An Estimation-Theoretic Framework for Spatially Scalable Video Coding

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

***Abstract*— This paper focuses on prediction optimality in spatially scalable video coding. It draws inspiration from an estimation-theoretic prediction framework for quality (SNR) scalability earlier developed by our group, which achieved optimality by fully accounting for relevant information from the current base layer (e.g., quantization intervals) and the enhancement layer, to efficiently calculate the conditional expectation that forms the optimal predictor. It was central to that approach that all layers reconstruct approximations to the same original transform coefficient. In spatial scalability, however, the layers encode different resolution versions of the signal. To approach optimality in enhancement layer prediction, this paper departs from existing spatially scalable codecs that employ pixel domain resampling to perform interlayer prediction. Instead, it incorporates a transform domain resampling technique that ensures that the base layer quantization intervals are accessible and usable at the enhancement layer despite their differing signal resolutions, which in conjunction with prior enhancement layer information, enable optimal prediction. A delayed prediction approach that complements this framework for spatial scalable video coding is then provided to further exploit future base layer frames for additional enhancement layer coding performance gains. Finally, a low-complexity variant of the proposed estimation-theoretic prediction approach is also devised, which approximates the conditional expectation by switching between three predictors depending on a simple condition involving information from both layers, and which retains significant performance gains. Simulations provide experimental evidence that the proposed approaches substantially outperform the standard scalable video codec and other leading competitors.**

***Index Terms*— Spatial scalability, scalable video coding, estimation-theoretic prediction, transform domain resampling, delayed prediction.**

I. Introduction

In SCALABLE video coding (SVC), the video sequence is encoded into a single bit-stream consisting of multiple layers with progressively higher spatial, temporal, or quantization resolutions, thus enhancing streaming flexibility, without retaining multiple independent bit-streams of different quality levels. The higher resolution layers will typically benefit from differential coding from lower layers, based on inter-layer prediction, which results in significant bit-rate reduction. Thus SVC is an attractive solution for multimedia streaming in modern network infrastructures serving decoders of diverse display resolutions and channel capacities [1]. Of the various features of SVC, the focus of this paper is on spatial scalability. For simplicity of exposition, we present our discussion in the context of a two-layered spatial SVC codec, while emphasizing that the proposed approach is extensible to multiple layers, and provide coding performance evaluation in both contexts.

A spatial SVC scheme consists of downsampling a high resolution video sequence to lower resolutions, and coding these resolutions into separate layers. The lowest resolution signal is coded by a base layer coder, which is essentially a single-layer coder, while the enhancement layers encode information necessary to reconstruct the sequence at progressively higher spatial resolutions, up to its original form. At the enhancement layer, the current video frame can be predicted from a combination of its reconstruction at the base layer, and a motion compensated reference from prior enhancement layer coded frames. For instance, in the single-loop design, the prediction could either be the enhancement layer motion compensated reference, or its linear combination with the quantized base-layer residual. More details on existing spatial SVC approaches are provided in Section II, and also available in [2]. The inter-layer prediction is commonly performed in the pixel-domain: the base layer reconstructed pixels (or base layer reconstructed residuals) are upsampled via interpolation to the enhancement layer resolution prior to prediction, and the resultant prediction residuals are then transformed and coded. Substantial earlier research has focused on the quality of such interpolation, which impacts the prediction accuracy, and hence coding performance [2], [3].

The ad hoc nature of the above schemes, which switch between, or *linearly* combine, reconstructions from different sources, strongly motivates the search for a true estimation-theoretic (ET) approach to spatial SVC, where all the information available to the enhancement layer coder is fully and optimally exploited. Inspiration is drawn from an ET technique proposed earlier by our group in [4] for a special setting of SVC: *quality* (SNR) scalability, where the *same* original sequence is coded by all the layers but at different quantization resolutions. Thus, the true value of a transform coefficient must lie in the interval determined by base layer quantization. This observation effectively captures all the information provided by the base layer, and is the central postulate of the ET approach in [4]. A conditional probability density

Manuscript received April 12, 2013; revised December 31, 2013; accepted June 2, 2014. Date of publication June 18, 2014; date of current version July 9, 2014. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Charles Boncelet.

J. Han and K. Rose are with the Department of Electrical and Computer Engineering, University of California at Santa Barbara, 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 at Santa Barbara, Santa Barbara, CA 93106 USA. He is now with 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.2014.2331761

function (pdf), truncated by (and normalized to) the base layer quantization interval, was employed to compute the exact conditional expectation that forms the optimal prediction. The ET approach of [4] was further enhanced by allowing delayed prediction [5], and incorporating resilience to packet loss [6], all in the setting of quality scalability.

Challenges arise in the general case of *spatial* scalability, where the base layer encodes a *downsampled* version of the sequence encoded by the enhancement layer. This means that different layers quantize different transform coefficients. Consequently, the precise quantization interval and other related information from the base layer cannot be used directly to optimize prediction at the enhancement layer. We hence develop a unified ET framework that is tailored to enable full exploitation of base layer information, in conjunction with regular inter-frame motion compensation, for optimal enhancement layer prediction. In order to render base layer quantizer intervals accessible and relevant to the enhancement layer codec, the proposed method departs from regular pixel domain resampling techniques, and generates the downsampled base layer in the transform domain. It discards high frequency transform coefficients of a larger transform applied to the original signal and rebuilds the downsampled signal from the remaining low frequency coefficients, thus providing a direct correspondence between coefficients of the two layers. Note that the base-layer video sequence that the end-user receieves and which is used for enhancement layer prediction is the downsampled *and* encoded sequence. Subjective experiments evidence that the reconstructed base layer sequences when pixel and transform domain downsamplers are employed have similar quality, affirming that although the choice of the transform domain downsampler was motivated by the need for ET prediction of the enhancement layer, it has no significant impact on the base layer coding performance.

This unified ET framework opens the door to further incorporate future base layer information in enhancement layer prediction for additional performance gains. We note that a fundamental property of the SVC paradigm is that the base layer is coded independently of the enhancement layers, to ensure the worst case availability of coarse reconstruction. Hence the coding at the enhancement layer can in principle be 'delayed' to exploit the reconstruction of future base layer frames, which potentially provide useful information to calibrate the enhancement layer prediction. The proposed ET approach in this work hence integrates, in the transform domain, the three disparate sources of information – quantization intervals from the current base layer frame, and motion compensated information from both prior enhancement layer and *future* base layer frames – in a conditional pdf, the expectation over which constitutes the optimal enhancement layer prediction given a coding delay.

Related prior work includes [7] which proposed a rate-distortion optimized selection between three modes of enhancement layer prediction, all computed in the pixel domain: inter-frame prediction using enhancement layer motion compensation alone, a linear combination of the enhancement layer motion compensation with the interpolated base layer quantized residual (sometimes referred to as pyramid prediction, see [7]), and a linear combination of a high-pass filtered version of the enhancement layer motion compensation with the interpolated base layer reconstruction (sometimes referred to as subband prediction). A theoretical analysis of the rate-distortion performance was also derived under the assumption of a stationary Gaussian model in [8]. While in [7] the prediction mode was chosen on a per-enhancement layer block basis, earlier work by Tan et al [9] introduced a conditional replacement technique, which similar to our approach here operates in the transform domain, and where low frequency transform coefficients of an enhancement layer block could be individually predicted via either pyramid prediction or subband prediction, and the choice is based on a heuristic: it is conditioned on whether the residual of the corresponding transform coefficient in the base-layer was quantized to zero. While we defer details to Section II, it can be argued that, unlike the ET prediction approach proposed here, neither [7] nor [9], optimally exploits all available information, i.e., the base layer quantization interval and the enhancement layer reference, simultaneously, for optimal prediction. Experiments demonstrate the considerable enhancement layer coding gains achieved by the proposed ET framework, over standard H.264 extensions for spatial SVC and other leading competitors. We note that while the proposed approach was implemented and tested in the H.264/AVC Scalable Video Coding Extension reference framework, the principle is generally applicable to other motion compensation based predictive codecs including VP9 [10] and HEVC [11].

Some preliminary results were reported in our recent work [12] and [13] to provide initial validation of the potential benefits of the ET approach, where a few base layer coding heuristics, e.g., skip mode, forcing zero-coefficient (see Section VI) that were empirically known to provide compression performance benefits, were temporarily disabled due to their interference with the base layer quantization intervals. The proposed scheme in this work eschews such limitations by employing a switch mechanism that selects the prediction sources depending on the collocated base layer block coding mode. As will be discussed in Sec III-B, the derived predictor is formulated as a non-linear conditional expectation, which involves numerical computations of exponentials. To overcome this complexity intricacy, a low-complexity variant of the original ET framework is further devised in this paper that only involves simple arithmetic operations and requires no pre-assumption of the statistics of video signal, while retaining superior enhancement layer coding performance.

II. Background

This section provides relevant background on the H.264/AVC SVC extension and its variants. The standard SVC coder spatially downsamples the original input sequence, and the resultant lower dimension frames are coded by a standard single-layer codec into the base layer. The choice of the down-sampler is not standardized, and commonly employed strategies include the windowed sinc filter, pixel decimation, etc.. The enhancement layer prediction of the

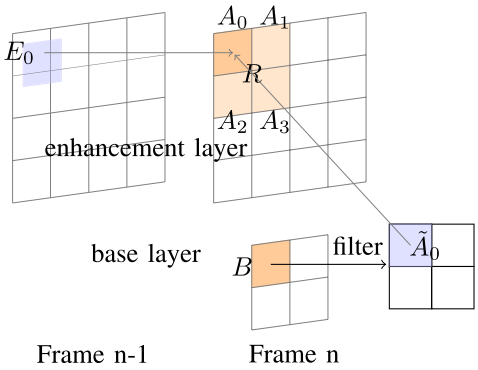


Fig. 1. An example of pixel-domain enhancement layer prediction in spatial SVC.

standard codec follows the single-loop design [2], [14], where the prediction modes include inter-frame motion compensation, a sum of the motion-compensated reference and the upsampled reconstructions of base layer residual, as well as the upsampled base layer reconstructions (if the collocated base layer block is intra-coded). As an aside, note that the second of these prediction modes is the same as the pyramid prediction mode described in Section I in the context of [7]. The encoder selects amongst all the possible modes the one that minimizes the rate-distortion cost, per macroblock. An illustration of the process is provided by Fig. 1. To encode block A_0 at the enhancement layer, the coder starts with motion search from previously reconstructed frames in the same layer to generate a motion-compensated reference block E_0. It then calculates the position of the base layer block B obtained by downsampling the region R. A separable four-tap polyphase interpolation filter, in conjunction with the deblocking operation, is employed in the standard to upsample the base layer reconstruction of B to a block of the same spatial dimension as R. The subblock $\tilde{A}_0$ in the resultant interpolation is collocated with A_0. Either E_0 or $\tilde{A}_0$ could be used as the enhancement layer prediction, and both are tested by the encoder to find the one that minimizes the rate-distortion cost. Here, for the purpose of illustration, we have implicitly assumed that the base layer block B is intra-coded. If B was instead inter-coded, the decoded residuals for the block would be interpolated and summed up with E_0 to obtain the prediction for A_0. A more detailed reference on the single-loop design can be found in [2] and [14].

Another popular alternative is the multi-loop design [15] where, in addition to the modes available in the single-loop design, the base layer reconstructed pixels could be used for enhancement layer prediction even when the base layer block is inter-coded. In other words, the multi-loop design requires full reconstruction of the base layer at the decoder, while the single-loop design could forgo various base layer operations if only the enhancement layer reconstruction is desired. In [7] a variant of the multi-loop design was proposed where enhancement layer prediction employs one of the following modes: inter-frame prediction from a motion compensated enhancement layer reference, pyramid prediction, or subband prediction (a linear combination of the high-pass filtered motion-compensated enhancement layer reference and the upsampled base layer reconstruction). Effectively, the subband prediction mode uses the base layer reconstruction as prediction for low frequency transform coefficients, and the motion-compensated enhancement layer reference as prediction for high frequency transform coefficients. The approach in [7] is reported to provide notable gains over single-loop prediction. However, none of the prediction modes in this approach (nor in single-loop design) fully utilize all the information available for enhancement layer prediction. For instance, these prediction modes do not exploit the quantization interval information available from the base layer, which encapsulates all base layer information on the transform coefficient, and hence all the information made available by the base layer for enhancement layer prediction. Note, in particular, that downsampling, upsampling, and prediction are performed in the pixel domain, thus precluding any attempt to optimally utilize such interval information, which is only accessible in the transform domain.

In [9], an enhancement layer prediction approach is proposed, which operates in the transform domain. Similar to the approach proposed here, downsampling is accomplished by discarding high frequency transform coefficients of the enhancement layer block, i.e., brick-wall filtering. The upsampling/interpolation in the reverse direction is accomplished by filling in zeros for the missing higher frequencies. Then, a conditional replacement approach to transform domain prediction was described where, on a per-transform coefficient basis, the enhancement layer encoder chooses either the base layer reconstruction of that coefficient, or the corresponding transform coefficient in the motion-compensated enhancement layer reference block. The decision is based on a heuristic: If the base layer quantizes the residual to zero and hence provides no correction to its reference, the encoder prefers the enhancement layer reference; otherwise, if the base layer does provide an update, its reconstruction is preferred. Note that in this approach the prediction for a transform coefficient is either solely derived from the base layer or solely from the enhancement layer, and thus the two sources of information are not jointly utilized. In contrast, the focus of the proposed work is to achieve exactly such joint utilization: for each enhancement layer transform coefficient, the base layer quantizer interval information is optimally combined with the corresponding enhancement layer motion compensated prediction in an ET framework.

Other related spatial SVC work includes a substantial volume of research devoted to designing the pixel-domain decimation and interpolation filters, and evaluation of performance benefits [2]. A notable method was proposed in [3], where the upsampling filter is derived to match the downsampling operation while accounting for the quantization noise in the base layer reconstructed pixels.

III. The Unified Estimation-Theoretic Framework for Resampling and Prediction

As noted in Section I, the prevalent ad hoc approach to enhancement layer prediction in spatial SVC combines

base layer reconstructed pixels (or residuals) with motion-compensated enhancement layer reference, and does not guarantee optimal utilization of *all* available information. This motivates the ET approach described in this section, which jointly optimizes the framework for downsampling, upsampling, and enhancement layer prediction to maximally utilize the information extractable from the base and enhancement layers. In the discussion that follows, each base layer block is of dimension $M \times M$, and is obtained by downsampling a block of size $N \times N$ at the resolution of the enhancement layer.

A. Transform Domain Resampling

We assume separability of the 2D transform, i.e., it is accomplished by applying 1D operations sequentially along the vertical and horizontal directions. Hence, for clarity of exposition, we first present the main ideas in the framework of a 1D transform. Consider a vector of pixels $\underline{a} = [a_0, a_1, \ldots, a_{N-1}]^T$, with inter-pixel correlation close to 1. Here the superscript T denotes transposition. The optimal approach to convert $\underline{a}$ into a vector of dimension $M(< N)$ is to apply the Karhunen-Loeve transform (KLT) to fully decorrelate the samples and discard the lower energy $N - M$ coefficients. It is well known that, for certain Markov processes, DCT exhibits decorrelation and energy compaction properties approaching that of the KLT, and is commonly adopted as a substitute due to its low implementation complexity. Let T_N denote the N-point DCT matrix, and hence $\underline{\alpha}_N = T_N\underline{a}$ is the DCT representation of vector $\underline{a}$. Define continuous functions

$$f_0(t) = \sqrt{\frac{1}{N}}; \quad f_j(t) = \sqrt{\frac{2}{N}} cos(j\pi t), \quad j = 1, \ldots, N-1, \tag{1}$$

as cosine functions with periods that are sub-multiples of the time interval $[0, 1]$. Thus, the j^{th} basis function (row) of T_N can be generated by sampling $f_j(t)$ at time instances $t = \frac{1}{2N}, \frac{3}{2N}, \ldots, \frac{2N-1}{2N}$. Consider the continuous-time signal $a(t) = \sum_{j=0}^{N-1} \alpha_j f_j(t)$, where α_j is the j^{th} transform coefficient in $\underline{\alpha}_N$. Sampling $a(t)$ at the rate $\frac{1}{N}$ with initial shift $\frac{1}{2N}$ yields exactly the discrete-time signal $\underline{a}$.

Now define,

$$g_0(t) = \sqrt{\frac{1}{M}}, \quad g_j(t) = \sqrt{\frac{2}{M}} cos(j\pi t), \quad j = 1, \ldots, M-1, \tag{2}$$

the analog cosine functions which when sampled at rate $\frac{1}{M}$ yield the basis functions for a DCT of dimension M. The approximation (in mean squared error sense) to the signal $a(t)$ using only M of the N transform coefficients in $\underline{\alpha}_N$ is obtained by choosing the M coefficients of lowest frequency:

$$\tilde{a}(t) \approx \sum_{j=0}^{M-1} \alpha_j f_j(t) = \sum_{j=0}^{M-1} \left(\sqrt{\frac{M}{N}} \alpha_j \right) g_j(t). \tag{3}$$

This implies that the N-point pixel vector $\underline{a}$ can be downsampled by a factor of $\frac{N}{M}$ to $\underline{b}$ as:

$$\underline{b} = \sqrt{\frac{M}{N}}\ T_M^T \begin{pmatrix} I_M & 0_M \end{pmatrix} T_N \underline{a}, \tag{4}$$

where I_M and 0_M denote the identity and null matrices, respectively, of dimension $M \times M$. Conversely, the up-sampling from the M-point pixel vector $\underline{b}$ to an N-tuple can be accomplished by inserting zeros as high frequency coefficients:

$$\hat{\underline{a}} = \sqrt{\frac{N}{M}}\ T_N^T \begin{pmatrix} I_M \\ 0_M \end{pmatrix} T_M \underline{b}. \tag{5}$$

Under the assumption that DCT closely approximates KLT in performance, the resultant $\hat{\underline{a}}$ has minimum mean squared distance from the original vector $\underline{a}$, and downsampling to $\underline{b}$ maximally preserves the information in $\underline{a}$. Related material on DCT domain resampling can be found in [16] and [17]. Although we described this resampling in the 1D framework, the extension to 2D pixel blocks is straightforward. The downsampling (or upsampling) can be sequentially applied to the vertical and horizontal directions. Subjective experiments described in Section VI indicated that this transform domain resampling approach can in general serve as an alternative to the pixel-domain downsampling and interpolation traditionally employed in spatial SVC, by demonstrating their perceptual equivalence. However, as discussed next, this resampling method is particularly advantageous for the proposed ET spatial SVC paradigm.

B. Optimal Enhancement Layer Prediction

We now describe the ET approach to prediction at the enhancement layer. Similar to the standard approach, each frame (at the spatial resolution of the enhancement layer) is partitioned into macroblocks (say, of size 16x16), and each macroblock is coded with inter-layer or inter-frame prediction, or in intra mode. Transforms may be applied to the prediction residual at sub-macroblock resolution (say, 4x4 and 8x8)[1], followed by quantization and entropy coding. Windowing and cropping operations, e.g., "pan and scan" technique, are employed to tailor the frame size of each layer to fit the block-wise operations, which also provide flexibility in the choice of transform dimensions to perform the downsampling. We hence assume the block (transform) dimension used for encoding the base layer is $M \times M$ as used by the DCT employed for downsampling.

Consider encoding the enhancement layer blocks $\{A_i, i = 0, \ldots, 3\}$ in frame n (Fig. 2). The entire region R is mapped into block B in the base layer frame via the *transform domain downsampling* previously described in Section III-A. Let $x_n^e(i, j)$, where $i, j \in \{0, \ldots, N-1\}$, denote the value of the transform coefficient at frequency (i, j) obtained by applying a DCT of size $N \times N$ to R. Using (3), the first $M \times M$ transform coefficients of the resultant DCT are scaled appropriately to yield $x_n^b(i, j)$, $i, j \in \{0, \ldots, M-1\}$, the transform coefficients of the base layer block B:

$$x_n^b(i, j) = \frac{M}{N} x_n^e(i, j), i, j \in \{0, \ldots, M-1\}. \tag{6}$$

These coefficients may be subjected to an $M \times M$ inverse DCT to yield the downsampled, base layer pixel block B, which

[1]Recent standardization efforts propose to use larger macroblock size and transform dimension for high definition or higher resolution video.

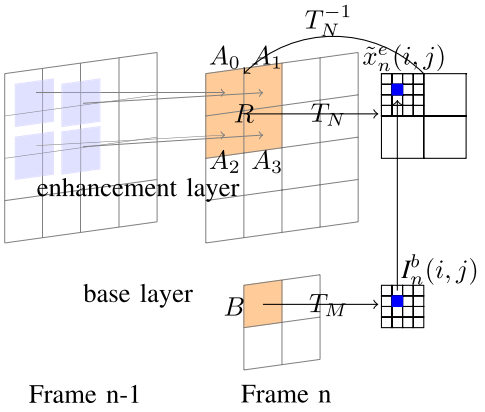


Fig. 2. Estimation-theoretic enhancement layer prediction in spatially scalable video coding: the enhancement layer block R is predicted in the transform domain. The optimal prediction of transform coefficient at frequency (i, j), denoted by $\tilde{x}_n^e(i, j)$, is formulated as the centroid of pdf conditioned on information of both layers (11).

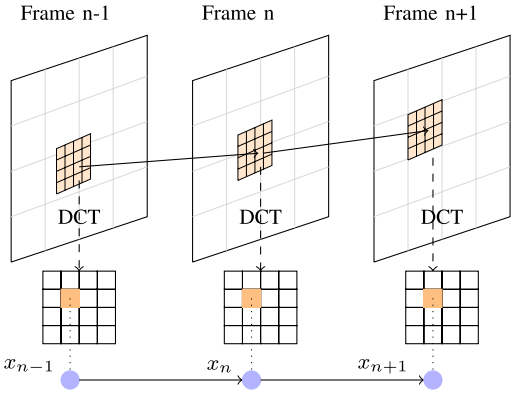


Fig. 3. Transform domain perspective: blocks of DCT coefficients along a motion trajectory are modeled as an AR process per frequency.

is effectively compressed as usual by the base layer coder. The base layer quantization process essentially prescribes an interval $I_n^b(i, j)$ that contains the true value of $x_n^b(i, j)$. This interval summarizes all the information provided by the base layer about the transform coefficient $x_n^b(i, j)$.

The traditional course of action would now be to upsample the base layer reconstruction of block B. In accordance with Section III-A this would entail zero-padding the $M \times M$ DCT of the reconstruction of block B to yield an $N \times N$ block of transform coefficients, which is then appropriately scaled by the inverse of the scaling applied in (6), and inverse transformed to get a pixel domain approximation of block R in Fig. 2. This could then be combined in the pixel domain with the enhancement layer reconstruction of earlier frames, and used for prediction in the current frame. However, such an approach that combines reconstructions in the pixel domain suffers from significant under-utilization of the information provided by the base layer.

Let:

$$I_n^e(i, j) = \left\{ \frac{N}{M} x | x \in I_n^b(i, j) \right\}, \; i, j \in \{0, \ldots, M-1\} \quad (7)$$

i.e., $I_n^e(i, j)$ is the interval obtained by scaling the base-layer quantization interval $I_n^b(i, j)$ by the factor $\frac{N}{M}$. Due to transform domain resampling, the following relation holds:

$$x_n^e(i, j) \in I_n^e(i, j), \; i, j \in \{0, \ldots, M-1\}, \quad (8)$$

which implies that the base layer quantization interval directly translates into information about transform coefficients at the enhancement layer. We note that such information only emerges in the *transform domain*, and cannot be expressed in the pixel domain, due to the fact that quantization is a highly non-linear operation performed in the transform domain. The ET prediction approach described next improves coding performance by exploiting all the available information in its purest form.

We model blocks of DCT coefficients along the same motion trajectory as an auto-regressive (AR) process per frequency (Fig. 3). Thus, $x_n^e(i, j)$ and the corresponding transform coefficient, $x_{n-1}^e(i, j)$, in the (uncoded) motion-compensated reference of block R, conform to the first order AR recursion:

$$x_n^e(i, j) = \rho x_{n-1}^e(i, j) + z_n(i, j), \quad (9)$$

where $z_n(i, j)$ denotes the independent and identically distributed (i.i.d.) innovation term drawn from probability density function (pdf) $p_Z(z_n(i, j))$. Following the implicit assumption in conventional pixel domain motion compensation, but without loss of generality, we set the correlation coefficient $\rho \approx 1$ at all frequencies. Assuming that the enhancement layer reference approximates the original coefficient, $\hat{x}_{n-1}^e(i, j) \approx x_{n-1}^e(i, j)$, we approximate the conditional pdf

$$p(x_n^e(i, j) | \hat{x}_{n-1}^e(i, j)) \approx p_Z(x_n^e(i, j) - \hat{x}_{n-1}^e(i, j)).$$

In the absence of additional base layer information, the best prediction of $x_n^e(i, j)$ would simply be $\hat{x}_{n-1}^e(i, j)$, the default inter-frame estimate. But the base layer indicates that $x_n^e(i, j) \in I_n^e(i, j)$ for $i, j \in \{0, \ldots, M-1\}$, which refines the conditional pdf of $x_n^e(i, j)$ to

$$p(x_n^e(i, j) | \hat{x}_{n-1}^e(i, j), I_n^e(i, j)) = \begin{cases} \dfrac{p_Z(x_n^e(i, j) - \hat{x}_{n-1}^e(i, j))}{\int_{I_n^e(i,j)} p_Z(x_n^e - \hat{x}_{n-1}^e(i, j)) dx_n^e} & x_n^e(i, j) \in I_n^e(i, j), \\ 0 \quad else. \end{cases} \quad (10)$$

Note that this is equivalent to centering the innovation pdf at $\hat{x}_{n-1}^e(i, j)$, restricting it to the interval $I_n^e(i, j)$ (a highly non-linear operation), and then normalizing to obtain a valid pdf (Fig. 4). The optimal predictor at the enhancement layer is now given by

$$\tilde{x}_n^e(i, j) = E\{x_n^e(i, j) | \hat{x}_{n-1}^e(i, j), I_n^b(i, j)\} = \begin{cases} E\{x_n^e(i, j) | \hat{x}_{n-1}^e(i, j), I_n^e(i, j)\}, & i, j \in \{0, \ldots, M-1\}, \\ \hat{x}_{n-1}^e(i, j), & else. \end{cases} \quad (11)$$

The above equation describes the transform coefficients of the enhancement layer prediction for the entire $N \times N$ region R in Fig. 2. The prediction residual at the enhancement layer could be directly calculated in the transform domain and quantized. However, in practice a hybrid transform coder that

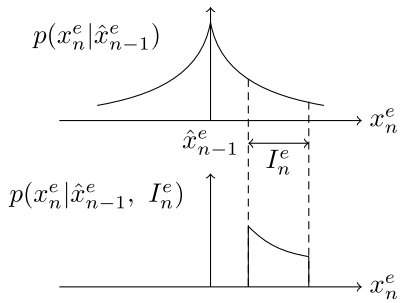


Fig. 4. The probability density function of x_n^e conditioned on both base and enhancement layer information is effectively the innovation pdf centered at $\hat{x}_{n-1}^e$, truncated by and normalized to interval I_n^e provided by the base layer.

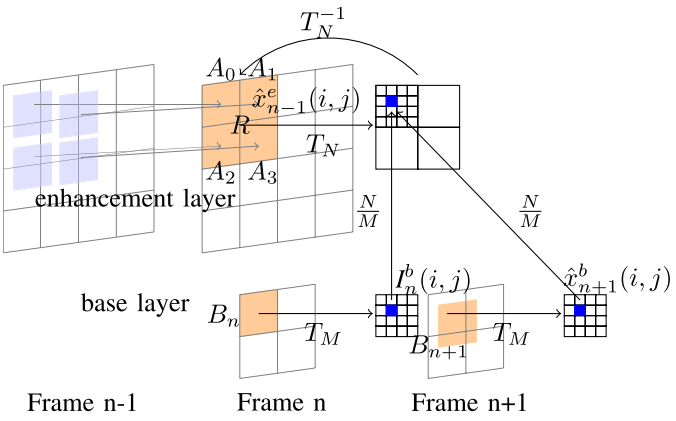


Fig. 5. Estimation-theoretic enhancement layer delayed prediction.

allows multiple transform block sizes to optimize the trade off between the coding performance of stationary signal and the changes in statistic is known to provide certain coding advantage. To preserve such flexibility, this transform domain prediction of R is therefore inverse transformed to generate the pixel domain prediction for each individual block A_i. Subsequently, as in the standard codec, the pixel domain prediction residuals of block A_i are calculated and then transformed, quantized, and entropy coded.

In the implementation we will assume that the innovation pdf is Laplacian [18]–[22], i.e.,

$$p_Z(z_n) = \frac{1}{2}\lambda e^{-\lambda|z_n|}, \tag{12}$$

where the parameter λ is frequency dependent in accordance with our earlier work [4], [23], and [24]. In the experiments of Section VI, this frequency dependent Laplacian parameter λ was estimated from a separate set of training sequences using maximum-likelihood estimate and fixed throughout all the test cases. Given outcomes z_0, z_1, $\cdots$, z_{L-1} of L independent draws of random variable Z, the maximum-likelihood estimate of Laplacian parameter is

$$\lambda = \frac{L}{\sum_{i=0}^{L-1}|z_i|}. \tag{13}$$

Ideally, one would need to obtain the innovations at each frequency from the original video signal, and substitute in (13) to estimate the corresponding Laplacian parameter. In our experiments, we further assumed that in high bit-rate cases the reconstructed prediction errors closely approximate these innovations, and hence simply substituted them in (13) to obtain the requisite λ per frequency. We note that the actual value of Laplacian parameter could also vary over time and spatial location. A motion-trajectory adaptive approach along the lines of [23] could potentially provide more precise parameter estimate and hence improved coding performance, which is beyond the scope of this work.

IV. Delayed Enhancement Layer Prediction

An important feature of scalable coding is that the base layer can be decoded independently of enhancement layers, which allows the enhancement layer coder potential access to information about future base layer frames, at a given coding latency relative to the base layer. This section considers means to exploit such future information, in addition to the current base layer and prior enhancement layer information, in a scheme that complements the above ET framework, to further refine the enhancement layer prediction, and thereby achieve considerable performance gains on top of the non-delayed ET prediction.

For simplicity, let us focus on the setting where the enhancement layer sequence encoding is delayed by one frame relative to the base layer, i.e., when the enhancement layer coder encodes frame n, it has access to base layer information for frame $(n+1)$. The proposed scheme in this case is illustrated using Fig. 5, where the enhancement layer block R in frame n corresponds to base layer block B_n. The motion trajectory that includes this base layer block is continued into the future frame $n+1$ to arrive at base layer block B_{n+1}, in other words we first consider the construction of the motion compensated future reference for B_n in frame $n+1$. One approach is to run a full motion search with reconstructed frame $n+1$ as the reference frame. Note however that this would mandate that the same motion search also be conducted at the decoder in order to generate an enhancement layer prediction that is identical to the one used at the encoder, with obvious implications on decoder complexity. We thus propose a simpler alternative that exploits the already available base layer motion vector information for frame $n+1$, that maps on-grid blocks in frame $n+1$ to their (potentially) off-grid reference blocks in frame n. The coder first identifies the locations of these reference blocks in frame n for all the inter-frame coded blocks in frame $(n+1)$. Then for each on-grid block B_n in frame n, the coder finds one of such reference blocks that maximally overlaps with it, reverses the associated motion vector of frame $n+1$, which relative to the position of B_n on the grid identifies the required reference block B_{n+1}. A similar inverse motion search method has been employed in our earlier work on an optimal delayed decoding scheme for predictively encoded video sequences in a regular single-layer coder [23], [24]. Let the transform coefficients of B_{n+1} denoted by $\hat{x}_{n+1}^b(i,j)$, as shown in Fig. 5. We now describe the proposed ET approach that improves coding performance by specifically utilizing $I_n^e(i,j)$ and $\hat{x}_{n+1}^b$,

in conjunction with inter-frame prediction at the enhancement layer.

Following the above transform domain modeling of (9), and adding to the previously assumed $x^e_{n-1}(i,j) \approx \hat{x}^e_{n-1}(i,j)$, a complementary approximation $x^e_{n+1}(i,j) \approx \frac{N}{M}\hat{x}^b_{n+1}(i,j)$, the pdf of x^e_n conditioned on the previous enhancement layer motion-compensated reference $\hat{x}^e_{n-1}(i,j)$, current quantization interval $I^e_n(i,j)$, and *future* base layer reference $\hat{x}^b_{n+1}(i,j)$ is thus obtained[2]:

$$\begin{aligned} &p(x^e_n(i,j)|\hat{x}^e_{n-1}(i,j), I^e_n(i,j), \hat{x}^b_{n+1}(i,j)) \\ &\approx \frac{p(x^e_n|\hat{x}^e_{n-1}, I^e_n)\cdot p(\hat{x}^b_{n+1}|x^e_n)}{\int_{I^e_n} p(x^e_n|\hat{x}^e_{n-1}, I^e_n)\cdot p(\hat{x}^b_{n+1}|x^e_n)dx^e_n} \\ &\approx \begin{cases} \dfrac{p_Z(x^e_n-\hat{x}^e_{n-1})\cdot p_Z(\frac{N}{M}\hat{x}^b_{n+1}-x^e_n)}{\int_{I^e_n} p_Z(x^e_n-\hat{x}^e_{n-1})\cdot p_Z(\frac{N}{M}\hat{x}^b_{n+1}-x^e_n)dx^e_n}, & x^e_n \in I^e_n, \\ 0, \qquad else. \end{cases} \end{aligned} \quad (14)$$

Here we have applied the Bayes rule, followed by the Markov property of AR process of (9): given the current sample $x^e_n(i,j)$, a future sample $x^e_{n+1}(i,j)$, or approximately $\frac{N}{M}x^b_{n+1}(i,j)$, is conditionally independent of the past sample $x^e_{n-1}(i,j)$ (see Appdx. VII for proof). We note that in the above equation, the *causal* pdf of $x^e_n(i,j)$, i.e., $p(x^e_n(i,j)|\hat{x}^e_{n-1}(i,j), I^e_n(i,j))$, is weighted by $p(\hat{x}^b_{n+1}(i,j)|x^e_n(i,j))$, the probability density of the *known future outcome* to obtain the one-sample delayed pdf of (14), which incorporates all available information at the enhancement layer coder, at up to one frame coding delay. The overall conditional pdf is then truncated by and re-normalized to the quantization interval, the centroid of which forms the optimal predictor at the enhancement layer with one frame coding delay.

In practice, we impose the restriction that the overlap area between B_n and the reference block it maximally overlaps through motion compensation should be greater than a prescribed threshold to justify the assignment of the inverse motion vector. Thus it is possible that, occasionally, the block B_n will not be matched with any inverse motion compensated reference B_{n+1} in frame $(n+1)$. In such cases, (14) degenerates to the non-delayed pdf of (10), due to the absence of reliable future base layer information.

Further note that for high frequency coefficients where i or $j \in \{M, M+1, \ldots, N-1\}$, both $I^e_n(i,j)$ and $\hat{x}^b_{n+1}(i,j)$ are not available, and the best prediction of $x^e_n(i,j)$ is simply $\hat{x}^e_{n-1}(i,j)$, the default inter-frame motion compensated estimate. In summary, the optimal predictor at the enhancement layer is given by

$$\tilde{x}^e_n(i,j) = \begin{cases} E\{x^e_n(i,j)|\hat{x}^e_{n-1}(i,j), I^e_n(i,j), \hat{x}^b_{n+1(i,j)}\}, \\ \qquad\qquad i,j \in \{0,\ldots,M-1\}, \\ \hat{x}^e_{n-1}(i,j), \qquad else. \end{cases} \quad (15)$$

The above equations describe the transform coefficients prediction at the enhancement layer for the entire $N \times N$ region R in Fig. 2. Similar to the non-delayed ET prediction scheme, the region R is now inversely transformed to generate the pixel-domain prediction, and the prediction residual for each individual block A_i is then coded. While in the above a delay of one frame has been prescribed, the approach can be generalized to exploit multiple future base layer frames along the line of our recent work on the problem of delayed *decoding* in single layer video codecs [23].

V. Switched Prediction Approximation to the Estimation-Theoretic Scheme

Under assumptions of a Laplacian innovation pdf in (10), closed form expressions for the ET prediction of (11) and (15) were derived. These closed form expressions in turn lead to a low-complexity approximation of the optimal ET prediction proposed in this section, which only involves simple arithmetic, but largely retains the enhancement layer coding performance gains. This approach is inspired by earlier work by our group in the context of quality (PSNR) scalability [25]. For the simplicity of exposition, we focus our discussion on the causal ET prediction (11).

Assuming scalar quantization of the coefficients in the base-layer, we denote the interval of interest

$$I^e_n = [a,b) = \left[\frac{N}{M}a', \frac{N}{M}b'\right),$$

where $I^b_n = [a', b')$ is the base layer interval. Note that the limits of the base layer interval a' and b' can be determined by the base layer reconstruction $\hat{x}^b_n$ and quantizer index i^b_n. Modeling the pdf of the innovation process $\{z_n\}$ as Laplacian, the computation of $\tilde{x}^e_n(i,j) = E\{x^e_n(i,j)|\hat{x}^e_{n-1}(i,j), I^e_n(i,j)\}$, $i,j \in \{0,\ldots,M-1\}$, in (11) can be classified into the following cases[3].

Case 1: $\hat{x}^e_{n-1} < a$

In this case we can evaluate (11) as

$$\begin{aligned} \tilde{x}^e_n &= \frac{\int_{I^e_n} x_n e^{-\lambda(x_n-\hat{x}^e_{n-1})}dx_n}{\int_{I^e_n} e^{-\lambda(x_n-\hat{x}^e_{n-1})}dx_n} \\ &= \frac{\int_{I^e_n} x_n e^{-\lambda x_n}dx_n}{\int_{I^e_n} e^{-\lambda x_n}dx_n} \\ &= \frac{1}{\lambda} + a + (b-a)\frac{e^{-\lambda(b-a)}}{e^{-\lambda(b-a)}-1} \\ &= \frac{1}{\lambda} + a + \Delta\frac{e^{-\lambda\Delta}}{e^{-\lambda\Delta}-1}, \end{aligned} \quad (16)$$

where $\Delta = \frac{N}{M}(b'-a')$ is the size of the interval of interest I^e_n (solely a function of the index i^b_n), and λ is a frequency dependent Laplacian model parameter defined in (12). Note that under the assumption of Laplacian innovations, given λ and the quantization index i^b_n the above expression can be employed to calcuate the prediction exactly. Moreover, the third term in this expression can be pre-calculated for each quantization index and for a number of values of the λ parameter, and can be stored in the form of a table. Rather than evaluating the third term in the above equation exactly for the specific transmitted/estimated λ, it could be approximated

[2]To avoid cumbersome expressions, the frequency index (i,j) is omitted throughout the equation.

[3]The frequency index (i,j) is omitted in the discussion when there is no risk of confusion.

by the entry in the table for the nearest tabulated λ. A typical choice of the base layer quantizer is the deadzone uniform quantizer, in which case all cells, except the one containing the origin, are of the same size. Let $i_n^b = 0$ refer to this center cell of the quantizer. It is then necessary to store the reconstructions for only two conditions on the quantization index, $i_n^b = 0$ and $i_n^b \neq 0$. This look-up table approach thus significantly reduces the complexity of the estimation process, while requiring very little memory.

Case 2: $\hat{x}_{n-1}^e > b$

Similar to the prior case, the conditional expectation can be simplified as

$$\tilde{x}_n^e = \frac{\int_{I_n^e} x_n e^{\lambda(x_n - \hat{x}_{n-1}^e)} dx_n}{\int_{I_n^e} e^{\lambda(x_n - \hat{x}_{n-1}^e)} dx_n} = -\frac{1}{\lambda} + a + \Delta \frac{e^{\lambda\Delta}}{e^{\lambda\Delta} - 1}, \quad (17)$$

Similar to Case 1 above, given λ and the quantization index i_n^b this expression provides the prediction exactly, and a close low-complexity approximation can be derived that employs a table look-up.

Case 3 $\hat{x}_{n-1}^e \in I_n^e$

Again, in this case, it is possible to derive an exact expression for the prediction in terms of λ and the boundaries of the interval I_n^e. However,the interval I_n^e contains the center (peak) of Laplacian pdf, which typically dominates the centroid calculation and thus an approximation for the prediction is:

$$\tilde{x}_n^e = \hat{x}_{n-1}^e.$$

Thus, conditioned on where the motion compensated value $\hat{x}_{n-1}^e$ falls relative to the interval I_n^e, the enhancement layer prediction module *switches* between the above three simple predictors. We refer to this low-complexity approximation of the optimal ET approach as switched prediction.

VI. Results

We now describe experiments and results that compare the enhancement layer compression performance of the proposed spatial SVC approaches to existing methods. However, we first provide evidence to support the use of the unconventional tranform domain downsampler to obtain the base layer sequence.

A. Validation of the Use of the Transform-Domain Downsampler

Since the focus of this paper is on enhancement layer coding performance we will restrict the base layer coded sequence for all the competing codecs to be the same – and obtained via transform domain downsampling and encoding the result. Note that the transform domain downsampler is mandatory to facilitate the proposed ET approaches to SVC, while the standard SVC encoding process is somewhat agnostic to which downsampling approach is used. Since one could argue that the standard SVC encoder has been "tuned" to pixel domain downsampling, we justify our usage of the transform domain downsampler thus: (a) we present subjective test results that show that transform domain downsampling is a viable alternative as far as base-layer quality goes, (b) the enhancement layer coding performance (PSNR vs bit-rate) of a standard encoder is essentially invariant to which of the two downsamplers is used to obtain the base layer sequence.

TABLE I
PERCEPTUAL TEST RATING

Score	Quality
100	Identical
80	Excellent
60	Good
40	Acceptable
20	Poor
0	Gray scene

It is reasonable to expect that a downside of employing the transform domain resampler of Section III-A is the possibility of blocking artifacts due to brickwall filtering. The proposed ET approach continues to use the motion compensation reference in the previous frame (not subject to any brickwall filtering) to generate the prediction for the current frame, and it *refines* the low frequency components of this prediction by exploiting quantization interval information from the base layer: thus blocking-artifacts are not present in this prediction itself. The potential for blocking artifacts, however, does exist in the downsampled base layer sequence. Motivated by the observation that what the end-user recieves in the base-layer is the *encoded* low-resolution sequence, where quantization artifacts might in fact dominate ringing artifacts due to brickwall filtering, we conducted a subjective test to assess the quality of the base layer coded sequences when the standard pixel-domain and proposed transform-domain downsamplers are used respectively.

The subjective test required the viewer to visually compare two downsampled and coded versions of the same original video sequence (one using the pixel domain downsampler and the other using the transform domain downsampler) against a reference uncoded version. Since pixel domain downsampling is generally accepted as the standard, we employed the uncoded pixel domain downsampled sequence as the reference. The two coded base layer sequences are obtained at similar bit-rates (subject to small differences due to encoder constraints) – Table II provides the bit-rates for the two coded versions of different test clips featured in the test. To avoid fluctuation in perceptual quality due to rate control, $IPPP$ format with fixed quantization parameter was employed across the entire sequence. The sequences were coded at frame rate of 30 Hz. A relatively high target bit-rate was chosen to ensure that the decoded clips closely resembled the source sequences. The codec allowed normal intra/inter prediction modes selected in a rate-distortion optimization framework, optional 8×8 transform, single reference frame, context-based binary arithmetic coding, and the default in-loop deblocking filter.

The subjective test was blind and the coded sequences for each test clip were randomly ordered. All clips were shown at equal distance to the viewer, who could place the clips in any

TABLE II
THE RESOLUTIONS AND BIT-RATES OF CODED SEQUENCES IN THE SUBJECTIVE TEST

Test Clip	Frame Size	Bit Rate (kbps)	
		pixel domain	transform domain
bus	176×144	758	738
foreman	176×144	207	218
mobile	176×144	1087	1115
city	352×288	1347	1328
harbour	352×288	1915	2103
soccer	352×288	1161	1181
oldtown	960×512	2355	2269
parkjoy	960×512	12844	12995

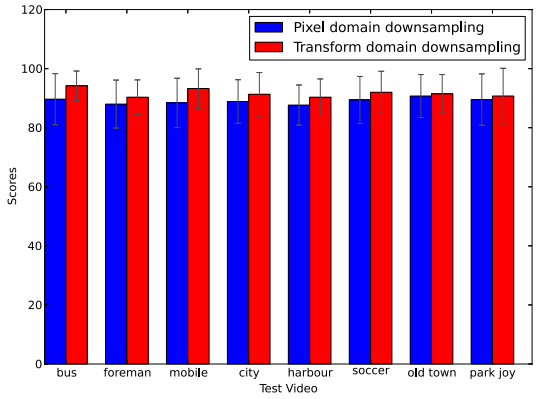


Fig. 6. Subjective test results demonstrating the perceptual equivalence of the encoded transform-domain and pixel-domain downsampled base layer sequences. The error-bars denote the standard deviation around the mean.

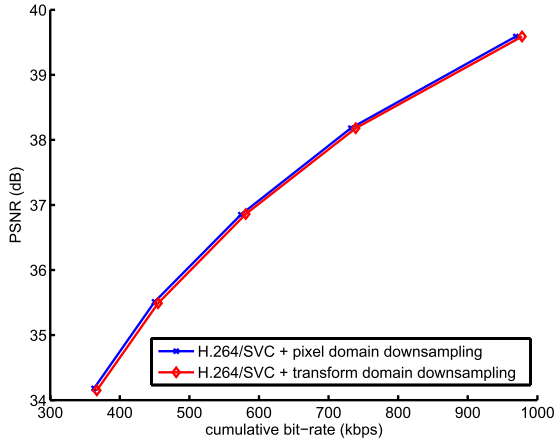


Fig. 7. H.264/SVC enhancement layer coding performance comparison for the two downsampling methods. Original sequence is *foreman* at *CIF* resolution. Base layer sequences generated at *QCIF* Resolution by (i) pixel domain and (ii) transform domain downsampling, were coded at 170 *kbps* and 179 *kbps*, respectively.

desired relative position, e.g., side-by-side, overlapping, etc., playback at any frame rates, and pause at any frame instance. The viewer was required to carefully compare the two coded clips, and grade them relative to the reference on a scale of 0-to-100, where a general guideline for mapping perceptual quality to a test score is provided by Table I. A total of 15 viewers participated the perceptual experiments.

The mean subjective scores, and standard deviation around the mean, for both base-layer codecs/downsamplers are summarized per-test clip in Fig. 6. As evident, the two coded versions have statistically identical perceptual quality indicating no visible degradation attributable specifically to transform domain downsampling. The reader can access the test clips and viewing instructions at [26].

We next consider enhancement layer coding performance of the H.264/SVC reference codec, which employs single-loop pixel domain enhancement layer prediction, with base layer sequences generated using the two downsampling methods. In the experiments, both layers were coded at frame rate 30 Hz in the $IPPP$ format. A constant quantization parameter per layer was applied across the entire sequence. We fixed base layer quantization parameter and varied that of the enhancement layer to obtain the operational points. The codec employed default intra/inter prediction modes, regular motion compensated prediction, single reference frame, optional 8×8 transform, etc. The H.264/SVC enhancement layer coding performance associated with pixel and transform domain downsampled base layer was compared over various sequences and no measurable difference in performance was found. For example the performance comparison on *foreman* is presented in Fig. 7. Clearly, transform-domain downsampling causes no degradation of the standard SVC's enhancement layer compression performance.

B. Spatial SVC Codecs Compression Performance Evaluation

The proposed SVC approaches as well as the competitors were implemented in the H.264/AVC Scalable Video Coding Extension JSVM 9.19 framework. Since the focus is on enhancement layer prediction efficiency, and given the perceptual equivalence of pixel-domain and transform-domain downsampling followed by encoding as described in the previous section, in order to guarantee fairness of comparison of the different enhancement layer coding approaches we have decided to utilize the same base layer bitstream in the case of all codecs. Further, given the context of spatial scalability, we restricted our comparison to encoding at frame rate 30 Hz and in $IPPP$ format for all layers - bi-directional prediction (B-frames) typically encountered in the context of temporal scalability was disabled in the experiments.

All competing codecs employed regular quarter-pixel motion search for inter-frame prediction, single reference frame for motion compensation, dead-zone quantization of

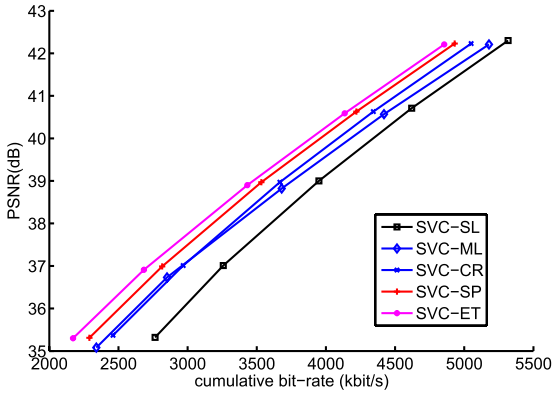


Fig. 8. Comparison of the coding performance of the competing spatial SVC approaches: The test sequence is $harbour$ at CIF resolution. The base layer is at $QCIF$ resolution, and is coded at 1200 $kbit/s$.

residuals, context-base adaptive binary arithmetic coding for syntax elements, intra/inter mode selection, and optional 8×8 transform for enhancement layer. The encoding decisions were made within a rate-distortion optimization framework. The quantization step sizes were fixed per layer, and varied for each consecutive run to obtain different operating points.

The following enhancement layer codecs are compared in the results shown in this section:

1) **SVC-ET**: The proposed ET approach (Section III-B) for optimal causal enhancement layer prediction implemented in the H.264/SVC framework, and consequently employs the transform domain upsampling of Section III-A. In addition to the ET prediction mode, a inter-frame prediction mode is also allowed where the prediction for all transform coefficients of an enahncement layer block is the corresponding enhancement layer motion compensation reference.
2) **SVC-ET-DP**: The proposed ET approach (Section IV) for optimal delayed prediction at the enhancement layer implemented in the H.264/SVC framework. The remaining details are similar to SVC-ET.
3) **SVC-SP**: The proposed switched predictor (Section V). The look-up tables are derived for the fixed values of λ employed in SVC-ET, and thus SVC-SP deviates from SVC-ET only due to the Case 3 prediction described in Section V.
4) **SVC-SL**: Standard H.264/SVC employing single-loop enhancement layer prediction (a competing codec). Incorporates the three prediction modes described for this approach in Section II. The base layer residuals in the pyramid prediction mode (when base layer is inter-coded) or base layer reconstructions (in intra-mode) are interpolated in pixel-domain via a 4-tap poly-phase filter and deblocking operations.
5) **SVC-ML**: A competing multi-loop codec that closely emulates [7], derived from SVC-SL with the replacement of the inter layer residual prediction mode with a subband prediction mode as stipulated by [7]. The high-pass filtering of the enhancement layer motion compensation, however, is implemented using the transform domain upsampling scheme: the scaled reconstructions of the base layer transform coefficients form the low frequency transform coefficients of the prediction at the enhancement layer, while the reminder high frequency coefficients are predicted by the corresponding coefficients of the enahancement layer motion compensation.
6) **SVC-CR**: A competing codec derived from SVC-SL, with the replacement of the pyramid prediction mode with a prediction based on [9]. The transform domain upsampling of Section III-A is employed in this mode, and on a per-transform coefficient basis at the enhancement layer conditional replacement is employed to calculate the prediction. Although this codec borrows the conditional replacement of [9], we note that inclusion of this mode in a rate-distortion optimization scheme in the H.264/SVC framework has not been implemented in prior work.

We note that a common practice employed in many video encoder implementations is to force certain quantized coefficients to zero value, if it happens that only a single coefficient in the block is quantized to level +/- 1 while all others are at zero level. It is empirically known that such a strategy improves the overall base layer coding performance. However in the ET framework such an approach in the base layer coder results in erroneous interval information about the modified coefficient at the enhancement layer, i.e., the quantization interval available to the enhancement layer does not contain the original value of the coefficient. Given that the decoder needs to replicate the same prediction, the enhancement layer codec needs to work without the knowledge of which coefficient is modified to zero level. However, it does know that such a modification would have been made only if the all base layer coefficients have been quantized to zero. In our implementations (SVC-ET, SVC-ET-DP, and SVC-SP) we allow such a 'zero-forcing' operation at the base layer, and overcome its impact on later enhancement layer coding by using inter-frame motion-compensated prediction alone when the collocated base layer block has all-zero coefficients. Equivalently, the base layer interval for all coefficients in this case is assumed to be the entire real line[4]. Similarly, if a base layer block is coded in *skip* mode, no base layer interval information is available for the enhancement layer. Again, in this setting, the enhancement layer only employs the motion-compensated reference from the previously reconstructed frame at the same layer for prediction.

C. Discussion of Results

We first discuss the performance of the proposed non-delayed codecs in comparison with the competition. The enhancement layer coding performance for the corresponding five codecs for the sequence $harbour$ at CIF resolution is

[4]A more advanced approach to recouping the base layer information in this scenario, is to use a finitely bound interval around zero assembled by intervals associated with levels +/-1 and 0.

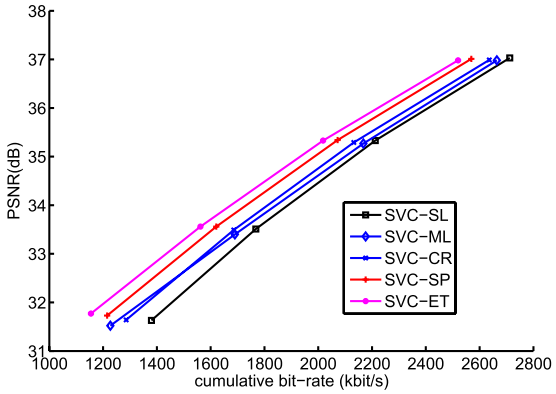


Fig. 9. Comparison of the coding performance of the competing spatial SVC approaches: The test sequence is *harbour* at CIF resolution. The base layer is at $QCIF$ resolution, and is coded at 590 $kbit/s$.

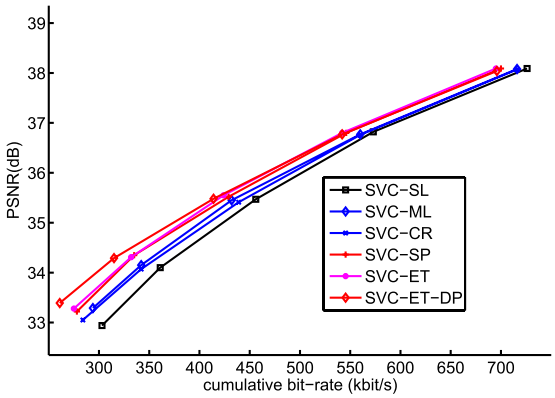


Fig. 12. Comparison of the coding performance of the competing spatial SVC approaches: The test sequence is *foreman* at CIF resolution. The base layer is at $QCIF$ resolution, and is coded at 185 $kbit/s$.

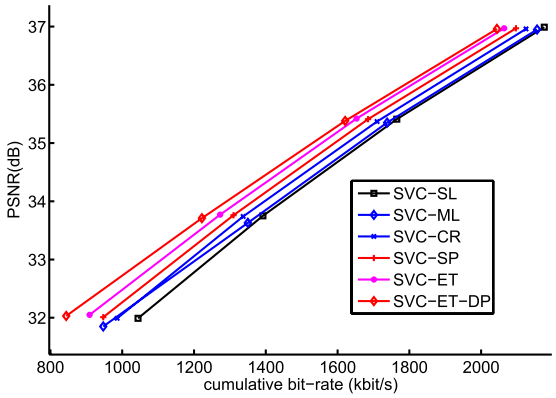


Fig. 10. Comparison of the coding performance of the competing spatial SVC approaches: The test sequence is *coastguard* at CIF resolution. The base layer is at $QCIF$ resolution, and is coded at 360 $kbit/s$.

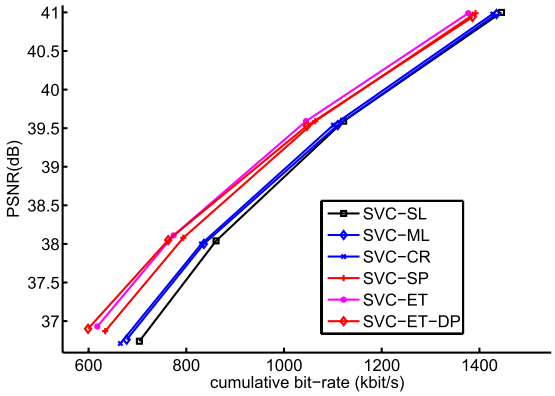


Fig. 13. Comparison of the coding performance of the competing spatial SVC approaches: The test sequence is *foreman* at CIF resolution. The base layer is at $QCIF$ resolution, and is coded at 350 $kbit/s$.

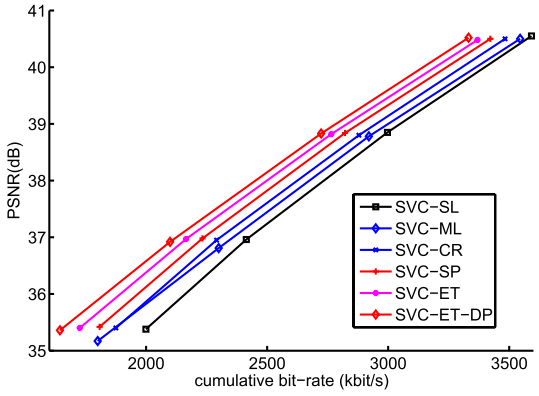


Fig. 11. Comparison of the coding performance of the competing spatial SVC approaches: The test sequence is *coastguard* at CIF resolution. The base layer is at $QCIF$ resolution, and is coded at 700 $kbit/s$.

shown in Fig. 8. Clearly the proposed unified ET approach for optimal prediction - SVC-ET - significantly outperforms in the competition. The approximate SVC-SP is not far behind SVC-ET and continues to outperform the competitors at all rates. The conditional replacement of [9] (SVC-CR) and the subband prediction mode of [7] (SVC-ML) enable improved performance over the standard SVC-SL, as expected. Similar enhancement layer performance improvements were obtained in the settings of different base layer bit-rates as shown in Fig. 9.

Fig. 10 compares both proposed (delayed and non-delayed) optimal prediction approaches versus the competition. The utility of delayed prediction at the enhancement layer is evidenced by the improved performance of SVC-ET-DP compared to SVC-ET. We note that in theory the advantage of utilizing future information is associated with base layer quantization settings (and hence bit-rate) in two competing aspects. A finer quantization allows higher quality of future frame reconstruction, thereby rendering future base layer reference ($\frac{N}{M}\hat{x}^b_{n+1}$) better approximation of the original coefficient (x^e_{n+1}) in (14) for more precise prediction. On the other hand, a coarse quantization results a relative large base layer interval, which leaves more uncertainty in the causal prediction to be further removed by use of future reference. Therefore, we observed a mixed results in the performance gains of SVC-ET-DP over SVC-ET versus base layer bit-rates, depending on the statistical nature of the test sequences. For example, the gap between SVC-ET-DP and SVC-ET on *coastguard* at CIF slightly increases with base layer bit-rates in Fig. 10 and 11, while it decreases on *foreman* at CIF in Fig. 12 and 13.

TABLE III

PERFORMANCE COMPARISON OF THE PROPOSED SWITCHED PREDICTION, ET PREDICTION, AND ET DELAYED PREDICTION, DENOTED BY SP, ET, AND ET-DP, RESPECTIVELY, AGAINST THE H.264/SVC WITH SINGLE-LOOP PREDICTION (SL), THE MULTI-LOOP APPROACH OF [7] (ML), AND THE CONDITIONAL REPLACEMENT PREDICTION OF [9] (CR): THE ORIGINAL TEST SEQUENCE IS $city$ AT CIF RESOLUTION. THE BASE LAYER IS CODED AT $QCIF$ RESOLUTION

Base layer bit-rate	Enhancement layer bit-rate ($kbit/s$)	PSNR (dB)					
		SVC-SL	SVC-ML	SVC-CR	SVC-SP	SVC-ET	SVC-ET-DP
230	410	32.30	32.41	32.46	32.66	32.75	32.78
	550	33.90	33.97	33.98	34.24	34.32	34.33
	700	35.42	35.46	35.45	35.65	35.70	35.71
	850	36.61	36.64	36.64	36.81	36.82	36.81
500	900	35.35	35.39	35.48	35.82	35.99	36.02
	1200	37.26	37.30	37.34	37.58	37.77	37.78
	1500	38.81	38.84	38.83	38.97	39.15	39.16
	1800	40.05	40.07	40.08	40.14	40.34	40.33

TABLE IV

PERFORMANCE COMPARISON OF THE PROPOSED SWITCHED PREDICTION, ET PREDICTION, AND ET DELAYED PREDICTION, DENOTED BY SP, ET, AND ET-DP, RESPECTIVELY, AGAINST THE H.264/SVC WITH SINGLE-LOOP PREDICTION (SL), THE MULTI-LOOP APPROACH OF [7] (ML), AND THE CONDITIONAL REPLACEMENT PREDICTION OF [9] (CR): THE ORIGINAL TEST SEQUENCE IS $sheriff$ AT 704×480 RESOLUTION. THE BASE LAYER IS CODED AT 352×240 RESOLUTION

Base layer bit-rate	Enhancement layer bit-rate ($kbit/s$)	PSNR (dB)					
		SVC-SL	SVC-ML	SVC-CR	SVC-SP	SVC-ET	SVC-ET-DP
460	950	34.35	34.89	34.84	35.01	35.09	35.10
	1600	36.81	36.97	36.94	37.08	37.16	37.18
	2200	38.35	38.43	38.42	38.59	38.60	38.60
	3100	40.03	40.02	40.06	40.18	40.18	40.17
1000	2000	37.30	37.38	37.41	37.63	37.71	37.72
	2800	38.82	38.87	38.90	39.13	39.22	39.23
	3600	40.14	40.18	40.22	40.41	40.49	40.50
	4600	41.52	41.56	41.60	41.77	41.84	41.83

TABLE V

PERFORMANCE COMPARISON OF THE PROPOSED SWITCHED PREDICTION, ET PREDICTION, AND ET DELAYED PREDICTION, DENOTED BY SP, ET, AND ET-DP, RESPECTIVELY, AGAINST THE H.264/SVC WITH SINGLE-LOOP PREDICTION (SL), THE MULTI-LOOP APPROACH OF [7] (ML), AND THE CONDITIONAL REPLACEMENT PREDICTION OF [9] (CR): THE ORIGINAL TEST SEQUENCE IS $husky$ AT 704×480 RESOLUTION. THE BASE LAYER IS CODED AT 352×240 RESOLUTION

Base layer bit-rate	Enhancement layer bit-rate ($kbit/s$)	PSNR (dB)					
		SVC-SL	SVC-ML	SVC-CR	SVC-SP	SVC-ET	SVC-ET-DP
3280	5200	30.39	30.54	30.43	30.55	30.55	30.67
	8000	32.92	33.21	33.01	33.22	33.23	33.39
	10000	34.69	35.02	34.82	35.04	35.04	35.22
	12100	36.43	36.84	36.59	36.84	36.84	36.99
4780	8800	33.72	34.11	33.80	33.92	33.93	34.04
	12000	36.14	36.56	36.28	36.55	36.57	36.65
	15000	38.26	38.80	38.47	38.75	38.79	38.90
	17500	39.94	40.53	40.21	40.46	40.52	40.64

Additional results comparing all six codecs are provided in Fig. 12, Fig. 13, and Table III-VI.

The enhancement layer performance gains were also evaluated in the setting of multiple layer coding and the results provided in Table VII. We note that the PSNR values of Layer 0 and 1 were calculated with respect to the *Downsampled* version of the original sequence, since the distortion due to discarding high frequency coefficients typically exceeds the lossy quantization error. Further, these layers are usually targeted for lower resolution display. It is experimentally shown that the overall ET framework consistently outperforms the existing pixel domain competitors across a wide range of bit-rates.

TABLE VI

PERFORMANCE COMPARISON OF THE PROPOSED SWITCHED PREDICTION, ET PREDICTION, AND ET DELAYED PREDICTION, DENOTED BY SP, ET, AND ET-DP, RESPECTIVELY, AGAINST THE H.264/SVC WITH SINGLE-LOOP PREDICTION (SL), THE MULTI-LOOP APPROACH OF [7] (ML), AND THE CONDITIONAL REPLACEMENT PREDICTION OF [9] (CR): THE ORIGINAL TEST SEQUENCE IS *park_joy* AT 1920 × 1024 RESOLUTION. THE BASE LAYER IS CODED AT 960 × 512 RESOLUTION

Base layer bit-rate	Enhancement layer bit-rate ($kbit/s$)	PSNR (dB)					
		SVC-SL	SVC-ML	SVC-CR	SVC-SP	SVC-ET	SVC-ET-DP
10200	15000	32.66	33.08	33.13	33.36	33.46	33.63
	19200	33.81	34.13	34.21	34.48	34.60	34.68
	25800	35.17	35.37	35.43	35.72	35.81	35.92
	39000	36.54	36.67	36.76	36.86	37.12	37.14
16000	27000	35.79	36.42	36.37	36.69	36.79	36.85
	36000	37.30	37.68	37.68	37.98	38.06	38.12
	45000	38.57	38.95	38.96	39.24	39.31	39.38
	54000	39.71	40.00	40.00	40.29	40.35	40.40

TABLE VII

PERFORMANCE COMPARISON OF THE PROPOSED SWITCHED PREDICTION, ET PREDICTION, AND ET DELAYED PREDICTION, DENOTED BY SP, ET, AND ET-DP, RESPECTIVELY, AGAINST THE H.264/SVC WITH SINGLE-LOOP PREDICTION (SL), THE MULTI-LOOP APPROACH OF [7] (ML), AND THE CONDITIONAL REPLACEMENT PREDICTION OF [9] (CR) IN THE SETTINGS OF MULTIPLE LAYER CODING. THE ORIGINAL TEST SEQUENCE IS *harbour* AT $4CIF$ RESOLUTION. THE PSNR VALUES OF LAYER 0 AND 1 ARE CALCULATED WITH RESPECT TO THE ORIGINAL *downsampled* SEQUENCES

Layer	Resolution	Cumulative bit-rate ($kbit/s$)	PSNR (dB)					
			SVC-SL	SVC-ML	SVC-CR	SVC-SP	SVC-ET	SVC-ET-DP
0	$QCIF$	450	33.94	33.94	33.94	33.94	33.94	33.94
1	CIF	1600	32.89	32.92	33.08	33.34	33.53	33.54
2	$4CIF$	3380	33.18	33.45	33.49	33.86	34.15	34.27
0	$QCIF$	800	37.88	37.88	37.88	37.88	37.88	37.88
1	CIF	2900	36.45	36.60	36.67	36.96	37.13	37.16
2	$4CIF$	6900	36.34	36.65	36.67	37.15	37.40	37.52
0	$QCIF$	1200	41.60	41.60	41.60	41.60	41.60	41.60
1	CIF	4900	40.21	40.50	40.51	40.88	40.95	40.97
2	$4CIF$	12600	39.56	40.08	40.15	40.73	40.94	41.11

VII. CONCLUSIONS

This paper proposes a novel unified framework for resampling and estimation-theoretic enhancement layer prediction in spatial SVC. Aided by unconventional transform domain resampling, the ET prediction approach maximally utilizes information from the enhancement layer reconstruction of the previous frames and both current and available future base layer information. All the information is combined into an appropriate conditional pdf, the expectation over which then forms the optimal enhancement layer prediction. Considerable and consistent coding gains are obtained by using the proposed unified framework, in comparison to standard H.264/SVC and its variants. A switched prediction approximation to the ET scheme is also devised that greatly reduces the codec complexity, while retaining major coding performance gains.

APPENDIX
PROOF OF (14) IN SECTION IV

Let $\{x_n\}$ denote the first order AR process of the enhancement layer transform coefficient at a particular frequency (i, j), and let I_n be the interval that contains the current sample x_n. We consider the conditional pdf of x_n given x_{n-1}, x_{n+1}, and I_n. Specifically, in (14), the requisite conditions are given or approximated by

$$\begin{aligned} I_n &= I_n^e(i, j), \\ x_{n-1} &\approx \hat{x}_{n-1}^e(i, j), \\ x_{n+1} &\approx \frac{N}{M}\hat{x}_{n+1}^b(i, j). \end{aligned}$$

Claim: The pdf of x_n conditioned on I_n, x_{n-1}, and x_{n+1} can be decomposed as:

$$\begin{aligned} &p(x_n|x_{n-1}, I_n, x_{n+1}) \\ &= \begin{cases} \dfrac{p(x_n|x_{n-1})\ p(x_{n+1}|x_n)}{\int_{I_n} p(x_n|x_{n-1})\ p(x_{n+1}|x_n)dx_n}, & x_n \in I_n \\ 0, & otherwise. \end{cases} \end{aligned} \tag{18}$$

Proof: Since $x_n \in I_n$, the conditional pdf $p(x_n|x_{n-1}, I_n, x_{n+1})$ can be written as:

$$p(x_n|x_{n-1}, I_n, x_{n+1}) = \begin{cases} \dfrac{p(x_n|x_{n-1}, x_{n+1})}{\int_{I_n} p(x_n|x_{n-1}, x_{n+1})dx_n}, & x_n \in I_n \\ 0, & otherwise. \end{cases} \tag{19}$$

Note that the above is equivalent to truncating $p(x_n|x_{n-1}, x_{n+1})$ by the interval I_n, and normalizing to

obtain a valid pdf. The Markov property of (9) implies that:

$$p(x_{n+1}|x_n, x_{n-1}) = p(x_{n+1}|x_n). \tag{20}$$

Applying Bayes rule and (20) to $p(x_n|x_{n-1}, x_{n+1})$, we can obtain

$$\begin{aligned} p(x_n|x_{n-1}, x_{n+1}) &= \frac{p(x_{n+1}|x_n, x_{n-1})\ p(x_n, x_{n-1})}{p(x_{n-1}, x_{n+1})} \\ &= \frac{p(x_{n+1}|x_n)\ p(x_n|x_{n-1})}{p(x_{n+1}|x_{n-1})}. \end{aligned} \tag{21}$$

Plugging (21) in the numerator and denominator of (19), one obtains (18).

REFERENCES

[1] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the H.264/AVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 9, pp. 1103–1120, Sep. 2007.

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

[3] C. A. Segall and A. Katsaggelos, "Resampling for spatial scalability," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2006, pp. 181–184.

[4] 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.

[5] J. Han, V. Melkote, and K. Rose, "Estimation-theoretic approach to delayed prediction in scalable video coding," in *Proc. 17th IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2010, pp. 1289–1292.

[6] J. Han, V. Melkote, and K. Rose, "A unified framework for spectral domain prediction and end-to-end distortion estimation in scalable video coding," in *Proc. 18th IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2011, pp. 3278–3281.

[7] R. Zhang and M. L. Comer, "Efficient inter-layer motion compensation for spatially scalable video coding," *IEEE Trans. Circuits Syst., Video Technol.*, vol. 18, no. 10, pp. 1325–1334, Oct. 2008.

[8] R. Zhang and M. L. Comer, "Rate distortion analysis for spatially scalable video coding," *IEEE Trans. Image Process.*, vol. 19, no. 11, pp. 2947–2957, Nov. 2010.

[9] T. K. Tan, K. K. Pang, and K. N. Ngan, "A frequency scalable coding scheme employing pyramid and subband techniques," *IEEE Trans. Circuits Syst., Video Technol.*, vol. 4, no. 2, pp. 203–207, Apr. 1994.

[10] (2012, Dec.). *WebM Project* [Online]. Available: http://www.webmproject.org/code/

[11] G. J. Sullivan, J. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[12] J. Han, V. Melkote, and K. Rose, "An estimation-theoretic approach to spatially scalable video coding," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 817–820.

[13] J. Han, V. Melkote, and K. Rose, "An estimation-theoretic framework for spatially scalable video coding with delayed prediction," in *Proc. 19th Int. Packet Video Workshop (PV)*, May 2012, pp. 167–172.

[14] H. Schwarz, T. Hinz, D. Marpe, and T. Wiegand, "Constrained inter-layer prediction for single-loop decoding in spatial scalability," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2005, pp. 870–873.

[15] F. Wu, S. Li, and Y.-Q. Zhang, "A framework for efficient progressive fine granularity scalable video coding," *IEEE Trans. Circuits Syst., Video Technol.*, vol. 11, no. 3, pp. 332–344, Mar. 2001.

[16] J. M. Adant, P. Delogne, E. Lasker, B. Macq, L. Stroobants, and L. Vandendorpe, "Block operations in digital signal processing with application to TV coding," *Signal Process.*, vol. 13, no. 4, pp. 385–397, Dec. 1987.

[17] S. A. Martucci, "Image resizing in the discrete cosine transform domain," in *Proc. Int. Conf. Image Process. (ICIP)*, Oct. 1995, pp. 244–247.

[18] 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, no. 6, pp. 477–488, Nov. 1992.

[19] 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.

[20] 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.

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

[22] 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.

[23] J. Han, V. Melkote, and K. Rose, "An estimation-theoretic approach to delayed decoding of predictively encoded video sequences," *IEEE Trans. Image Process.*, vol. 22, no. 3, pp. 1175–1185, Mar. 2013.

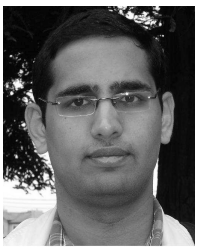
[24] J. Han, V. Melkote, and K. Rose, "Estimation-theoretic delayed decoding of predictively encoded video sequences," in *Proc. Data Compress. Conf. (DCC)*, Mar. 2010, pp. 119–128.

[25] K. Rose and S. L. Regunathan, "Towards optimal scalability in predictive video coding," in *Proc. Int. Conf. Image Process. (ICIP)*, Oct. 1998, pp. 965–976.

[26] (2013, Dec.). *Coded Base Layer Clips* [Online]. Available: http://www.scl.ece.ucsb.edu/spatial-et-svc/base_layer.rar

Jingning Han (S'10–M'12) received the B.S. degree in electrical engineering from Tsinghua University, Beijing, China, in 2007, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of California at Santa Barbara, Santa Barbara, CA, USA, in 2008 and 2012, respectively. He is currently with the WebM Codec Team, Google Inc., Mountain View, CA, USA, where he is involved in video compression, processing, and related technologies. His research interests include video coding and computer architecture.

Dr. Han was a recipient of the Outstanding Teaching Assistant Awards from the Department of Electrical and Computer Engineering, University of California at Santa Barbara, in 2010 and 2011, the Dissertation Fellowship in 2012, and 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 IIT Madras, Chennai, India, in 2005, and the M.S. and Ph.D. degrees in electrical and computer engineering from the University of California at Santa Barbara, Santa Barbara, CA, USA, in 2006 and 2010, respectively. He is currently with the Sound Technology Research Group, Dolby Laboratories, Inc., San Francisco, CA, USA, where he is involved in audio compression and related technologies. His other research interests include video compression and estimation theory. He interned with the Multimedia Codecs Division, Texas Instruments, Bangalore, India, in Summer 2004, and with the Audio Systems Group, Qualcomm, Inc., San Diego, CA, USA, 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, CA, USA, in 1991.

He then joined the Department of Electrical and Computer Engineering, University of California at Santa Barbara, Santa Barbara, CA, USA, where he is currently a Professor. His main research activities are in the areas of information theory and signal processing, rate-distortion theory, source and source-channel coding, audio and video coding and networking, pattern recognition, and nonconvex optimization. He is interested in the relations between information theory, estimation theory, statistical physics, and their potential impact on fundamental and practical problems in diverse disciplines.

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