ESTIMATION-THEORETIC APPROACH TO DELAYED PREDICTION IN 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

Scalable video coding (SVC) employs inter-frame prediction at the base and/or the enhancement layers. Since the base layer can be encoded/decoded independent of the enhancement layers, we consider here the potential gains when prediction at the enhancement layers is delayed to accumulate and incorporate additional future information from the base layer. We build on two basic estimation-theoretic (ET) approaches developed by our group: an ET approach for enhancement layer prediction that optimally combines current base layer with prior enhancement layer information, and our recent ET approach for delayed decoding. The proposed technique fully exploits all the available information from the base layer, including any future frame information, and past enhancement layer information. It achieves considerable gains over zero-delay techniques including both standard SVC, and SVC with optimal ET prediction (but with zero encoding delay).

***Index Terms*—** estimation-theoretic, prediction, scalable video coding, quality scalability, delayed encoding, delayed decoding

1. INTRODUCTION

Advances in video coding technology and novel network infrastructures enable video applications such as multi-cast that cater to multiple receivers over diverse communication channels which support different bit-rates. Scalable video coding (SVC), where the video is encoded in multiple layers, some of which may be dropped to adjust the transmission rate, is an attractive solution for these applications [1]. It eliminates the need for redundant versions of the video, coded at different bit-rates, to be stored on the server, and allows rate adjustment decisions to be made at intermediate network nodes.

In general, SVC employs inter-frame prediction to reduce temporal redundancies, at the base as well as enhancement layers. Throughout this paper, for exposition simplicity, we consider only inter-prediction (P-frames) and a two-layered bit-stream, although the concepts are extensible to more layers and other types of temporal prediction, such as bidirectional prediction (B-frames). The base layer consists of regular, single layer video coding, and is decodable independent of the enhancement layer. The enhancement layer coding of a video frame may, in general, take into account all base layer information up to the current frame, and prior enhancement layer information. Standard approaches that target SNR scalability (the focus of this paper) usually obtain the enhancement layer prediction as a linear combination of prior enhancement layer reconstruction with the current base layer prediction error (residual) reconstruction (in the so called single loop design), or reconstructed base layer pixels (in multi loop design) [1]. In contrast, a non-linear but *optimal* enhancement layer prediction approach was proposed in [2], which efficiently combines base layer quantization information with motion compensated, prior enhancement layer reconstruction, in an estimation-theoretic (ET) framework. This ET approach substantially outperformed the standard linear combination.

Here we take this ET framework a step further. Since the base layer is encoded (and decoded) independent of enhancement layers, encoding the enhancement layer for the current frame could also be accomplished after base layer coding of future frames. Can such enhancement layer encoding delay, and the additional accumulated future base layer information, improve the enhancement layer prediction for the current frame? We propose here an ET framework to optimally combine not only prior enhancement layer reconstructions, and current base layer information, but also future information, to further improve the enhancement layer estimate (i.e., prediction) for the current frame. Such a framework builds on the optimal prediction method in [2], as well as a more recent ET framework for optimal delayed decoding of predictively encoded signals [3] and its extension to video coding [4]. The latter work exploits the fact that, although at high bit-rates motion compensated prediction results in encoding of only the innovation (or non-redundant information) in a video frame, at low bit-rates the prediction errors are indeed correlated across frames, so that accumulating information from future frames, at the cost of decoding delay, can aid the reconstruction of the current frame. The ET delayed enhancement layer encoding approach proposed herein will be shown by experiments to provide substantial gains as compared to the non-delayed ET method in [2], and standard SVC.

2. SCALABLE PREDICTIVE VIDEO CODING

We provide here background information about SVC as implemented in the JSVM reference, and the ET approach in [2] for scalable video coding with optimal prediction at the enhancement layer.

2.1. The Standard Approach

In the H.264/AVC SVC extension framework [1] the base layer is coded as usual with inter-prediction. Every block of the current frame is predicted from prior reconstructed frames via motion compensation, and the residual block transformed using the discrete cosine transform (DCT). The DCT coefficients are quantized and encoded into the base layer. In coding the enhancement layer, the standard starts with motion compensated prediction from a prior enhancement layer reconstructed block. The pixel domain residual is calculated and transformed. Then the standard adaptively switches between simply quantizing and coding this residual (i.e., no base

This work was supported in part by the NSF under grant CCF-0917230, and Qualcomm Inc.

layer information is used), or further subtracting from this residual the base layer *prediction error reconstruction* to generate a second level residual, and then quantizing it. This coding scheme is referred to in the video coding literature as a single-loop design, in which the decoder that targets a specific layer does not need to buffer its base layer reconstructed frames. Earlier standards such as H.263 (Annex O) and MPEG-4 (part 2) employed as enhancement layer prediction, a weighted combination of the base layer reconstruction and enhancement layer motion compensation, or adaptively switched between the two in a rate-distortion sense. Such methods that require to buffer base layer reconstructions of preceding frames are referred to as multi-loop designs. It has been shown in [5] that multi-loop design offers better coding performance than the single-loop approach, but the gain is minimal. Although we restrict our comparison here to the current standard (SVC in single-loop design), we note that in [2] substantial gains over even the multi-loop design were obtained by use of the ET optimal prediction approach.

2.2. Optimal Enhancement Layer Prediction in SVC

Note that the ad hoc linear combination of base layer residual, and prior enhancement layer reconstructions employed by the standard SVC cannot guarantee the optimal prediction for the enhancement layer. In [2] an ET approach for optimal prediction at the enhancement layer was proposed which we briefly describe here.

Let x_n denote the DCT coefficient of a particular frequency in a 4×4 block of the current frame. Since the DCT is unitary, motion compensated prediction followed by residual transformation is also equivalent to transforming both, the original block and its motion compensated reference, to the DCT domain and then subtracting the transform coefficients of the latter from that of the former. Let $\hat{x}^b_{n-1}$ denote the reconstructed DCT coefficient of the same frequency as x_n, but of the base layer motion compensated reference. Thus the operation of the standard base layer encoder is equivalent to 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. Thus, $x_n \in \mathcal{I}^b_n = [\hat{x}^b_{n-1} + a_n, \hat{x}^b_{n-1} + b_n)$ captures all the information on x_n provided by the base layer.

When coding the enhancement layer of x_n, the encoder can access enhancement layer information of previous frames too. In other words, it has access to the transform coefficient $\hat{x}^e_{n-1}$ of the motion compensated reference obtained using all information up to the enhancement layer. In [2], an approach is proposed to combine the prior enhancement layer information $\hat{x}^e_{n-1}$, with the base layer interval $\mathcal{I}^b_n$ to obtain the optimal enhancement layer prediction for the coefficient x_n. Note that although the information $\hat{x}^e_{n-1}$ can be equivalently viewed in terms of the corresponding pixel domain inverse transform, the quantization interval $\mathcal{I}^b_n$ cannot be represented easily in the pixel domain.

Traditionally, 4×4 blocks of pixels along the same motion trajectory in consecutive frames of the video are modeled as an autoregressive (AR) process, and motion compensation employed to align these pixel blocks, and pixel domain subtraction (prediction) removes temporal redundancies. Instead in [2], the equivalent viewpoint (due to the unitarity of the transform) that corresponding blocks of DCT coefficients form an AR process is adopted. Thus x_n (at any frequency) and the corresponding transform coefficient x_{n-1} in its *uncoded* motion compensated reference conform to the first order AR model:

$$x_n = \rho x_{n-1} + z_n \tag{1}$$

where z_n are independent and identically distributed (i.i.d) innovations of the process. The innovation probability density function (pdf) denoted by $p_Z(z)$ is assumed to be Laplacian in keeping with prior work [6], i.e.,

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

The parameter λ is itself frequency dependent. Since the pixel domain motion compensation is simply subtracted from the current block, in the following we assume that the correlation coefficient $\rho = 1$ at all frequencies. The above transform domain AR process perspective provides the advantage that the motion compensation $\hat{x}^e_{n-1}$, and the quantization interval $\mathcal{I}^b_n$, can now be combined in an ET framework.

Assuming that $\hat{x}^e_{n-1} \approx x_{n-1}$, the pdf of x_n given $\hat{x}^e_{n-1}$ is simply

$$p(x_n|\hat{x}^e_{n-1}) \approx p_Z(x_n - \hat{x}^e_{n-1}) \tag{3}$$

In the absence of additional base layer information, the best prediction of x_n would just be $\hat{x}^e_{n-1}$, the minimum mean squared error (MMSE) estimate with respect to above pdf. But the base layer indicates that $x_n \in \mathcal{I}^b_n$, given which the conditional pdf of x_n is

$$p(x_n|\hat{x}^e_{n-1}, \mathcal{I}^b_n) \approx \begin{cases} \frac{p_Z(x_n - \hat{x}^e_{n-1})}{\int_{\mathcal{I}^b_n} p_Z(x_n - \hat{x}^e_{n-1}) dx_n} & x_n \in \mathcal{I}^b_n \\ 0 & else \end{cases} \tag{4}$$

Note that the above is equivalent to centering the Laplacian pdf at $\hat{x}^e_{n-1}$, retaining it (cutting it) only in the interval $\mathcal{I}^b_n$ (a very non-linear operation), and then re-normalizing the new function. The optimal predictor $\tilde{x}^e_n$ at the enhancement layer is given by [2]

$$\tilde{x}^e_n = E[x_n|\hat{x}^e_{n-1}, \mathcal{I}^b_n] \tag{5}$$

the centroid of the above pdf in the interval $\mathcal{I}^b_n$. Note that the Laplacian innovations imply that a closed form of the above expectation can be derived. The residual $x_n - \tilde{x}^e_n$ is then quantized and encoded in the enhancement layer.

3. DELAYED DECODING OF PREDICTIVELY ENCODED VIDEO

Now consider a single layer video codec, say just the base layer described in Sec. 2.2. We follow the same notation here. At high rates the prediction error $x_n - \hat{x}^b_{n-1}$ is almost equal to the innovation z_n, which are i.i.d. Hence the indices $\{i^b_n\}$ obtained by quantization of the prediction errors are themselves almost independent across frames. In this case future indices $\{i^b_l\}_{l>n}$ provide no additional information on the current sample x_n. But in practice bit-rates are limited and such approximations do not hold, in which case future indices do contain information about the current sample x_n which can be appropriately exploited by a *non-causal* decoder. An optimal delayed decoding algorithm was proposed in [3] for generic AR sources encoded by differential pulse code modulation, which recursively calculated the conditional pdf of x_n given all past and available future information (up to a fixed delay), and obtained the reconstruction as the appropriate conditional expectation. In [4], we adapted the ET approach in [3] for delayed decoding of predictively encoded video, via suitable approximations. We describe here this approximate ET delayed decoder.

Consider the pdf of x_n conditioned on the prediction $\hat{x}^b_{n-1}$ and the current index i^b_n, *as well as one future index* i^b_{n+1}. Note that these indices provide the information that $x_n \in \mathcal{I}^b_n$, and $x_{n+1} \in \mathcal{I}^b_{n+1}$.

$$p(x_n|\hat{x}^b_{n-1}, \mathcal{I}^b_n, \mathcal{I}^b_{n+1}) = \frac{p(x_n|\hat{x}^b_{n-1}, \mathcal{I}^b_n) p(\mathcal{I}^b_{n+1}|x_n)}{\int p(x_n|\hat{x}^b_{n-1}, \mathcal{I}^b_n) p(\mathcal{I}^b_{n+1}|x_n) dx_n} \tag{6}$$

which we obtain by Bayes' rule and the Markov property of the process (1): given x_n the probability of the future event $\mathcal{I}^b_{n+1}$ is independent of any other past information. In the above equation the *zero-delay pdf* of x_n, $p(x_n|\hat{x}^b_{n-1}, \mathcal{I}^b_n)$, is weighed by the probability $p(\mathcal{I}^b_{n+1}|x_n)$ of the known future outcome to obtain the *1-sample delayed pdf* on the LHS of (6), that incorporates all known information at the decoder up to a delay of 1 sample. Now,

$$p(\mathcal{I}^b_{n+1}|x_n) = \int_{\mathcal{I}^b_{n+1}} p_Z(x_{n+1} - x_n) dx_{n+1} \qquad (7)$$

$$\approx p_Z(\hat{x}^b_{n+1} - x_n)(b_{n+1} - a_{n+1}) \, . \qquad (8)$$

The above approximation is obtained by assuming that the integrand in (7) is almost a constant with value $p_Z(\hat{x}^b_{n+1} - x_n)$, which is indeed true at high bit-rates. The pdf $p(x_n|\hat{x}^b_{n-1}, \mathcal{I}^b_n)$ in (6) is well approximated by

$$p(x_n|\hat{x}^b_{n-1}, \mathcal{I}^b_n) \approx \begin{cases} \frac{p_Z(x_n - \hat{x}^b_{n-1})}{\int_{\mathcal{I}^b_n} p_Z(x_n - \hat{x}^b_{n-1}) dx_n} & x_n \in \mathcal{I}^b_n \\ 0 & else \end{cases} \qquad (9)$$

Note that the above is very similar in form to (4), but with $\hat{x}^e_{n-1}$ replaced by $\hat{x}^b_{n-1}$. Thus, by (6) and (8), we obtain:

$$p(x_n|\hat{x}^b_{n-1}, \mathcal{I}^b_n, \mathcal{I}^b_{n+1}) \approx \begin{cases} \frac{p_Z(x_n - \hat{x}^b_{n-1}) p_Z(\hat{x}^b_{n+1} - x_n)}{\int_{\mathcal{I}^b_n} p_Z(x_n - \hat{x}^b_{n-1}) p_Z(\hat{x}^b_{n+1} - x_n) dx_n} & x_n \in \mathcal{I}^b_n \\ 0 & else \end{cases} \qquad (10)$$

The optimal *1-sample delayed* reconstruction of x_n is now given by

$$\hat{\hat{x}}^b_n = E[x_n|\hat{x}^b_{n-1}, \mathcal{I}^b_n, \mathcal{I}^b_{n+1}] \qquad (11)$$

which is nothing but the expectation over the above pdf. We use the notation $\hat{\hat{x}}^b_n$ to distinguish from the regular non-delayed reconstruction $\hat{x}^b_n$.

The video decoder in [4] waits to collect future information and incorporates it into (10) to obtain the delayed reconstruction of the transform coefficients of each 4×4 on-grid block in the frame. For every such block of frame n it requires the corresponding reconstructed block in the previous frame, that is part of the AR process/motion trajectory. This past block is nothing but the regular motion compensated prediction. The past DCT coefficient $\hat{x}^b_{n-1}$ in (10) is just obtained by DCT of this pixel domain prediction. In addition to this information, (10) also needs the subsequent block, i.e., in frame $n+1$, of the motion trajectory. But there are no motion vectors for the current frame that point to such future blocks. Nevertheless, since the decoder incorporates a one frame delay, it has motion vectors of the subsequent frame $n + 1$ that map on-grid blocks of that frame to potentially off-grid motion compensation blocks in the current frame. The delayed video decoder in [4] determined, for every on-grid block of the current frame, one of such motion compensations of the frame $n + 1$ that best overlaps with it. It then reversed the corresponding motion vectors, and re-assigned them to the on-grid blocks of the current frame. These reversed motion vectors then point every on-grid block in the current frame to a 4×4 block in frame $n + 1$. The regular, zero-delay decoder already provides a coarse estimate of the pixels in these future blocks. These coarse reconstructions are then transformed to obtain the required $\hat{x}^b_{n+1}$ in (10). For more details of this motion trajectory construction see [4].

Note that for every frame in the video bit-stream, quantization indices i^b_n are available only for the on-grid blocks of the frame. Since future blocks obtained by the above reverse motion mapping can be potentially off-grid, the future index i^b_{n+1}, and hence the interval $\mathcal{I}^b_{n+1}$, need not be available at all. This is actually the reason why the approximation in (8) is necessitated. As mentioned previously, the required future information $\hat{x}^b_{n+1}$ is now obtained via DCT of future, reconstructed blocks.

4. ENCODING THE ENHANCEMENT LAYER WITH DELAYED INFORMATION FROM THE BASE LAYER

We now combine the ET approach in Sec. 2.2 for optimal (zero-delay) enhancement layer prediction with ideas borrowed from the ET delayed (single layer) decoding approach in Sec. 3, to obtain a more accurate prediction for the enhancement layer at both encoder and decoder. Note that the base layer can be encoded/decoded independent of the enhancement layer. Hence we consider encoding the enhancement layer for frame n after encoding the base layer of frame $n + 1$. The motion vectors of frame $n + 1$, along with the base layer pixel domain reconstruction of that frame, can now indicate for every DCT coefficient x_n in the current frame, a future base layer reconstruction $\hat{x}^b_{n+1}$. Thus in contrast to the situation in Sec. 2.2, the available information to encode the enhancement layer for the DCT coefficient x_n includes not just $\hat{x}^e_{n-1}$, and $\mathcal{I}^b_n$ (see (4)), but also $\hat{x}^b_{n+1}$. Therefore the conditional pdf of x_n given all the above information is just,

$$p(x_n|\hat{x}^e_{n-1}, \mathcal{I}^b_n, \hat{x}^b_{n+1}) \approx \begin{cases} \frac{p_Z(x_n - \hat{x}^e_{n-1}) p_Z(\hat{x}^b_{n+1} - x_n)}{\int_{\mathcal{I}^b_n} p_Z(x_n - \hat{x}^e_{n-1}) p_Z(\hat{x}^b_{n+1} - x_n) dx_n} & x_n \in \mathcal{I}^b_n \\ 0 & else \end{cases} \qquad (12)$$

The above is equivalent to substitution of $\hat{x}^e_{n-1}$ in place of $\hat{x}^b_{n-1}$ in (10). The former, enhancement layer reconstruction is a better estimate of x_{n-1} than the latter, base layer estimate. Thus the prediction for the enhancement layer of x_n is now the expectation over the above pdf:

$$\tilde{\tilde{x}}^e_n = E[x_n|\hat{x}^e_{n-1}, \mathcal{I}^b_n, \hat{x}^b_{n+1}] \qquad (13)$$

Given this *delayed* prediction we obtain the prediction error $x_n - \tilde{\tilde{x}}^e_n$, quantize it, and correspondingly generate the enhancement layer reconstruction $\hat{x}^e_n$. Since the prediction now uses more information (i.e., from one future base layer index), it is expected to be more accurate than $\tilde{x}^e_n$ of (5). Indeed the latter can be viewed as a special case of the former, when no future information is available. The improved estimate $\tilde{\tilde{x}}^e_n$ should reduce the variance of the prediction error, thus increasing the prediction gain, and resulting in bit-savings. The above scheme induces an end-to-end enhancement layer latency of one frame.

In summary, the above delayed enhancement layer encoding procedure entails the following steps to encode the enhancement layer for frame n:

- Encode base layer of frame n and frame $n+1$, and obtain the pixel domain reconstruction of frame $n + 1$.
- Invert motion vectors of frame $n + 1$ to map every on-grid block in frame n to a future reconstructed block.
- Apply DCT on these future reconstructed blocks, as well as on the motion compensated predictions that employ past enhancement layer reconstructions, to obtain $\hat{x}^b_{n+1}$ and $\hat{x}^e_{n-1}$, respectively, corresponding to every DCT coefficient of all on-grid blocks in frame n.

- Use the base layer information of frame n to obtain the intervals $\mathcal{I}_n^b$.
- Combine the above in (12), and obtain the enhancement layer prediction $\tilde{\tilde{x}}_n^e$ via (13).
- Quantize and encode the prediction error $x_n - \tilde{\tilde{x}}_n^e$.
- Reconstruct these coefficients, apply inverse DCT, and deblock, to obtain the enhancement layer pixel domain reconstruction of frame n

5. SIMULATION RESULTS

We implemented the proposed delayed ET enhancement layer prediction scheme of Sec. 4 in the JSVM 9.18 reference software framework. We henceforth refer to this implementation as ET-SVC-1D, to explicitly indicate that encoding of the enhancement layer is delayed to incorporate base layer information from one future frame. The implementation of the corresponding non-delayed ET prediction scheme of Sec. 2.2 in the JSVM 9.18 framework will be addressed as ET-SVC-0D. We compare in Fig. 1 the performance of ET-SVC-1D and ET-SVC-0D against standard H.264/SVC (Sec. 2.1), in the context of two-layered scalable coding of the $coastguard$ sequence at QCIF resolution. The difference between base and enhancement layer QP values is 2 at all points on the graph. Note that all three methods use the same base layer coding method (H.264/AVC compatible). They differ only in their enhancement layer prediction. The PSNR corresponds to that of the enhancement layer reconstruction. The proposed ET-SVC-1D provides gains as high as 1.4 dB compared to ET-SVC-0D, which itself substantially outperforms the standard H.264/SVC. The performance of conventional (non-delayed) single layer H.264/AVC is also provided as a benchmark. This benchmark represents an upper bound on the performance of all (non-delayed) scalable codecs in the H.264/SVC framework.

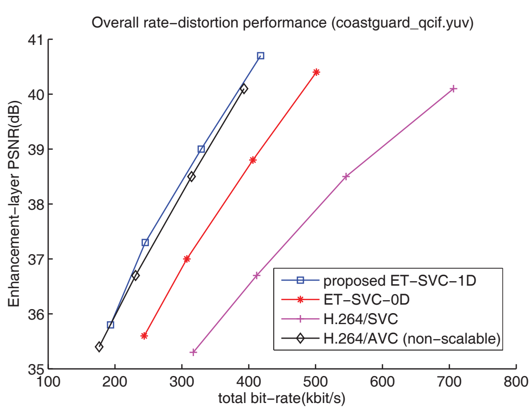

Fig. 1. Comparison of the performance of H.264/SVC, ET-SVC-0D and the proposed ET-SVC-1D on $coastguard$ at QCIF resolution. Also included is the performance of conventional non-delayed, non-scalable, H.264/AVC

We now compare the performance of ET-SVC-0D and ET-SVC-1D when the base layer resolution is fixed (i.e., base layer bit-rate is fixed), and the enhancement layer bit-rate is varied. Table. 1 compares their performance on $mobile$ at CIF resolution with base layer QP fixed at 30 (base layer bit-rate and PSNR, are respectively, 1708 kb/s, and 33.4 dB). As the enhancement layer QP is decreased (higher enhancement layer rate), the gains due to ET-SVC-1D diminish as expected. As noted in Sec. 3, delayed information is more useful at lower bit-rates, as the prediction errors across frames are then more correlated.

	ET-SVC-0D		proposed ET-SVC-1D	
QP	PSNR	Enhancement layer bit-rate(kb/s)	PSNR	Enhancement layer bit-rate(kb/s)
28	35.3	707.7	35.5	441.8
27	36.2	963.0	36.1	669.6
26	36.8	1112.3	36.7	862.7
24	38.6	1572.6	38.6	1411.4

Table 1. Comparison of the performance of ET-SVC-0D and ET-SVC-1D on *mobile* at CIF resolution, when base layer resolution is fixed, and enhancement layer rate (QP) is varied

6. CONCLUSION

We propose in this paper an estimation-theoretic approach for delayed encoding at the enhancement layer in scalable predictive video coding, which builds on prior work in our research group on ET approaches for optimal enhancement layer prediction in SVC, and optimal delayed decoding of predictively encoded video sequences. The motivation for this work stems from the realization that as the base layer is independently decodable, it is possible to exploit future base layer information to improve the enhancement layer prediction of the current frame. Considerable and consistent gains are obtained by encoding the enhancement layer with the proposed delayed ET method, in comparison to both standard SVC, and the zero delay ET-SVC method.

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

[3] V. Melkote and K. Rose, "Optimal delayed decoding of predictively encoded sources," in *Proc. IEEE ICASSP*, Mar 2010.

[4] J. Han, V. Melkote, and K. Rose, "Estimation-theoretic delayed decoding of predictively encoded video sequences," in *Proc. IEEE DCC*, Mar 2010.

[5] H. Schwarz, T. Hinz, D. Marpe, and T. Wiegand, "Constrained inter-layer prediction for single-loop decoding in spatial scalability," *Proc. IEEE ICIP*, pp. 870–873, September 2005.

[6] F. Bellifemine, A. Capellino, A. Chimienti, R. Picco, and R. Ponti, "Statistical analysis of the 2D-DCT coefficients of the differential signal for images," *Sig. Proc.: Img. Comm.*, pp. 477–488, May 1992.