A RECURSIVE OPTIMAL SPECTRAL ESTIMATE OF END-TO-END DISTORTION IN VIDEO COMMUNICATIONS

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

End-to-end distortion estimation is critical to effective error-resilient video coding. The recursive optimal per-pixel estimate (ROPE) is a known approach to compute up to second moments of decoder-reconstructed pixels, and thereby optimally estimate the distortion. ROPE accurately accounts for encoding/decoding operations that are recursive in the pixel domain, and their interaction with packet loss and decoder concealment. The premise of this work is that considerable gains could be recouped by a dual estimation technique that would *perform its recursion in the transform domain*. This opens the door to accurate distortion estimation in conjunction with estimation-theoretic source coding approaches that involve transform domain operations, including improved prediction in both single-layer and scalable video coding. We present a novel recursive optimal estimate that operates entirely in the transform domain, namely, the *spectral coefficient-wise optimal recursive estimate* (SCORE). The method overcomes intricacies due to motion compensation from "off-grid" blocks. We first demonstrate that its accuracy matches ROPE in the usual setting where ROPE is known to be optimal. Then we consider an enhanced encoding scenario involving spectral operations that cannot be accurately tracked by ROPE, but for which SCORE still maintains optimality and hence enables substantial end-to-end performance gains over a large range of packet loss rates.

Index Terms— recursive estimate, end-to-end distortion, ROPE, mode decision, optimization

1. INTRODUCTION

Most current video coders employ motion compensated prediction to exploit temporal redundancies, at the cost of increased sensitivity to packet loss, due to temporal and spatial error propagation via the prediction loop. Many error resilience tools and paradigms have been employed to mitigate this problem, including forward error correction, intra refresh, multiple description coding, macroblock retransmission, etc., (see [1] for an overview of some relevant techniques). Since error-resilience typically incurs additional bit-rate costs, the fundamental optimization problem that underlies the coder is formulated in terms of the tradeoff between bit rate and the distortion perceived at the decoder, also referred to as end-to-end distortion (EED). Clearly, optimization of encoding decisions depends directly on the encoder's ability to accurately estimate the EED, while accounting for all factors, including compression, packet loss and error propagation due to the prediction loop, and concealment at the decoder. The recursive optimal per-pixel estimate (ROPE) [2], which originated in our lab, is an efficient and effective approach to optimally estimate the EED. Since packet losses are random, the encoder must treat the decoder reconstruction of a pixel as a random variable. The main idea of ROPE is to recursively calculate the first and second moments of reconstructed pixels, via update equations that explicitly account for motion compensated prediction, packet loss rate, and concealment at the decoder. The optimal EED estimate is then directly obtained from the first and second moments of the reconstructed pixels (details in Sec. 2.1). The basic version of ROPE [2] was extended in [3] to better comply with current standard options by accounting for operations such as sub-pixel motion compensation, deblocking, rounding, etc., which involve inter-pixel correlation terms.

Basic ROPE and its extensions have been successfully incorporated into various methods for error-resilient video coding, including for example [4]-[6]. We note, however, that it is inherently restricted to account for error propagation due to recursive operations performed in the pixel domain. This is not a significant limitation for many or most current video coding applications, where both prediction and error concealment are either actually performed in the spatial (pixel) domain, or are equivalent to such spatial operation. However, there are source coding approaches of significant interest that involve operations that are recursive in the transform domain rather than the pixel domain. The need to provide such applications with a ROPE-like EED estimate for effective error resilience is a main motivation for this work. In particular, [7]-[10] propose estimation-theoretic approaches for video source encoding/decoding that offer substantial compression gains, by recursively operating in the transform domain, typically the discrete cosine transform (DCT). Specifi-

The work was supported in part by the NSF under grant CCF-091723.

cally, these approaches view the sequence of DCT coefficients at a given spatial frequency, from blocks along a motion trajectory across consecutive frames, as an autoregressive (AR) process, and exploit this per-coefficient AR model to estimate the coefficients of a given block, either for prediction [7, 8, 10], or for reconstruction [9]. Let us focus specifically on a transform domain motion-compensated prediction (TD-MCP) scheme that was proposed in [8] which largely eliminates spatial correlations before spectral components (transform coefficients) are independently predicted. This technique incorporates the true temporal correlations that only emerge after signal decomposition, and which vary considerably from low to high frequency components. This precise nature of temporal dependencies is entirely masked by pixel domain prediction of standard video codecs, by the uniformly high correlation coefficient ($\rho \approx 1$) imposed on all pixels due to the dominance of low frequency components. Considerable coding gains were obtained by TD-MCP over the standard H.264 video codec.

Recent work in the context of distributed source coding already faced the need to estimate EED for DCT coefficients. In [11] this was achieved by exploiting the linearity of the transform to perform approximate conversion of pixel domain moments obtained by basic ROPE to the DCT domain, aided by the calculation of some inter-pixel correlation terms. On the other hand, [12] developed a recursive calculation of transform domain moments, which is in the spirit of the general approach we will propose here, but in order to circumvent the main complications due to "off-grid" reference blocks, the authors simply approximated motion compensation with motion vectors that point to on-grid blocks. This assumption yields substantially sub-optimal EED estimates as was demonstrated in [11]. It is in fact the reason why [11] reverted to applying basic ROPE in the pixel domain, and then converted the moments to the DCT domain. This is a feasible solution in certain applications but it is not general enough. In the case of TD-MCP, although the unitarity of the transform ensures that the DCT-domain distortion in a block equals the pixel-domain distortion, basic ROPE is nevertheless mismatched because it calculates a wrong pixel domain distortion, due to its inability to account for transform domain weighting (temporal correlations and corresponding prediction coefficients vary across frequencies).

Having established the need for a ROPE-like technique capable of accounting for error propagation due to recursive operations in the transform domain, this paper proposes the *spectral coefficient-wise optimal recursive estimate* (SCORE). We derive SCORE in a general setting to recursively calculate the moments of each transform (in practice DCT) coefficient of blocks in a frame, and account for general transform domain operations. The efficacy of this EED estimate is demonstrated in the setting of the TD-MCP video codec described in [8]. It is first shown that SCORE provides an accurate EED estimate when the TD-MCP codec is deployed over a lossy network, while basic ROPE is mismatched due to its inability to account for DCT domain recursions. Since standard pixel domain prediction is a special case of TD-MCP, it is experimentally verified that SCORE and ROPE coincide in this case, i.e., SCORE effectively subsumes basic ROPE. We finally exploit the EED estimates for improved (intra-inter) mode decisions. It is demonstrated that indeed the estimation accuracy of SCORE translates into improved rate-distortion performance of video transmission over a lossy network.

2. RELEVANT BACKGROUND

This section provides a brief review of ROPE and the TD-MCP approach of [8].

2.1. The recursive optimal per-pixel estimate

Consider point-to-point video communication, with encoder access to some statistical information about the network condition. For simplicity (but without implied loss of generality) assume that packet loss is statistically uniformly distributed, and let the packet loss rate (PLR), denoted p, be available to the encoder. Clearly, for optimal performance, the encoder must optimize its decisions with respect to the reconstructed video quality *at the decoder*. However, the decoder reconstruction is a random process as far as the encoder is concerned, with the ultimate effect of channel loss greatly complicated by error propagation through the prediction loop, error concealment efforts at the decoder, etc.

Let f_n^i denote the original value of pixel i in frame n, and let $\hat{f}_n^i$ denote its *encoder* reconstruction. The reconstructed value at the *decoder*, possibly after error concealment, is denoted by $\tilde{f}_n^i$, which is a random variable for the encoder. The overall expected distortion (in the mean squared sense) for this pixel is

$$E\{(f_n^i - \tilde{f}_n^i)^2\} = (f_n^i)^2 - 2f_n^i E\{\tilde{f}_n^i\} + E\{(\tilde{f}_n^i)^2\}. \quad (1)$$

Observe that evaluating this distortion only requires the first and second moments of the decoder reconstructed pixel $\tilde{f}_n^i$. ROPE employs the following recursion formulas, developed separately for the two cases of intra- and inter- coding, sequentially to compute these two moments for each pixel.

Intra-coding: The packet containing pixel i is received correctly with probability $1-p$, producing $\tilde{f}_n^i = \hat{f}_n^i$. If the packet is lost, we set the motion vector estimate to zero, and conceal as $\tilde{f}_n^i = \tilde{f}_{n-1}^i$ with probability p [1]. The first and second moments of $\tilde{f}_n^i$ for an intra-coded pixel are computed

[1] Although more sophisticated error concealment schemes can be handled in the ROPE framework, for simplicity of exposition we employ here the 'slice copy' concealment technique.

as:

$$E\{\tilde{f}_n^i\}(I) = (1-p)(\hat{f}_n^i) + pE\{\tilde{f}_{n-1}^i\}\,, \quad (2)$$

$$E\{(\tilde{f}_n^i)^2\}(I) = (1-p)(\hat{f}_n^i)^2 + pE\{(\tilde{f}_{n-1}^i)^2\}\,. \quad (3)$$

For simplicity we assume throughout this paper that intra-prediction (from spatially neighboring blocks) available in the H.264 standard is disabled. Thus, all the intra-coded macroblocks are self-contained and serve as instantaneous refresh points if received by the decoder.

Inter-coding: Let pixel i be predicted from pixel j in the previous frame, i.e., the encoder generates the prediction error

$$e_n^i = f_n^i - \hat{f}_{n-1}^j\,. \quad (4)$$

The prediction errors in a block are spatially transformed, quantized, encoded, and transmitted together with the motion vector. We denote by $\hat{e}_n^i$ the effective reconstruction of the prediction error at the encoder. Even if the current packet is correctly received, the decoder must use for prediction the *decoder's* reconstruction of pixel j in the previous frame, $\tilde{f}_{n-1}^j$, which is potentially different from $\hat{f}_{n-1}^j$ used by the encoder. Thus the first and second moments of $\tilde{f}_n^i$ for an inter-coded pixel are:

$$E\{\tilde{f}_n^i\}(P) = (1-p)(\hat{e}_n^i + E\{\tilde{f}_{n-1}^j\}) + pE\{\tilde{f}_{n-1}^i\}\,, \quad (5)$$

$$\begin{aligned} E\{(\tilde{f}_n^i)^2\}(P) &= (1-p)E\{(\hat{e}_n^i + \tilde{f}_{n-1}^j)^2\} + pE\{(\tilde{f}_{n-1}^i)^2\} \\ &= (1-p)((\hat{e}_n^i)^2 + 2\hat{e}_n^i E\{\tilde{f}_{n-1}^j\} \\ &\quad + E\{(\tilde{f}_{n-1}^j)^2\}) + pE\{(\tilde{f}_{n-1}^i)^2\}\,. \end{aligned} \quad (6)$$

Once the first and second moments are calculated, (1) provides the EED of the pixel. Employing ROPE to optimize inter/intra mode and quantization step selection within a rate-EED framework [2] has been demonstrated to provide substantial gains over heuristic methods for EED calculation.

2.2. Transform-domain motion-compensated prediction

Conventional motion-compensated prediction inherently assumes that a sequence of pixels (from consecutive frames) along a motion trajectory forms a temporal AR process, and effectively assumes that such sequences are independent of each other. Thus, *inter-pixel (spatial) correlation within each block is ignored during temporal (motion compensated) prediction*. In [8], we instead modeled the pair of transform coefficients, denoted by (x_n, x_{n-1}), at the same frequency of an inter-coded block and its motion compensated reference, as two successive samples of a scalar AR process with

$$x_n = \rho x_{n-1} + z_n \quad (7)$$

where the innovations z_n are zero-mean, independent, and identically distributed. We henceforth assume that the transform is DCT with block size restricted to the 4×4 option available in H.264. By running regular pixel domain motion search to get matched pairs of blocks between an (uncoded) frame and its (uncoded) preceding frame, and for multiple frame pairs, the correlation coefficient ρ at each frequency coefficient can be calculated by averaging pairwise correlations over all matched blocks. Provided in Table. 1 is the matrix of 16 correlation coefficients in the case of $coastguard\_qcif.yuv$. Note that the correlation is close to 1 for DC, but quite different otherwise. Such characteristics are also exhibited by other video sequences. The variation in

0.9998	0.9946	0.9916	0.9470
0.9893	0.9424	0.9068	0.8056
0.9807	0.9215	0.8696	0.7717
0.9680	0.9015	0.8309	0.7317

Table 1. Matrix of correlation coefficients for the 16 DCT coefficients in $coastguard\_qcif.yuv$

temporal correlation across frequencies, as observed above, motivated the transform domain motion-compensated prediction (TD-MCP) approach of [8].

Unlike the conventional approach that applies spatial transformation on the residual pixel domain block, in TD-MCP each block and its motion compensated reference are individually transformed, the DCT coefficients of the latter are weighted by frequency-appropriate correlation coefficients, and the prediction residue directly calculated in the transform domain. In other words, in inter-mode, the TD-MCP codec encodes the transform domain prediction error

$$y_n = x_n - \rho\hat{x}_{n-1} \quad (8)$$

at each frequency in every block of frame n. Here $\hat{x}_{n-1}$ is the corresponding motion compensated transform coefficient from the previous frame, and ρ is the correlation coefficient appropriate to that frequency. Note that by linearity of DCT, conventional pixel domain prediction (4) is equivalent to employing $\rho = 1$ at all frequencies, and is thus a special case of TD-MCP. Performance improvement as high as 1dB in PSNR was observed when TD-MCP was substituted into the H.264 codec (in place of the standard pixel domain motion compensated prediction).

It is important to emphasize that, in [8] the performance of TD-MCP was demonstrated in the setting of lossless transmission. The question of whether such gains can be maintained despite transmission over lossy networks illustrates both the motivation and focus of this paper. TD-MCP is therefore a representative example for enhanced source coding techniques that require new EED estimation tools. It is useful to note that the standard H.264 codec and the TD-MCP codec differ only in inter-mode coding, and not in intra-mode. Further, for the purpose of simplifying the presentation of this paper we assume that decoders in both cases employ the same 'slice copy' concealment scheme described in Sec. 2.1.

3. LIMITATIONS OF PIXEL DOMAIN ESTIMATION

Consider employing ROPE for EED estimation in the TD-MCP encoder. The update equations (2) and (3) are still valid: TD-MCP does not differ from the standard encoder in terms of intra-coding, the slice copy concealment scheme is retained, and the transform is linear. But the inter mode update equations of ROPE, (5) and (6), are no longer valid for the TD-MCP codec. Clearly, the transform domain weighting in (8) cannot be accounted for via these per-pixel recursions.

One could view the transform domain weighting (i.e., multiplication) involved in TD-MCP as the application of an equivalent 2-D linear filter on the corresponding pixel domain motion compensation block (i.e., convolution). The filtered output is then employed as the pixel domain prediction block. But it can be easily shown that accounting for any type of pixel filtering operations in ROPE requires, in addition to first and second order moments, the recursive calculation of cross-correlations between all pixel pairs within the frame, which tremendously increases complexity and memory requirements. This is a well known difficulty of ROPE, with various approximative solutions including, e.g., [3] where ROPE was extended to approximately account for sub-pixel motion estimation, an operation that involves interpolation (i.e., filtering) between pixels. This difficulty is further exacerbated if the objective is to account for transform domain operations that are also non-linear in nature, such as some of the estimation-theoretic techniques in [7], [10], and [9].

4. SPECTRAL COEFFICIENT-WISE OPTIMAL RECURSIVE ESTIMATE

4.1. The method

The previous observations indicate the requirement for a ROPE-like technique for EED estimation that works directly in the transform domain, and hence can efficiently account for operations in that domain. This motivates the proposed SCORE approach for EED estimation described below. Rather than calculate moments and distortion of individual pixels as ROPE does, SCORE directly tracks the moments and distortion of individual transform coefficients. To concretize the presentation, we describe SCORE in conjunction with the TD-MCP codec.

We expand on the notation of Sec. 2.2 to define $x_n^{k,m}$ as the unquantized value of transform coefficient m in block k of frame n. In keeping with the convention in Sec. 2.1, $\hat{x}_n^{k,m}$ and $\tilde{x}_n^{k,m}$ denote the encoder and decoder reconstructions of this coefficient, respectively. Note that this block may not be predicted from an on-grid reference block in the previous frame. Let $u_n^{k,m}$ denote the unquantized value of coefficient m in this (possibly off-grid) reference block.[2] The encoder and decoder reconstructions of this coefficient are denoted, as $\hat{u}_n^{k,m}$ and $\tilde{u}_n^{k,m}$, respectively. The encoder considers $\tilde{x}_n^{k,m}$ and $\tilde{u}_n^{k,m}$ as random variables due to the stochastic nature of packet loss. The correlation coefficient at coefficient frequency m of block k is denoted $\rho_n^{k,m}$. The TD-MCP prediction error (8) is thus rewritten as

$$y_n^{k,m} = x_n^{k,m} - \rho_n^{k,m}\hat{u}_n^{k,m} \tag{9}$$

Let $\hat{y}_n^{k,m}$ denote the quantized transform domain prediction error, whose value is encoded and transmitted to the decoder. The notation $\rho_n^{k,m}$ admits variation of the frequency-dependent correlation coefficient across frames, as well as blocks within a frame, i.e., TD-MCP could involve adaptation to temporal and spatial variations in temporal correlation.

The expected distortion at coefficient $x_n^{k,m}$ is

$$\begin{aligned}\delta_n^{k,m} &= E\{(x_n^{k,m} - \tilde{x}_n^{k,m})^2\} \\ &= (x_n^{k,m})^2 - 2x_n^{k,m}E\{\tilde{x}_n^{k,m}\} + E\{(\tilde{x}_n^{k,m})^2\}. \end{aligned}\tag{10}$$

The computation of $\delta_n^{k,m}$ only requires the first and second moments of the decoder reconstruction $\tilde{x}_n^{k,m}$. SCORE employs the following recursion functions, developed separately for the two cases of intra- and inter-coding, to sequentially compute these two moments for each transform coefficient in a frame.

Intra-coding: The recursions are practically the same as in ROPE, albeit with transform coefficients replacing pixels. Since the assumed concealment is "slice copy", if $\hat{x}_n^{k,m}$ is unavailable due to packet loss, it is concealed as $\tilde{x}_{n-1}^{k,m}$, i.e., it is equivalent to copying in the pixel domain.

$$E\{\tilde{x}_n^{k,m}\}(I) = (1-p)(\hat{x}_n^{k,m}) + pE\{\tilde{x}_{n-1}^{k,m}\}\,, \tag{11}$$

$$E\{(\tilde{x}_n^{k,m})^2\}(I) = (1-p)(\hat{x}_n^{k,m})^2 + pE\{(\tilde{x}_{n-1}^{k,m})^2\}. \tag{12}$$

Inter-coding: Following arguments similar to ROPE it is easily shown that,

$$E\{\tilde{x}_n^{k,m}\}(P) = (1-p)(\hat{y}_n^{k,m} + \rho_n^{k,m}E\{\tilde{u}_n^{k,m}\}) + pE\{\tilde{x}_{n-1}^{k,m}\}, \tag{13}$$

$$\begin{aligned}E\{(\tilde{x}_n^{k,m})^2\}(P) &= (1-p)E\{(\hat{y}_n^{k,m} + \rho_n^{k,m}\tilde{u}_n^{k,m})^2\} + pE\{(\tilde{x}_{n-1}^{k,m})^2\} \\ &= (1-p)((\hat{y}_n^{k,m})^2 + 2\rho_n^{k,m}\hat{y}_n^{k,m}E\{\tilde{u}_n^{k,m}\} \\ &\quad +(\rho_n^{k,m})^2E\{(\tilde{u}_n^{k,m})^2\}) + pE\{(\tilde{x}_{n-1}^{k,m})^2\}\,. \end{aligned}\tag{14}$$

It is obvious that the SCORE update equations, (11)-(14), are very similar to that of ROPE, except for the important fact that SCORE has a natural ability to incorporate transform domain weighting as is evident in (13) and (14). Note that these equations also involve the first and second moments of transform coefficients of the motion compensated block, which is potentially off the grid. We thus propose a complementary method to extract the required moments of off-grid blocks from the already available moments of on-grid blocks in frame $n-1$.

[2] Note that while $u_n^{k,m}$ is indexed by n and k to indicate the location on the current frame it provides a reference for, it is in fact a function of pixels in frame $n-1$.

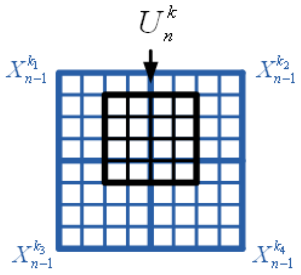


Fig. 1. Each off-grid block in a frame overlaps with 4 on-grid blocks. Here the blue blocks are on-grid, and the black off-grid block is employed for motion compensated prediction in the subsequent frame.

Any off-grid block in a frame overlaps with at most four on-grid blocks (Fig. 1). Let block U_n^k shown in the figure be the reference block for the current block k in frame n. This block, located in frame $n-1$, overlaps with on-grid blocks $X_{n-1}^{k_i}$ in the frame. The decoder reconstruction of block U_n^k is associated with coefficients $\tilde{u}_n^{k,m}$. Since we assume a linear transformation (e.g., DCT), there exist constants $a_{i,m}$ such that,

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

These constants purely depend on the position of U_n^k relative to the on-grid blocks[3]. Thus, the first moment of $u_n^{k,m}$ is simply

$$E\{\tilde{u}_n^{k,m}\} = \sum_{i=1}^{4} \sum_{m=0}^{15} a_{i,m} E\{\tilde{x}_{n-1}^{k_i,m}\} \, . \tag{16}$$

The second moment of $u_n^{k,m}$ is more complicated, and involves cross-correlations of DCT coefficient pairs of the on-grid blocks:

$$E\{(\tilde{u}_n^{k,m})^2\} = \sum_{i=1}^{4} \sum_{j=1}^{4} \sum_{m=0}^{15} \sum_{l=0}^{15} a_{i,m} a_{j,l} E\{\tilde{x}_{n-1}^{k_i,m} \tilde{x}_{n-1}^{k_j,l}\} \, . \tag{17}$$

In Sec. 3 it was observed that a limitation of extending ROPE to account for transform domain weighting is the necessity of calculating inter-pixel correlation terms that appear as a result of the implied pixel domain filtering. But (17) suggests that the calculation of cross-correlations cannot be avoided even if moments are updated directly in the transform domain. However, there is a major advantage to the transform domain if we assume a largely *decorrelating* transformation as is indeed sought in compression applications, such as DCT in video coding. Specifically, the following approximation of 'uncorrelatedness' holds well in the DCT domain:

$$E\{\tilde{x}_n^{k_i,m} \tilde{x}_n^{k_j,l}\} = E\{\tilde{x}_n^{k_i,m}\} E\{\tilde{x}_n^{k_j,l}\} \text{ when } j \neq i \text{ or } l \neq m. \tag{18}$$

On the other hand, the analogous pixel domain approximation:

$$E\{\tilde{f}_n^i \tilde{f}_n^j\} = E\{\tilde{f}_n^j\} E\{\tilde{f}_n^j\} \;\; j \neq i \tag{19}$$

has been demonstrated to be inaccurate [3]. Results presented in Sec. 4.2 support the approximation (18). Substituting (18) in (17), and subsequent use of (16), yields the required transform domain first and second moments of the motion compensated blocks as a simple linear combination of the moments of on-grid blocks in that frame. We now summarize the update procedure of SCORE.

SCORE update steps

Given the transform domain first and second moments of coefficients of on-grid blocks in frame $n-1$:

1. Identify the (motion compensated) reference block U_n^k in frame $n-1$ for each on-grid block k in frame n.
2. Compute the transform domain first and second moments of U_n^k via (16) - (18).
3. Compute the transform domain first and second moments of on-grid blocks in frame n via (11)-(14).

4.2. SCORE accuracy results

We first compare the EED estimation accuracy of SCORE and ROPE in the setting of the TD-MCP codec described in Sec. 2.2. Recall that this codec is simply the standard H.264 codec with transform domain weighted prediction replacing pixel domain prediction in inter mode coding. The current implementation does not adapt the correlations over time or across blocks. Some error resilience is incorporated into the codec via the 'random intra' technique: in each frame 10% of the macroblocks are randomly selected to be intra-coded. Both SCORE and ROPE are embedded in the encoder to obtain respective EED estimates assuming a certain PLR p. But it must be emphasized that neither estimate influences the encoder's decisions in any way, i.e., these estimates are calculated solely for the purpose of evaluating their accuracy. In other words, both approaches provide a corresponding estimated EED for the same coded video sequence. The transmission of this video sequence is now simulated over 100 different realizations of the lossy channel. In the simulation, each packet, assumed to contain a row of macroblocks, was deemed lost with probability p. The distortion of each frame in the video sequence is averaged over realizations, and converted to a PSNR value for the frame. In the case of ROPE, the per-pixel EED estimate is averaged across pixels in a frame, whereas in the case of SCORE the average is over DCT coefficients within a frame. Fig. 2 compares the PSNR of different frames obtained via simulation with its estimate obtained via ROPE and

[3]Without loss of generality, the top-left corner of U_n^k is one of the 16 pixel locations in block $X_{n-1}^{k_1}$. Each of these positions has an associated set of constants $a_{i,m}$.

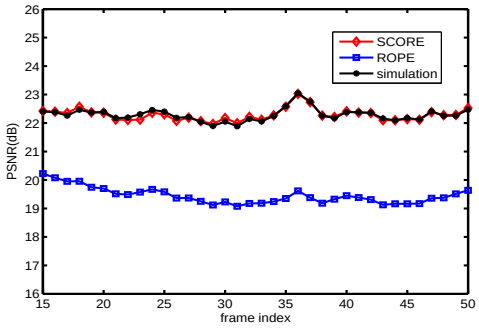


Fig. 2. Comparison of simulated and estimated PSNRs for the $mobile\_cif$ sequence encoded by TD-MCP: bit-rate is $800kbps$, frame rate $30f/s$, and PLR $p = 5\%$.

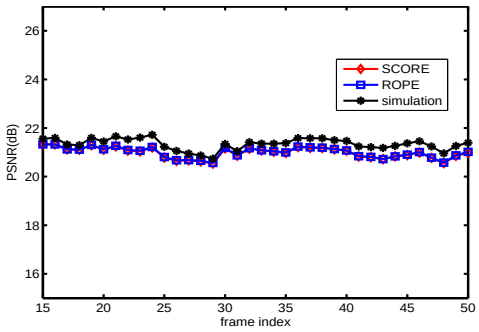


Fig. 3. Comparison of simulated and estimated PSNRs for the $mobile\_cif$ sequence encoded by standard H.264: bit-rate is $800kbps$, frame rate $30f/s$, and PLR $p = 5\%$.

SCORE. It is evident that SCORE provides a very accurate estimate of EED while ROPE is mismatched to the working of the encoder, i.e., to the DCT domain weighted prediction.

We next compare SCORE and ROPE estimates in the framework of the standard H.264 codec that employs regular pixel domain motion compensated prediction. Note that, as described in Sec. 2.2, this codec is merely a special case of the TD-MCP codec with ρ assumed to be uniformly unity at all frequencies, and the SCORE update equations are still valid. Random intra again provides some error resilience. Fig. 3 compares the PSNRs by simulation and estimation under the same conditions as the previous experiment. Note that SCORE and ROPE estimates of PSNR practically match for each frame, and in this case both estimates are very close to the value obtained by simulation. Thus, SCORE subsumes in it at least the functionality of basic ROPE. Further, the accuracy of SCORE in both these experiments support the 'uncorrelatedness' approximation of (18).

5. CODING PERFORMANCE

5.1. Optimal mode decisions

In this section we compare SCORE and ROPE in terms of the performance obtained when the estimates are employed to optimize the coding mode (Intra/Inter). The EED and bit costs incurred in encoding macroblock k of frame n in coding mode μ and quantization parameter q are denoted $D_n^k(q,\mu)$ and $B_n^k(q,\mu)$, respectively. The optimization entails the following minimization per block (given the quantization parameter):

$$\mu_n^k(\lambda, q_n) = \arg\min_{\mu}\{D_n^k(q_n,\mu) + \lambda B_n^k(q_n,\mu)\}, \qquad (20)$$

and the per frame optimization:

$$q_n(\lambda) = \arg\min_{q} \sum_k D_n^k(q,\mu_n^k) + \lambda B_n^k(q,\mu_n^k)\,, \qquad (21)$$

where λ is a Lagrange parameter whose value is fixed for all frames in the simulation. Varying λ provides an operational rate-distortion curve. The above can be performed for either SCORE or ROPE, given an assumed packet loss rate p. Multiple realizations of the lossy channel (i.e., instances of the packet loss sequence) are now simulated, where a packet is randomly deemed lost with probability p. The encoded video sequence is decoded over each channel realization, and distortion is averaged over the entire video sequence as well as different channel realizations, and converted to a PSNR value, which is coupled with the bit rate to produce a point on the curve.

5.2. Performance results

We consider two encoders:

1) The standard H.264 encoder with pixel domain motion compensated prediction where macroblock mode decisions and frame QPs are optimized via the above algorithm but with EED calculated using ROPE, i.e., EED of a macroblock is defined as the sum of the estimated distortion for each pixel in the macroblock. Note that ROPE does provide the optimal estimate of EED in this case. We refer to this codec as H.264-ROPE.

2) The pixel domain motion compensated prediction in the standard encoder is now replaced by TD-MCP. The correlations at each DCT frequency are calculated and employed for transform domain weighting of the motion compensation. The mode decisions are now optimized with EED defined by SCORE, which is optimal in this case, i.e., the EED of each macroblock is defined as the sum of estimated distortion of DCT coefficients in its 4×4 on-grid sub-blocks. We refer to this codec as TD-MCP-SCORE.

The decoders for both encoders employ the slice-copy concealment technique. Fig. 4 and Fig. 5 compares the

performance of H.264-ROPE and TD-MCP-SCORE via rate-PSNR plots at PLRs of 1% and 5% for the video sequences $mobile$ in CIF resolution, and $coastguard$ in QCIF resolution. The plots for 0% PLR are included in each case as a reference to demonstrate the gains of TD-MCP over standard H.264. Substantial performance gains over H.264-ROPE are obtained via TD-MCP-SCORE, in particular at the lower PLR. The decrease in gains with increasing PLR is attributed to the fact that intra mode is chosen more frequently when PLR is high, and TD-MCP differs from standard H.264 only in inter mode coding. Further note that at higher PLRs concealment plays a bigger role, and the employment of the same slice-copy technique (a purely pixel domain operation) in both codecs marginalizes the gains due to exploiting the true temporal correlations in the DCT domain.

6. GENERALIZATION OF SCORE

Although the update equations in Sec. 4 were presented in the context of the TD-MCP framework, the SCORE concept itself is fairly general and can accommodate other transform domain operations. For instance, instead of the simple 'slice copy' technique, alternate concealment schemes that exploit transform domain correlations could be employed, and accounted for in SCORE. Such a concealment scheme is also expected to further improve the performance of TD-MCP-SCORE compared to H.264-ROPE at high PLRs. While this paper focused exclusively on single layer video coding, methods such as the estimation-theoretic enhancement layer prediction scheme proposed in [7] for scalable video coding, and its extension in [10] to exploit transform domain correlations, are also compatible with the SCORE concept. These latter methods involve non-linear transform domain operations (as opposed to the linear weighting involved in TD-MCP), and appropriate linearizations lead to update equations similar in spirit with (11)-(14).

7. CONCLUSIONS

A technique to find the optimal per-spectral coefficient estimate of end-to-end distortion is proposed. This approach, called SCORE, is motivated by the need to account for coding operations that are recursive in the transform domain, rather than in the pixel domain. It operates via update equations that recursively calculate the first and second moments of decoder reconstructed transform coefficients. The efficacy of the approach is demonstrated in comparison with the well established ROPE technique that only accounts for pixel domain operations. The two end-to-end distortion estimation techniques are employed in appropriate encoders to perform macroblock coding mode decision optimization, and substantial coding gains are observed for the SCORE-based encoder.

8. REFERENCES

[1] Y. Wang, S. Wenger, J. Wen, and A. K. Katsaggelos, "Error resilient video coding techniques," *IEEE Sig. Proc. Mag.*, vol. 17, no. 4, pp. 61–82, Jul 2000.

[2] 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.

[3] H. Yang and K. Rose, "Advances in recursive per-pixel end-to-end distortion estimation for robust video coding in H.264/AVC," *IEEE Trans. Circ. Sys. Video Tech.*, vol. 17, pp. 845–856, July 2007.

[4] A. Leontaris and P. C. Cosman, "Video compression for lossy packet networks with mode switching and a dual-frame buffer," *IEEE Trans. Img. Proc.*, vol. 13, no. 7, pp. 885897, Jul 2004.

[5] A. R. Reibman, L. Bottou, and A. Basso, "DCT-based scalable video coding with drift," in *IEEE ICIP*, Oct 2001.

[6] B. A. Heng, J. G. Apostolopoulos, and J. S. Lim, "End-to-end rate-distortion optimized mode selection for multiple description video coding," in *IEEE ICASSP*, Apr 2005.

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

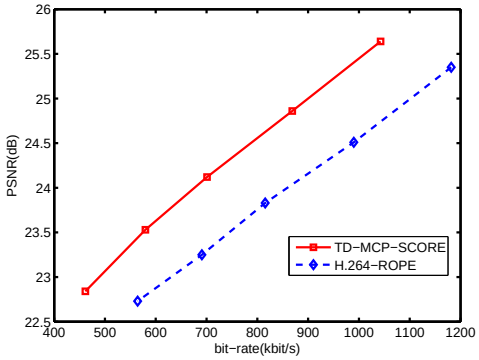
[8] J. Han, V. Melkote, and K. Rose, "Transform-domain temporal prediction in video coding: exploiting correlation variation across coefficients," in *IEEE ICIP*, Sep 2010.

[9] J. Han, V. Melkote, and K. Rose, "Estimation-theoretic delayed decoding of predictively encoded video sequences," in *Proc. IEEE DCC*, Mar 2010.

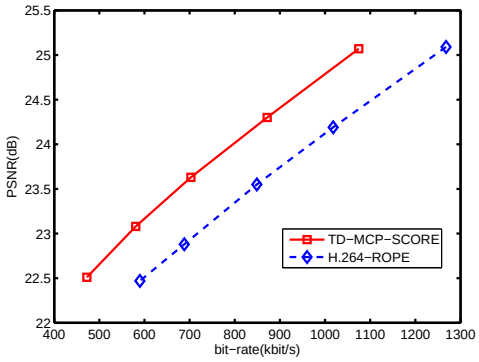
[10] J. Han, V. Melkote, and K. Rose, "Estimation-theoretic approach to delayed prediction in scalable video coding," in *IEEE ICIP*, Sep 2010.

[11] M. Fumagalli, M. Tagliasacchi, and S. Tubaro, "Improved bit allocation in an error-resilient scheme based on distributed source coding," in *IEEE ICASSP*, May 2006.

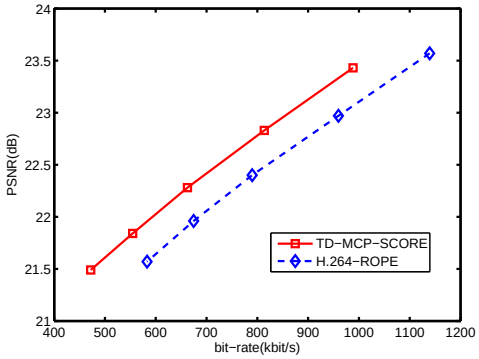
[12] A. Majumdar, J. Wang, and K. Ramchandran, "Drift reduction in predictive video transmission using a distributed source coded side-channel," in *ACM Multimedia*, Oct 2004.



(a) $p = 0\%$

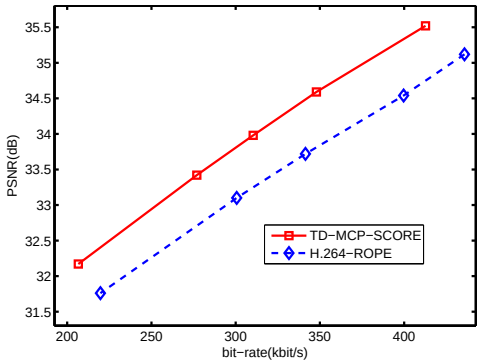


(b) $p = 1\%$

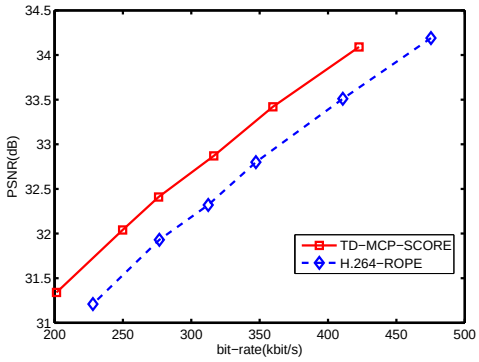


(c) $p = 5\%$

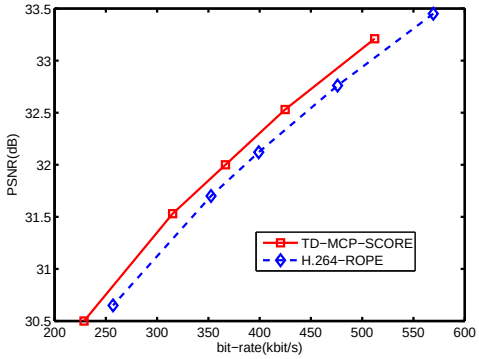
Fig. 4. Results of mode decision optimization for the sequence $mobile\_cif$ via SCORE and ROPE in terms of rate-PSNR curves at different PLRs. The case $p = 0\%$ is provided as a reference to indicate the gains due to TD-MCP over standard H.264.



(a) $p = 0\%$



(b) $p = 1\%$



(c) $p = 5\%$

Fig. 5. Results of mode decision optimization for the sequence $coastguard\_qcif$ via SCORE and ROPE in terms of rate-PSNR curves at different PLRs. The case $p = 0\%$ is provided as a reference to indicate the gains due to TD-MCP over standard H.264.