ON GENERALIZING THE ESTIMATION-THEORETIC FRAMEWORK TO SCALABLE VIDEO CODING WITH QUADTREE STRUCTURED BLOCK PARTITIONS

Shunyao Li, Tejaswi Nanjundaswamy, Bohan Li and Kenneth Rose

Department of Electrical and Computer Engineering, University of California Santa Barbara, CA 93106
E-mail: {shunyao_li,tejaswi,bohan_li,rose}@ece.ucsb.edu

ABSTRACT

Scalable video coding suffers from the under-utilization of base layer information, where usually only the reconstruction in the base layer is used for enhancement layer prediction. Prior work from our lab proposed an optimal estimation-theoretic (ET) approach for quality scalable coding, wherein the estimates are obtained by utilizing all the available information from base layer quantization interval and enhancement layer distribution for transform coefficients. While this approach was proposed for fixed block size encoding, modern codecs employ variable block size quadtree structured partitioning, which results in different partitions at base layer and enhancement layer based on the rate-distortion trade-off, thus makes the base layer information not directly usable in the enhancement layer. Other new tools such as hybrid transform and the rate-distortion optimized quantizer (RDOQ) also have an impact on the information available for optimal estimation. In this paper, we generalize the ET framework for quality scalable video coding to account for the quadtree structured partitioning, hybrid transform and the RDOQ adjustment. Experimental evidence is provided for consistent coding gains over standard SHVC.

Index Terms**—** Scalable video coding, prediction, transform, quantization, quadtree partitioning

1. INTRODUCTION

Modern video applications (streaming, broadcasting, etc.) operate on RTP/IP [1] networks for real-time services, which is characterized by a broad range of connection qualities and receiving devices. To adapt to the differences in the end-user devices' capabilities and network conditions, scalable video coding (SVC) was proposed and adopted as extensions to video coding standards of H.264 [2] and HEVC (SHVC) [3]. SVC allows the video sequences to be encoded "progressively", i.e., a video sequence encoded at one quality can be enhanced to a higher quality by adding a refinement bitstream, successively any number of times. In this hierarchical structure, even if the top refinement bitstreams are lost due to temporary constraints in the network, the rest would still be a valid decodable bitstream. Specifically, bitstream at the lowest quality is referred to as the base layer (BL), while bitstreams at higher qualities are referred to as the enhancement layers (EL). In addition to quality scalability, there is also spatial and temporal scalability where the resolution and frame rate varies between layers.

A critical challenge that limits the practical use for SVC is how to exploit the BL information effectively in the ELs, especially in the EL prediction. In the SVC standards [2, 3], the BL reconstruction can be used as an additional reference frame for EL motion compensation, and the BL motion vectors can be used to predict EL motion vectors. In [4], a pyramid approach is proposed where the interpolated BL residual is used to predict the EL residual. In [5, 6], a subband coding approach is proposed where the different resolutions of subband data are obtained from different layers. A linear combination of EL and BL with three additional weighting types is introduced in [7]. A rate-distortion (RD) optimized selection between the above approaches is proposed in [8, 9]. While all the prior approaches try to exploit the BL reconstruction for EL prediction, none of them utilize the BL quantizer information in the transform domain, which gives the exact region that the original value lies in. Prior work from our lab [10] exploits this quantization interval information and combines it with the transform coefficients distribution information in EL from an estimation-theoretic (ET) viewpoint, which provides the theoretically optimal EL prediction. A follow-up work on this [11, 12] significantly extended it to the spatially scalable video coding with considerable coding gains.

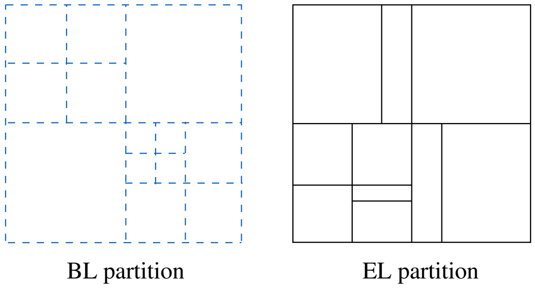


Fig. 1. An example of different partitions at base and enhancement layer.

However, as more advanced tools are being developed for video coding, the ET approach has not been updated to account for them. One critical new tool to be supported is the quadtree block partition [13], which provides significant flexibility and hence commonly used in modern video coders. With this tool, the BL and ELs are generally partitioned differently due to their different rate-distortion trade-offs, as shown in Fig. 1. This mismatch in partitions obviously carries over to the transform domain, and thus the BL interval information cannot be directly combined with the EL distribution information as proposed in ET prediction. The hybrid transform adopted by HEVC [13] and proposed for the next generation video codec JVET [14] greatly expand the family of transform kernels and leads to more variations in the distribution of transform coefficients, which needs to be accounted for properly. The rate-distortion optimized quantizer (RDOQ) [13], where the quantized index is adjusted to achieve better rate-distortion performance, results in erroneous interval information, i.e., the quantization interval does not always contain the true value of the coefficients, which also needs to be taken into account.

In this paper, we generalize the ET framework to account for the advanced tools of quadtree partitioning, hybrid transform, and RDOQ for quality scalable video coding. To account for partitioning mismatch, we first use EL prediction parameters to generate transform coefficients distribution at BL transform unit size, then combine this with BL quantization interval information as in the standard ET approach, and finally transform this to generate final ET prediction in the EL prediction unit size. To account for the various types of transform, we train the distribution parameters for DCT, ADST and transform skip (TS) at different target bitrates separately, and apply them in tandem for hybrid transform. We adjust the quantizer interval information accounting for the RDOQ to avoid inaccurate interval information. The proposed approach is implemented in SHVC and compatible with all the existing features with no additional overhead and negligible additional complexity. Consistent gains across video sequences at different resolutions are presented to prove the efficacy of the approach.

2. BACKGROUND: ET PREDICTION

The ET approach for EL prediction is formulated as an estimation problem of the current sample given all the available information. Without loss of generality, we assume there are only two layers, which are coded in a quality scalable encoder. For each sample in the EL, there are two sources of information available: EL reconstruction of prior samples, and the parameters (reconstruction, prediction, compressed residual, quantization parameters, etc.) associated with the BL coder for the same sample.

In a single-layer coder where only one information source is available, the prediction, $\tilde{x}$, can be derived via motion compensation or intra prediction. The residual, $x - \tilde{x}$, is then transformed, quantized, and sent to the decoder. It has been shown in prior work [15, 16, 17] that the DCT coefficients of the residual, ϵ, can be approximated by a Laplacian distribution centered at zero, $\frac{\lambda}{2} exp(-\lambda|\epsilon|)$. Therefore, the DCT coefficients, x^T, of the actual pixel value, x, would follow the same Laplacian distribution centered at the prediction in the transform domain $\tilde{x}^T$, i.e.,

$$f(x^T \mid \tilde{x}^T) = \frac{\lambda}{2} exp(-\lambda|x^T - \tilde{x}^T|). \tag{1}$$

The modern quantizer is hence designed as an uniform dead-zone quantizer based on the distribution and the quantization parameters (QP).

In a two-layer SNR scalable coder where two sources of information are available from BL and EL, the BL reconstruction is usually used as an additional reference and combined with EL prediction, either linearly as in the standard [3], or via other suboptimal approaches [4, 5, 6]. However there is more information available from the BL prediction and QP. Given quantized residual index i^b and QP, we know the exact interval (a, b) associated with i^b. If $\tilde{x}^{bT}$ is the BL prediction in the transform domain, we have

$$\epsilon^b = x^T - \tilde{x}^{bT} \in (a, b), \tag{2}$$

$$x^T \in (\tilde{x}^{bT} + a, \tilde{x}^{bT} + b). \tag{3}$$

Similar to (1), the EL prediction, $\tilde{x}^{eT}$, provides the distribution information, $f(x^T \mid \tilde{x}^{eT}) = \lambda/2 \; exp(-\lambda|x^T - \tilde{x}^{eT}|)$, and (3) provides the interval information from BL that indicates the region the original value would fall in. Together we have a truncated Laplacian distribution, as shown in Fig. 2, the centroid of which would be the best estimation for x^T(also referred to as ET prediction in the rest of the paper),

$$\begin{aligned} \tilde{x}^{eT}_{ET} &= E(x^T \mid x^T \in (\tilde{x}^{bT} + a, \tilde{x}^{bT} + b), \tilde{x}^{eT}) \\ &= \frac{\int_{\tilde{x}^{bT}+a}^{\tilde{x}^{bT}+b} x^T f(x^T \mid \tilde{x}^{eT}) d(x^T)}{\int_{\tilde{x}^{bT}+a}^{\tilde{x}^{bT}+b} f(x^T \mid \tilde{x}^{eT}) d(x^T)}. \end{aligned} \tag{4}$$

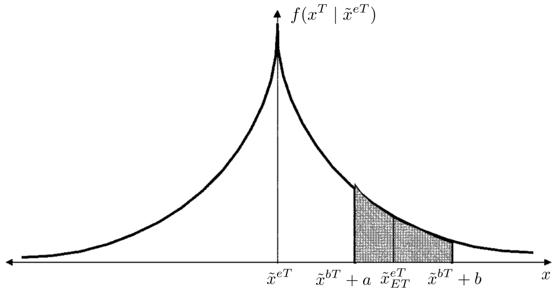


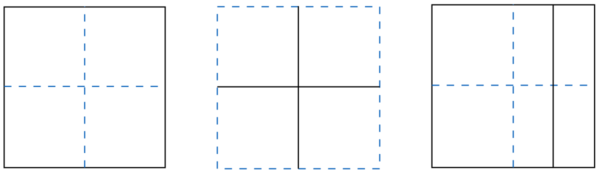
Fig. 2. The distribution for transform coefficients (the centroid of the shaded region is its optimal ET prediction)

3. ET PREDICTION WITH PARTITIONING MISMATCH BETWEEN LAYERS

In modern video codecs such as HEVC [13], each video frame is divided into 64×64 blocks (referred to as CTU), then each of them can be further split recursively into different sizes of coding units (CU) in a quadtree structure. At each leaf node of the quadtree, a CU can be further partitioned (rectangularly) into different prediction unit (PU), each associated with a motion vector or intra prediction mode. After the prediction, each CU is further split recursively in a similar quadtree method to different sizes of transform units (TU). In general, finer partition leads to better coding quality (less distortion) but at a higher bitrate. Depending on the target bitrate (or target quality) requirement, the encoder makes the partition decision based on the rate-distortion cost, which is a Lagrangian formula defined for the rate-distortion trade-off. ELs and BL are coded at different qualities, thus usually have different partition decisions across layers (as an example shown in Fig. 1).

As described in Section 2, in the ET approach, the BL interval information lies in the transform domain thus is determined by the BL TU sizes. However, if the BL TU is not aligned with EL PU, this interval information cannot be directly used with the transform domain distributions for EL prediction. Although, as a naive approach, we can make them compatible by performing a linear transform (from one size to another) either on the interval information or on the distribution information, it is neither effective or practical. Performing a linear transform on the interval information means finding the overall support region of a linear objective function, $y = \mathbf{c}^T\mathbf{x}$, with $x_i \in (a_i, b_i]$. This overall region would be, $y \in \bigcup R_i$, where $R_i = (c_i a_i, c_i b_i]$ if $c_i \geq 0$, and $R_i = [c_i b_i, c_i a_i)$ if $c_i < 0$. This would result in a much larger interval for y, and sometimes lead to meaningless information of $(-\infty, \infty)$ whenever any of the variables in **x** has no interval information available (e.g., due to RDOQ as explained later). Performing the linear transform on the distribution involves a set of convolution operations, which are too complicated to be practical.

Instead, we propose an elegant and optimal solution, where we exploit the linearity property of expectations. From Section 2, we know the ET prediction for the EL, as in (4), is the centroid (a.k.a.



(a) Case 1 (b) Case 2 (c) Case 3

Fig. 3. Three cases of the partition mismatch between the EL PU (black line) and BL TU (blue dotted line)

the expectation) of the distribution for each transform coefficient, which by linearity property follows,

$$E(\mathbf{Y}) = E(\mathbf{A}^T\mathbf{X}\mathbf{A}) = \mathbf{A}^T E(\mathbf{X})\mathbf{A}. \tag{5}$$

Therefore, instead of performing the linear transform on the interval information or the distribution information, we directly work on the expectations, which retains the optimality of ET prediction even after any linear operation. To preserve flexibility for further TU partitioning after the prediction, we transform the ET prediction back to the pixel domain. Hence for our framework, $\mathbf{X}$ represents the transform domain coefficients in BL TU size, $\mathbf{Y}$ represents its corresponding pixels, and $\mathbf{A}$ corresponds to the inverse transform kernel. Specifically, we first transform the EL prediction in the same size as $\mathbf{X}$ to obtain $\tilde{x}^{eT}$, then calculate the optimal ET prediction $E(\mathbf{X})$ via (4), and finally calculate $E(\mathbf{Y})$ using (5). Transforming the EL prediction (in a PU) in the size of the BL TU involves blocks merging, extending and cropping. Depending on the mismatch of partitions, we consider the following three different cases:

- *Case 1: The BL TUs are inside the EL PU (see Fig. 3(a)).* The EL PU is divided into small blocks that are aligned with BL TU. For each small block, the EL prediction is converted into transform domain to compute the optimal ET prediction using (4). Then we get the corresponding pixel domain ET prediction for each small block using (5), and finally merge them together to get the optimal ET predicted EL PU.
- *Case 2: Part of the BL TU is outside the EL PUs (see Fig. 3(b)).* We could simply merge the EL predictions in different PUs, but this would introduce delay since the prediction of all the required PUs might not be available at the same instant. Instead, for each EL PU, we extend the EL prediction to the BL TU size using the same motion vector, and transform it to get the optimal ET prediction using (4). Then we get the corresponding pixel domain ET prediction using (5), and copy the corresponding prediction to the EL PU region. Also in this case, if it is intra predicted in the EL, we skip ET prediction due to lack of boundary information for extending beyond EL PU region.
- *Case 3: An EL PU covers multiple BL TUs, some of which are fully inside the EL PU and some extend outside (see Fig. 3(c)).* We transform such an EL PU using the same size as BL TUs via both division and extension. For the BL TUs fully inside an EL PU, we divide the EL PU into the same sizes as the BL TUs as we do in case 1; for those partly outside an EL PU, we extend the prediction to the size of BL TU as we do in case 2. The optimal ET prediction in pixel domain for all the divisions and extensions are merged together as the overall prediction.

The full block diagram of the EL prediction framework with ET scheme in scalable video coding with quadtree structured partitioning is shown in Fig. 4. In the traditional EL prediction without the ET scheme (the red block), the only information exploited from the BL is the motion vector and the reconstruction, while with the ET scheme we are also using the partition and quantization interval information from the BL. For each effective EL prediction via motion compensation or intra direction, we enhance it using the ET approach, and compare it with the BL reconstruction and use the best one as prediction. Note that although in principle the BL reconstruction is contained within the interval information, we noticed that directly referencing from base layer sometimes yields better rate-distortion performance due to savings in side information. The residual is then transformed and quantized using the most optimal TU quadtree structure. One of the future research directions will be to further exploit and account for the BL partition information while optimizing the CU/TU partition in EL.

It has been shown that DCT is not always the best separable transform to approximate KLT. Hence, modern video coders, such as HEVC, employ hybrid transform (DCT, ADST) for better decorrelation under certain conditions, and transform skip (TS), where quantization is done directly in pixel domain. Though the Laplacian distribution assumption for transform coefficients is usually only valid for DCT, we extend it to ADST and TS, and train the λ for the three different transform types, following the maximum-likelihood estimation, with N number of samples as,

$$\lambda = \frac{N}{\sum_{i=0}^{N-1} |x_i^T - \tilde{x}_i^{eT}|}. \tag{6}$$

Since statistics of prediction, $\tilde{x}^{eT}$, also depend on QP and transform block size, we train separate λ for different range of QPs and for each block size. We then employ these λ adaptively according to the QP and block size chosen by the encoder.

To improve the overall performance, rate-distortion optimized quantizer (RDOQ) was introduced in recent video coding standards. For each residual transform coefficient, in addition to its correct quantizer magnitude L, the encoder also considers two additional magnitudes $L - 1$ and 0, and chooses the one with the lowest RD cost. Similarly, the encoder also has the option to eliminate a whole coefficient group (which is usually 4 × 4) if it is cost effective. The skip mode (where the residual of the whole block is set to 0) is also used quite often when the bitrate budget is low. All of these techniques contribute significantly in improving the RD performance, but they also result in inaccurate interval information if derived solely from the quantization index. [11] dealt with a simpler variation of RDOQ in H.264 by disabling the ET prediction for a certain corner case. But in SHVC, we need a more robust approach to address the problem. Let's denote the quantizer interval associated with index i^b as $I_{i^b} = (a_{i^b}, b_{i^b}]$. Since we employ regular quantizers, the intervals of neighboring indices are consecutive, i.e., $a_{i+1} = b_i$. We propose a more robust rule to account for RDOQ by expanding the BL interval information in the following way:

- If $i^b = 0$, set the interval as $I_{-1} \bigcup I_0 \bigcup I_1 = (a_{-1}, b_1]$
- If $i^b > 0$, set the interval as $I_{i^b-1} \bigcup I_{i^b} = (a_{i^b-1}, b_{i^b}]$
- If $i^b < 0$, set the interval as $I_{i^b} \bigcup I_{i^b+1} = (a_{i^b}, b_{i^b+1}]$
- If $i^b = 0$ for all the transform coefficients in the block, then this block is very likely to be coded in skip mode, where no interval information is available, i.e., the interval is $(-\infty, \infty)$, and thus the ET prediction $\tilde{x}_{ET}^{eT}$ is the same as the EL prediction $\tilde{x}^{eT}$.

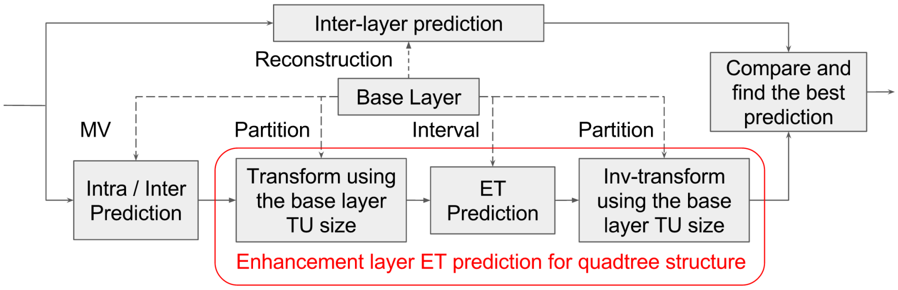

Inter-layer prediction
Reconstruction
Base Layer
MV
Partition
Interval
Partition
Compare and find the best prediction
Intra / Inter Prediction
Transform using the base layer TU size
ET Prediction
Inv-transform using the base layer TU size
Enhancement layer ET prediction for quadtree structure

Fig. 4. The EL framework block diagram in SHVC with ET prediction

4. EXPERIMENTAL RESULTS

To evaluate the performance, the proposed ET framework is implemented in SHM 8.0, and is compared to standard SHVC with two-layer quality scalability. Eleven test sequences were tested in the lowdelay P configuration, each with four bitrate points: BL QPs (25, 30, 35, 40) combined with an EL QP offset of -3 (which results in a BL bitrate of about half of the total bitrate). Similar to [11], we use a look-up table to store the centroid offset within the interval, which significantly reduces the complexity of the ET framework.

We conducted three sets of experiments to show the effectiveness of the proposed ET framework. In our first experiment, we evaluate the prediction gain purely from the blocks that have valid ET prediction in EL (i.e., these blocks have valid base layer interval information available). As shown in Table 1, the ET prediction framework provides an average 2.47dB gain in prediction (equivalent to 45% reduction in prediction error). However, in practice, only 3% to 15% of the blocks (depending on the bitrate) have a valid interval information from base layer, which largely dilutes the overall gain. In our second experiment, we compare the overall RD performance of the SHVC with ET framework and the standard SHVC, and get an average of 3.3% reduction in bdrate [18], as shown in the “Standard ET-SHVC” column of Table 2. This dilution also suggests a future research direction of jointly optimizing BL and EL, where interval information is introduced in BL so as to benefit the ET prediction in EL.

To show that we have effectively tackled the challenges due to the quadtree structured partitioning, hybrid transform, and RDOQ, we conducted a third experiment where ET framework is applied on a constrained SHVC where none of the above tools are enabled. In this third experiment, all the block sizes are forced to be 8 × 8, DCT is used as the only transform and RDOQ is disabled. The performance of the original ET framework in this limited version of SHVC over the baseline is shown in the “Constrained ET-SHVC” column of Table 2, with an average of 3.18% reduction in bdrate. We achieve a similar and consistent gain in our proposed framework with all the tools enabled, which proves its effectiveness in practice.

Table 1. Prediction gains for blocks with valid ET prediction in EL

	Prediction Gain
BQMall (480p)	2.67 dB
BasketballDrill (480p)	1.69 dB
Keiba (480p)	3.00 dB
FourPeople (720p)	2.07 dB
Johnny (720p)	0.75 dB
Vidyo1 (720p)	1.29 dB
Cactus (1080p)	2.77 dB
BasketballDrive (1080p)	2.89 dB
BQTerrace (1080p)	1.78 dB
Kimono (1080p)	5.05 dB
ParkScene (1080p)	3.20 dB
AVERAGE	**2.47 dB**

Table 2. Overall bitrate reduction of the ET framework

	Standard ET-SHVC	Constrained ET-SHVC
BQMall (480p)	3.43%	3.99%
BasketballDrill (480p)	4.21%	2.43%
Keiba (480p)	2.13%	0.29%
FourPeople (720p)	3.88%	4.83%
Johnny (720p)	2.92%	3.64%
Vidyo1 (720p)	3.23%	4.43%
Cactus (1080p)	4.41%	3.92%
BasketballDrive (1080p)	2.84%	1.41%
BQTerrace (1080p)	2.45%	4.13%
Kimono (1080p)	3.12%	1.12%
ParkScene (1080p)	3.94%	4.78%
AVERAGE	**3.32%**	**3.18%**

5. CONCLUSION

This paper generalizes the ET framework to manage the mismatch in quadtree structured partitioning at different layers, by exploiting the linearity property of estimations to convert information between different partitions. The parameter of the transform coefficient distribution is separately trained for different types of transform, block sizes and QPs. And a more robust way of exploiting the BL quantization interval information is proposed to avoid erroneous information due to the RDOQ adjustment. Experimental results demonstrate the effectiveness of the proposed technique with consistent gains over standard SHVC. Future research directions include the joint optimization of BL and EL, and further exploitation of BL partition information.

6. REFERENCES

[1] V. Jacobson, R. Frederick, S. Casner, and H. Schulzrinne, “RTP: A transport protocol for real-time applications,” *RFC 1889*, Jan. 1996.

[2] H. Schwarz, D. Marpe, and T. Wiegand, “Overview of the

scalable video coding extension of the H. 264/AVC standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 17, no. 9, pp. 1103–1120, 2007.

[3] J. M. Boyce, Y. Ye, J. Chen, and A. K. Ramasubramonian, "Overview of SHVC: scalable extensions of the high efficiency video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 1, pp. 20–34, 2016.

[4] M. L. Comer, "A new approach to motion compensation in spatially scalable video coding," in *Electronic Imaging*. International Society for Optics and Photonics, 2006, pp. 60770L–60770L.

[5] R. Zhang and M. L. Comer, "Subband motion compensation for spatially scalable video coding," in *Proceedings of the SPIE Conference on Visual Communications and Image Processing*, 2007, vol. 6508, p. 65082v.

[6] R. Xiong, J. Xu, and F. Wu, "In-scale motion compensation for spatially scalable video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 2, pp. 145–158, 2008.

[7] X. Li, J. Chen, K. Rapaka, and M. Karczewicz, "Generalized inter-layer residual prediction for scalable extension of HEVC," in *IEEE International Conference on Image Processing (ICIP)*, 2013, pp. 1559–1562.

[8] R. Zhang and M. L. Comer, "Efficient inter-layer motion compensation for spatially scalable video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 10, pp. 1325–1334, 2008.

[9] T. K. Tan, K. K. Pang, and K. N. Ngan, "A frequency scalable coding scheme employing pyramid and subband techniques," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 4, no. 2, pp. 203–207, 1994.

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

[11] J. Han, V. Melkote, and K. Rose, "An estimation-theoretic framework for spatially scalable video coding," *IEEE Transactions on Image Processing*, vol. 23, no. 8, pp. 3684–3697, 2014.

[12] J. Han, V. Melkote, and K. Rose, "An estimation-theoretic approach to spatially scalable video coding," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 817–820.

[13] G. J. Sullivan, J. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[14] V. Lorcy, Philippe P., Biatek T., Zhao X., Seregin V., and Karczewicz M., "EE2: adaptive primary transform improvement," *Doc. JVET-D0065 ITU-T SG 16 WP3, Chengdu, CN, 15-21 Oct 2016*.

[15] G. J. Sullivan, "Efficient scalar quantization of exponential and laplacian random variables," *IEEE Transactions on Information Theory*, vol. 42, no. 5, pp. 1365–1374, 1996.

[16] J. W. Kang and C. S. Kim, "On DCT coefficient distribution in video coding using quad-tree structured partition," in *Asia-Pacific Signal and Information Processing Association, Annual Summit and Conference (APSIPA)*. IEEE, 2014, pp. 1–4.

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

[18] G. Bjontegaard, "Calcuation of average PSNR differences between RD-curves," *Doc. VCEG-M33 ITU-T Q6/16, Austin, TX, USA, 2-4 April 2001*.