard SVC optimized by ROPE, under various settings of packet loss and bit rates of base and enhancement layers. We note that in the special case of guaranteed loss-free base layer, the uncertainty of base layer reconstructions vanishes and the update recursion of ET prediction subsumes to the simplified scheme discussed in [5], where the error concealment simply consists of "upward" replacement using the base layer reconstruction. Note that the proposed framework employs ET concealment when feasible, and that its effects are fully taken into account by SCORE at the encoder to jointly optimize the coding decisions, thereby it fully exploits the potential of the ET approach.

Other relevant work includes allowing the base layer to be predicted from prior enhancement layer reconstructions to improve the expected quality of point-to-point video transmission, e.g., [12, 13], for the setting where both layers are received albeit at different packet loss rates. In this paper, we focus on the common broadcast settings, where multiple users are served at different resolution layers, hence the base layer itself should be coded/optimized as a self sufficient layer, which can be decoded at its prescribed quality without access to the enhancement layers. We note that while the proposed scheme is implemented in H.264/SVC reference framework to demonstrate its efficacy, the basic principles are more generally applicable to other predictive scalable video coders, such as VP8 and HEVC.

2. BACKGROUND: STANDARD SCALABLE VIDEO CODERS

For exposition simplicity, we consider the two-layer quality scalable setting throughout this paper, although the proposed concepts are extensible to more layers and other types of scalability [14]. The H.264/SVC coder compresses the base layer as a single bit-stream, and employs a single-loop design to code the enhancement layer, where the decoder need not buffer its base layer reconstruction to produce the enhancement layer signal. Particularly, the enhancement layer coder starts with motion compensation from previously reconstructed frames in the same layer to generate a prediction residual block. It then adaptively decides whether to further subtract the base layer reconstructed residual from this residual block before transformation and quantization [1, 2]. In earlier standards such as H.263++ the enhancement layer prediction switches between (motion compensated) prior enhancement layer reconstruction and current base layer reconstruction, or a linear combination thereof, in what is referred to as a multi-loop design [15]. It has been recognized that multi-loop design performs better than single-loop at the expense of more decoding complexity [2]. Since the main focus of this paper is on optimality in coding performance, the H.264/SVC codec is modified to the better performing multi-loop design, while retaining the other advanced components and capabilities, such as sub-pixel motion compensation, context adaptive binary arithmetic coding, etc. ROPE is then incorporated in this framework to optimize SVC encoding decisions as explained in [7].

3. ET-SVC BUILDING BLOCKS

The principles underlying the ET approach originally appeared in [3], which we briefly summarize here for enhancement layer prediction and concealment, respectively.

3.1. Estimation-Theoretic Prediction

Let x_n denote the value of a particular transform coefficient in a block of the current frame. Since the prediction is initially performed at the *encoder*, the notation in this subsection will always refer to encoder entities, noting that as long as the channel is lossless, they will be perfectly reproducible at the decoder. Let $\hat{x}^b_{n-1}$ denote the transform coefficient of the same frequency as x_n, in the base layer motion compensated reference, which is obtained from the reconstruction of the previous frame. Thus the operation of the standard base layer encoder is equivalent to a quantization of $(x_n - \hat{x}^b_{n-1})$ to produce the index i^b_n. Let $[a_n, b_n)$ be the quantization interval associated with index i^b_n. Clearly, the statement $x_n \in \mathcal{I}^b_n = [\hat{x}^b_{n-1} + a_n, \hat{x}^b_{n-1} + b_n)$ captures all the information on x_n provided by the base layer, namely, it specifies the interval in which x_n must reside. When encoding the enhancement layer of x_n, the encoder also has access to transform coefficient $\hat{x}^e_{n-1}$ of the motion-compensated reference block, generated from the previously reconstructed frame. In [3], the prior enhancement layer information $\hat{x}^e_{n-1}$ is combined with the base layer interval $\mathcal{I}^b_n$, in an estimation-theoretic framework to obtain the optimal prediction for coefficient x_n. It is important to note that although the enhancement layer reconstruction information ($\hat{x}^e_{n-1}$) can be equivalently expressed in the spatial pixel domain, the quantization interval $\mathcal{I}^b_n$ does not map to the spatial domain in a simple and useful way. Thus, ad hoc spatial domain linear combinations of base layer residual or reconstruction, with prior enhancement layer reconstruction, as employed by current and prior standard SVCs cannot achieve optimal enhancement layer prediction.

Motion-compensated predictive video coders typically model

blocks of pixels along the same motion trajectory in consecutive frames as an autoregressive (AR) process. Motion compensation is employed to align these pixel blocks, and pixel domain subtraction removes temporal redundancies. An equivalent alternative viewpoint that the DCT coefficients of corresponding blocks form an AR process per coefficient or frequency, was adopted in [3]. Thus x_n (at a given frequency) and the corresponding motion-compensated transform coefficient from previous frame x_{n-1} conform to the first order AR model: $x_n = \rho x_{n-1} + z_n$, where $\{z_n\}$ are independent and identically distributed (i.i.d) innovations of the process with probability density function (pdf) $p_Z(z)$. To mimic what is implicitly assumed by pixel domain motion-compensated prediction, we will arbitrarily (and for simplicity) assume the maximum correlation coefficient $\rho \approx 1$ at all frequencies. The above transform domain AR process perspective provides the advantage that the motion compensated $\hat{x}^e_{n-1}$, and the quantization interval $\mathcal{I}^b_n$, can now be combined to produce the optimal estimate.

Assuming that $\hat{x}^e_{n-1} \approx x_{n-1}$, we obtain the conditional pdf $p(x_n|\hat{x}^e_{n-1}) \approx p_Z(x_n - \hat{x}^e_{n-1})$. In the absence of additional base layer information, the best prediction of x_n would just be $\hat{x}^e_{n-1}$, the default enhancement layer inter-frame prediction. But the base layer indicates that $x_n \in \mathcal{I}^b_n$, which refines the conditional pdf of x_n to

$$p(x_n|\hat{x}^e_{n-1}, \mathcal{I}^b_n) \approx \begin{cases} \frac{p_Z(x_n - \hat{x}^e_{n-1})}{\int_{\mathcal{I}^b_n} p_Z(x_n - \hat{x}^e_{n-1}) dx_n} & x_n \in \mathcal{I}^b_n \\ 0 & else \end{cases} \quad (1)$$

Note that the above is equivalent to centering $p_Z(z)$ at $\hat{x}^e_{n-1}$, restricting it to the interval $\mathcal{I}^b_n$ (a non-linear operation), and then normalizing to obtain a valid pdf. The optimal predictor at the enhancement layer is given by [3]

$$f(\hat{x}^e_{n-1}, \mathcal{I}^b_n) = E[x_n|\hat{x}^e_{n-1}, \mathcal{I}^b_n], \quad (2)$$

the centroid of the above pdf in the interval $\mathcal{I}^b_n$. The residual $(x_n - f(\hat{x}^e_{n-1}, \mathcal{I}^b_n))$ is then quantized and encoded in the enhancement layer.

3.2. Estimation-Theoretic Concealment

To reproduce the ET prediction (2), the decoder needs information from both base and enhancement layer packets. Whenever either is lost, the decoder has to conceal the missing blocks. Since the base layer packet loss rate is typically much lower than that of enhancement layer, it is reasonable to assume that the drift effect on base layer reconstruction is smaller than enhancement layer.

Case 1: Enhancement layer packet is lost while base layer packet is received

In this case, the decoder does not know the prediction mode of the enhancement layer macroblock, and the concealment operation is completely dependent on the base layer conditions. If the macroblock is inter-coded at the base layer, the decoder has access to the quantization index i^b_n and the motion information (of base layer) to perform ET concealment; otherwise, the upward replacement of base layer reconstruction will be used as concealment. We inherit the basic notation from the previous subsection for encoder quantities, but must additionally denote reconstruction at the decoder, which may differ due to loss. Let $\tilde{x}^b_n$ be the decoder base layer reconstruction of transform coefficient x_n. The quantization interval $[a_n, b_n)$ associated with index i^b_n is identical to that of the encoder, but is now shifted by $\tilde{x}^b_n$ to produce the reconstruction interval $\tilde{\mathcal{I}}^b_n = [\tilde{x}^b_n + a_n, \tilde{x}^b_n + b_n)$. The motion-compensated reference $\tilde{x}^c_{n-1}$ is generated from the decoder enhancement layer reconstruction of the prior frame, using motion information from the current base layer[1]. The ET concealment is the constructed as $f(\tilde{x}^c_{n-1}, \tilde{\mathcal{I}}^b_n) = E[x_n|\tilde{x}^c_{n-1}, \tilde{\mathcal{I}}^b_n]$, where the conditional pdf is defined by (1).

Case 2: Enhancement layer packet is received but base layer packet is lost

The enhancement layer motion information is known to the decoder, and can be used to generate a motion-compensated reference from prior decoded frames, i.e., $(\tilde{x}^e_{n-1} + \hat{r}^e_n)$, where $\hat{r}^e_n$ denotes the quantized residual, which is then employed as concealment.

Case 3: Both packets lost

This event is of significantly lower probability. The decoder has no choice but to use upward replacement with base layer reconstruction, as the enhancement layer concealment.

In the implementations, but without loss of theoretical generality, we will assume that the innovation pdf is Laplacian, i.e., $p_Z(z_n) = \frac{1}{2}\lambda e^{-\lambda|z_n|}$, where the parameter λ is frequency dependent. It is useful to note that the Laplacian distribution assumption offers an easily derived closed form of the above expectation, due to its memoryless property.

4. SPECTRAL COEFFICIENT-WISE OPTIMAL RECURSIVE ESTIMATE IN ET-SVC

Errors due to packet loss generally propagate in time and across layers through the prediction loop. A natural tool to enhance error-resilience is to provide the option to occasionally cut off temporal prediction, via intra-, inter-layer prediction, etc. Encoding decisions, including the prediction mode and quantization parameters, should optimize the tradeoff between rate and EED, and hence critically depend on accurate estimation of EED. We therefore extend the basic SCORE approach [11] to encompass the ET-SVC setting. We assume that the base and enhancement layer packet loss rates, p_b and p_e, are known to the encoder.

4.1. Base Layer

The base layer of an SVC is essentially the same as the regular single layer coder. Thus, the SCORE update recursions are akin to those discussed in [11], which we briefly summarize next.

Let $x^{k,m}_n$ denote the original value of transform coefficient m in block k of frame n. Denote the encoder and decoder base layer reconstructions by $\hat{x}^{k,m}_{n,b}$ and $\tilde{x}^{k,m}_{n,b}$, respectively. Similarly let $\hat{r}^{k,m}_{n,b}$ be the quantized transform coefficient residual, whose value is coded and transmitted to the decoder. The motion-compensated reference block is potentially 'off-grid' in the prior frame. We use $\hat{u}^{k,m}_{n,b}$ and $\tilde{u}^{k,m}_{n,b}$ to denote the encoder and decoder reconstructions of this coefficient. Note that while $\hat{u}^{k,m}_{n,b}$ and $\tilde{u}^{k,m}_{n,b}$ are indexed by n and k to indicate the location on the current frame they provide reference for, they are in fact determined by the encoder and decoder reconstructions of frame $n-1$. As far as the encoder is concerned, $\tilde{x}^{k,m}_{n,b}$ and $\tilde{u}^{k,m}_{n,b}$ are random variables, due to the stochastic nature of packet

[1] We use $\tilde{x}^c_{n-1}$ to denote that this reference uses potentially different motion information than the enhancement layer would normally use to produce $\tilde{x}^e_{n-1}$.

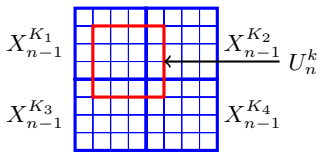

Fig. 1. An off-grid block (red) overlaps 4 on-grid blocks (blue).

loss. Hence the encoder estimates the EED of this transform coefficient at the base layer as the expectation:

$$\begin{aligned}\delta_{n,b}^{k,m} &= E\{(x_n^{k,m} - \tilde{x}_{n,b}^{k,m})^2\} \\ &= (x_n^{k,m})^2 - 2x_n^{k,m}E\{\tilde{x}_{n,b}^{k,m}\} + E\{(\tilde{x}_{n,b}^{k,m})^2\}. \quad (3)\end{aligned}$$

The computation of $\delta_{n,b}^{k,m}$ requires the first and second moments of the decoder reconstruction $\tilde{x}_{n,b}^{k,m}$. SCORE recursively evaluates these moments for every transform coefficient of on-grid blocks in the frame, via update recursions depending on the prediction modes.

Intra-Mode: The packet containing the current block is received with probability $1 - p_b$, producing $\tilde{x}_{n,b}^{k,m} = \hat{x}_{n,b}^{k,m}$. If the packet is lost at probability p_b, the decoder uses 'slice copy' concealment, i.e., $\tilde{x}_{n,b}^{k,m} = \hat{x}_{n-1,b}^{k,m}$. The moments are thus computed as:

$$\begin{aligned}E\{\tilde{x}_{n,b}^{k,m}\}(I) &= (1-p_b)(\hat{x}_{n,b}^{k,m}) + p_b E\{\tilde{x}_{n-1,b}^{k,m}\}\,, \\ E\{(\tilde{x}_{n,b}^{k,m})^2\}(I) &= (1-p_b)(\hat{x}_{n,b}^{k,m})^2 + p_b E\{(\tilde{x}_{n-1,b}^{k,m})^2\}. \quad (4)\end{aligned}$$

Inter-Mode: The packet contains motion information and the residual $\hat{r}_{n,b}^{k,m}$. If the packet arrives, the decoder uses the motion information to generate $\tilde{u}_{n,b}^{k,m}$ from the previous *decoded* frame, which is potentially different from the decoder's $\hat{u}_{n,b}^{k,m}$. Therefore,

$$\begin{aligned}E\{\tilde{x}_{n,b}^{k,m}\}(P) &= (1-p_b)(\hat{e}_{n,b}^{k,m} + E\{\tilde{u}_{n,b}^{k,m}\}) + p_b E\{\tilde{x}_{n-1,b}^{k,m}\}\,, \\ E\{(\tilde{x}_{n,b}^{k,m})^2\}(P) &= (1-p_b)((\hat{r}_{n,b}^{k,m})^2 + \hat{r}_{n,b}^{k,m}E\{\tilde{u}_{n,b}^{k,m}\} \\ &\quad + E\{(\tilde{u}_{n,b}^{k,m})^2\}) + p_b E\{(\tilde{x}_{n-1,b}^{k,m})^2\}. \quad (5)\end{aligned}$$

The above implies that the required decoder reconstruction moments can be computed as long as the moments $E\{\tilde{u}_{n,e}^{k,m}\}$ and $E\{(\tilde{u}_{n,e}^{k,m})^2\}$ of the potentially off-grid motion compensation reference are available. We thus provide a complementary method to derive off-grid moments from the available moments of on-grid blocks in frame $n-1$. An off-grid block overlaps at most four on-grid blocks (Fig. 1). Let block U_n^k shown in the figure be the reference block for the block k in the current frame n. This block, located in frame $n-1$, overlaps with on-grid blocks $X_{n-1}^{k_i}$ in the frame. The decoder base layer reconstruction of block U_n^k is associated with coefficients $\tilde{u}_{n,b}^{k,m}$. The linearity of the transform implies that there exists a set of constants $a_{i,m}$ named *construction constants*, such that

$$\tilde{u}_{n,b}^{k,m} = \sum_{i=1}^{4}\sum_{m=0}^{15} a_{i,m}\tilde{x}_{n-1,b}^{k_i,m}\,.$$

The construction constants only depend on the relative position of U_n^k in this four block grid. The first moment of $\tilde{u}_{n,b}^{k,m}$ is given by

$$E\{\tilde{u}_{n,b}^{k,m}\} = \sum_{i=1}^{4}\sum_{m=0}^{15} a_{i,m}E\{\tilde{x}_{n-1,b}^{k_i,m}\}\,.$$

Computation of the second moment of $\tilde{u}_{n,b}^{k,m}$ involves cross-correlation terms for pairs of transform coefficients in on-grid blocks:

$$E\{(\tilde{u}_{n,b}^{k,m})^2\} = \sum_{i=1}^{4}\sum_{j=1}^{4}\sum_{m=0}^{15}\sum_{l=0}^{15} a_{i,m}a_{j,l}E\{\tilde{x}_{n-1,b}^{k_i,m}\tilde{x}_{n-1,b}^{k_j,l}\}\,.$$

The computationally intensive calculation of these cross-correlation terms is circumvented by the 'uncorrelatedness' approximation for DCT coefficients, whose validity has been demonstrated in [11]: $E\{\tilde{x}_{n,b}^{k_i,m}\tilde{x}_{n,b}^{k_j,l}\} \approx E\{\tilde{x}_{n,b}^{k_i,m}\}E\{\tilde{x}_{n,b}^{k_j,l}\}, j \neq i$ or $l \neq m$. Thus the recursions of (5) are complete.

Once these moments are known, the EED can be computed via (3), and employed to select base layer coding mode and parameters so as to optimize the rate-EED cost. We note that ideally a joint optimization of bit-allocation across layers might further improve the overall SVC performance, at the expense of significant increment in encoder complexity. Since this paper is mainly focused on the enhancement layer optimization, the base layer is optimized as a single layer without consideration of other layers. In the our experiments, all competing SVC schemes will use an identical base layer coder.

4.2. Enhancement Layer

Let $\hat{x}_{n,e}^{k,m}$ and $\tilde{x}_{n,e}^{k,m}$ denote the encoder and decoder enhancement layer reconstructions of $x_n^{k,m}$, respectively. Also let $\hat{r}_{n,e}^{k,m}$ denote the quantized enhancement layer prediction residual. The enhancement layer motion-compensated reference generated by the encoder and decoder are denoted by $\hat{u}_{n,e}^{k,m}$ and $\tilde{u}_{n,e}^{k,m}$, respectively. The EED of this transform coefficient is thus given by

$$\delta_{n,e}^{k,m} = E\{(x_n^{k,m} - \tilde{x}_{n,e}^{k,m})^2\}\,.$$

Again the computation of $\delta_{n,e}^{k,m}$ only requires the first and second moments of the decoder enhancement layer reconstruction $\tilde{x}_{n,e}^{k,m}$. We derive SCORE recursion formulae for the two additional prediction modes employed by the enhancement layer. Note that the base layer reconstruction moments are available to the enhancement layer.

Inter-Layer Mode: The packet containing the quantized prediction residual $\hat{r}_{n,e}^{k,m}$ is received with probability $1 - p_e$, allowing the reconstruction $\tilde{x}_{n,e}^{k,m} = \tilde{x}_{n,b}^{k,m} + \hat{r}_{n,e}^{k,m}$. When it is lost, with probability p_e, the decoder conceals the missing block, conditioned on the base layer prediction mode and packet arrival as discussed in Sec. 3.2. Hence, if the base layer macroblock is intra-coded,

$$\begin{aligned}E\{\tilde{x}_{n,e}^{k,m}\}(IL) &= (1-p_e)(E\{\tilde{x}_{n,b}^{k,m}\} + \hat{r}_{n,e}^{k,m}) + p_e E\{\tilde{x}_{n,b}^{k,m}\}\,, \\ E\{(\tilde{x}_{n,e}^{k,m})^2\}(IL) &= (1-p_e)(E\{(\tilde{x}_{n,b}^{k,m})^2\} + 2\hat{r}_{n,e}^{k,m}E\{\tilde{x}_{n,b}^{k,m}\} \\ &\quad + (\hat{r}_{n,e}^{k,m})^2) + pE\{(\tilde{x}_{n,b}^{k,m})^2\}\,.\end{aligned}$$

If the base layer macroblock is inter-coded, then with probability $(1 - p_b)$ the decoder receives the base layer packet, uses its motion information to generate an enhancement layer motion-compensated reference $\tilde{x}_{n,c}^{k,m}$, and performs ET concealment. The recursion in this

case is thus:

$$\begin{aligned}E\{\tilde{x}_{n,e}^{k,m}\}(IL) &= (1-p_e)(E\{\tilde{x}_{n,b}^{k,m}\}+\hat{r}_{n,e}^{k,m}) \\ &\quad +p_e((1-p_b)E\{f(\tilde{\mathcal{I}}_n^b,\tilde{x}_{n-1,c}^{k,m})\}+p_bE\{\tilde{x}_{n,b}^{k,m}\})\,, \\ E\{(\tilde{x}_{n,e}^{k,m})^2\}(IL) &= (1-p_e)(E\{(\tilde{x}_{n,b}^{k,m})^2\}+2\hat{r}_{n,e}^{k,m}E\{\tilde{x}_{n,b}^{k,m}\} \\ &\quad +(\hat{r}_{n,e}^{k,m})^2)+p_e((1-p_b)E\{(f(\tilde{\mathcal{I}}_n^b,\tilde{x}_{n-1,c}^{k,m}))^2\} \\ &\quad +p_bE\{(\tilde{x}_{n,b}^{k,m})^2\})\,.\end{aligned}$$

ET Prediction Mode: The decoder requires both base and enhancement layer packets to reconstruct the ET-coded coefficient as $(f(\tilde{\mathcal{I}}_n^b,\tilde{x}_{n-1,e}^{k,m})+\hat{r}_{n,e}^{k,m})$. If the base layer packet is lost but enhancement layer packet is received, the decoder will reproduce $(\tilde{x}_{n-1,e}^{k,m}+\hat{r}_{n,e}^{k,m})$. Otherwise, the decoder will choose ET concealment or upward replacement, depending on the base layer coding mode. Therefore, the update recursions of ET prediction mode are stated as follows. For an intra-coded base layer macroblock:

$$\begin{aligned}E\{\tilde{x}_{n,e}^{k,m}\}(ET) &= (1-p_e)(\hat{r}_{n,e}^{k,m}+(1-p_b)E\{f(\tilde{\mathcal{I}}_n^b,\tilde{u}_{n,e}^{k,m})\} \\ &\quad +p_bE\{\tilde{x}_{n-1,e}^{k,m}\})+p_eE\{\tilde{x}_{n,b}^{k,m}\} \\ E\{(\tilde{x}_{n,e}^{k,m})^2\}(ET) &= (1-p_e)((\hat{r}_{n,e}^{k,m})^2 \\ &\quad +2\hat{r}_{n,e}^{k,m}((1-p_b)E\{f(\tilde{\mathcal{I}}_n^b,\tilde{u}_{n,e}^{k,m})\}+p_bE\{\tilde{x}_{n-1,e}^{k,m}\}) \\ &\quad +(1-p_b)E\{(f(\tilde{\mathcal{I}}_n^b,\tilde{u}_{n,e}^{k,m}))^2\}+p_bE\{(\tilde{x}_{n-1,e}^{k,m})^2\}) \\ &\quad +p_eE\{(\tilde{x}_{n,b}^{k,m})^2\}\,. \qquad (6)\end{aligned}$$

For an inter-coded base layer macroblock:

$$\begin{aligned}E\{\tilde{x}_{n,e}^{k,m}\}(ET) &= (1-p_e)(\hat{r}_{n,e}^{k,m}+(1-p_b)E\{f(\tilde{\mathcal{I}}_n^b,\tilde{u}_{n,e}^{k,m})\} \\ &\quad +p_bE\{\tilde{x}_{n-1,e}^{k,m}\})+p_e(p_bE\{\tilde{x}_{n,b}^{k,m}\} \\ &\quad +(1-p_b)E\{f(\tilde{\mathcal{I}}_n^b,\tilde{u}_{n,e}^{k,m})\}) \\ E\{(\tilde{x}_{n,e}^{k,m})^2\}(ET) &= (1-p_e)((\hat{r}_{n,e}^{k,m})^2 \\ &\quad +2\hat{r}_{n,e}^{k,m}((1-p_b)E\{f(\tilde{\mathcal{I}}_n^b,\tilde{u}_{n,e}^{k,m})\}+p_bE\{\tilde{x}_{n-1,e}^{k,m}\}) \\ &\quad +(1-p_b)E\{(f(\tilde{\mathcal{I}}_n^b,\tilde{u}_{n,e}^{k,m}))^2\}+p_bE\{\tilde{x}_{n-1,e}^{k,m}\}). \\ &\quad +p_e((1-p_b)E\{(f(\tilde{\mathcal{I}}_n^b,\tilde{u}_{n,e}^{k,m}))^2\} \\ &\quad +p_bE\{(\tilde{x}_{n,b}^{k,m})^2\})\,. \qquad (7)\end{aligned}$$

The off-grid moments of $\tilde{u}_{n,e}^{k,m}$ and $\tilde{x}_{n-1,e}^{k,m}$ can be generated as shown earlier for the base layer. The above update equations also involve first and second moments of $f(\tilde{\mathcal{I}}_n^b,\tilde{u}_{n,e}^{k,m})$, a non-linear function whose *exact* evaluation via recursive update equations is highly complex. Note, however, that $\tilde{\mathcal{I}}_n^b$ is linear in $\tilde{x}_{n,b}^{k,m}$. Therefore, we approximate $f(x,u)$ by its Taylor series expansion about $(E\{\tilde{x}_{n,b}^{k,m}\},E\{\tilde{u}_{n,e}^{k,m}\})$, retaining only up to the second order terms:

$$\begin{aligned}f(x,u) &\approx f(E\{\tilde{x}_{n,b}^{k,m}\},E\{\tilde{u}_{n,e}^{k,m}\}) \\ &+ (u-E\{\tilde{u}_{n,e}^{k,m}\})\frac{df(x,u)}{du}|_{(E\{\tilde{x}_{n,b}^{k,m}\},E\{\tilde{u}_{n,e}^{k,m}\})} \\ &+ (x-E\{\tilde{x}_{n,b}^{k,m}\})\frac{df(x,u)}{dx}|_{(E\{\tilde{x}_{n,b}^{k,m}\},E\{\tilde{u}_{n,e}^{k,m}\})} \\ &+ \frac{(u-E\{\tilde{u}_{n,e}^{k,m}\})^2}{2}\frac{d^2f(x,u)}{du^2}|_{(E\{\tilde{x}_{n,b}^{k,m}\},E\{\tilde{u}_{n,e}^{k,m}\})} \\ &+ \frac{(x-E\{\tilde{x}_{n,b}^{k,m}\})^2}{2}\frac{d^2f(x,u)}{dx^2}|_{(E\{\tilde{x}_{n,b}^{k,m}\},E\{\tilde{u}_{n,e}^{k,m}\})} \\ &+(u-E\{\tilde{u}_{n,e}^{k,m}\})(x-E\{\tilde{x}_{n,b}^{k,m}\})\frac{d^2f(x,u)}{dxdu}|_{(E\{\tilde{x}_{n,b}^{k,m}\},E\{\tilde{u}_{n,e}^{k,m}\})} \qquad (8)\end{aligned}$$

For the example of the Laplace-Markov model (see Sec. 3), $f(x,u)$ can be written in closed form, and thus its first and second order partial derivatives can be explicitly evaluated. Taking expectation of both sides of (8) and plugging $x=\tilde{x}_{n,b}^{k,m}$ $u=\tilde{u}_{n,e}^{k,m}$ yields first moment of $f(\tilde{\mathcal{I}}_n^b,\tilde{u}_{n,e}^{k,m})$,

$$\begin{aligned}E\{f(\tilde{\mathcal{I}}_n^b,\tilde{u}_{n,e}^{k,m})\} &\approx f(E\{\tilde{x}_{n,b}^{k,m}\},(E\{\tilde{u}_{n,e}^{k,m}\}) \\ &+ \frac{E\{(u-E\{\tilde{u}_{n,e}^{k,m}\})^2\}}{2}\frac{d^2f(x,u)}{du^2}|_{(E\{\tilde{x}_{n,b}^{k,m}\},E\{\tilde{u}_{n,e}^{k,m}\})} \\ &+ \frac{E\{(x-E\{\tilde{x}_{n,b}^{k,m}\})^2\}}{2}\frac{d^2f(x,u)}{dx^2}|_{(E\{\tilde{x}_{n,b}^{k,m}\},E\{\tilde{u}_{n,e}^{k,m}\})} \\ &+E\{(u-E\{\tilde{u}_{n,e}^{k,m}\})(x-E\{\tilde{x}_{n,b}^{k,m}\})\}\frac{d^2f(x,u)}{dxdu}|_{(E\{\tilde{x}_{n,b}^{k,m}\},E\{\tilde{u}_{n,e}^{k,m}\})}\,,\end{aligned}$$

where the first three terms are readily obtainable from the known first and second moments of $\tilde{x}_{n,b}^{k,m}$ and $\tilde{u}_{n,e}^{k,m}$, the last term however involves the cross correlation of the two. Since both of them are highly correlated with the reference sample of $\tilde{x}_{n,b}^{k,m}$ in the base layer reconstruction of prior frame, we simply assume the maximum correlation between them (from Schwarz inequality) and approximate

$$\begin{aligned}&E\{(u-E\{\tilde{u}_{n,e}^{k,m}\})(x-E\{\tilde{x}_{n,b}^{k,m}\})\} \\ &\approx \sqrt{E\{(u-E\{\tilde{u}_{n,e}^{k,m}\})^2\}}\sqrt{E\{(x-E\{\tilde{x}_{n,b}^{k,m}\})^2\}}\,. \qquad (9)\end{aligned}$$

The value of $E\{(f(\tilde{\mathcal{I}}_n^b,\tilde{u}_{n,e}^{k,m})^2\}$ can be obtained similarly. Therefore, the update recursions of ET prediction mode are complete.

5. SIMULATION RESULTS

Having established the ET prediction and concealment building blocks and the SCORE approach for tracking the EED, we now evaluate the end-to-end performance obtained when SCORE's EED estimates are employed to optimize the (enhancement layer) coding decisions of ET-SVC. Let $D_{n,e}^k(q,\mu)$ and $B_{n,e}^k(q,\mu)$ denote the EED and bit costs incurred in encoding macroblock k of frame n at the enhancement layer with quantization parameter (QP) q and prediction mode μ. All macroblocks in the frame share the same QP, denoted by $q_{n,e}$. The optimization problem is formulated as the per-macroblock minimization:

$$\mu_{n,e}^k(\lambda,q)=\arg\min_{\mu}\{D_{n,e}^k(q,\mu)+\lambda B_{n,e}^k(q,\mu)\},$$

and the subsequent per-frame minimization:

$$q_{n,e}(\lambda)=\arg\min_{q}\sum_k D_{n,e}^k(q,\mu_{n,e}^k)+\lambda B_{n,e}^k(q,\mu_{n,e}^k),$$

where λ is a Lagrangian multiplier whose value is fixed for the entire sequence in our simulation. Varying λ provides an operational rate-distortion curve. The proposed ET-SVC codec whose coding decisions are optimized by SCORE is denoted ET-SVC-SCORE. We also modified the H.264/SVC reference to employ multi-loop design, while retaining its advanced coding tools, e.g., sub-pixel motion compensation, context adaptive binary arithmetic coding, etc., and whose decisions are optimized using EED estimates provided by ROPE [7]. The overall reference system is denoted by H.264/MLOOP-ROPE. For fair comparison, the same base layer optimized by ROPE is shared by both SVC schemes. Note that ET-SVC-SCORE which normally would not use ROPE also embeds

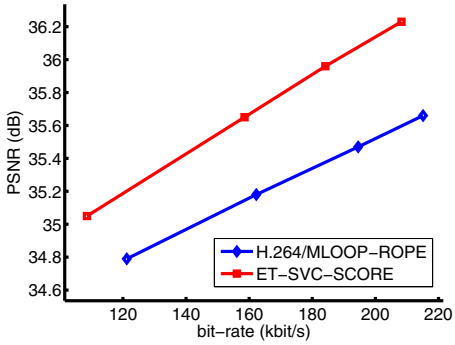

Fig. 2. End-to-end performance versus enhancement layer bit rate, on sequence $foreman$ at $QCIF$ resolution: the base layer is encoded at $128kbps$ and is transmitted with packet loss rate 1%. The enhancement layer packet loss rate is 5%.

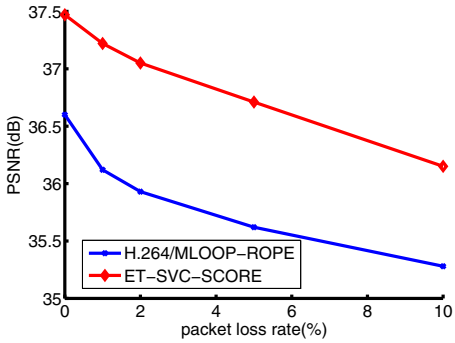

Fig. 3. End-to-end performance versus enhancement layer packet loss rate, on sequence $coastguard$ at $QCIF$ resolution: the base layer bit-rate is $170kbps$, and is transported at packet loss rate 1%; the enhancement layer bit rate is $340kbps$.

SCORE in the base layer to capture the moments needed for enhancement layer use, but without affecting the coding decisions of base layer.

The rate-distortion performance of sequence $foreman$ at $QCIF$ resolution is shown in Fig. 2. To demonstrate the coding performance under various channel conditions, sequence $coastguard$ at $QCIF$ resolution is encoded with fixed enhancement layer bitrate and is evaluated at different packet loss rates. The performance shown in Fig.3 demonstrates that the proposed ET-SVC-SCORE scheme substantially outperforms the competition across a wide range of packet loss rates. We note that similar coding gains are also observed on other sequences.

6. CONCLUSION

A novel error-resilient SVC scheme is proposed that achieves two optimality goals. It incorporates optimal (non-linear) enhancement layer prediction and concealment that exploit all available information from both the base and enhancement layers. It complements this with a recursive end-to-end distortion estimate that necessarily operates in the spectral domain, and which accounts for compression, packet loss, error propagation, and concealment. Simulations provide evidence for substantial performance gains of the overall SVC system.

7. REFERENCES

[1] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Trans. Circ. Sys. Video Tech.*, vol. 17, pp. 1103–1120, Sep. 2007.

[2] C. A. Segall and G. J. Sullivan, "Spatial scalability within the H.264/AVC scalable video coding extension," *IEEE Trans. Circ. Sys. Video Tech.*, vol. 17, no. 9, pp. 1121–1135, Sep. 2007.

[3] K. Rose and S. L. Regunathan, "Toward optimality in scalable predictive coding," *IEEE Trans. Img. Proc.*, vol. 10, no. 7, pp. 965–976, July 2001.

[4] R. Zhang, S. L. Regunathan, and K. Rose, "Optimal estimation for error concealment in scalable video coding," *Proc. Asilomar Conf. Signals, Systems, and Computers*, pp. 1374–1378, Oct. 2000.

[5] J. Han, V. Melkote, and K. Rose, "A unified framework for spectral domain prediction and end-to-end distortion estimation in scalable video coding," *Proc. IEEE ICIP*, pp. 3278–3281, Sep. 2011.

[6] R. Zhang, S. L. Regunathan, and K. Rose, "Video coding with optimal inter/intra-mode switching for packet loss resilience," *IEEE Jrnl. Sel. Areas Comm.*, vol. 18, pp. 966–976, June 2000.

[7] S. L. Regunathan, R. Zhang, and K. Rose, "Scalable video coding with robust mode selection," *Sig. Proc.: Image Comm.*, vol. 16, pp. 725–732, May 2001.

[8] F. Wu, S. Li, R. Yaw, X. Sun, and Y.-Q. Zhang, "Efficient and universal scalable video coding," *Proc. IEEE ICIP*, vol. 2, pp. 37–40, Sep. 2002.

[9] A. Leontaris and P. C. Cosman, "Drift-resistant SNR scalable video coding," *IEEE Trans. Img. Proc.*, vol. 15, pp. 2191–2197, Aug. 2006.

[10] Y. Guo, Y. Chen, Y.-K. Wang, H. Li, M. M. Hannuksela, and M. Gabbouj, "Error resilient coding and error concealment in scalable video coding," *IEEE Trans. Circ. Sys. Video Tech.*, vol. 19, 2009.

[11] J. Han, V. Melkote, and K. Rose, "A recursive optimal spectral estimate of end-to-end distortion in video communications," *Proc. Packet Video*, pp. 94–101, 2010.

[12] A. R. Reibman, L. Bottou, and A. Basso, "Scalable video coding with managed drift," *IEEE Trans. Circ. Sys. Video Tech.*, vol. 13, no. 2, pp. 131–140, Feb. 2003.

[13] H. Yang, R. Zhang, and K. Rose, "Drift management and adaptive bit rate allocation in scalable video coding," *Proc. IEEE ICIP*, vol. 2, pp. 49–52, Sep. 2002.

[14] J. Han, V. Melkote, and K. Rose, "An estimation-theoretic approach to spatially scalable video coding," *Proc. IEEE ICASSP*, Mar. 2012.

[15] F. Wu, S. Li, and Y.-Q. Zhang, "A framework for efficient progressive fine granularity scalable video coding," *IEEE Trans. Circ. Sys. Video Tech.*, vol. 11, 2001.