An Estimation-Theoretic Framework for Spatially Scalable Video Coding with Delayed Prediction

Jingning Han, Vinay Melkote*, and Kenneth Rose
Department of Electrical and Computer Engineering
University of California Santa Barbara, CA 93106
Email: {jingning,melkote,rose}@ece.ucsb.edu

***Abstract*—A novel estimation-theoretic (ET) approach is developed for optimal enhancement layer prediction, in *spatially* scalable video coding (SVC), which incorporates motion compensation at the enhancement layer, with both current and *future* information from the base layer. It is inspired by the early ET framework (originated in our group) for *quality* (SNR) scalability, which achieved optimal enhancement layer prediction by fully accounting for information from the current base layer (e.g., the quantization intervals) and the enhancement layer, to efficiently calculate the conditional expectation that forms the optimal predictor. Central to that approach was the fact that all layers reconstruct approximations to the same original transform coefficient. This, however, is not the case in spatial scalability, where the layers encode different resolution versions of the signal. To approach optimal enhancement layer prediction, the current work departs from existing spatial SVC schemes that employ pixel-domain resampling and causal prediction. Instead, it integrates a transform domain resampling technique that makes the base layer quantization intervals and reconstructions accessible to and usable at the enhancement layer. The approach is extended for an SVC framework that allows delay in enhancement layer coding relative to the base layer, and achieves optimal *delayed* prediction, in conjunction with spatial SVC. Simulations provide experimental evidence that the overall proposed approach substantially outperforms existing spatially scalable coders.**

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 fidelity resolutions. The higher resolution layers will typically benefit from differential coding from lower layers, based on inter-layer prediction, which results in significantly reduced bit-rate as well as enhanced streaming flexibility, without retaining multiple independent bit-streams, each of a different quality level. 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 restrict our discussion to a two-layered spatial SVC codec, while emphasizing that the proposed approach is extensible to more layers.

A spatial SVC scheme consists of downsampling a high resolution video sequence to a lower resolution, and coding the two resolutions into separate layers. The lower resolution signal is coded by a base layer coder, which is essentially a single-layer coder, while the enhancement layer encodes information necessary to reconstruct the sequence at its original higher spatial resolution. 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 multi-loop design frequently employed in standard codecs, the prediction mode is selected amongst the two sources such that the rate-distortion cost is minimized. More details on existing spatial SVC approaches are provided in Sec.II, and also available in [2]. The inter-layer prediction is commonly performed in the pixel-domain, where the base-layer reconstructed pixels are upsampled via interpolation to the enhancement layer resolution prior to prediction, and the resultant 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 (see e.g., [2], [3]). A notable approach is proposed in [4], where an additional prediction mode that is formed as a linear combination of inter-layer and motion compensated (inter-frame) predictions is introduced, and significantly improves the enhancement layer coding performance.

The ad hoc nature of the above approaches, which *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 [5] for the very different setting of *quality* (SNR) scalability, where the *same* original sequence is coded by all the layers but at different quantization resolution. Thus, the true value of a transform coefficient must lie in the interval determined by its quantization at the base layer. This observation effectively captures all the information provided by the base layer, and is the central postulate of the ET approach in [5], which employs a conditional probability density function (pdf), truncated (and normalized) to the base layer quantization interval, and computes the exact conditional expectation that forms the optimal prediction for the transform coefficient. The ET approach was further enhanced by allowing delayed prediction [6], and extended to incorporate resilience to packet

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

loss [7], all in the setting of quality scalability.

Challenges arise in the *spatial* scalability case we focus on here, 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 quantization interval and other 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 (including future) 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 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.

We note that a fundamental property of 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 additional useful information for 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 with certain coding delay. Experiments provide evidence for considerable enhancement layer coding gains achieved by the proposed ET framework, over standard H.264 extensions for spatial SVC and other leading competitors that employ pixel-domain prediction methods. Further, examination of the base layer reconstruction indicates that these gains are achieved at no degradation to the base layer quality, due to the unconventional transform domain resampling. Some preliminary results of this ET approach to spatial SVC were reported in our recent work [8], albeit without the framework extension to enable exploitation of coding delay which is central to this paper. We note that while the proposed approach was implemented and tested in the H.264/AVC SVC extension framework, the principle is generally applicable to other motion compensation based predictive codecs including VP8 and HEVC.

II. Background

We provide related background information on standard spatial SVC and its variants. The H.264/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.. In this paper the enhancement layer prediction of the standard codec is modified to follow the multi-loop design [9], where the prediction switches between the motion-compensated reference from prior frames at the same layer, and the current base layer reconstruction (upsampled via pixel filtering), so that the rate-distortion cost is minimized. We note that the actual standard follows the alternate single-loop design [2]. Nevertheless, it has been recognized that the multi-loop design performs slightly better than the single-loop approach at the expense of increasing decoder complexity [2], and is employed here as the focus is on optimality in coding performance. The modified H.264/SVC codec retains the original techniques for inter-frame prediction, e.g., sub-pixel motion compensation, deblocking filter, intra coding, etc.

The modified standard encoder works as follows. To encode block A_0 (see Fig.1) 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, and is used in computing the enhancement layer prediction for that block. Both prediction modes in the multi-loop design are tested by the encoder to find the one that minimizes rate-distortion cost. A significant amount of study has been devoted to designing the interpolation filter, and to determine whether supporting additional filters would be beneficial. However, no clear winner was identified [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. In [4], an additional mode that generates the prediction as a linear combination of E_0 and $\tilde{A}_0$ is proposed for more efficient enhancement layer coding, where the weight coefficients are derived as a function of the resampling operations.

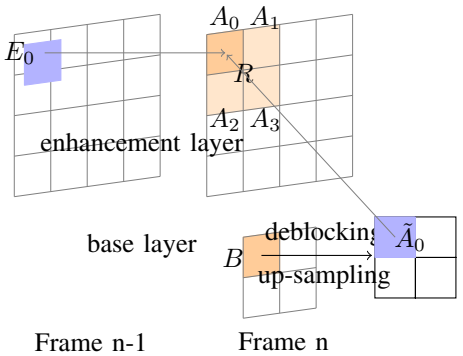

Fig. 1: Pixel-domain enhancement layer prediction in spatial SVC (multi-loop structure).

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

As noted earlier in Sec.I, the existing ad hoc approach to enhancement layer prediction in spatial SVC, which combines base layer reconstructed pixels (or residuals) with the enhancement layer motion compensated reference, does not guarantee optimal utility of all available information. Moreover, independent base layer coding provides the enhancement layer with the possibility of exploiting future base layer information, at the expense of limited coding delay. These motivate the ET approach described in this section, which jointly optimizes the framework in terms of transform domain resampling and enhancement layer prediction, to maximally utilize all (including future) information extractable from the base layer sequence, as well as motion compensated information from the enhancement layer itself. 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 2-D transform, i.e., it is accomplished by applying 1-D operations sequentially along the vertical and horizontal directions. Hence, for clarity of exposition, we first present the main ideas of this transform domain resampling approach in the framework of a 1-D transform. Consider a vector of pixels $\underline{a} = [a_0, a_1, \cdots, a_{N-1}]^T$, with inter-pixel correlation close to unity. Here the superscript T denotes transposition. The optimal approach to compress $\underline{a}$ 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 the 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 $\underline{\alpha}_N = T_N \underline{a}$ is the DCT of vector $\underline{a}$. Define:

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

analog cosine functions with a period that is a sub-multiple 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 $t = \frac{1}{2N}, \frac{3}{2N}, \cdots, \frac{2N-1}{2N}$. Consequently, 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 $\underline{\alpha}_N$, when sampled at the rate $\frac{1}{2N}$ yields exactly the vector $\underline{a}$. Now define,

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

the analog cosine functions which when sampled at rate $\frac{1}{2M}$ yield the basis functions for a DCT of dimension M. The best approximation (in mean squared error sense) for the signal $a(t)$ using only M of the N transform coefficients in $\underline{\alpha}_N$ is that provided by choosing the M coefficients of lowest frequency:

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

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

$$\underline{b} = \sqrt{\frac{M}{N}}\, T_M^T \begin{pmatrix} I_M & 0_M \end{pmatrix} T_N \underline{a}, \quad (2)$$

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

$$\hat{\underline{a}} = \sqrt{\frac{N}{M}}\, T_N^T \begin{pmatrix} I_M \\ 0_M \end{pmatrix} T_M \underline{b}. \quad (3)$$

Under the assumption that the DCT has performance very close to the KLT, 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 e.g., [10], [11]. While we described this resampling in the 1-D framework, the extension to pixel blocks is straightforward. The downsampling (or upsampling) can be sequentially applied to the vertical and horizontal directions. This transform domain resampling approach can in general serve as an alternative to the pixel-domain downsampling and interpolation traditionally employed in spatial SVC. However, as discussed next, this resampling method is of particular advantage to the proposed ET spatial SVC paradigm.

B. Optimal Enhancement Layer Delayed Prediction

We now describe the estimation-theoretic approach to *delayed* prediction at the enhancement layer. Similar to the standard approach, each frame (at the spatial resolution of the enhancement layer) is partitioned into macroblocks (usually of size 16x16), and each macroblock is coded with inter-layer, inter-frame prediction, or in intra mode. Transforms are applied at sub-macroblock resolution (typically 4x4 and 8x8) to the prediction residuals, which are then quantized and entropy coded. Windowing and cropping operations, e.g., "pan and scan" technique, are performed 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 identical to the $M \times M$ DCT employed for downsampling.

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

$x_n^b(i,j)$, $i,j \in \{0,\cdots,M-1\}$, the transform coefficients of the base layer block B_n:

$$x_n^b(i,j) = \frac{M}{N} x_n^e(i,j), i,j \in \{0,\cdots,M-1\} \quad (4)$$

These coefficients are then transformed by an $M \times M$ inverse DCT to generate the base layer pixel block B_n, which is coded as usual by the base layer coder. Since the choice of spatial transform applied to the base layer prediction residual block (either in intra-mode or inter-mode) is assumed to be same as that for downsampling, it can be easily shown that the base layer coding process essentially prescribes a quantization interval $I_n^b(i,j)$ that contains $x_n^b(i,j)$ for all $i,j \in \{0,1,\cdots,M-1\}$. This interval summarizes all the information provided by the base layer *at time instance* n about the transform coefficient $x_n^b(i,j), \forall i,j \in \{0,1,\cdots,M-1\}$. Note that this interval does not exist for other high frequency coefficients, i.e., the base layer provides no information on $x_n^b(i,j)$ if i or $j \in \{M, M+1,\cdots,N-1\}$, since they have been discarded during downsampling to produce the lower spatial resolution representation.

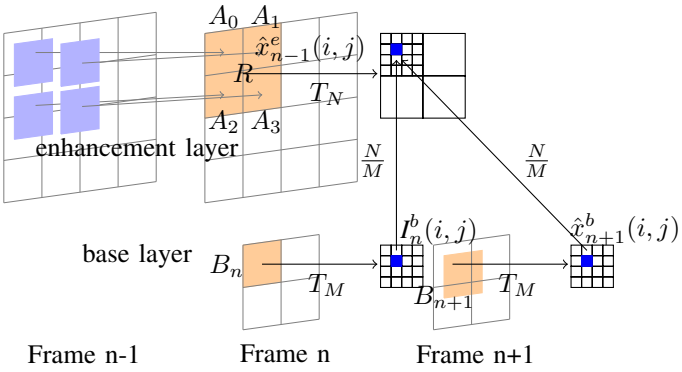

Fig. 2: Estimation-theoretic enhancement layer delayed prediction.

The traditional course of action would now be to upsample the base layer reconstruction of block B_n. In accordance with Sec.III-A, this would entail zero-padding the $M \times M$ DCT of the reconstruction of block B_n to yield an $N \times N$ block of transform coefficients, which is then appropriately scaled by the inverse scaling factor applied in (4), and inversely transformed to get a pixel domain approximation of block R in Fig.2. This could then be combined in pixel-domain with the enhancement layer motion compensated references from previously reconstructed frames, as the prediction for $\{A_i, i = 0,\cdots,3\}$ in frame n.

However, such an approach suffers from significant under-utilization of the information provided by the base layer, mainly in two aspects: the quantization interval of the current sample, and the reconstructions of future samples. In particular, on account of the transform domain resampling the following relation holds:

$$x_n^e(i,j) \in \boldsymbol{I}_n^e(i,j) = \frac{N}{M} I_n^b(i,j), \; i,j \in \{0,\cdots,M-1\}, \quad (5)$$

which implies that the information in the base layer quantization intervals directly translates into information about transform coefficients at the enhancement layer, both at time instance n. Note that since the quantization is performed per transform coefficient, this information emerges only in transform domain, instead of pixel domain.

We next consider the construction of motion compensated reference for B_n in future frames. For simplicity, let us restrict to the setting where the enhancement layer sequence encoding is delayed by one frame relative to the base layer, i.e., the enhancement layer coder has access to the coding information of frame $(n+1)$ at base layer, when encoding frame n. To find the motion compensated reference in the future frame, the coder first identifies the locations of reference blocks (potentially off-grid) in frame n for all the inter-frame coded blocks in frame $(n+1)$. Then for each on-grid block in frame n, the motion vector of the reference block that overlaps this on-grid block most will be inversed, which is then used to generate the required motion compensated reference in frame $(n+1)$. Similar inverse motion search method has been adopted in an optimal delayed decoding scheme for predictively encoded sequences by a regular single-layer coder [12]. Let B_{n+1} denote the corresponding reference block for B_n in frame $(n+1)$, whose transform coefficients are hence denoted by $\hat{x}_{n+1}^b(i,j)$, as shown in Fig.2. 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.

We model blocks of DCT coefficients along the same motion trajectory as an auto-regressive (AR) process per frequency. 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:

$$x_n^e(i,j) = \rho x_{n-1}^e(i,j) + z_n(i,j), \quad (6)$$

where $z_n(i,j)$ denotes the independent and identically distributed (i.i.d.) innovation of the process with probability density function (pdf) $p_Z(z_n(i,j))$. Following the implicit assumption in conventional pixel domain motion compensation, we set the correlation coefficient $\rho = 1$ at all frequencies. Assuming that the transform coefficient of the reconstructed motion compensation reference $\hat{x}_{n-1}^e(i,j) \approx x_{n-1}^e(i,j)$, and similarly $\frac{N}{M}\hat{x}_{n+1}^b(i,j) \approx x_{n+1}^e(i,j)$, the pdf of x_n^e conditioned on the previous enhancement layer (motion compensated) reference $\hat{x}_{n-1}^e(i,j)$, current quantization interval $I_n^e(i,j)$, and *future* base layer reference $\hat{x}_{n+1}^b(i,j)$ is thus[1]

$$\begin{aligned} & p(x_n^e(i,j)|\hat{x}_{n-1}^e(i,j), \boldsymbol{I}_n^e(i,j), \hat{x}_{n+1}^b(i,j)) \\ \approx & \frac{p(x_n^e|\hat{x}_{n-1}^e, \boldsymbol{I}_n^e) \cdot p(\hat{x}_{n+1}^b|x_n^e)}{\int_{I_n^e} p(x_n^e|\hat{x}_{n-1}^e, \boldsymbol{I}_n^e) \cdot p(\hat{x}_{n+1}^b|x_n^e) dx_n^e} \\ \approx & \begin{cases} \frac{p_Z(x_n^e - \hat{x}_{n-1}^e) \cdot p_Z(\frac{N}{M}\hat{x}_{n+1}^b - x_n^e)}{\int_{I_n^e} p_Z(x_n^e - \hat{x}_{n-1}^e) \cdot p_Z(\frac{N}{M}\hat{x}_{n+1}^b - x_n^e) dx_n^e}, & x_n^e \in \boldsymbol{I}_n^e, \\ 0, & else, \end{cases} \end{aligned} \quad (7)$$

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

which is obtained by applying the Markov property of the AR process (6): given the current sample $x_n^e(i,j)$, a future sample $x_{n+1}^e(i,j)$ (or equivalently $\frac{N}{M}x_{n+1}^b(i,j)$) is conditionally independent of the past. We note that in the above equation, the *causal* pdf of $x_n^e(i,j)$, i.e., $p(x_n^e(i,j)|\hat{x}_{n-1}^e(i,j), I_n^e(i,j))$, is weighted by $p(\hat{x}_{n+1}^b(i,j)|x_n^e(i,j))$, the probability density of the *known future outcome* to obtain the one-sample delayed pdf of (7), which incorporates all available information at the enhancement layer coder, at up to one frame coding delay. The overall conditional pdf is then truncated to the quantization interval of the current sample, the centroid of which is thus the optimal predictor at the enhancement layer, at one frame delay.

In practice, the minimum overlap area between B_n and the reference blocks in frame n is thresholded to allow the use of inverse motion vector, which connects B_n and B_{n+1}. Thus it is possible that, occasionally, the block B_n will not find an inverse motion compensated reference B_{n+1} in frame $(n+1)$. In such cases, the $p(\hat{x}_{n+1}^b(i,j)|x_n^e(i,j))$ term will cancel out due to the absence of future base layer information, and (7) specializes to the non-delayed pdf as discussed in [8]:

$$p(x_n^e(i,j)|\hat{x}_{n-1}^e(i,j), I_n^e(i,j)) \quad (8)$$
$$= \begin{cases} \frac{p_Z(x_n^e(i,j)-\hat{x}_{n-1}^e(i,j))}{\int_{I_n^e(i,j)} p_Z(x_n(i,j)-\hat{x}_{n-1}^e(i,j))dx_n}, & x_n^e(i,j) \in I_n^e(i,j), \\ 0, & else. \end{cases}$$

which is equivalent to centering the innovation pdf at $\hat{x}_{n-1}^e(i,j)$, restricting it to the interval $I_n^e(i,j)$, and then normalizing to obtain a valid pdf. Further note that for high frequency coefficients where i or $j \in \{M, M+1, \cdots, N-1\}$, both $I_n^e(i,j)$ and $\hat{x}_{n+1}^b(i,j)$ are not available, and the best prediction of $x_n^e(i,j)$ is simply $\hat{x}_{n-1}^e(i,j)$, the default inter-frame motion compensated estimate. In summary, the optimal predictor at the enhancement layer is given by

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

The above equations describe the transform coefficients prediction at the enhancement layer for the entire $N \times N$ region R in Fig.2. This transform domain prediction of R is now inversely transformed to generate the pixel domain prediction for each individual block A_i. Subsequently, as in the standard codec, the pixel-domain prediction residual for each block A_i is calculated, spatial transformation applied, and the resultant transform coefficients quantized and coded.

In the implementation we will assume that the innovation pdf is Laplacian, i.e., $p_Z(z_n) = \frac{1}{2}\lambda e^{-\lambda|z_n|}$, where the parameter λ is frequency dependent in accordance with our earlier work [5]. It is interesting to note that the memoryless property of Laplacian distribution offers closed form expressions of the above conditional expectations.

IV. Simulation Results

We implemented the proposed unified ET approach in the JSVM reference framework, with one frame delay in enhancement layer coding relative to the base layer, which is denoted by ET-SVC-1DP. As a special case where no future base layer information is exploited, our proposed scheme degenerates to our ET causal enhancement layer prediction, which we refer to as ET-SVC-0DP. One competing codec was created by modifying standard H.264/SVC to support multi-loop inter-layer prediction, using the 4-tap polyphase filter and deblocking operations for up-sampling, in addition to the inter-frame prediction, which is henceforth referred to as H.264/SVC-ML. The scheme that allows an additional mode, where the prediction is formed as a linear combination of inter-layer and inter-frame predictions [4], was also implemented in the modified H.264/SVC framework, and is referred to as H.264/SVC-LC. All the codecs employ regular pixel domain motion estimation at quarter-pixel resolution, and all use the same base layer coder.

The enhancement layer coding performance of the four codecs for the sequence $coastguard$ at CIF resolution is shown in Fig.3, where the base layer is coded at bit-rate of 368 $kbit/s$ and PSNR of 35.0 dB. Clearly, ET-SVC-0DP outperforms either pixel domain competitor, H.264/SVC-LC or H.264/SVC-ML, across a wide range of bit rates, while ET-SVC-1DP offers substantial further performance improvements on top of ET-SVC-0DP, by incorporating enhancement layer coding delay. Similar performance improvements were also obtained for other sequences as shown in Fig.4-6. To reduce clutter in subsequent plots, we demonstrate the overall achievable coding gains provided by the proposed ET approach.

A potential downside of employing an unconventional resampling technique is the possibility of blocking artifacts in the base-layer. The visual comparison of the reconstructed base layer frames generated using DCT domain downsampling and pixel domain decimation, respectively, is provided in [8]. Tests on typical sequences reveal that both methods provide smooth and sharp representations at lower spatial resolution, where no strong blocking artifacts are visible. Further objective evaluation suggests that the difference in the rate-distortion performance of coding the two base layer sequences is indeed negligible. Thus, the ET scheme offers major gains in enhancement layer performance at no discernible degradation of the base layer.

V. Conclusion

This paper proposes a novel unified framework for resampling and estimation-theoretic enhancement layer delayed prediction in spatial SVC. Aided by unconventional transform domain resampling, the ET prediction approach maximally utilizes information from the prior enhancement layer reconstructions and both current and available future base layer information, and combines them into an appropriate conditional pdf. The enhancement layer prediction is then obtained as the corresponding conditional expectation. Considerable and consistent coding gains are obtained by using the proposed unified framework, in comparison to standard H.264/SVC and its variants.

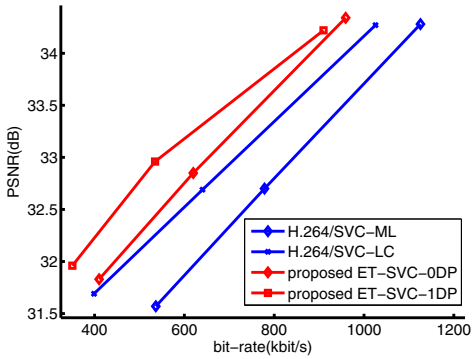

Fig. 3: Comparison of the enhancement layer coding performance of the four spatial SVC approaches: The test sequence is $coastguard$ at CIF resolution. The base layer is at $QCIF$ resolution, and is coded at 368 $kbit/s$ with reconstruction quality 35.0 dB (with respect to the downsampled version of original sequence).

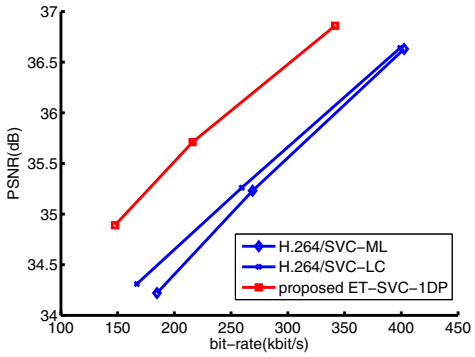

Fig. 4: Comparison of the enhancement layer coding performance of the proposed ET versus the two competing spatial SVC approaches: The test sequence is $foreman$ at CIF resolution. The base layer is at $QCIF$ resolution, and is coded at 238 $kbit/s$ with reconstruction quality 36.7 dB (with respect to the downsampled version of original sequence).

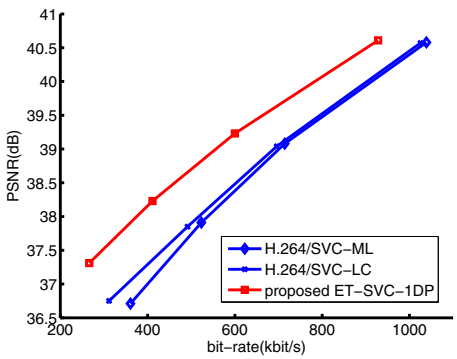

Fig. 5: Comparison of the enhancement layer coding performance of the three competing spatial SVC approaches: The test sequence is $foreman$ at CIF resolution. The base layer is at $QCIF$ resolution, and is coded at 408 $kbit/s$ with reconstruction quality 39.7 dB (with respect to the downsampled version of original sequence).

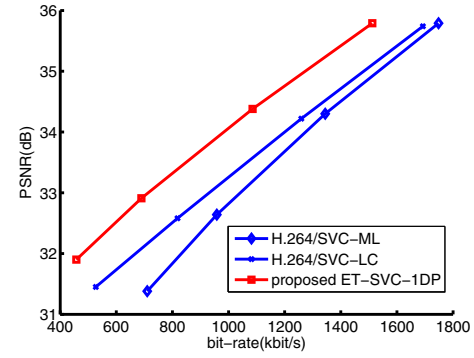

Fig. 6: Comparison of the enhancement layer coding performance of the three competing spatial SVC approaches: The test sequence is $harbour$ at CIF resolution. The base layer is at $QCIF$ resolution, and is coded at 680 $kbit/s$ with reconstruction quality 34.6 dB (with respect to the downsampled version of original sequence).

REFERENCES

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

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

[3] C. A. Segall and A. Katsaggelos, "Resampling for spatial scalability," *Proc. IEEE ICIP*, pp. 181–184, Oct. 2006.

[4] R. Zhang and M. Comer, "Efficient inter-layer motion compensation for spatially scalable video coding," *IEEE Trans. Circ. Sys. Video Tech.*, vol. 18, pp. 1325–1334, Oct. 2008.

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

[6] J. Han, V. Melkote, and K. Rose, "Estimation-theoretic approach to delayed prediction in scalable video coding," *Proc. IEEE ICIP*, pp. 1289–1292, Sep. 2010.

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

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

[9] F. Wu, S. Li, and Y.-Q. Zhang, "A framework for efficient progressive fine granularity scalable video coding," *IEEE Trans. Circ. Sys. Video Tech.*, vol. 11, pp. 332–344, Mar. 2001.

[10] 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 Processing*, vol. 13, pp. 385–397, Dec. 1987.

[11] S. A. Martucci, "Image resizing in the discrete cosine transform domain," *Proc. IEEE ICIP*, vol. 13, pp. 244–247, Oct. 1995.

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