# TRANSFORM-DOMAIN TEMPORAL PREDICTION IN VIDEO CODING: EXPLOITING CORRELATION VARIATION ACROSS COEFFICIENTS

*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

Temporal prediction in standard video coding is performed in the spatial domain, where each pixel is predicted from a motion-compensated reconstructed pixel in a prior frame. This paper is premised on the realization that such standard prediction treats each pixel independently and ignores underlying spatial correlations, while transform-domain prediction would eliminate much of the spatial correlation before signal components (transform coefficients) are independently predicted. Moreover, the true temporal correlations emerge after signal decomposition, and vary considerably from low to high frequency components. This precise nature of the temporal dependencies is entirely masked in spatial domain prediction by the high temporal correlation coefficient ($\rho \approx 1$) imposed on all pixels by the dominant low frequency components. We derive optimal transform-domain per-coefficient predictors for three main settings: basic inter-frame prediction; bi-directional prediction; and enhancement-layer prediction in scalable coding. Experimental results provide evidence for substantial performance gains in all settings.

***Index Terms***— Inter-frame prediction, correlation coefficient, bidirectional prediction, scalable video coding

## 1. INTRODUCTION

Modern video coding methods, such as H.264, exploit the inherent temporal correlation in the video sequence via inter-frame prediction [1]. Motion compensation is optimized per block in the current frame, via motion search over previously encoded frames available in the buffer, and involves *pixel* domain block matching. The resulting motion compensated block is subtracted directly from the original to produce the residual, which is then subjected to spatial transformation, typically by the discrete cosine transform (DCT), and the transform coefficients are quantized and coded. There is a considerable volume of prior research on accurate motion compensation with focus on various issues including the use of a long-term buffer [2], overlapped filter [3], variable size partition [4], etc.

The application of motion compensated prediction assumes that blocks of pixels along a motion trajectory form an autoregressive (AR) sequence. The reason for the direct subtraction of the prediction from the current block is that the temporal correlation coefficient, as calculated between pixel blocks, typically approaches one, $\rho \approx 1$. An alternative viewpoint that transform coefficients of the blocks form a scalar AR process, at each spatial frequency, was adopted in both [5] where we proposed an estimation-theoretic (ET) approach to delayed video decoding, and [6] where an ET approach for scalable predictive coding was proposed. Such a model is congruent with the pixel domain AR model due to the unitarity of the spatial transform applied. But we note that in [5], or [6], such a viewpoint was necessitated primarily because quantization interval information exploited by the ET framework was readily available only in the transform domain. The innovations of the scalar AR process at each frequency was assumed to be Laplacian (see prior work on the statistics of transform coefficients of residual blocks after motion compensated prediction [7]). In both, [5] and [6], the temporal correlation coefficient of each AR process at different spatial frequencies was assumed to approach the value one, as is common practice for the pixel domain. But an analysis of the actual temporal correlations at each frequency (see Sec. 2) reveals that only in the case of DC, and certain lower frequencies, this approximation is valid, and that this correlation decreases noticeably at higher frequencies. The typical concentration of energy in low frequency transform coefficients results in their high temporal correlation dominating any pixel domain calculation of the temporal correlation coefficient, which leads to the common assumption of $\rho \approx 1$.

To exploit the non-uniformity in temporal correlation at different frequencies we consider in this paper application of motion compensated prediction in the DCT domain, with scaling of the predicted transform coefficients by the appropriate correlation coefficient. When applied to standard P- (inter) and B-frame (bidirectionally predicted) coding modes substantial improvements in coding performance is obtained. We then proceed to demonstrate its potential in scalable video coding (SVC), particularly in the case of quality (SNR) scalability. The ET approach for optimal enhancement layer prediction proposed in [6] is chosen as the framework for the SVC implementation. In contrast to the adaptive switched prediction adopted in the current SVC standard [8], the technique in [6] optimally combines both prior enhancement layer information, and current base layer information, thus significantly outperforming the standard. Since this ET approach explicitly required motion compensated prediction to be considered in the transform domain, this makes an ideal setting for exploiting the inherent variation in temporal correlations across transform coefficients. The ET approach is modified to account for this more accurate modeling of temporal correlations, and considerable performance gains are obtained.

## 2. STATISTICAL MODEL

Motion-compensated prediction is employed under the assumption that blocks along a motion trajectory form a temporal AR source. We instead consider the pair of transform (DCT) coefficients, denoted by $(x_n, x_{n-1})$, of the same frequency of an inter-coded block and its motion compensated reference, as two successive samples of a

This work was supported in part by Qualcomm Inc.

| | | | |
|---|---|---|---|
| 0.9998 | 0.9946 | 0.9916 | 0.9470 |
| 0.9893 | 0.9424 | 0.9068 | 0.8056 |
| 0.9807 | 0.9215 | 0.8696 | 0.7717 |
| 0.9680 | 0.9015 | 0.8309 | 0.7317 |

**Table 1**. Matrix of temporal correlation coefficients for the 16 DCT coefficients in $coastguard\_qcif.yuv$

| | | | |
|---|---|---|---|
| 27246 | 2454 | 1091 | 76 |
| 1533 | 233 | 102 | 23 |
| 890 | 170 | 72 | 18 |
| 340 | 79 | 30 | 8 |

**Table 2**. Matrix of variances of the 16 DCT coefficients in $coastguard\_qcif.yuv$

scalar AR process with

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

where the innovations $z_n$ are zero-mean, independent and identically distributed with the Laplacian probability density function (pdf) [7]:

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

The parameter $\lambda$ is itself frequency dependent. We run regular pixel domain motion search to get matched pairs of blocks between an (uncoded) frame and its (uncoded) preceding frame, and for multiple frame pairs. The transform block size is restricted to the $4 \times 4$ option available in H.264. The temporal correlation coefficient $\rho$ at each of the 16 frequencies can now be calculated, by averaging pairwise temporal correlations at the same frequency over all matched blocks. Provided in Table. 1 is the matrix of these 16 temporal correlation coefficients in the case of $coastguard\_qcif.yuv$. Note that the correlation is close to 1 for DC, but quite different otherwise. Table. 2 provides the variance of the transform coefficients at different frequencies. The DC component has a substantially higher variance than the rest, and hence its temporal correlation is the dominant component in any calculation of the temporal correlation in the pixel domain. Such characteristics are also exhibited by other video sequences.

We emphasize that the conventional AR model inherently assumes pixels of the blocks form *independent* scalar temporal AR processes, i.e., the model *completely ignores inter-pixel (spatial) correlation within each block during temporal (motion compensated) prediction*. However in the proposed model, spatial correlation is first largely eliminated via DCT, and then the DCT coefficients (which are almost uncorrelated) are modeled as a temporal AR process for prediction purposes.

It should be noted that the zero-mean innovations in (1) inherently imply that $x_n$ is itself zero-mean, whenever $|\rho| < 1$ (i.e., any non-zero means during initialization of the process are eventually damped down by $\rho$). It was indeed observed in the above experiment that the mean of DCT coefficients at any AC frequency was always nearly zero. The DC coefficient in general is not zero-mean because pixel values are always positive. Formally, one would need a correction constant term in the model of (1) but this correction term is negligible in practice since $\rho \approx 1$ in the DC case.

## 3. PREDICTION IN THE TRANSFORM DOMAIN

The variation in temporal correlation across frequencies, as observed in the preceding section, motivates performing motion-compensated prediction in the transform domain, where the prediction at each frequency is weighted by the appropriate correlation coefficient. In other words, instead of performing a transform on the residual pixel domain block, each block and its motion compensated reference are individually transformed, and after suitably weighting the transform coefficients of the latter, the residue is directly calculated in the frequency domain.

In case of P-frames, the optimal prediction for each frequency coefficient $x_n$ is simply

$$\tilde{x}_n^P = \rho \hat{x}_{n-1} \,, \tag{3}$$

where $\hat{x}_{n-1}$ is the corresponding frequency coefficient of the motion compensation, and $\rho$ is the temporal correlation coefficient appropriate to that frequency. Conventional pixel domain prediction is equivalent to employing $\rho = 1$ at all frequencies.

In bidirectional prediction (B-frames), let us consider $IPBPBP$ coding mode with single reference frame from each side for simplicity. We assume that the current block along with its two reference blocks (one from the past and one from the future) form a motion trajectory. We consider transform coefficients $x_{n-1}$, $x_n$, and $x_{n+1}$ from consecutive blocks on this motion trajectory, with their relation modeled by (1). In this case, when coding $x_n$ the reconstructions $\hat{x}_{n-1}$ and $\hat{x}_{n+1}$ are already available, and the optimal estimation of $x_n$ given $\hat{x}_{n-1}$ and $\hat{x}_{n+1}$ is the minimum mean squared error (MMSE) estimate

$$\tilde{x}_n^B = E[x_n|\hat{x}_{n-1}, \hat{x}_{n+1}] \tag{4}$$

where expectation is over the conditional pdf $p(x_n|\hat{x}_{n-1}, \hat{x}_{n+1})$. The conditional pdf of $x_n$ given the actual samples $x_{n-1}$ and $x_{n+1}$ is

$$p(x_n|x_{n-1}, x_{n+1}) = \frac{p(x_n|x_{n-1})p(x_{n+1}|x_n)}{\int p(x_n|x_{n-1})p(x_{n+1}|x_n)dx_n} \tag{5}$$

$$= \frac{p_Z(x_n - \rho x_{n-1})p_Z(x_{n+1} - \rho x_n)}{\int p_Z(x_n - \rho x_{n-1})p_Z(x_{n+1} - \rho x_n)dx_n} \tag{6}$$

The first of the above equalities is obtained by Bayes' rule and the Markov property of the process (1): given $x_n$, the pdf of $x_{n+1}$ is independent of any other information from the past, i.e., $x_{n-1}$. The second is the result of the innovation $z_n$ being independent of $x_{n-1}$. Unless otherwise specified all integrals are over the real line. Assuming that the reconstructions $\hat{x}_{n-1}$ and $\hat{x}_{n+1}$ are close to the corresponding actual sample values we obtain:

$$p(x_n|\hat{x}_{n-1}, \hat{x}_{n+1}) \approx \frac{p_Z(x_n - \rho\hat{x}_{n-1})p_Z(\hat{x}_{n+1} - \rho x_n)}{\int p_Z(x_n - \rho\hat{x}_{n-1})p_Z(\hat{x}_{n+1} - \rho x_n)dx_n} \tag{7}$$

The bidirectional prediction $\tilde{x}_n^B$ is now the expectation of $x_n$ over the above pdf, with the definition of $p_Z(\cdot)$ given by (2). We can show that when $\rho = 1$, it specializes to

$$\tilde{x}_n^B = \frac{\hat{x}_{n-1} + \hat{x}_{n+1}}{2} \tag{8}$$

corresponding to conventional biprediction in the pixel domain.

These simple modifications to the motion compensated prediction in P- and B- modes, respectively incur, one, or two additional DCTs for each 4x4 block of a frame, which introduces a moderate increment in computational complexity.

## 4. OPTIMAL PREDICTION IN SCALABLE CODING

We now consider the ET approach proposed in [6] for optimal prediction in SVC. Only inter prediction (P-frames) is considered here although the scheme can be extended to scalable bidirectional prediction too. Let $x_n$, $\hat{x}_n^b$ and $\hat{x}_n^e$ denote a particular transform coefficient, its base and enhancement layer reconstructions, respectively. Let $\tilde{x}_n^b$ and $\tilde{x}_n^e$ be the corresponding predictions at each layer. The optimal base layer prediction is just $\tilde{x}_n^b = \rho\hat{x}_{n-1}^b$, where $\hat{x}_{n-1}^b$ is the transform coefficient at the same frequency of the motion compensated reference obtained when using only base layer information to reconstruct previous frames. Conventional base layer prediction is equivalent to employing $\rho = 1$ at all frequencies. At the base layer, $\tilde{x}_n^b$ is subtracted from $x_n$ and the residue is quantized as index $i_n^b$. Let $[a_n, b_n)$ be the quantization interval associated with index $i_n^b$. Thus, the statement $x_n \in [\tilde{x}_n^b + a_n, \tilde{x}_n^b + b_n)$ captures *all the information* provided by the base layer on $x_n$.

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 $\hat{x}_{n-1}^e$ of the motion compensation obtained using all information up to the enhancement layer. In this case, assuming that $\hat{x}_{n-1}^e \approx x_{n-1}$, the pdf of $x_n$ given $\hat{x}_{n-1}^e$ is simply

$$p(x_n|\hat{x}_{n-1}^e) \approx p_Z(x_n - \rho\hat{x}_{n-1}^e) \tag{9}$$

In the absence of any base layer information, the best prediction of $x_n$ would just be $\rho\hat{x}_{n-1}^e$, the MMSE estimate with respect to above pdf. But the base layer indicates that $x_n \in [\tilde{x}_n^b + a_n, \tilde{x}_n^b + b_n)$, given which the conditional pdf of $x_n$ is

$$p(x_n|\hat{x}_{n-1}^e, x_n \in [\tilde{x}_n^b + a_n, \tilde{x}_n^b + b_n)) \approx \begin{cases} \frac{p_Z(x_n - \rho\hat{x}_{n-1}^e)}{\int_{\tilde{x}_n^b + a_n}^{\tilde{x}_n^b + b_n} p_Z(x_n - \rho\hat{x}_{n-1}^e)dx_n} & x_n \in [\tilde{x}_n^b + a_n, \tilde{x}_n^b + b_n) \\ 0 & else \end{cases} \tag{10}$$

Therefore the optimal predictor $\tilde{x}_n^e$ at the enhancement layer is given by [6]

$$\tilde{x}_n^e = E[x_n | x_n \in [\tilde{x}_n^b + a_n, \tilde{x}_n^b + b_n), \hat{x}_{n-1}^e] \tag{11}$$

the MMSE estimate of $x_n$ with respect to the pdf in (10). In [6] $\rho$ was assumed to be uniformly unity at all spatial frequencies. The objective here though is to demonstrate the additional gains due to the proposed temporal correlation coefficient scaling. Note that the Laplacian innovations imply that a closed form of the above expectation can be derived (see [6]). The residual $x_n - \tilde{x}_n^e$ is then quantized and encoded in the enhancement layer.

We contrast the above with the standard method in H.264/SVC. The standard starts off with the prior enhancement layer reconstructed block as the motion compensated prediction for the current pixels. The residual is calculated in the pixel domain, and then transformed. Note that this is equivalent to calculating $x_n - \hat{x}_{n-1}^e$, at all 16 frequencies. Then the standard adaptively switches between simply quantizing and coding this residual (i.e., no base layer information is used), or further subtracting from this residual the base layer prediction error reconstruction to generate a second level residual (equivalently $x_n - \hat{x}_{n-1}^e - \hat{e}_n^b$ where $\hat{e}_n^b$ is the reconstruction associated with the index $i_n^b$), and then quantizing it. This adaptive switching scheme is naturally sub-optimal compared to the optimal prediction in [6]. This coding scheme is called 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) employ as enhancement layer prediction, a weighted combination of the base layer reconstruction and enhancement layer motion compensation, or adaptively switch 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 [9] that multi-loop design offers better coding performance than single-loop, but the gain is minimal. Although we restrict our comparison here to the current standard (SVC in single-loop design), we note that in [6] substantial gains over even the multi-loop design were obtained by use of the ET optimal prediction approach.

## 5. SIMULATION RESULTS

We first applied the proposed transform domain prediction with correlation coefficient scaling to inter-mode blocks in the JM 16.0 reference software framework. The coding mode is set as $IPPP$ with regular inter-frame motion search in the pixel domain. The 16 correlation coefficients are calculated a priori, and are assumed unchanged for the entire sequence. Thus the additional side information required to be sent to the decoder is negligible. Since quantization settings have minimal influence on the motion search decision [10], at least at medium to high bit-rates, we assume that the same correlation coefficients apply at all QP values. The rate-distortion performance of the competing methods for the sequence $mobile\_cif$ is shown in Fig.1. Gains of about 1dB at different bit-rates are evident. Similar performance was observed with other video sequences, with different degrees of motion.

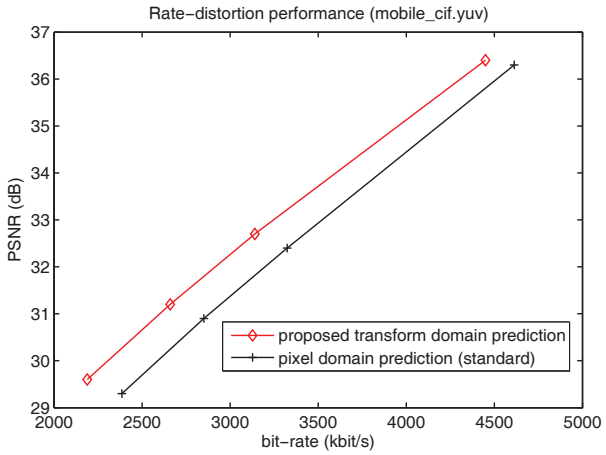


**Fig. 1**. Comparison of the performance of standard H.264 and proposed transform domain prediction with $IPPP$ coding of *mobile* at CIF resolution

Table. 2 demonstrates the gains obtained when instead of regular pixel domain bidirectional prediction, the proposed transform domain bidirectional prediction was employed. The encoder is configured as $IPBPBP$ with normal pixel domain motion search. Now the $\lambda$ values for the 16 frequencies (required in (7)) are also deduced from the original video, and fixed throughout the sequence. Each row in the table corresponds to a specific QP configuration pair for P- and B-frames, hence by both standard and proposed methods the PSNR is almost the same. Gains can be elicited in terms of the rate savings compared to standard biprediction.

We next implemented the optimal prediction for SVC [6] in the

| Standard biprediction | | | Proposed transform domain biprediction | | |
|---|---|---|---|---|---|
| PSNR | Total rate (kb/s) | B-frame rate (kb/s) | PSNR | Total rate (kb/s) | B-frame rate (kb/s) |
| 39.3 | 626.9 | 160.6 | 39.4 | 585.3 | 153.1 |
| 36.3 | 388.6 | 63.5 | 36.3 | 366.1 | 59.9 |
| 33.3 | 218.1 | 20.3 | 33.4 | 201.9 | 19.0 |
| 32.2 | 163.5 | 10.4 | 32.3 | 151.5 | 9.6 |

**Table 3**. Comparison of bit-rates due to standard and proposed methods for *Coastguard* at QCIF resolution when coded in $IPBPB$ mode

JSVM 9.18 framework. This corresponds to the ET approach described in Sec. 4 with no scaling (i.e., $\rho = 1$) of the prediction in the transform domain. We will refer to this implementation as ET-SVC. We denote by ET-SVC-RCM the proposed modification to ET-SVC that takes into account the refined correlation model (i.e., the true temporal correlations at different frequencies in the DCT domain). We restrict ET-SVC-RCM to employ this modification *only for enhancement layer prediction*, i.e., ET-SVC and ET-SVC-RCM share the same conventional ($\rho = 1$ at all frequencies) base layer. The coding performance of both approaches compared to H.264/SVC (also the same base layer) is shown in Fig. 2. ET-SVC provides substantial gains over standard SVC. These gains are further amplified by the proposed modification that takes into account the true temporal correlations.

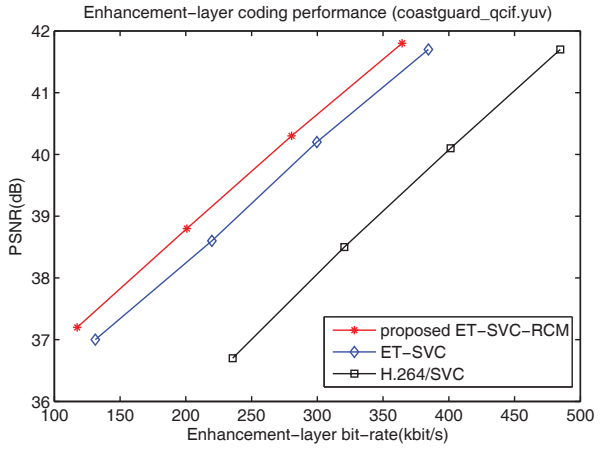


**Fig. 2**. Comparison of the performance of H.264/SVC, ET-SVC and the proposed ET-SVC-RCM on *coastguard* at QCIF resolution. The same base layer is shared by all codecs.

## 6. CONCLUSION

We propose here a transform domain motion compensated prediction approach for video coding that accounts for the true temporal correlation coefficients in the underlying AR process at different spatial frequencies. Such correlations are hidden from perspective of standard codecs that perform motion compensated prediction in the pixel domain, due to the pixel values being mostly inundated by the dominant DC component whose temporal correlation coefficient is close to unity. The proposed approach transforms the regular motion compensation, and scales the so obtained transform coefficients by the appropriate temporal correlations, and subsequently employs them as the prediction for the transform coefficients in the current block. Substantial gains are demonstrated by application of the proposed approach in inter-prediction, biprediction, and predictive SVC.

## 7. REFERENCES


[1] T. Wiegand, G. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circ. Sys. Video Tech.*, vol. 13, pp. 560–576, July 2003.

[2] T. Wiegand, X. Zhang, and B. Girod, "Long term memory motion compensated prediction for video coding," *IEEE Trans. Circ. Sys. Video Tech.*, vol. 9, pp. 70–84, Feb 1999.

[3] M. T. Orchard and G. J. Sullivan, "Overlapped block motion compensation an estimation-theoretic approach," *IEEE Trans. on Image Processing*, vol. 3, pp. 693–699, Sep 1994.

[4] M. H. Chan, Y. B. Yu, and A. G. Constantinides, "Variable size block matching motion compensation with applications to video coding," *IEEE Trans. Image Proc.*, vol. 137, pp. 205–212, Aug 1990.

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

[6] K. Rose and S. L. Regunathan, "Toward optimality in scalable predictive coding," *IEEE Trans. Image Proc.*, vol. 10, pp. 965–976, July 2001.

[7] 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.

[8] 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.

[9] 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.

[10] O. G. Guleryuz and M. T. Orchard, "Rate-distortion based temporal filtering for video compression," *Proc. IEEE DCC*, pp. 122–131, Jan 1996.