TOWARDS PREDICTOR, QUANTIZER AND ENTROPY CODER OPTIMALITY IN SCALABLE VIDEO CODING

Jingning Han and Kenneth Rose

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

ABSTRACT

A novel coding paradigm is proposed to jointly optimize the prediction, quantization, and entropy coding modules, thereby approaching optimality in scalable video coding. It departs from conventional video coding schemes that consider prediction, transformation, quantization, and entropy coding, as largely separate sequential functional components. The method draws inspiration from an early estimation-theoretic approach, developed by our group for enhancement layer prediction, which efficiently combines all the information available to the enhancement layer coder, to produce the optimal prediction. The framework is significantly expanded here to also incorporate optimization of entropy-constrained quantization and arithmetic coding, while fully accounting for hitherto ignored relevant factors, inherent to predictive scalable coding, including information from the base layer quantization operation, and from the enhancement layer motion compensated reference. Experimental evidence is provided for substantial coding gains over conventional scalable video coding.

Index Terms— Scalable video coding, estimation-theoretic framework, entropy-constrained quantization, CABAC

1. INTRODUCTION

In scalable video coding (SVC), the base layer consists of information about the video sequence that can be decoded independently to obtain a reconstruction of coarse quality, whereas the enhancement layers' information allows a decoder to successively refine the reconstruction. Enhancement layer packets may be dropped on-the-fly at intermediate network nodes to adjust the transmission rate, while retaining a baseline decoding quality. Hence, SVC is an attractive approach for applications that cater to receivers with varying reception bandwidth or for transmission over networks with diverse communication links [1, 2]. For simplicity of exposition, we restrict our discussion throughout this text to a two-layered quality SVC codec, while emphasizing that the proposed approach is extensible to more layers and other scalability types [3, 4].

The base layer encoding is essentially the same as single layer video coding, where macroblocks are predicted using either motion compensation, or intra-frame interpolation, and the resultant residuals are spatially transformed, quantized, and then entropy coded. The enhancement layer coder, however, has access to information from both the base and enhancement layers. Conventional designs of the enhancement layer coder typically inherit the base layer coder structure, while allowing utilization of the additional base layer information to improve the *prediction* quality, and hence the coding performance therein. For instance, standard approaches [1] perform the enhancement layer prediction in the pixel domain by selecting amongst the available prediction modes the one that minimizes the rate-distortion cost, which are inherently suboptimal since they cannot fully exploit information from both base and enhancement layers (see Sec. 2 for more details). An estimation-theoretic (ET) prediction approach was proposed in [3], where the enhancement layer motion compensated reference is efficiently combined with base layer quantization information, in a suitably derived conditional expectation framework, for optimal enhancement layer prediction per transform coefficient. The ET prediction approach was experimentally demonstrated to substantially outperform existing pixel domain prediction methods, and was further enhanced by allowing delayed prediction [5] and optimized for deployment over lossy networks [6]. This principle was also successfully implemented in the context of Wyner-Ziv scalable coding [7]. The prediction residual then undergoes quantization and entropy coding, both of which are typically performed without recourse to the additional information that was exploited for prediction. An alternative perspective will be pursued here.

The quantization of a random variable given its probability density function (pdf), is effectively a partition of its support into several mutually exclusive cells, each represented by the corresponding centroid and associated with its probability of containing the outcome. The fundamental design problem can be formulated as a trade-off between the expected reconstruction distortion (the average squared reconstruction error), and the rate spent to specify the cell or representative (whose probability mass function is determined by the source and partition), which is approximated by the entropy, as is justified in the case of arithmetic coding [8]. The design of the optimal quantizer that minimizes the rate-distortion cost has been intensively studied over decades [9]. In particular, it was shown that for a Laplacian process, which is commonly used for modeling the temporal innovations in transform domain of video sequences, the deadzone quantizer can achieve coding performance fairly close to the optimum [10]. The deadzone quantizer and its variants are widely adopted in single-layer video encoders that employ motion-compensated prediction to exploit temporal correlations, and were also "inherited" by the SVC codecs. However, as we will show, the enhancement layers of SVC have additional access to base layer information, conditioned on which the effective pdf may differ significantly from the Laplacian distribution, thereby casting doubts about the efficacy of deadzone quantizers here. In this paper, we approach this problem by first deriving the conditional probability distribution, given information from both the base and enhancement layers, based on which an optimum entropy-constrained quantizer can be selected for lossy compression, per transform coefficient. Related prior work, where adaptive quantizers are employed, includes a scalar quantizer design approach that exploits the previously coded local texture information for image coding [11] . In [12], a coding scheme that switches the quantizers depending on the base layer information was proposed for scalable audio coding which, unlike the video coding case, often

does not exploit inter frame correlation, i.e., the coding scheme is akin to inter layer coding in SVC.

We observe that the derived quantizer, known to both encoder and decoder, also provides the probability associated with each cell (i.e., each quantization index), which is vital to the efficacy of arithmetic coding. We hence develop a quantizer-adaptive m-ary arithmetic coding (QAMAC) for 2-D blocks of quantization indices at the enhancement layer, instead of using the context-based adaptive binary arithmetic coding (CABAC) inherited from single-layer coding [13]. Related research focused on improving entropy coding for SVC includes [14], where a more sophisticated probability estimation method was devised for the context models of regular CABAC for successive bit-plane coding, i.e., motion compensated prediction was precluded from the enhancement layer coding.

While the proposed approach is implemented in H.264/AVC scalable video coding extension reference framework to demonstrate its efficacy, its basic principles are generally applicable to most predictive coding systems, e.g., VP8, HEVC, etc.

2. BACKGROUND

We briefly describe related materials on standard SVC and its variants. The H.264/SVC coder compresses the base layer as a single bitstream, and employs a single-loop design to code the enhancement layer, where the decoder need not buffer its base layer reconstruction to produce the enhancement layer signal. Particularly, the enhancement layer coder starts with motion compensation from previously reconstructed frames in the same layer to generate a prediction residual block. It then adaptively decides whether to further subtract the base layer reconstructed residual from this residual block before transformation and quantization [1, 2]. In earlier standards such as H.263++ the enhancement layer prediction switches between (motion compensated) prior enhancement layer reconstruction and current base layer reconstruction, or a linear combination thereof, in what is referred to as a multi-loop design [15]. It has been recognized that multi-loop design performs better than single-loop at the expense of more decoding complexity [2]. Since the main focus of this paper is on optimality in coding performance, the H.264/SVC codec is modified to the better performing multi-loop design, while retaining the other advanced components and capabilities, such as sub-pixel motion compensation, CABAC, etc.

3. ESTIMATION-THEORETIC SCALABLE VIDEO CODING ENGINE

We present the proposed SVC coding scheme that efficiently combines information from both base and enhancement layers in an appropriately derived ET framework, to jointly optimize the quantization and entropy coding.

3.1. Optimal Entropy-Constrained Predictive Quantizer

Let x_n denote the value of a particular transform coefficient in a block of the current frame. Let $\hat{x}^b_{n-1}$ denote the transform coefficient of the same frequency as x_n, in the base layer motion compensated reference block, which is obtained from the reconstruction of the previous frame. Thus the operation of the standard base layer encoder is equivalent to a quantization of $(x_n - \hat{x}^b_{n-1})$ to produce the index i^b_n. Let $[a_n, b_n)$ be the quantization interval associated with index i^b_n. Clearly, the fact $x_n \in \mathcal{I}^b_n = [\hat{x}^b_{n-1} + a_n, \hat{x}^b_{n-1} + b_n)$ captures all the information on x_n provided by the base layer, namely, it specifies the interval in which x_n must reside. When encoding the enhancement layer of x_n, the encoder also has access to transform coefficient $\hat{x}^e_{n-1}$ of the motion-compensated reference block, generated from the previously reconstructed frame. The prior enhancement layer information $\hat{x}^e_{n-1}$ can then be combined with the base layer interval $\mathcal{I}^b_n$, in an estimation-theoretic framework to obtain the optimal prediction for coefficient x_n. It is important to note that although the enhancement layer reconstruction $\hat{x}^e_{n-1}$ can be equivalently expressed in the spatial pixel domain, the quantization interval $\mathcal{I}^b_n$ emerges only in *transform* domain. Thus, ad hoc pixel domain linear combinations of base layer residual or reconstruction, with prior enhancement layer reconstruction, as employed by current and prior standard SVCs cannot enable the efficient combination of both sources to achieve optimality in SVC.

Predictive video coders typically model blocks of pixels along the same motion trajectory in consecutive frames as an autoregressive (AR) process. Motion compensation is employed to align these blocks, and pixel domain subtraction removes temporal redundancies. The alternative viewpoint that the DCT coefficients of corresponding blocks form an AR process per frequency coefficient, was adopted in [3]. Here x_n (at a given frequency) and the corresponding motion compensated transform coefficient from previous frame x_{n-1} conform to the first order AR model: $x_n = \rho x_{n-1} + z_n$, where $\{z_n\}$ are independent and identically distributed (i.i.d) innovations of the process with pdf $p_Z(z_n)$. The implicit assumption of the standard pixel domain motion compensated prediction that the temporal correlation coefficient is $\rho \approx 1$, will be made here for simplicity at all frequencies (this assumption can be relaxed to enhance the accuracy). The above transform domain AR process perspective provides the advantage that the motion compensated $\hat{x}^e_{n-1}$, and the quantization interval $\mathcal{I}^b_n$, can now be combined to derive the conditional pdf of x_n.

Assuming that $\hat{x}^e_{n-1} \approx x_{n-1}$, we obtain the conditional pdf $p(x_n|\hat{x}^e_{n-1}) \approx p_Z(x_n - \hat{x}^e_{n-1})$. The base layer further indicates that $x_n \in \mathcal{I}^b_n$, which refines the conditional pdf of x_n to

$$p(x_n|\hat{x}^e_{n-1}, \mathcal{I}^b_n) \approx \begin{cases} \frac{p_Z(x_n - \hat{x}^e_{n-1})}{\int_{\mathcal{I}^b_n} p_Z(x_n - \hat{x}^e_{n-1}) dx_n} & x_n \in \mathcal{I}^b_n \\ 0 & else \end{cases} . \quad (1)$$

Note that the above is equivalent to centering $p_Z(z_n)$ at $\hat{x}^e_{n-1}$, restricting it to the interval $\mathcal{I}^b_n$ (a non-linear operation), and then normalizing to obtain a valid pdf. 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]-[6]. The optimal predictor at the enhancement layer is

$$\tilde{x}^e_n = E[x_n|\hat{x}^e_{n-1}, \mathcal{I}^b_n], \quad (2)$$

the centroid of the above pdf over the *entire* interval $\mathcal{I}^b_n$ [3].

If the prediction and quantization operations are separately performed, then the traditional course of action would now be to quantize the residual $(x_n - \tilde{x}^e_n)$ via a deadzone quantizer and encode the index (typically) using CABAC. However, such separate treatment of the prediction and quantization suffers from significant underutilization of the available information. In particular, in anticipation of the optimally matched entropy coder to be discussed next, the optimal entropy-constrained quantizer for x_n, given the conditional pdf $p(x_n|\hat{x}^e_{n-1}, \mathcal{I}^b_n)$ can be obtained via a variant of the Lloyd-Max algorithm [16]. Consider a scalar quantizer of N levels. Let the decision or boundary points of the partition be denoted by $\{t_i | i = 0, 1, \cdots, N\}$, and the reproduction levels by $\{r_i | i = 1, 2, \cdots, N\}$. The base layer quantization interval $\mathcal{I}^b_n$ bounds the effective support

of the signal considered by the enhancement layer quantizer and effectively sets $t_0 = \hat{x}^b_{n-1} + a_n$ and $t_N = \hat{x}^b_{n-1} + b_n$. The necessary conditions for optimality of an entropy-constrained quantizer of N levels were specified in [9]:

$$\beta\, log_2(\frac{P_{i+1}}{P_i}) = (r_{i+1} - r_i)(r_{i+1} + r_i - 2t_i),\ \forall i = 1,\ 2, \cdots,\ N-1, \tag{3}$$

where P_i is the probability of the i^{th} region, and β is the Lagrangian multiplier whose value may be varied to obtain the desired point on the operational rate-distortion curve. The necessary condition of (3), leads to an entropy-constrained Lloyd-Max quantizer design. The variant of the design algorithm for the enhancement layer quantizer is derived in a straightforward manner and the pseudo-code is given in Fig.1, where ϵ determines the convergence test. Upon

Initialize
$t_0 = \hat{x}^b_{n-1} + a_n$
$t_N = \hat{x}^b_{n-1} + b_n$
for $i = 1$ to N **do**
  $t_i \leftarrow t_0 + i \cdot \frac{t_N - t_0}{N}$
  $r_i \leftarrow \frac{\int_{t_{i-1}}^{t_i} x_n\ p(x_n|\hat{x}^e_{n-1}, \mathcal{I}^b_n) dx_n}{\int_{t_{i-1}}^{t_i} p(x_n|\hat{x}^e_{n-1}, \mathcal{I}^b_n) dx_n}$
end for
$c \leftarrow r_N$

repeat
  $r_N \leftarrow c$
  for $i = 1$ to $(N-1)$ **do**
    $r_i \leftarrow \frac{\int_{t_{i-1}}^{t_i} x_n\ p(x_n|\hat{x}^e_{n-1}, \mathcal{I}^b_n) dx_n}{\int_{t_{i-1}}^{t_i} p(x_n|\hat{x}^e_{n-1}, \mathcal{I}^b_n) dx_n}$
    $P_i \leftarrow \int_{t_{i-1}}^{t_i} p(x_n|\hat{x}^e_{n-1}, \mathcal{I}^b_n) dx_n$
    $P_{i+1} \leftarrow \int_{t_i}^{t_{i+1}} p(x_n|\hat{x}^e_{n-1}, \mathcal{I}^b_n) dx_n$
    $t_i \leftarrow \frac{1}{2}(r_i + r_{i+1}) - \frac{\beta}{2} \frac{log_2(P_{i+1}/P_i)}{r_{i+1} - r_i}$
  end for
  $c \leftarrow \frac{\int_{t_{N-1}}^{t_N} x_n\ p(x_n|\hat{x}^e_{n-1}, \mathcal{I}^b_n) dx_n}{\int_{t_{i-1}}^{t_i} p(x_n|\hat{x}^e_{n-1}, \mathcal{I}^b_n) dx_n}$
until $|c - r_N| < \epsilon$

Fig. 1: Pseudo code for the enhancement layer entropy-constrained predictive quantizer design.

convergence, the rate-distortion cost associated with this N-level scalar quantizer can be calculated. We then vary the value of positive integer N to find the one that provides overall minimum rate-distortion cost as the optimum quantizer for x_n, given conditional pdf $p(x_n|\hat{x}^e_{n-1}, \mathcal{I}^b_n)$.

The Laplacian memoryless property allows for a set of generic quantizers to be pre-calculated and stored during the initial stage of coding process. The encoder can then simply fetch the needed quantizer, conditioned on the motion compensated reference and base layer information, eliminating the need to redesign quantizers. Hence the overall increment in computational complexity is modest.

3.2. Quantizer-Adaptive M-ary Arithmetic Coding

Entropy coding typically consists of two steps: converting the 2-D block of quantizer indices into a 1-D sequence, and coding the index at each position. The H.264/AVC standard and the SVC extension, employ the coded block flag, a bit that indicates whether or not there are nonzero (quantized) transform coefficients to encode. If this flag is zero, no further information needs to be transmitted; otherwise, the encoder scans the 2-D block in a zag-zig manner from low to high frequency components, and sends a binary-valued significance map to specify the positions of nonzero coefficients. The absolute values and the signs of nonzero coefficients are then encoded in reverse scanning order, which enables better exploitation of inter symbol correlation, in terms of choosing the proper context models according to the number and absolute values of nonzero coefficients previously coded in the same block. The underlying hypothesis of such design is that DCT tends to compact the energy into lower frequency components. To better adapt to the probability models, a binarization scheme is employed in CABAC that maps a given non-binary syntax element (e.g., the absolute value of nonzero coefficient) into a unique binary sequence, called bin string. Depending on the position of binary decision in this bin string, a set of context models will be assigned per bin to represent probability distributions for arithmetic coding, conditioned on priorly coded symbols. Significant rate reductions were achieved by CABAC over other existing variable length based coding methods, in the context of both single-layer and scalable video coding [13].

The entropy coder we propose for SVC here exploits additional available information to achieve significant gains. It is premised on the fact that the optimal quantizer (see Sec.3.1) *explicitly* provides the probability mass function, which can be used to optimize the dimension conversion and index coding, namely, quantizer-adaptive m-ary arithmetic coding (QAMAC). In particular, we assign to the most probable cell of each transform coefficient the index zero, and use coded block flag to indicate whether there are any nonzero indices, i.e., there are "significant coefficients". If the flag is on, the encoder scans the 2-D block in descending order of coefficient total probability of significance (instead of the traditional zig-zag order), and generates a binary-valued significance map which is coded by the binary arithmetic coder. To encode the significant coefficients, the QAMAC employs an m-ary arithmetic coder, the recursive interval subdivisions of which are conditioned on the quantizers for the significant coefficients. Suppose a transform coefficient is coded by an N-level quantizer, where the k^{th} region is the most probable one. The significance map indicates that the true value of this coefficient does not fall into the most probable region, which eliminates r_k (indexed zero) from the sample space and refines the probability mass function of this significant coefficient as

$$\tilde{P}_i = \frac{P_i}{1 - P_k},\ \forall i \neq k,\ i \in \{1, 2, \cdots, N\}. \tag{4}$$

The internal range of the arithmetic coder is thus divided into $(N-1)$ subintervals, the i^{th} of which has a length proportional to $\tilde{P}_i$. Depending on the observed symbol value, the corresponding subinterval will be chosen as the new current interval. The binary expansion of the number pointing into this interval effectively represents the sequence of symbols coded so far, and hence forms the coded bit stream. We note that in the implementation the intervals are approximately represented by integers in $[0, 2^{10}]$, which enables a table-based m-ary arithmetic coding, akin to the table-based binary approach in CABAC [13], to reduce the codec complexity.

4. SIMULATION RESULTS

The proposed ET coding engine is implemented in H.264/AVC scalable video coding extension reference framework, and is referred to as ETQ-SVC. The standard SVC codec was only modified to include multi-loop prediction, and is denoted by H.264/SVC-ML. The

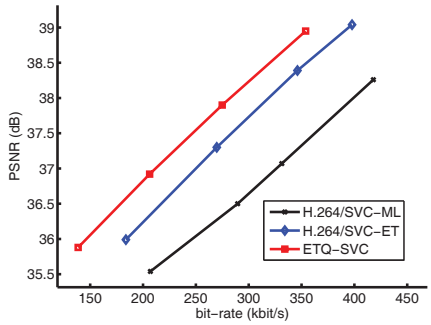

Fig. 2: Comparison of the enhancement layer coding performance. The test sequence is $coastguard$ at $QCIF$ resolution. The base layer is coded at 200 $kbit/s$ with reconstruction quality of 34.3 dB.

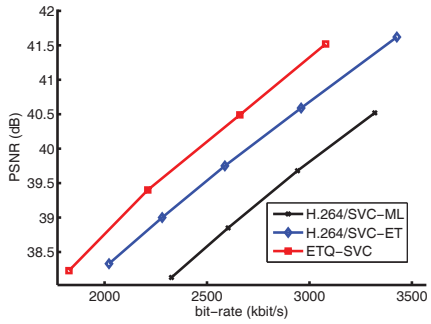

Fig. 3: Comparison of the enhancement layer coding performance. The test sequence is $mobile$ at CIF resolution. The base layer is coded at 1680 $kbit/s$ with reconstruction quality of 34.0 dB.

original ET prediction approach developed by our group to improve prediction but not quantizer and entropy encoder, was also included as H.264/SVC-ET. The three SVC codecs use the same base layer coder. All layers employ regular quarter pixel resolution motion search and single reference frame for inter-frame motion compensated prediction. We emphasize that other more sophisticated inter-frame motion compensated prediction methods can be directly incorporated in the propose framework. The test sequences are coded in $IPPP$ format at frame rate of 30 f/s. The enhancement layer coding performance for sequence $coastguard$ at $QCIF$ resolution is presented in Fig.2, where we fix the base layer coding configurations and vary the Lagrangian multiplier β to generate enhancement layer operational points. Clearly, the proposed ET coding engine consistently outperforms other competing schemes. Similar coding performance gains are achieved for sequence $mobile$ at CIF resolution as shown in Fig.3. We note that the experiments also suggest similar improvements for other test sequences, across a wide range of bit rates.

5. CONCLUSIONS

This paper proposes a novel enhancement layer coding engine for optimal compression in SVC. The approach efficiently combines all relevant information from both base and enhancement layers in an ET framework to derive the conditional pdf, which enables derivation of the optimal entropy-constrained predictive quantizer and conditional arithmetic coder. Considerable and consistent coding gains are obtained by using the proposed ET coding engine, in comparison to standard H.264/SVC as well as our research group's earlier ET approach to SVC that focused on enhancement layer prediction.

6. 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] 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.

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

[5] 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.

[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," *Proc. IEEE ICIP*, pp. 3278–3281, Sep. 2011.

[7] H. Wang, N. M. Cheung, and A. Ortega, "A framework for adaptive scalable video coding using Wyner-Ziv techniques," *EURASIP Jrnl. App. Sig. Proc.*, Jan. 2006.

[8] J. Rissanen and G. G. Langdon, "Universal modeling and coding," *IEEE Trans. Inform. Theory*, vol. 27, no. 1, pp. 12–23, Jan. 1981.

[9] T. Berger, "Minimum entropy quantizers and permutation codes," *IEEE Trans. Inform. Theory*, vol. 28, no. 2, pp. 149–157, Mar. 1982.

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

[11] A. Ortega and M. Vetterli, "Adaptive scalar quantization without side information," *IEEE Trans. Img. Proc.*, vol. 6, no. 5, pp. 665–676, May 1997.

[12] A. Aggarwal, S. L. Regunathan, and K. Rose, "Efficient bit-rate scalability for weighted squared error optimization in audio coding," *IEEE Trans. Audio, Speech, and Language Proc.*, vol. 14, no. 4, pp. 1313–1327, July 2006.

[13] D. Marpe, H. Schwarz, and T. Wiegand, "Context-based adaptive binary arithmetic coding in the H.264/AVC video compression standard," *IEEE Trans. Circ. Sys. Video Tech.*, vol. 13, no. 7, pp. 620–636, July 2003.

[14] S.-H. Kim and Y.-S. Ho, "Fine granular scalable video coding using context-based binary arithmetic coding for bit-plane coding," *IEEE Trans. Circ. Sys. Video Tech.*, vol. 17, no. 10, pp. 1301–1310, Oct. 2007.

[15] 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.

[16] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inform. Theory*, vol. 28, no. 2, pp. 129–137, Mar. 1982.