Estimation-Theoretic Approach to Delayed Decoding of Predictively Encoded Video Sequences

Jingning Han, *Student Member, IEEE,* Vinay Melkote, *Member, IEEE,* and Kenneth Rose, *Fellow, IEEE*

***Abstract*— Current video coders employ predictive coding with motion compensation to exploit temporal redundancies in the signal. In particular, blocks along a motion trajectory are modeled as an auto-regressive (AR) process, and it is generally assumed that the prediction errors are temporally independent and approximate the innovations of this process. Thus, zero-delay encoding and decoding is considered efficient. This paper is premised on the largely ignored fact that these prediction errors are, in fact, temporally dependent due to quantization effects in the prediction loop. It presents an estimation-theoretic delayed decoding scheme, which exploits information from future frames to improve the reconstruction quality of the current frame. In contrast to the standard decoder that reproduces every block instantaneously once the corresponding quantization indices of residues are available, the proposed delayed decoder efficiently combines all accessible (including any future) information in an appropriately derived probability density function, to obtain the optimal delayed reconstruction per transform coefficient. Experiments demonstrate significant gains over the standard decoder. Requisite information about the source AR model is estimated in a spatio-temporally adaptive manner from a bit-stream conforming to the H.264/AVC standard, i.e., no side information needs to be sent to the decoder in order to employ the proposed approach, thereby compatibility with the standard syntax and existing encoders is retained.**

***Index Terms*— Delayed decoding, differential pulse code modulation, estimation-theoretic prediction, motion trajectory, predictive coding.**

I. Introduction

THE EARLY approaches to predictive coding focused on differential pulse code modulation (DPCM) [1]–[4]. Predictive coding was subsumed in video coders in the form of motion-compensated prediction [5]. The efficacy of such schemes in achieving considerable compression gains is premised on the assumption that blocks of the video signal along a motion trajectory form an auto-regressive source, where the temporal redundancy can be largely removed by predicting the present samples from prior reconstructions, thus only the residuals are coded as innovative information to the decoder. While the emphasis of this paper is on video compression, let us start by considering the simpler case of a generic first order stationary auto-regressive (AR) process $\{x_n\}$ that consists of a sequence of zero-mean, real-valued random variables satisfying the recursion

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

where ρ is the correlation coefficient of consecutive samples, and the driving innovation variables denoted by $\{z_n\}$ are independent and identically distributed with probability density function $p_Z(z)$. A DPCM encoder (Fig. 1) generates a prediction $\tilde{x}_n$, based on previously reconstructed samples, and subtracts it from the current sample x_n to produce the prediction error e_n, which is then quantized into an index i_n of the codebook, which is entropy coded and transmitted to the decoder. The reconstruction of x_n is $\hat{x}_n = \tilde{x}_n + \hat{e}_n$, where $\hat{e}_n$ is the reconstruction of the prediction error indexed by i_n. At high bit-rates, $\hat{x}_{n-1} \approx x_{n-1}$, and the optimal predictor is $\tilde{x}_n = \rho \hat{x}_{n-1}$. This form of predictor is often used at medium and low bit-rates as well. In the AR source model, x_n is correlated with preceding samples, and independent of subsequent innovations, $\{z_l\}_{l>n}$. At high bit-rates, the prediction error $e_n \approx z_n$, hence $\{i_n\}$ are approximately i.i.d.. In this case, future quantization indices $\{i_l\}_{l>n}$ convey no additional information on the current sample x_n. In the practical, limited bit-rate scenarios, however, the reconstruction error of $\hat{x}_{n-1}$ affects the prediction loop and introduces correlations between prediction errors $\{e_n\}$, and correspondingly $\{i_n\}$. Therefore, future indices indeed contain information about x_n and could potentially be used to improve its reconstruction at the decoder. Naturally this would entail decoding delay.

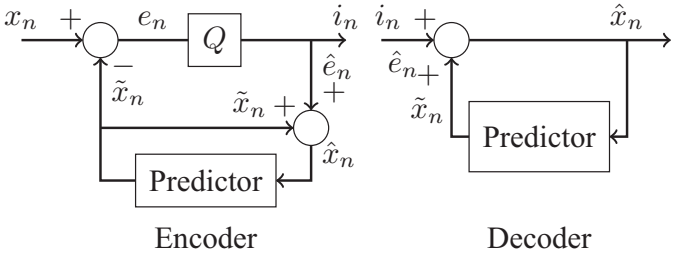

Fig. 1. Prototype DPCM codec for an AR sequence $\{x_n\}$ composed of zero-mean and real-valued random variables.

Manuscript received May 11, 2012; revised August 23, 2012; accepted October 16, 2012. Date of publication November 16, 2012; date of current version January 28, 2013. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Béatrice Pesquet-Popescu.

J. Han and K. Rose are with the Department of Electrical and Computer Engineering, University of California Santa Barbara, CA 93106 USA (e-mail: jingning@ece.ucsb.edu; rose@ece.ucsb.edu).

V. Melkote was with the Department of Electrical and Computer Engineering, University of California, Santa Barbara, CA 93106 USA. He is now with the Dolby Laboratories, Inc., San Francisco, CA 94103 USA (e-mail: melkote@ece.ucsb.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2012.2227773

Modern video coding schemes such as H.264/AVC [5] employ DPCM in the form of inter-frame predictive coding to exploit temporal redundancy in the video sequences. Typically, an encoder partitions a frame into an array of blocks, and predicts each of them using motion compensated reference blocks from previously reconstructed frames.[1] The prediction residuals are spatially transformed by a 2-D discrete cosine transform (DCT), whose coefficients are then quantized and entropy coded. Upon receiving the motion information and quantized transform coefficients, the decoder applies an inverse transform to reproduce the residuals, adds them to the motion-compensated prediction, and thereby reconstructs the block. Making abstraction of the spatial transform, the similarity with DPCM is evident. Hence it becomes obvious that information about future frames could be employed to enhance the reconstruction quality of the current frame. In particular, whenever decoding delay is acceptable, motion vectors of future frames can potentially be used to extend the motion trajectory of every block in the current frame, and the coded information of future blocks will then be exploited to improve the reconstruction of the current frame.

In the context of DPCM, decoding delay has been previously considered in [6] and [7], both of which apply filtering to 'smooth' the output of a typical DPCM, $\{\hat{x}_n\}$, with a suitable non-causal post-filter. More recently, an estimation-theoretic (ET) approach was developed by our group in [8], which effectively accounts for all the information available to a DPCM decoder, at a given decoding delay, in an appropriately derived conditional expectation framework, for optimal reconstruction. Central to that ET approach was the postulate that the true value of each sample in the sequence $\{x_n\}$ must reside in an interval determined by the quantizer index. It was shown to substantially outperform the smoothing methods of [6] and [7], and provided evidence for significant gains achievable by delayed decoding, which motivate the delayed decoder for video signals proposed herein. We note that this approach, in light of combining temporal correlation with quantization information, drew inspiration from an approach to optimal prediction in scalable video coding developed earlier by our group in [9].

Unlike the case of the synthetic scalar AR model, however, major challenges arise in video decoding due to the combination of motion compensated prediction and spatial transformation. In particular, while only on-grid blocks of pixels are coded, the reference blocks are potentially off-grid (i.e., they might not have been coded as separate blocks in previous and future frames). Furthermore, the quantization is performed in the transform domain per each on-grid block, which implies that the quantizer intervals of samples in the sequence, as required by the ET approach in [8], are not directly accessible. To overcome such difficulties, we propose a video decoding scheme that constructs a motion trajectory for each on-grid block in the current frame, going from prior frames into future ones, and calculates the probability density function (pdf), per transform coefficient, conditioned on information from subsequent blocks lying on the same motion trajectory, in addition to the current quantizer interval. The optimal reconstruction of the transform coefficient is then obtained as the appropriate conditional expectation. Keeping in mind the complexity concerns of video decoding, practical design strategies and critical modifications to the generic ET approach of [8] will also be discussed.

Highly relevant prior work includes preprocessing operations proposed in [10]–[12], which incorporated encoder delay to exploit correlation with future frames, using motion compensated temporal filters. A related approach was developed, in conjunction with spatial filter, as a postprocessing algorithm to reduce blocking artifacts and mosquito noise for DCT-based video/image coder [13], and was applied to subband filtering [14]. We note that unlike the long understood blocking artifacts and mosquito noise caused by *spatial transform coding*, the artifacts effectively addressed by the delayed video decoding scheme proposed herein are mainly due to *temporal predictive coding*, which have become the focus of recent studies of video quality assessment [15]–[17]. An analytic quantitative characterization of the trade-off between performance and decoding latency for the special case of scalar Gauss-Markov sources was derived in [18], where the potential gains of delayed decoding were analyzed from an information-theoretic perspective. It is noteworthy that the pdf of the temporal innovations in video sequences are better modeled by Laplacian process [19]–[23]. In our implementation of the proposed ET delayed video decoder, the Laplacian assumption was adopted and the corresponding parameters were learned from the coded bit-stream in accordance with H.264/AVC standard, i.e., compatibility with the standard syntax and existing encoders is retained. Some of our preliminary results under simplified assumptions were reported in [24] to validate the potential benefits of delayed decoding, where the motion compensation was performed at full pixel accuracy, a single-frame decoding delay was supported, and a spatially stationary model for video signal was assumed, i.e., the Laplacian parameters were fixed per frame, while ignoring statistical variations across motion trajectories. The proposed scheme in this paper eschews such limitations by employing techniques including an enhanced motion trajectory construction method, a spatio-temporally adaptive model estimate, and a recursive frame refinement approach for multi-frame decoding delay. We note that while the proposed approach was implemented in H.264/AVC reference framework to demonstrate its efficacy, the basic principle is generally applicable to other motion compensated predictive video codecs, such as VP8 [25] and HEVC [26].

II. Generic ET Delayed Decoding for Scalar DPCM

We briefly review the ET delayed decoding approach proposed in [8], in light of the first order AR process (1) and the DPCM scheme described in Sec. I. The mean squared error (MSE) is employed as the distortion metric throughout this paper. Hence, the optimal reconstruction of the sample x_n,

[1] Prediction from available reconstructed boundary pixels in the same frame, i.e., Intra-mode, is also an option in most video codecs, but inter-mode generally offers better compression and is most frequently selected by the encoder [5].

given all the information available to the decoder for a fixed decoding delay L, i.e., indices $\{i_l\}_{l\le n+L}$, is the minimum MSE estimate

$$\hat{x}_n^* = E[x_n|\{i_l\}_{l\le n+L}] \tag{2}$$

the expectation over the conditional pdf $p(x_n|\{i_l\}_{l\le n+L})$, the derivation of which is considered next. We use the streamlined notation $p(\cdot)$ to denote any pdf or probabilities, and add a subscript when the interpretation is not obvious from the context.

Let the quantization index i point to a quantizer interval $[a(i), b(i))$. Thus, at time l the index i_l along with the prediction $\tilde{x}_l(=\rho\hat{x}_{l-1})$ determine an interval $I_l = [\tilde{x}_l + a(i_l), \tilde{x}_l + b(i_l))$, in which the true value of x_l must reside. The statement $\{x_l \in I_l | l \le n\}$ effectively captures all the information provided by the indices $\{i_l\}_{l\le n}$ to the DPCM decoder of Fig. 1. The conditional pdf $p(x_n|\{i_l\}_{l\le n+L})$ can then be written as (see Appendix I for proof)

$$\begin{aligned} p(x_n|\{i_l\}_{l\le n+L}) &= p(x_n|\{I_l\}_{l\le n+L}) \\ &= \begin{cases} \frac{p(x_n|\{I_l\}_{l<n})\ p(\{I_l\}_{n<l\le n+L}|x_n)}{\int_{I_n} p(x_n|\{I_l\}_{l<n})\ p(\{I_l\}_{n<l\le n+L}|x_n)dx_n}, & x_n \in I_n \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{3}$$

In the interest of notational brevity, in the above equations (and elsewhere when such meaning is obvious) the symbol I_l has been abused to denote the event $\{x_l \in I_l\}$. Therefore $p(x_n|\{I_l\}_{l\le n})$ is the pdf of x_n conditioned on all information provided by the quantization indices $\{i_l\}_{l\le n}$up to the current time instance n, the expectation over which provides the optimal non-delayed estimate of x_n. The optimal *delayed* decoder further weights it with $p(\{I_l\}_{n<l\le n+L}|x_n)$, representing the conditional probability of the known future outcomes given x_n, to produce the composite pdf $p(x_n|\{I_l\}_{l\le n+L})$ in (3). Note that (3) incorporates all the known information up to a fixed delay L.

Two recursions are employed to obtain the requisite conditional probabilities used in (3) [8]. One recursion updates the causal pdf $p(x_{n-1}|\{I_l\}_{l\le n-1})$ employed at time instance $(n-1)$, to the corresponding pdf $p(x_n|\{I_l\}_{l\le n})$ at time n, namely the forward recursion. In particular, the pdf of x_n conditioned on all $prior$ information, $\{I_l\}_{l\le n-1}$, is obtained by applying the total probability theorem

$$\begin{aligned} p(x_n|\{I_l\}_{l\le n-1}) &= \int_{I_{n-1}} p(x_n, x_{n-1}|\{I_l\}_{l\le n-1})dx_{n-1} \\ &= \int_{I_{n-1}} p(x_n|x_{n-1}, \{I_l\}_{l\le n-1})\ p(x_{n-1}|\{I_l\}_{l\le n-1})dx_{n-1} \\ &= \int_{I_{n-1}} p_Z(x_n - \rho x_{n-1})\ p(x_{n-1}|\{I_l\}_{l\le n-1})dx_{n-1} \end{aligned} \tag{4}$$

where the third equality exploits the Markov property of (1). The current interval I_n additionally specifies a range that x_n must lie in. Thus, incorporating I_n further refines the above pdf of x_n as

$$p(x_n|\{I_l\}_{l\le n}) = \begin{cases} \frac{p(x_n|\{I_l\}_{l\le n-1})}{\int_{I_n} p(x_n|\{I_l\}_{l\le n-1})dx_n}, & x_n \in I_n \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

which completes the forward recursion of $p(x_n|\{I_l\}_{l\le n})$ from $p(x_{n-1}|\{I_l\}_{l\le n-1})$.

A second recursion yields the conditional probability $p(\{I_l\}_{n<l\le n+L}|x_n)$ of the L future intervals, given the value of x_n. Given the probability $p(\{I_l\}_{n+m<l\le n+L}|x_{n+m})$ of future quantization intervals $\{I_l\}_{n+m<l\le n+L}$ conditioned on x_{n+m}, where $m < L$, the probability $p(\{I_l\}_{n+m-1<l\le n+L}|x_{n+m-1})$ can be derived via applying Markov property of (1) as (see Appendix II for proof)

$$\begin{aligned} &p(\{I_l\}_{n+m-1<l\le n+L}|x_{n+m-1}) \\ &= \int_{I_{n+m}} p(\{I_l\}_{n+m<l\le n+L}|x_{n+m}) \cdot \\ &\quad p_Z(x_{n+m} - \rho x_{n+m-1})dx_{n+m}. \end{aligned} \tag{6}$$

Initializing

$$p(I_{n+L}|x_{n+L-1}) = \int_{I_{n+L}} p_Z(x_{n+L} - \rho x_{n+L-1})dx_{n+L}$$

the above recursive equation can be applied $L-1$ times to obtain the requisite probability $p(\{I_l\}_{n<l\le n+L}|x_n)$ of the known future outcomes conditioned on x_n, and is henceforth referred to as backward recursion. These two recursions, together with (3) and (2), provide the optimal reconstruction $\hat{x}_n^*$, given L future quantization intervals. This ET delayed decoding approach for DPCM was experimentally demonstrated in [8] to substantially outperform other existing filtering-based methods, in various settings of synthetic scalar source models.

III. ET Delayed Decoding of Compressed Video: From Theory to Practice

The efficacy of the ET delayed decoding approach for scalar DPCM motivates our proposed ET video decoder, which exploits future frame information (up to a certain delay) to enhance the reconstruction quality of the current frame. In this section, we present several critical modifications to the generic ET delayed decoding algorithm and the proposed complementary techniques, so as to overcome the intricacies due to the interaction of temporal prediction with spatial transform coding, prevalent in current standard video coding schemes [5]. For simplicity, we start by considering a video decoder with single-frame delay. The extension to incorporate multi-frame delay will be discussed in Sec. IV.

A. Motion Trajectory Construction

Delayed decoding of a block in the current frame requires reference blocks (both past and future) that lie on the same motion trajectory. Information on the statistics of the underlying AR process can then be exploited. Let on-grid block B in Fig. 2 be a block of interest in (the current) frame n. The motion vector of B points to the reference block A, in frame $n-1$, which is the predecessor of B in the AR process. The subsequent block C in this process, in frame $n+1$, also needs to be identified. Note that since decoding delay is allowed, the motion vectors of frame $n+1$ are available and can, in principle, be reversed to obtain future blocks in the regular non-delayed reconstruction of frame $n+1$, relevant to the current block of interest in frame n.

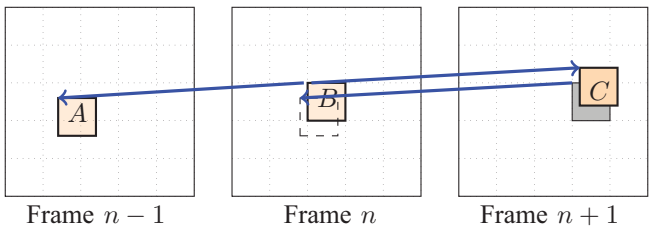

Fig. 2. Motion trajectory construction: blocks A, B, and C form a sequence in the underlying AR process. They are identified as a sequence using motion vectors available in the compressed video bit-stream.

Complications, however, arise as motion vectors are only assigned to on-grid blocks which are positioned on the prescribed grid dividing the frame into 4×4 blocks. Hence, the motion vectors of frame $n+1$ map its *on-grid* blocks to corresponding (potentially *off-grid*) reference blocks in frame n, while what the proposed algorithm requires is a mapping from on-grid blocks in frame n to blocks in frame $n+1$. This problem is resolved as follows: the motion vector of that on-grid block in frame $n+1$, whose motion-compensated reference block (located in frame n and potentially off-grid) maximally overlaps the current on-grid block of interest in frame n, is chosen. This motion vector is then reversed and applied to the on-grid block in frame n. This is illustrated in Fig. 2. The block bounded by dashed lines in frame n, which serves as reference for the gray-shaded on-grid block of frame $n+1$, is found to maximally overlap block B which is the on-grid block of interest in frame n. The motion vector associated with the gray-shaded block in frame $n+1$ is hence reversed and, given the position of block B, provides the location of the required future block C in frame $n+1$. This process of reversing already available motion vectors of the next frame, and applying them to on-grid blocks of the current frame, is a fast, low complexity alternative to a complete motion search to find a block in the reconstruction of the next frame that most resembles block B.

In our implementation, the decoder goes through the motion vectors of frame $n+1$ and hence the motion compensated reference blocks located in frame n, evaluates their overlapping areas with on-grid blocks in frame n at quarter-pixel resolution, and updates the maximum overlapping area (MOA) and the corresponding motion vector, both of which are maintained per on-grid block in frame n, all during the initial stage of decoding process. Whenever the MOA of an inter-coded block of the current frame is larger than half block area (e.g., 128 for a 4×4 block at quarter-pixel resolution), this block will be marked as ready for ET delayed decoding. For inter-coded blocks with insufficient MOA, a second round of motion search that involves block matching is employed to find their temporal references in frame $n+1$. A maximum absolute difference value is preset to threshold the eligibility of reference blocks in the next frame for motion trajectory reconstruction. This operation increases the availability of future reference and extends the applicability of ET delayed decoding to more current blocks, at the expense of an increment in decoding complexity. We note from experiments that this increment is moderate, since typically the majority of the current blocks have sufficient MOAs, and hence circumvent the need for motion search. For intra-coded blocks, the decoder performs standard non-delayed reconstruction.

B. Transform Domain Operation

In standard inter-frame coding mode, a block B (of original pixels) is predicted by A (from previously reconstructed frame $n-1$) to generate residual pixels, which are then transformed via a 2-D DCT to further remove the remaining spatial redundancy; and the resulting transform coefficients are quantized and entropy coded. Clearly, a DPCM scheme is effectively embedded in the system. However, a major difficulty arises in applying the ET delayed decoding of [8] to the pixel sequences along the temporal direction, since the quantization intervals that play a central role in this ET approach are available only in the transform domain. To circumvent this difficulty, an alternative perspective is adopted in this work, which models the transform coefficients of blocks on a given motion trajectory via an AR process per frequency. Due to the decorrelation property of DCT, this essentially decomposes the block sequence into a set of nearly uncorrelated scalar AR sequences, hence estimation can be performed separately for each frequency to exploit the quantization information readily available therein. We emphasize that while motion compensated prediction can be equivalently performed in either transform or pixel domain, the quantization, which is a highly non-linear operation, cannot be simply mapped into some equivalent pixel domain operation.

C. Statistical Model Estimation

Following the discussion in Sec. III-B, the evolution of transform coefficients along the motion trajectory is modeled by the AR process of (1), with x_n denoting a coefficient at a specific frequency in the current block, and x_{n-1} denoting the corresponding reference block coefficient at the same frequency, located in the previous frame. Much study has been devoted to establishing the probability distribution $p_Z(z_n)$ of the innovation term z_n, e.g., [19]– [23]. It is commonly recognized that this density is well approximated by the zero-mean Laplacian distribution

$$p_Z(z_n) = \frac{\lambda}{2} e^{-\lambda |z_n|} \tag{7}$$

whose statistical characteristics are determined by the model parameter λ. The maximum likelihood estimate of λ, given outcomes $z_0, \ldots, z_{N-1}$ of N independent draws of the random variable Z, is

$$\lambda_{ML} = \frac{N}{\sum_{i=0}^{N-1} |z_i|}. \tag{8}$$

Ideally, one would need to obtain the innovations of each motion trajectory, per frequency, from the original video signal, and substitute in (8) to estimate the corresponding Laplacian parameter. However, this approach involves a significant amount of side information. This can be avoided by instead estimating λ from information already available in

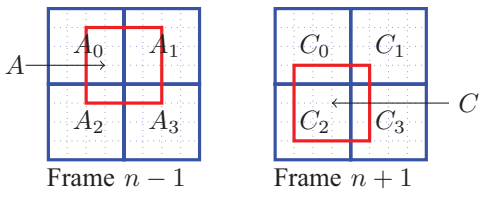

Fig. 3. Off-grid reference block overlaps up to four on-grid blocks.

the standard compatibly compressed video stream. Specifically, the encoded bit-stream contains the information needed to determine the reconstructed prediction errors of transform coefficients, which approximate the innovations of the transform domain AR processes. We hence propose a low-complexity approach that performs spatio-temporally adaptive estimation of the model parameter from the compressed bit-stream, and thereby accounts for statistical variations across the motion trajectories, from the compressed bit-stream.

A frame buffer is allocated for each decoded picture frame to store the mean absolute values of reconstructed prediction residuals of on-grid block transform coefficients. Let $\bar{r}_n^{B,m}$ denote this accumulative value of prediction residual at frequency m in block B of frame n. The reconstructed prediction residual is denoted by $\hat{r}_n^{B,m}$. Consider the computation of $\bar{r}_n^{B,m}$, at time instance n (Fig. 2). The decoder locates, in frame $n-1$, the positions of on-grid blocks $\{A_i\}$ that overlaps reference block A (Fig. 3), whose mean absolute values of prediction residuals at frequency m are known as $\{\bar{r}_{n-1}^{A_i,m}\}$. Similarly, the on-grid blocks of frame $n+1$ overlapping reference block C are identified as $\{C_i\}$, the reconstructed residuals of which are $\{\hat{r}_{n+1}^{C_i,m}\}$. The value of $\bar{r}_n^{B,m}$ is calculated by

$$\bar{r}_n^{B,m} = \frac{1}{9}\left(\sum_{i=0}^{3} \bar{r}_{n-1}^{A_i,m} + \sum_{i=0}^{3} \hat{r}_{n+1}^{C_i,m} + \hat{r}_n^{B,m}\right) \tag{9}$$

which effectively averages the prediction residuals along the same motion trajectory from past, current, and future frames. The model parameter is thus computed by

$$\lambda_n^{B,m} = \frac{1}{\bar{r}_n^{B,m}}. \tag{10}$$

We note that other more complicated algorithms (e.g., along the lines of [27]) may provide more precise estimation of the model parameter, at the expense of significant increment in decoding complexity.

D. Approximate ET Approach for Video Decoding With Single-Frame Delay

Having established the statistical model of temporal predictive coding in transform domain, we are now ready to apply to video decoding an ET delayed decoding approach that is compatible with the standard syntax. We restrict our discussion to the case of single frame latency in this section. An extension to incorporate multi-frame decoding delay will be considered in Sec. IV.

Consider the AR process $\{x_n\}$ of transform coefficients of one particular frequency along a motion trajectory. It is evident from (3) that optimal delayed decoding involves the pdf $p(x_n|\{I_l\}_{l\le n})$ of the transform coefficient in the current block, which is conditioned not only on the current interval I_n, but also $\{I_l\}_{l<n}$ corresponding to the same spatial frequency in all preceding blocks along the motion trajectory. Since the encoder transforms the prediction residuals of on-grid blocks, quantizes these transform coefficients, and encodes the resulting indices, the current interval I_n is readily available from the bit-stream. But this is generally not the case for the prior intervals $\{I_l\}_{l<n}$. Suppose the x_n of interest belongs to the block B in frame n of Fig. 2. Its preceding block A is not necessarily seated on the grid of frame $n-1$, and hence is not exactly the block that was transformed and coded as part of the bit-stream for that frame. Thus, the interval I_{n-1} is *not* available to the decoder. This is in general the case with other prior intervals as well. The issue is resolved by approximating $x_{n-1} \approx \hat{x}_{n-1}$, replacing the interval I_{n-1} with x_{n-1} (i.e, the DPCM decoder estimate of the sample x_{n-1} is assumed accurate, and uncertainty about x_{n-1} due to lossy quantization is neglected), and then appealing to the Markov property of the AR process to simplify (5) as

$$p(x_n|\{I_l\}_{l\le n}) \approx \begin{cases} \frac{p_Z(x_n-\hat{x}_{n-1})}{\int_{I_n} p_Z(x_n-\hat{x}_{n-1})dx_n}, & x_n \in I_n \\ 0, & \text{otherwise} \end{cases} \tag{11}$$

i.e., the intervals $\{I_l\}_{l\le n-1}$ do not provide any additional information about x_n when x_{n-1} is exactly known. In the above equation, the common assumption in motion compensated prediction that $\rho \approx 1$ is implicit.

The second requirement in (3) is the probability $p(I_{n+1}|x_n)$ of interval I_{n+1} in the next frame conditioned on x_n. A similar alignment problem is encountered here: the subsequent block is potentially off-grid in frame $n+1$, and thus the interval I_{n+1} might not be known to the decoder. This necessitates a second approximation. Note that the location of this block is already determined by the motion trajectory construction of Sec. III-A, and the regular standard decoder provides a coarse estimate of the pixels in this block. Transformation can now be applied to this pixel block. Denote $\hat{x}_{n+1}$ as the resulting transform coefficient at the same frequency as x_n. Hypothetically, if the interval I_{n+1} in which the true value of x_{n+1} resides was known, then the probability $p(I_{n+1}|x_n) = \int_{I_{n+1}} p_Z(x_{n+1} - \rho x_n)dx_{n+1}$. Since the interval I_{n+1} is unknown, we approximate

$$\begin{aligned} p(I_{n+1}|x_n) &\approx \int_{\hat{x}_{n+1}-\frac{\Delta}{2}}^{\hat{x}_{n+1}+\frac{\Delta}{2}} p_Z(x_{n+1} - \rho x_n)dx_{n+1} \\ &\approx p_Z(\hat{x}_{n+1} - x_n)\Delta \end{aligned} \tag{12}$$

with recourse to the assumption that the true value of x_{n+1} lies within an interval of length Δ around the coarse estimate $\hat{x}_{n+1}$, with its pdf conditioned on x_n nearly uniform on that interval, which is indeed the case at high bit-rates [28].

We note that while it will be experimentally demonstrated that such approximations provides significant gains as anticipated by the original ET algorithm of [8], a certain performance penalty would be introduced due to the use of future and past sample reconstructions instead of actual intervals, especially at low bit-rates, where the approximation

$x_{n-1} \approx \hat{x}_{n-1}$ and the high resolution quantizer assumption [28] employed in (11) and (12), respectively, are less precise. Such potential loss of optimality due to lack of access to exact future and past quantization intervals will be quantitatively evaluated in Sec. V.

Applying (11) and (12) to (3), one obtains

$$p(x_n|\{I_l\}_{l\le n+1}) = \begin{cases} \frac{p(x_n|\{I_l\}_{l<n})\ p(I_{n+1}|x_n)}{\int_{I_n} p(x_n|\{I_l\}_{l<n})\ p(I_{n+1}|x_n)dx_n}, & x_n \in I_n \\ 0, & \text{otherwise} \end{cases} \approx \begin{cases} \frac{p_Z(x_n-\hat{x}_{n-1})\ p_Z(\hat{x}_{n+1}-x_n)}{\int_{I_n} p_Z(x_n-\hat{x}_{n-1})\ p_Z(\hat{x}_{n+1}-x_n)dx_n}, & x_n \in I_n \\ 0, & \text{otherwise.} \end{cases} \quad (13)$$

Consequently, the optimal delayed reconstruction of x_n can be computed via (2). This procedure is performed for all transform coefficients in the current block, followed by the inverse transform to produce the pixel domain reconstruction. In our implementation, (2) is calculated as

$$E[x_n|\{I_l\}_{l\le n+1}] = \frac{\int_{I_n} x_n p(x_n|\{I_l\}_{l\le n+1})dx_n}{\int_{I_n} p(x_n|\{I_l\}_{l\le n+1})dx_n} \approx \frac{\int_{I_n} x_n p_Z(x_n-\hat{x}_{n-1})\ p_Z(\hat{x}_{n+1}-x_n)dx_n}{\int_{I_n} p_Z(x_n-\hat{x}_{n-1})\ p_Z(\hat{x}_{n+1}-x_n)dx_n} = \hat{x}_{n-1} + \frac{\int_{I'_n} z_n p_Z(z_n)\ p_Z(\hat{x}_{n+1}-\hat{x}_{n-1}-z_n)dz_n}{\int_{I'_n} p_Z(z_n)\ p_Z(\hat{x}_{n+1}-\hat{x}_{n-1}-z_n)dz_n} \quad (14)$$

where the approximation follows (13) and $I'_l = [a(i_l),\ b(i_l))$, which is equivalent to shifting the interval I_l to the left by $\hat{x}_{n-1}$. This is equivalent to centering the proposed ET framework around the previously reconstructed reference $\hat{x}_{n-1}$, and estimating the innovation related to it. Although mathematically equivalent to (2), the variant of (14) applies the ET approach to the innovation, instead of the AR process, and offers the advantage that the dynamic range of the intermediate computations is significantly reduced, thus requiring lower precision in hardware design to achieve the same numerical accuracy.

We summarize the reconstruction process of frame n by the proposed ET video decoder with single-frame delay as follows:

1) Decode as regular standard decoder does up to frame $n+1$.
2) Construct the motion trajectory as described in Sec. III-A, estimate the model parameters, and obtain $\hat{x}_{n-1}$ and $\hat{x}_{n+1}$ from regular reconstructions of the previous and subsequent frames, respectively.
3) Perform ET delayed decoding as given by (11)–(14) for each transform coefficient of every on-grid block in frame n, and employ inverse transform to produce the pixel domain representation.
4) Apply (optionally) normal spatial deblocking filter to the refined reconstruction of frame n to further remove blocking artifacts due to transform coding.

To provide an idea that is independent of processor or implementation, we indicate the complexity in terms of basic video coding modules, namely, block transformation and motion compensated reference generation, in Table I. We emphasize that the motion trajectory construction described in Sec. III-A makes the generation of reference blocks in future frames a low complexity operation, and completely circumvents the high complexity of standard motion estimation. It is also noteworthy that the construction enables parallelization of the additional computation in practical implementations.

TABLE I
DECODING COMPUTATIONAL COMPLEXITY PER BLOCK WITH SINGLE-FRAME DELAY

Functional Unit	H.264/AVC Decoder	ET Delayed Decoder
Forward DCT	0	2
Inverse DCT	1	1
Generating Reference Block	1	2

IV. MULTI-FRAME DELAYED DECODING

The backward recursion (6) of the generic ET delayed decoding allows exploitation of multiple future intervals, which is experimentally shown in [8] to provide additional performance gains for typical scalar AR sources. Future intervals are not directly available in video decoding, due to the problem of off-grid blocks as discussed in Sec. III, and the proposed ET video decoder in the single-frame delay setting already employs the reconstruction of future samples in (12) and (13) to approximate (3). Following the high resolution analysis of [28], the accuracy of this approximation largely depends on how closely a decoded future sample approximates its original value. Note that single-frame delayed decoding provides a refined reconstruction of each frame it is applied to, for instance, the reconstruction of frame $n+1$ is refined by using information from frame $n+2$. The improved reconstruction of frame $n+1$ can now be employed in a second instance of single-frame delayed decoding to refine the reconstruction of frame n. Thus, this second round of delayed-decoding improves the reconstruction of frame n beyond the single-frame delayed decoding, by effectively using information from both frame $n+1$ and $n+2$. Therefore, the single-frame delayed decoding approach of Sec. III can be recursively employed to implement multi-frame delayed decoding.

In particular, consider rebuilding frame n with L-frame latency. The decoder starts with regular non-delayed decoding up to frame $n+L$. It then runs single-frame delayed decoding to refine the reconstruction of transform coefficients of on-grid blocks in frame $n+L-1$, by using their quantization intervals and the corresponding reference from zero-delay reconstructions in frames $n+L-2$ and $n+L$. The refined transform coefficients of frame $n+L-1$ are inversely transformed to produce the refined pixel domain representation of frame $n+L-1$. This improved pixel domain version of frame $n+L-1$ implies improved estimates of transform coefficients of both on-grid and off-grid blocks of the frame, which are now employed as future reference reconstructions by a second instance of single-frame delayed decoding to improve the reconstruction of frame $n+L-2$. The decoder recursively performs this procedure L times, which successively enhances the reconstruction quality

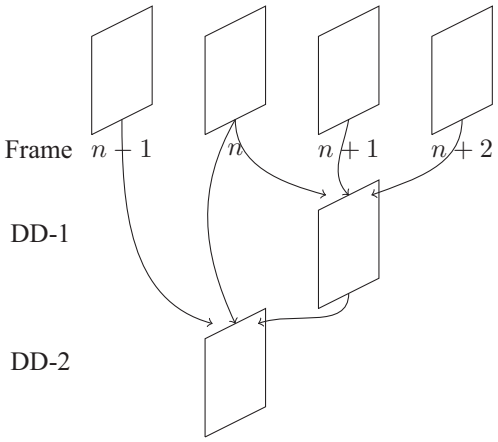

Fig. 4. ET video decoder with two-frame decoding delay: to estimate frame n, the decoder starts with regular decoding of the sequence up to frame $n+2$. It then performs single-frame delayed decoding of frame $n+1$, given its quantization intervals and the initially reconstructed reference frame n and $n+2$, during stage DD-1. The refined reconstruction of frame $n+1$ is then employed in stage DD-2, together with frame $n-1$ and the intervals of frame n, to rebuild frame n in a second single-frame delayed decoding framework.

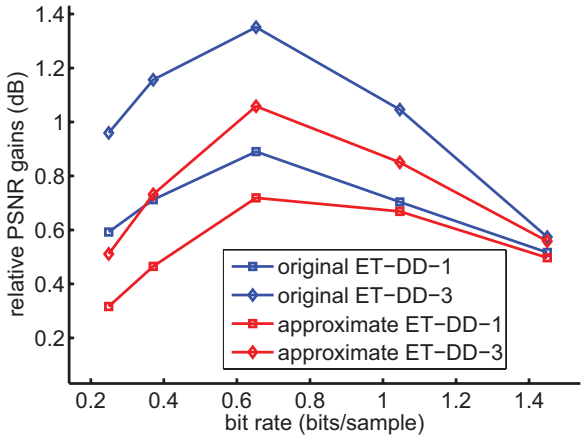

Fig. 5. Relative performance gains on synthetic data: scalar sequence forms a zero-mean AR process of correlation coefficient 0.95, whose innovations are i.i.d. with Laplacian pdf of unit variance. A dead-zone quantizer is employed to encode the sequence. The decoding performance of the original ET algorithm [8] and the approximate approach of Section III-D, which employs the reconstructions of previous and future samples, instead of the exact quantization intervals, is evaluated in terms of gains over the conventional nondelayed decoder.

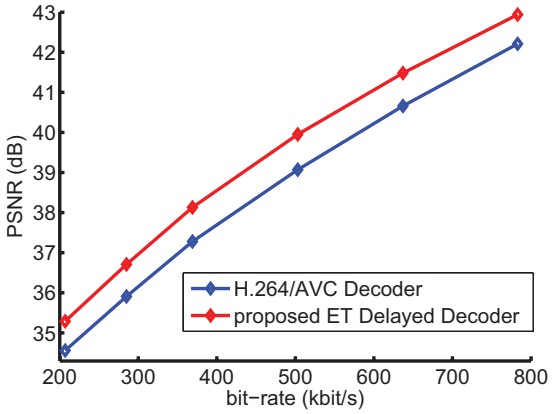

Fig. 6. Performance comparison of standard H.264/AVC decoder, and the proposed ET delayed decoder on test sequence *coastguard* at $QCIF$ resolution. A three-frame playback delay is used by the ET decoder.

of future frames from $n+L-1$ to $n+1$, and finally generates frame n where the future reference frame incorporates all accessible information up to frame $n+L$. Naturally this entails an increase in the decoder complexity by a factor of L. The ET video decoding process with two-frame latency is depicted in Fig. 4.

V. Simulation Results

A. Results With a Synthetic Source

We start with considering the performance of ET delayed decoding with a synthetic source model. Since a video decoder generally does not have access to the exact past and future quantization intervals, due to the interaction of motion compensation and spatial transform coding, the approximations of Sec. III-D are necessitated by the setting of video decoding. We first evaluate the typical loss due to such approximations in the context of a synthetic scalar AR model, and the DPCM setting of Fig. 1. Here the decoder does have access to all the quantization information about past and future samples, which enables the performance comparison of ET delayed decoding with and without using the exact quantization intervals. The zero-mean AR process is defined according to (1), where the driving innovation terms are i.i.d. Laplacian random variables of unit variance. A uniform threshold quantizer with a central dead-zone [20] is employed to encode the sequence. The rate is calculated as the first order entropy of the quantizer indices. Results compare the performance of the original ET delayed decoding and its approximate variant presented in Sec. III-D (and in Sec. IV for multiple samples delay) at decoding latency of one and three samples, denoted by ET-DD-1 and ET-DD-3, respectively. The performance is evaluated in terms of the reconstruction PSNR gains in dB relative to the conventional non-delayed decoder as depicted in Fig. 5. Clearly, both the original and approximate approaches attain superior decoding quality over non-delayed decoder, by exploiting future coding information. Both schemes achieve benefits from multiple future samples for additional performance gains, on top of those due to the use of single future sample. As expected in Sec. III-D, the gap between the original ET delayed decoder and its approximate variant proposed herein, i.e., the loss of optimality, is more pronounced at low bit-rates and tends to vanish at high bit-rates.

B. Results With Predictively Encoded Video Sequences

The proposed ET delayed decoder was then implemented within the H.264/AVC reference framework JM 16.2 for predictively encoded video sequences. The test video sequences were coded in $IPPP$ format at 30 fps by the standard recommended encoder operating at the extended profile, employing regular quarter-pixel motion search for inter frame prediction, single reference frame for motion compensation, dead-zone quantizers to the residuals, and context-based adaptive binary arithmetic coder for syntax elements. Two transform dimensions (4×4 and 8×8) were allowed. All encoding decisions, including intra-/inter-mode, macroblock partition

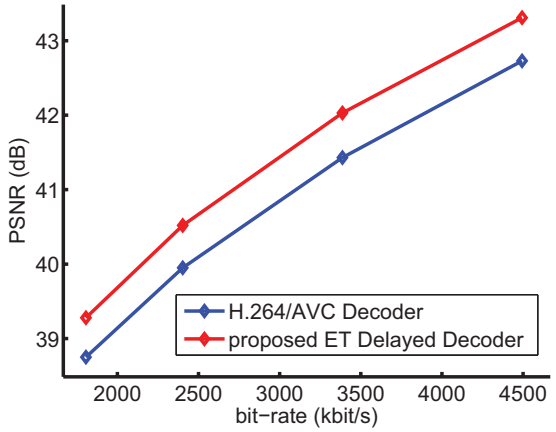

Fig. 7. Performance comparison of standard H.264/AVC decoder, and the proposed ET delayed decoder on test sequence *sheriff* at *SD* resolution. A three-frame playback delay is used by the ET decoder.

TABLE II

PERFORMANCE COMPARISON OF STANDARD H.264/AVC DECODER AND THE PROPOSED ET DELAYED DECODER FOR SEQUENCES AT QCIF RESOLUTION. ALL TEST SEQUENCES ARE ENCODED BY REGULAR STANDARD RECOMMENDED ENCODER AT EXTENDED PROFILE CONFIGURATIONS. THE PROPOSED ET DECODERS OPERATE AT SINGLE-FRAME, AND THREE-FRAME PLAYBACK DELAYS, DENOTED BY ET-DD-1 AND ET-DD-3, RESPECTIVELY

Test Sequence	Bit Rate (kbit/s)	PSNR (dB) Standard Decoder	ET-DD-1	ET-DD-3
foreman	414	42.08	42.46	42.55
	250	39.18	39.48	39.58
	200	38.00	38.30	38.40
	156	36.68	36.97	37.07
container	380	45.27	45.74	45.99
	224	42.11	42.43	42.56
	126	39.06	39.24	39.34
	98	37.76	37.89	37.98
bridge-far	334	43.19	43.51	43.64
	220	42.07	42.43	42.56
	132	41.04	41.32	41.41
	75	39.83	40.03	40.10
coastguard	636	40.66	41.46	41.48
	502	39.07	39.89	39.96
	368	37.28	38.01	38.14
	284	35.91	36.62	36.72
suzie	260	42.49	42.84	42.92
	200	41.16	41.48	41.58
	138	39.62	39.88	39.96
	106	38.40	38.63	38.69

TABLE III

PERFORMANCE COMPARISON OF STANDARD H.264/AVC DECODER AND THE PROPOSED ET DELAYED DECODER FOR SEQUENCES AT CIF RESOLUTION. ALL TEST SEQUENCES ARE ENCODED BY REGULAR STANDARD RECOMMENDED ENCODER AT EXTENDED PROFILE CONFIGURATIONS. THE PROPOSED ET DECODERS OPERATE AT SINGLE-FRAME, AND THREE-FRAME PLAYBACK DELAYS, DENOTED BY ET-DD-1 AND ET-DD-3, RESPECTIVELY

Test Sequence	Bit Rate (kbit/s)	PSNR (dB) Standard Decoder	ET-DD-1	ET-DD-3
harbour	3580	40.90	41.19	41.27
	2264	37.55	37.90	37.98
	1800	36.15	36.50	36.58
	1400	34.67	35.04	35.13
city	1950	41.27	41.57	41.72
	1086	38.15	38.43	38.57
	834	36.83	37.08	37.21
	622	35.46	35.69	35.78
bridge-close	2320	41.12	41.41	41.54
	1640	39.72	40.02	40.14
	1100	38.09	38.40	38.49
	796	36.89	37.16	37.23
waterfall	1466	41.42	41.84	41.99
	1080	39.96	40.35	40.50
	760	38.23	38.61	38.76
	568	37.01	37.30	37.43
soccer	1880	41.76	42.14	42.24
	1500	40.38	40.72	40.84
	1140	38.76	39.09	39.20
	910	36.20	36.49	36.57
galleon	1680	41.49	41.89	42.00
	1280	40.01	40.39	40.50
	946	38.12	38.46	38.58
	730	36.63	36.94	37.05
flower-garden	2360	40.11	40.48	40.53
	1528	36.64	37.08	37.17
	1210	35.00	35.47	35.59
	898	33.20	33.61	33.74

for motion search, transform block size, etc., were made in a rate-distortion optimization framework. The quantization parameters are fixed for encoding the entire sequence, and varied for each consecutive run to obtain multiple operating points. The in-loop deblocking filter was activated to remove blocking artifacts due to *spatial transform coding*. We note that our proposed ET delayed decoding effectively addresses the artifacts due to *temporal predictive coding*, in a suitably derived conditional expectation scheme.

The coded bit-stream was decoded by the ET delayed decoder, which employed the low-complexity adaptive model estimation and allowed up to three-frame decoding delay, to reconstruct the frame sequences. The same bit-stream was also decoded by the standard H.264/AVC decoder to generate a reconstructed sequence for comparison. We note that the standard syntax definitions are retained in the proposed decoding scheme, i.e., any standard compatible bit-stream can be decoded by the ET delayed decoder for enhanced reconstruction quality. The performance comparison for sequence *coastguard* at *QCIF* resolution is shown in Fig. 6. Clearly, the proposed ET delayed decoder achieves consistent gains over the standard H.264/AVC decoder, across a wide range of bit rates. Similar performance can be observed for sequence *sheriff* at *SD* resolution, as shown in Fig. 7.

Tables II-IV presents the performance of ET delayed decoding over test sequences of spatial resolutions ranging from *QCIF*, *CIF*, to *SD*, evaluated at typical operational points. It is worth noting that depending on the characteristics of video sequences, the operational bit-rate range of interest may differ quite considerably, e.g., sequences with intense motion activities tend to require higher bit rates than those stationary sequences do, in order to attain similar perceptual reconstruction quality. The ET decoder with single frame

TABLE IV
PERFORMANCE COMPARISON OF STANDARD H.264/AVC DECODER AND THE PROPOSED ET DELAYED DECODER FOR SEQUENCES AT SD RESOLUTION. ALL TEST SEQUENCES ARE ENCODED BY REGULAR STANDARD RECOMMENDED ENCODER AT EXTENDED PROFILE CONFIGURATIONS. THE PROPOSED ET DECODERS OPERATE AT SINGLE-FRAME, AND THREE-FRAME PLAYBACK DELAYS, DENOTED BY ET-DD-1 AND ET-DD-3, RESPECTIVELY

Test Sequence	Bit Rate (kbit/s)	PSNR (dB) Standard Decoder	ET-DD-1	ET-DD-3
stockholm	1316	43.34	43.75	43.89
	976	41.96	42.32	42.48
	764	40.78	41.11	41.26
	586	39.57	39.87	40.03
sheriff	4500	42.73	43.22	43.33
	2400	39.95	40.42	40.55
	1806	38.75	39.18	39.31
	1310	37.44	37.80	37.93
husky	6430	34.56	34.83	34.96
	4816	32.92	33.19	33.34
	3750	31.62	31.89	32.03
	2840	30.28	30.54	30.68

playback latency denoted by ET-DD-1 captures significant gains over the standard decoder, while further improvements in reconstruction quality can be achieved by trading off more latency, i.e., three-frame delay denoted by ET-DD-3 in the experiments. For perceptual quality comparison, a sample of reconstructed video clips is available for download at [29].

VI. CONCLUSION

A novel estimation-theoretic approach for delayed decoding of video sequences encoded via motion-compensated prediction is proposed. The approach models the temporal evolution of spatial transform coefficients of video blocks along a motion trajectory as an auto-regressive process, which is then exploited at the decoder to optimally reconstruct transform coefficients in the current frame by combining corresponding quantization intervals and information from both past and future frames in a minimum mean squared error estimator. Computationally efficient construction of the motion trajectory at the decoder is enabled by repurposing motion vectors of both current and future frames, already available in the bit-stream. The proposed approach adaptively estimates the model parameters from the regular non-delayed reconstruction and requires no additional side information, thereby retaining compatibility with existing video coding standards. Experiments provide evidence of significant performance gains over the regular non-delayed decoder for a wide variety of video sequences.

APPENDIX I
PROOF OF (3) IN SEC. II

Claim: The pdf of x_n conditioned on $\{I_l\}_{l\leq n+L}$ can be decomposed as

$$p(x_n|\{I_l\}_{l\leq n+L}) = \begin{cases} \frac{p(x_n|\{I_l\}_{l<n})p(\{I_l\}_{n<l\leq n+L}|x_n)}{\int_{I_n} p(x_n|\{I_l\}_{l<n})p(\{I_l\}_{n<l\leq n+L}|x_n)dx_n}, & x_n \in I_n \\ 0, & \text{otherwise.} \end{cases} \quad (15)$$

Proof: Let us denote $\{I_l\}_{l<n}$ and $\{I_l\}_{n<l\leq n+L}$ as events B and C, respectively. Hence, the conditional pdf $p(x_n|\{I_l\}_{l\leq n+L})$ can be rewritten as $p(x_n|B, I_n, C)$. The fact that $x_n \in I_n$ refines it as

$$p(x_n|B, I_n, C) = \begin{cases} \frac{p(x_n|B,C)}{\int_{I_n} p(x_n|B,C)dx_n}, & x_n \in I_n \\ 0, & \text{otherwise.} \end{cases} \quad (16)$$

Note that the above is equivalent to truncating $p(x_n|B, C)$ by the interval I_n, and normalizing to obtain a valid pdf. The Markov property of (1) indicates that given x_n, future intervals $\{I_l\}_{n<l\leq n+L}$ are independent of information preceding x_n (i.e., $\{I_l\}_{l<n}$)

$$p(C|x_n, B) = p(C|x_n). \quad (17)$$

Applying Bayes rule and (17) to $p(x_n|B, C)$, we can obtain

$$\begin{aligned} p(x_n|B, C) &= \frac{p(x_n, B, C)}{p(B, C)} \\ &= \frac{p(C|x_n, B)p(x_n, B)}{p(B, C)} \\ &= \frac{p(C|x_n)p(x_n|B)}{p(C|B)}. \end{aligned} \quad (18)$$

We then consider the denominator term $p(C|B) = p(B, C)/p(B)$. The total probability theorem states that $p(B, C) = \int_{I_n} p(x_n, B, C)dx_n$. Therefore

$$\begin{aligned} p(C|B) &= \frac{\int_{I_n} p(x_n, B, C)dx_n}{p(B)} \\ &= \frac{\int_{I_n} p(C|x_n)p(x_n, B)dx_n}{p(B)} \\ &= \int_{I_n} p(C|x_n)p(x_n|B)dx_n. \end{aligned} \quad (19)$$

Plugging (19) in (18), one can obtain

$$p(x_n|B, C) = \frac{p(C|x_n)p(x_n|B)}{\int_{I_n} p(C|x_n)p(x_n|B)dx_n}. \quad (20)$$

Taking (20) into (16), we have

$$p(x_n|\{I_l\}_{l\leq n+L}) = \begin{cases} \frac{p(C|x_n)p(x_n|B)}{\int_{I_n} p(C|x_n)p(x_n|B)dx_n}, & x_n \in I_n \\ 0, & \text{otherwise} \end{cases} \quad (21)$$

which implies that the conditional pdf of x_n can be indeed decomposed as

$$p(x_n|\{I_l\}_{l\leq n+L}) = \begin{cases} \frac{p(x_n|\{I_l\}_{l<n})p(\{I_l\}_{n<l\leq n+L}|x_n)}{\int_{I_n} p(x_n|\{I_l\}_{l<n})p(\{I_l\}_{n<l\leq n+L}|x_n)dx_n}, & x_n \in I_n \\ 0, & \text{otherwise.} \end{cases} \quad (22)$$

APPENDIX II
PROOF OF (6) IN SEC. II

Claim: The probability of $\{I_l\}_{n+m-1<l\leq n+L}$ conditioned on x_{n+m-1} can be decomposed as

$$\begin{aligned} p(\{I_l\}_{n+m-1<l\leq n+L}|x_{n+m-1}) &= \int_{I_{n+m}} p(\{I_l\}_{n+m<l\leq n+L}|x_{n+m}) \cdot \\ &\quad p_Z(x_{n+m} - \rho x_{n+m-1})dx_{n+m}. \end{aligned} \quad (23)$$

Proof: To verify this statement, we rewrite the left side of (23) as

$$\begin{aligned} &p(\{I_l\}_{n+m-1<l\le n+L}|x_{n+m-1}) \\ &\quad = p(I_{n+m}, \{I_l\}_{n+m<l\le n+L}|x_{n+m-1}) \\ &\quad = \int_{I_{n+m}} p(x_{n+m}, \{I_l\}_{n+m<l\le n+L}|x_{n+m-1}) d_{x_{n+m}} \end{aligned} \tag{24}$$

applying Markov property of (1) to which results in

$$\begin{aligned} &\int_{I_{n+m}} p(x_{n+m}, \{I_l\}_{n+m<l\le n+L}|x_{n+m-1}) d_{x_{n+m}} \\ &\quad = \int_{I_{n+m}} p(\{I_l\}_{n+m<l\le n+L}|x_{n+m}, x_{n+m-1}) \cdot \\ &\qquad p(x_{n+m}|x_{n+m-1}) d_{x_{n+m}} \\ &\quad = \int_{I_{n+m}} p(\{I_l\}_{n+m<l\le n+L}|x_{n+m}) \cdot \\ &\qquad p_Z(x_{n+m} - \rho x_{n+m-1}) d_{x_{n+m}}. \end{aligned} \tag{25}$$

Taking (25) into (24) gives

$$\begin{aligned} &p(\{I_l\}_{n+m-1<l\le n+L}|x_{n+m-1}) \\ &\quad = \int_{I_{n+m}} p(\{I_l\}_{n+m<l\le n+L}|x_{n+m}) \cdot \\ &\qquad p_Z(x_{n+m} - \rho x_{n+m-1}) dx_{n+m} \end{aligned} \tag{26}$$

which completes the proof.

REFERENCES

[1] L. D. Davisson, "Rate-distortion theory and application," *Proc. IEEE*, vol. 60, no. 7, pp. 800–808, Jul. 1972.

[2] D. S. Arnstein, "Quantization error in predictive coders," *IEEE Trans. Commun.*, vol. 23, no. 4, pp. 423–429, Apr. 1975.

[3] A. Hayashi, "Differential pulse code modulation of stationary Gaussian inputs," *IEEE Trans. Commun.*, vol. 26, no. 8, pp. 1137–1147, Aug. 1978.

[4] N. Farvardin and J. W. Modestino, "Rate-distortion performance of DPCM schemes for autoregressive sources," *IEEE Trans. Inf. Theory*, vol. 31, no. 3, pp. 402–418, May 1985.

[5] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003.

[6] M. L. Sethia and J. Anderson, "Interpolative DPCM," *IEEE Trans. Commun.*, vol. 32, no. 6, pp. 729–736, Jun. 1984.

[7] W.-W. Chang and J. D. Gibson, "Smoothed DPCM codes," *IEEE Trans. Commun.*, vol. 39, no. 9, pp. 1351–1359, Sep. 1991.

[8] V. Melkote and K. Rose, "Optimal delayed decoding of predictively encoded sources," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, Mar. 2010, pp. 3470–3473.

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

[10] E. Dubois and S. Aabri, "Noise reduction in image sequences using motion-compensated temporal filtering," *IEEE Trans. Commun.*, vol. 32, no. 7, pp. 826–831, Jul. 1984.

[11] O. G. Guleryuz and M. T. Orchard, "Rate-distortion based temporal filtering for video compression," in *Proc. IEEE Data Compress. Conf.*, Jan. 1996, pp. 122–131.

[12] O. G. Guleryuz and M. T. Orchard, "On the DPCM compression of Gaussian autoregressive sequences," *IEEE Trans. Inf. Theory*, vol. 47, no. 3, pp. 945–956, Mar. 2001.

[13] J. G. Apostolopoulos and N. S. Jayant, "Postprocessing for very low bit-rate video compression," *IEEE Trans. Image Process.*, vol. 8, no. 8, pp. 1125–1129, Aug. 1999.

[14] T.-S. Liu and N. Jayant, "Adaptive postprocessing algorithms for low bit rate video signals," *IEEE Trans. Image Process.*, vol. 4, no. 7, pp. 1032–1035, Jul. 1995.

[15] M. Barkowsky, J. Biakowski, B. Eskofier, R. Bitto, and A. Kaup, "Temporal trajectory aware video quality measure," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 266–279, Apr. 2009.

[16] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural video," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 335–350, Feb. 2010.

[17] A. K. Moorthy and A. C. Bovik, "Efficient video quality assessment along temporal trajectories," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 11, pp. 1653–1658, Nov. 2010.

[18] N. Ma and P. Ishwar, "On delayed sequential coding of correlated sources," *IEEE Trans. Inf. Theory*, vol. 57, no. 6, pp. 3763–3782, Jun. 2011.

[19] F. Bellifemine, A. Capellino, A. Chimienti, R. Picco, and R. Ponti, "Statistical analysis of the 2D-DCT coefficients of the differential signal for images," *Signal Process., Image Commun.*, vol. 4, pp. 477–488, Nov. 1992.

[20] G. J. Sullivan, "Efficient scalar quantization of exponential and Laplacian random variables," *IEEE Trans. Inf. Theory*, vol. 42, no. 5, pp. 1365–1374, Sep. 1996.

[21] H.-M. Hang and J.-J. Chen, "Source model for transform video coder and its application. I. Fundamental theory," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. 2, pp. 287–298, Apr. 1997.

[22] E. Lam and J. Goodman, "A mathematical analysis of the DCT coefficient distributions for images," *IEEE Trans. Image Process.*, vol. 9, no. 10, pp. 1661–1666, Oct. 2000.

[23] X. Li, N. Oertel, A. Hutter, and A. Kaup, "Laplace distribution based Lagrangian rate distortion optimization for hybrid video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 19, no. 2, pp. 193–205, Feb. 2009.

[24] J. Han, V. Melkote, and K. Rose, "Estimation-theoretic delayed decoding of predictively encoded video sequences," in *Proc. IEEE Data Compress. Conf.*, Mar. 2010, pp. 119–128.

[25] J. Bankoski, P. Wilkins, and Y. Xu, "Technical overview of VP8, an open source video codec for the web," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2011, pp. 1–6.

[26] D. Marpe, H. Schwarz, S. Bosse, B. Bross, P. Helle, T. Hinz, H. Kirchhoffer, H. Lakshman, T. Nguyen, S. Oudin, M. Siekmann, K. Suhring, M. Winken, and T. Wiegand, "Video compression using nested quadtree structures, leaf merging, and improved techniques for motion representation and entropy coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 20, no. 12, pp. 1676–1687, Dec. 2010.

[27] J. Han, V. Melkote, and K. Rose, "Transform domain temporal prediction in video coding with spatially adaptive spectral correlations," in *Proc. IEEE Multimedia Signal Process. Workshop*, Oct. 2011, pp. 1–6.

[28] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*. Norwood, MA: Kluwer, 1992.

[29] *sample Video Clips*. (2010) [Online]. Available: http://www.scl.ece.ucsb.edu/video/delayed_decoding/samples.rar

Jingning Han (S'10) received the B.S. degree in electrical engineering from Tsinghua University, Beijing, China, in 2007, and the M.S. degree in electrical and computer engineering from the University of California, Santa Barbara, in 2008, where he is currently pursuing the Ph.D.

He interned at Ericsson, Inc., Technicolor, Inc., and Google, Inc., in 2008, 2010, and 2012, respectively. His current research interests include video compression and networking.

Mr. Han was the recipient of the Outstanding Teaching Assistant Award from the Department of Electrical and Computer Engineering, University of California, Santa Barbara, in 2010 and 2011. He was a recipient of the Dissertation Fellowship in 2012. He was a recipient of the Best Student Paper Award at the IEEE International Conference on Multimedia and Expo in 2012.

Vinay Melkote (S'08–M'10) received the B.Tech. degree in electrical engineering from the Indian Institute of Technology Madras, Chennai, India, in 2005, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of California, Santa Barbara, in 2006 and 2010, respectively.

He is currently with the Sound Technology Research Group, Dolby Laboratories, Inc., San Francisco, CA, where he focuses on audio compression and related technologies. His current research interests include video compression and estimation theory. He interned with the Multimedia Codecs Division, Texas Instruments, India, in 2004, and with the Audio Systems Group, Qualcomm, Inc., San Diego, in 2006.

Dr. Melkote is an Associate Member of the Audio Engineering Society. He was a recipient of the Best Student Paper Award at the IEEE International Conference on Acoustics, Speech, and Signal Processing in 2009. He is a member of the IEEE Signal Processing Society's technical committee for Audio and Acoustic Signal Processing.

Kenneth Rose (S'85–M'91–SM'01–F'03) received the Ph.D. degree from the California Institute of Technology, Pasadena, in 1991.

He then joined the Department of Electrical and Computer Engineering, University of California, Santa Barbara, where he is currently a Professor. His main research activities are in the areas of information theory and signal processing, including rate-distortion theory, source and source-channel coding, audio and video coding and networking, pattern recognition, and non-convex optimization.

Dr. Rose was a co-recipient of the 1990 William R. Bennett Prize Paper Award of the IEEE Communications Society and the 2004 and 2007 IEEE Signal Processing Society Best Paper Award.