AN ESTIMATION-THEORETIC APPROACH TO SPATIALLY SCALABLE VIDEO CODING

Jingning Han, Vinay Melkote, and Kenneth Rose*

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

ABSTRACT

This paper focuses on prediction optimality in spatially scalable video coding. It is inspired by the earlier estimation-theoretic prediction framework developed by our group for quality (SNR) scalability, 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, the current work departs from existing spatially scalable codecs that employ pixel-domain resampling to perform inter-layer 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, which in conjunction with prior enhancement layer information, enable optimal prediction. Simulations provide experimental evidence that the proposed approach achieves substantial enhancement layer coding gains over the standard.

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

1. INTRODUCTION

In scalable video coding (SVC), the video sequence is encoded into a single bit-stream of multiple layers with progressively higher spatial, temporal, or quantizer resolution. The higher resolution layers are typically encoded differentially from the lower layers, i.e., inter-layer prediction is employed, which results in significantly reduced bit-rate compared to retaining multiple independent bit-streams, each of a different quality level. Thus SVC is an attractive solution for modern network infrastructure composed of decoders with multiple display resolutions and various channel capacities [1]. Of the various flavors of SVC, the focus of this paper is on spatial scalability. For simplicity of exposition, we restrict our discussion to a two-layered spatial codec, while emphasizing that the proposed approach is extensible to more layers.

Two-layered spatial SVC consists of downsampling a video sequence with high spatial resolution to a lower resolution, and coding it into a base-layer bitstream, while the enhancement layer codes information necessary to reconstruct the sequence at its original higher spatial resolution. At the enhancement layer, the current video frame is 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 so called multi-loop design frequently employed in standard codecs, this prediction is computed as a *linear* combination of the two different types of information. More details on existing spatial SVC approaches are provided in Sec.2, and also available in [1]. Typically, inter-layer prediction is performed in the pixel domain, the prediction residual transformed via DCT, and the transform coefficients quantized and coded. The base-layer reconstructed pixels are upsampled via interpolation to the enhancement layer resolution prior to prediction, and substantial past research has focused on the quality of such interpolation, which influences prediction accuracy, and hence coding performance [1, 2].

The ad hoc nature of the above approach, which linearly combines reconstructions from different sources, strongly motivates the search for a true estimation-theoretic (ET) approach to inter-layer prediction in spatial SVC, where all the information provided by the base layer is fully and optimally exploited. Inspiration is drawn from an ET technique proposed earlier by our group in [3] for the very different setting of *quality* (SNR) scalability, where the *same* sequence is coded by all the layers but at different quantization levels. 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 [3], which employs a conditional probability density function (pdf), truncated and normalized for the base layer quantization interval, to compute the exact conditional expectation that forms the optimal prediction for the transform coefficient. The ET approach was further enhanced by allowing delayed prediction [4], and was deployed over lossy channel [5].

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 the different layers quantize different transform coefficients. Consequently, the quantization interval in the base layer cannot be used directly to optimize prediction at the enhancement layer. 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. With this resampling framework in place the proposed ET approach combines, in the transform domain, the two disparate sources of information - quantizer intervals from the base layer and the motion compensated reference from the enhancement layer - in a conditional pdf, the expectation over which yields optimal enhancement layer prediction. Experiments provide evidence for considerable enhancement layer coding gains over standard H.264 spatial SVC and other leading competitors that employ pixel-domain resampling filters. Further, examination of the base layer reconstruction indicates that these gains are achieved at no degradation to the base layer quality, relative to the competition.

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

2. 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.. The standard employs a single-loop design to encode the enhancement layer, where the decoder need not buffer the base layer reconstruction to reproduce the desired layer signal. In particular, the coder starts with regular motion compensated prediction from previously reconstructed frames at the same layer to generate a residual block. It then adaptively decides whether to further subtract the base layer reconstructed residuals from this residual block before transformation and quantization [1]. In earlier standards such as H.263++, the enhancement layer prediction switches between the motion-compensated reference from prior enhancement layer frames, and the current base layer reconstruction (up-sampled via pixel filtering), in what is referred to as multi-loop design. It has been recognized that multi-loop design performs better than single-loop at the expense of more decoder complexity. Since this paper is focused on optimality in coding performance, the H.264/SVC codec is modified to better performing multi-loop design, while retaining other advanced coding techniques, e.g., sub-pixel motion compensation, intra coding, CABAC, 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 additional supporting filters would be beneficial. However, no clear winner was identified [1]. A notable approach was proposed in [2] where the upsampling filter is derived to match the downsampling operation while accounting for the quantization noise in the base layer reconstructed pixels. In [6], 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. Fairly significant improvements in coding performance were achieved by integrating this additional mode in the rate-distortion optimization framework of the modified H.264/SVC.

3. THE UNIFIED ESTIMATION-THEORETIC FRAMEWORK FOR RESAMPLING AND PREDICTION

As noted earlier in Sec.1, the existing ad hoc approach to enhancement layer prediction in spatial SVC that combines base layer reconstructed pixels (or residuals) with the enhancement layer motion-compensated reference does not guarantee optimal utility of *all* available information. This motivates the ET approach described in this section, that jointly optimizes the framework for downsampling, upsampling, and enhancement layer prediction to maximally utilize the information extractable from the base layer sequence, as well as that in the enhancement layer motion compensation. 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.

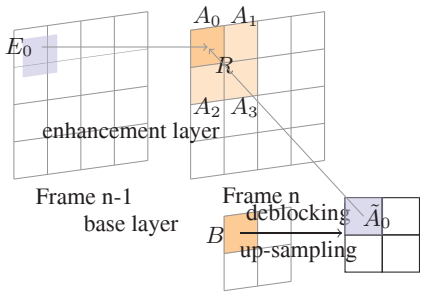

Fig. 1: Pixel domain enhancement layer prediction in spatial SVC.

3.1. 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 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 ≈ 1. Here the superscript T denotes matrix 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:

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

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,

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

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:

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

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

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

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:

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

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., [7]. 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.

3.2. The Optimal Enhancement Layer Prediction

We now describe the estimation-theoretic 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 (usually of size 16x16), and each macroblock is coded with inter-layer or inter-frame prediction, or in intra mode. Transforms are applied at sub-macroblock resolution (typically 4x4 and 8x8) to the prediction residuals, followed by quantization and entropy coding. 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 in the base layer frame *via the transform domain downsampling* previously described in Sec.3.1. 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 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 $x_n^b(i,j)$, $i,j \in \{0, \cdots, M-1\}$, the transform coefficients of the base layer block B:

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

These coefficients are subjected to an $M \times M$ inverse DCT to yield the base layer pixel block B, and coded as usual by the base layer codec. Since the choice of spatial transform applied to the base layer block (in intra mode) or its motion compensated prediction residual (in inter-frame or temporal prediction 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 $I_n^b(i,j)$ containing $x_n^b(i,j)$. 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 Sec.3.1 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 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. In particular, note that on account of the transform domain resampling the following relation holds:

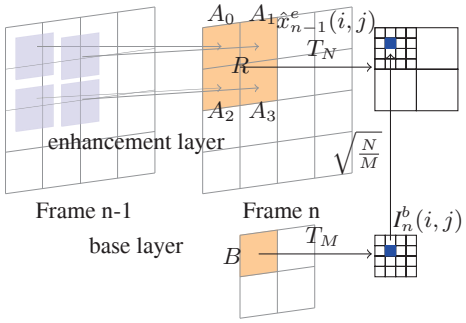

Fig. 2: Estimation-theoretic enhancement layer prediction.

$$x_n^e(i,j) \in I_n^e(i,j) = \frac{N}{M} I_n^b(i,j),\ i,j \in \{0, \cdots, M-1\}, \quad (7)$$

which implies that the information in the base layer quantization intervals directly translates into information about transform coefficients at the enhancement layer. This information cannot be utilized in the pixel domain. The ET prediction approach we now describe improves coding performance by specifically utilizing this interval information.

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: $x_n^e(i,j) = \rho x_{n-1}^e(i,j) + z_n(i,j)$, 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 $\hat{x}_{n-1}^e(i,j) \approx x_{n-1}^e(i,j)$, the conditional pdf $p(x_n^e|\hat{x}_{n-1}^e) \approx p_Z(x_n^e - \hat{x}_{n-1}^e)$. In the absence of additional base layer information, the best prediction of $x_n^e(i,j)$ is simply $\hat{x}_{n-1}^e(i,j)$, the default inter-frame estimate. But the base layer indicates that $x_n^e(i,j) \in I_n^e(i,j)$ when $i,j \in \{0, \cdots, M-1\}$, which refines the conditional pdf of $x_n^e(i,j)$ to

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

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. The optimal predictor at the enhancement layer is now given by

$$E\{x_n^e(i,j)\} \quad (9)$$
$$= \begin{cases} E\{x_n^e(i,j)|\hat{x}_{n-1}^e(i,j), I_n^e(i,j)\} & i,j \in \{0, \cdots, M-1\} \\ \hat{x}_{n-1}^e(i,j) & else. \end{cases}$$

The above equation describes the transform coefficients of the enhancement layer prediction for the entire $N \times N$ region R in Fig.2. This transform domain prediction of R is now inverse 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 [3].

4. SIMULATION RESULTS

We implemented the proposed unified ET approach in the JSVM reference framework. The 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 upsampling, in addition to the inter-frame prediction, which is henceforth referred to as H.264/SVC-ML. The matched upsampling filter proposed in [2] was further tested, which is denoted by H.264/SVC-MF. The scheme that allows an additional mode, where the prediction is formed as a linear combination of inter-layer and inter-frame predictions [6] was also implemented in the modified H.264/SVC framework, and is referred to as H.264/SVC-LC. Regular pixel domain motion estimation is enabled at quarter-pixel resolution for all four codecs.

Our experiments suggest that the base layer sequences generated by pixel and transform domain downsampling methods, respectively, render indistinguishable rate-distortion performance. Hence, for fair comparison, the transform domain downsampled sequence is used in all SVC codecs. The enhancement layer coding performance of the four codecs for the sequence $foreman$ at CIF resolution is shown in Fig.3. Clearly, the H.264/SVC-LC provides advantage compared to H.264/SVC-ML at relatively low bit-rate, while the proposed unified ET approach consistently provides substantial coding gains over either competing scheme. A potential downside of employing an unconventional resampling technique is the possibility of blocking artifacts in the base-layer. Fig.4 provides a visual comparion of a single reconstructed base layer frame of the $mobile$ sequence generated using DCT domain down-sampling and pixel domain decimation, respectively. Clearly both reconstructions are smooth and sharp, and no strong blocking artifacts are visible. Thus, the ET method offers major gains in enhancement layer performance at no discernible degradation of the base layer. Similar enhancement layer performance improvements were obtained for the sequence $harbour$ as shown in Fig.5, and for other test sequences.

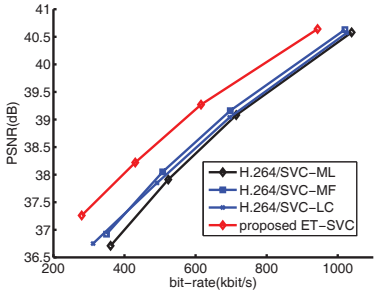

Fig. 3: Comparison of the coding performance of the competing spatial SVC approaches: The testing sequence is $foreman$ at CIF resolution. The base layer is at $QCIF$ resolution, and is coded at $408kbit/s$ with reconstruction quality $39.7dB$ (with respect to the downsampled sequence).

5. CONCLUSION

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 base layer and prior enhancement layer reconstructions, 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 one of its variants.

(a) proposed downsampling (b) pixel-domain decimation

Fig. 4: Comparison of the perceptual quality of reconstructed base layer frames with transform-domain (a) and pixel-domain (b) decimation: the base layer sequences are generated from $mobile$ at CIF resolution. Both versions were coded at $1200kbit/s$ and PSNR $38.3dB$. The frame shown is indexed 20.

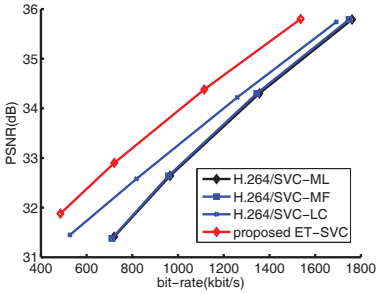

Fig. 5: Comparison of the coding performance of the competing spatial SVC approaches: The testing sequence is $harbour$ at CIF resolution. The base layer is at $QCIF$ resolution, and is coded at $680kbit/s$ with reconstruction quality $34.6dB$.

6. REFERENCES

[1] 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.

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

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

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

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

[6] 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.

[7] J. M. Adant et. al., "Block operations in digital signal processing with application to TV coding," *Signal Processing*, vol. 13, pp. 385–397, Dec 1987.