# TRANSFORM DOMAIN TEMPORAL PREDICTION WITH EXTENDED BLOCKS

*Shunyao Li, Tejaswi Nanjundaswamy, Kenneth Rose*

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

## ABSTRACT

Traditional temporal prediction relies on motion compensated pixel copying operations. Such per-pixel temporal prediction was shown in prior work to be sub-optimal since it ignores the underlying spatial correlation. Transform domain temporal prediction (TDTP) was thus proposed previously to first achieve spatial decorrelation within the block using DCT, then apply an optimal one-to-one prediction per frequency. However, for sub-pixel motion compensation, the low-pass filter used for interpolation interferes with TDTP. In this paper, we propose an extended block based TDTP, which: *i*) completely disentangles spatial and temporal correlations to fully account for the sub-pixel interpolation effect, and *ii)* fully exploits spatial correlations around reference block boundary. Experimental evidence is provided for substantial coding gains over standard HEVC.

***Index Terms—*** Temporal prediction, sub-pixel interpolation, spatial correlation, DCT, video coding

## 1. INTRODUCTION

Inter prediction [1] plays a critical role in video coders to exploit temporal dependencies. Current video coding standards, such as HEVC, employ pixel domain block matching to predict each block in a frame from a similar block in previously reconstructed frames. The prediction error is then transformed, quantized and sent to the decoder. In our past work [2, 3, 4], we have shown that this one-to-one pixel-copying approach is suboptimal, because it ignores the strong spatial correlation in pixel domain. In [2], we proposed transform domain temporal prediction (TDTP), where spatial decorrelation is (largely) achieved to allow one-to-one transform coefficient prediction. Moreover, TDTP also captured the variation in temporal correlation across frequencies, which is otherwise hidden in the pixel domain. In [4], an asymptotic closed-loop (ACL) design approach was proposed to overcome the design instability due to quantization error propagation.

A critical challenge for TDTP is its interference with sub-pixel interpolation filter. TDTP exploits the transform (DCT) domain temporal correlation for blocks along the motion trajectory, and forms a first-order AR process per frequency. As illustrated in Table 1, the temporal correlation tends to be close to 1 at low frequencies, and less than 1 at high frequencies, which implies high frequency components are scaled down more than low frequency components. This scaling for temporal prediction, is coincidentally similar to the frequency response of the low-pass filters used for sub-pixel interpolation, resulting in an interference between interpolation and prediction. In [4], the predictors were designed conditioned on the sub-pixel location (i.e., a particular combination of horizontal and vertical interpolation filters), and applied on the reference block obtained after interpolation. This approach effectively classifies the blocks based on the filters employed, but it does not disentangle the effect of interpolation filter on temporal prediction.

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 0.999 | 0.998 | 0.997 | 0.970 | 0.944 | 0.930 | 0.842 | 0.808 |
| 0.996 | 0.978 | 0.979 | 0.963 | 0.957 | 0.884 | 0.900 | 0.797 |
| 0.983 | 0.984 | 0.975 | 0.944 | 0.978 | 0.931 | 0.857 | 0.794 |
| 0.967 | 0.980 | 0.977 | 0.965 | 0.958 | 0.920 | 0.930 | 0.768 |
| 0.960 | 0.950 | 0.962 | 0.964 | 0.942 | 0.889 | 0.904 | 0.756 |
| 0.927 | 0.938 | 0.934 | 0.922 | 0.919 | 0.882 | 0.831 | 0.748 |
| 0.898 | 0.881 | 0.919 | 0.906 | 0.869 | 0.815 | 0.700 | 0.512 |
| 0.835 | 0.760 | 0.826 | 0.769 | 0.717 | 0.640 | 0.470 | 0.339 |

**Table 1**. Transform domain temporal correlation for 8x8 DCT coefficients for *mobile* sequence at QP=22

In this paper, we propose an extended block based TDTP (EB-TDTP) to fully disentangle the effect of interpolation, and appropriately account for the filters. The division of natural video data into blocks never strictly limits the spatial correlation to inside a block, and almost always, spatial correlation exists across a block boundary, as shown for an example in Fig. 1. Sub-pixel interpolation filter uses neighboring pixels outside the block boundary, and hence accounts for spatial correlation outside the block to some extent. However, employing the interpolation filter projects the pixels of a block and its neighbors along the boundary to a subspace. Transforming this block into frequency domain, will only achieve spatial decorrelation of information in this *subspace*. Thus employing transform domain interpolated block as a prediction reference for temporal prediction, does not disentangle the spatial and temporal correlation of the data completely, leading to suboptimality. In order to fully ex-

This work was supported in part by a gift from LG Electronics Inc.

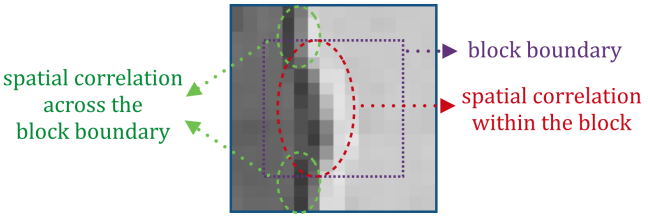


**Fig. 1**. An example of spatial correlation within a block and across its boundary

ploit the spatial correlation by completely separating it from temporal correlation, we propose employing TDTP with an extended block that is centered around the reference block and covers all the pixels required to generate the interpolated reference block. The sub-pixel interpolation filter is employed after this extended block is scaled for temporal prediction in transform domain. The prediction coefficients, conditioned on the sub-pixel location, are designed to minimize the final prediction error in the current block, using a two loop ACL design approach similar to [4]. Note that the EB-TDTP is beneficial even for motion compensated references at full-pixel locations where interpolation filters are not applied, as it allows fully exploiting spatial correlations around reference block boundary. Evaluation results demonstrate the proposed technique's reasonable gains over regular TDTP, which overall results in substantial performance improvements over standard HEVC. Our proposed disentanglement of TDTP and sub-pixel interpolation, also paves a path for future research on joint optimization of these modules for further performance improvements.

Other related work includes multi-tap pixel-domain filtering approaches [5, 6, 7] to account for spatial correlation in inter prediction. However, these techniques neglect the variation in temporal correlation across frequencies. In [8, 9], motion-compensated three-dimensional subband coding approaches is proposed to overcome similar challenges, but is limited by complexity and delay overhead. In [10], a sub-pixel interpolation in DCT domain is proposed, but this neglects the spatial correlation across block boundaries.

## 2. BACKGROUND

We assume the DCT coefficients of blocks along a motion trajectory form a first-order AR process per frequency. Let's denote by $x_n$, the DCT coefficients at a particular frequency, of blocks along a motion trajectory. The AR process can now be given as,

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

where, $\rho$ is the transform domain correlation coefficient, and $z_n$ is the innovation. The optimal TDTP predictor for each

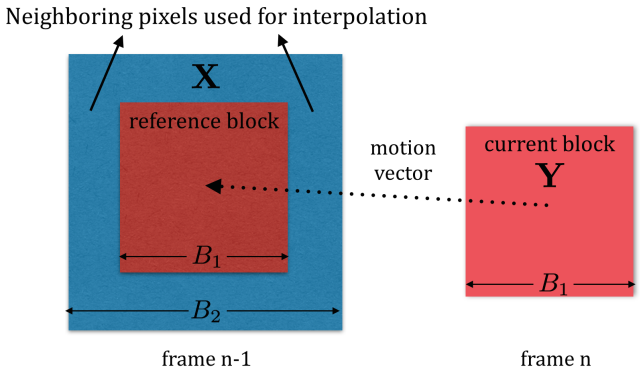


**Fig. 2**. Extended reference block with neighboring pixels

frequency coefficient is,

$$\tilde{x}_n = \rho \hat{x}_{n-1}, \tag{2}$$

where $\hat{x}_{n-1}$ is the reconstructed DCT coefficient. The optimal prediction coefficient $\rho$, which minimizes the mean squared prediction error, is given as,

$$\rho = \frac{E(x_n \hat{x}_{n-1})}{E(\hat{x}_{n-1}^2)}. \tag{3}$$

This forms the basic TDTP paradigm proposed in [2].

## 3. TRANSFORM DOMAIN TEMPORAL PREDICTION WITH EXTENDED BLOCKS

The basic TDTP described in Sec. 2 exploits the spatial correlation and temporal correlation between the motion compensated reference block and the current block. However, the spatial correlation with neighboring pixels outside the block is neglected. Here we propose EB-TDTP to incorporate all the relevant information by employing the DCT on a larger block, $\mathbf{X}$, which is centered around the motion compensated reference of current block, $\mathbf{Y}$, and covers an extended region in pixel domain, as shown in Fig. 2. The current block and the extended block are of size $B_1 \times B_1$ and $B_2 \times B_2$, respectively, with $B_2 > B_1$. $B_2$ is chosen such that all the pixels required for sub-pixel interpolation of reference block lies within $\mathbf{X}$. The vertical and horizontal interpolation filters in the matrix form are denoted as $\mathbf{F}_1$ and $\mathbf{F}_2$, which are of size $B_1 \times B_2$ and $B_2 \times B_1$, respectively. Therefore, the interpolated reference block is $\mathbf{F}_1\mathbf{X}\mathbf{F}_2$. The matrix operator of vertical 1D-DCT of size $b$ is denoted as $\mathbf{D}_b$. We also define the operator $\circ$ as element-by-element multiplication.

The traditional temporal prediction is the same as the interpolated reference block, i.e.,

$$\tilde{\mathbf{Y}} = \mathbf{F}_1\mathbf{X}\mathbf{F}_2. \tag{4}$$

The TDTP of $\mathbf{Y}$ as proposed in [2] can be formulated as,

$$\tilde{\mathbf{Y}} = \mathbf{D}'_{B_1}((\mathbf{D}_{B_1}\mathbf{F}_1\mathbf{X}\mathbf{F}_2\mathbf{D}'_{B_1}) \circ \mathbf{P}_{B_1})\mathbf{D}_{B_1}, \tag{5}$$

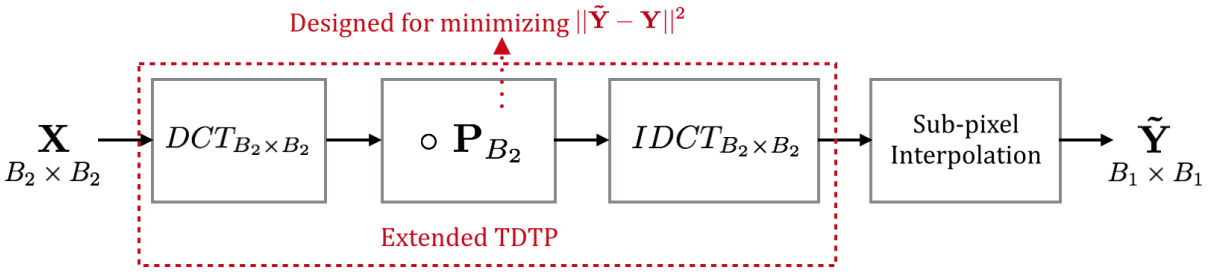


**Fig. 3**. Block diagram of the proposed TDTP framework employing extended blocks

where $\mathbf{P}_b$ is a $b \times b$ matrix with elements as the temporal prediction coefficients, $\rho$, corresponding to each frequency.

The block diagram of EB-TDTP is shown in Fig. 3. First the extended block is converted to DCT domain to achieve spatial decorrelation, then the prediction coefficients $\mathbf{P}_{B_2}$ are applied to the DCT domain prediction per frequency using (2), and finally the prediction is converted back to pixel domain for sub-pixel interpolation. The prediction coefficients $\mathbf{P}_{B_2}$ are designed to minimize the overall prediction error $J = ||\mathbf{Y} - \tilde{\mathbf{Y}}||^2$, to fully account for the sub-pixel interpolation effect. Therefore, the EB-TDTP can be formulated as,

$$\tilde{\mathbf{Y}} = \mathbf{F}_1 \mathbf{D}'_{B_2}((\mathbf{D}_{B_2}\mathbf{X}\mathbf{D}'_{B_2}) \circ \mathbf{P}_{B_2})\mathbf{D}_{B_2}\mathbf{F}_2. \quad (6)$$

To design the prediction coefficients $\mathbf{P}_{B_2}$, let us set $\mathbf{H}_1 = \mathbf{F}_1\mathbf{D}'_{B_2}$, $\mathbf{H}_2 = \mathbf{D}_{B_2}\mathbf{F}_2$, and $\mathbf{X}_T = \mathbf{D}_{B_2}\mathbf{X}\mathbf{D}'_{B_2}$, to simplify the cost as,

$$\begin{aligned} J &= ||\mathbf{Y} - \mathbf{H}_1(\mathbf{X}_T \circ \mathbf{P}_{B_2})\mathbf{H}_2||^2 \\ &\propto \sum_{m=1}^{B_1}\sum_{n=1}^{B_1}\Big[\mathbf{Y}(m,n) - \\ &\quad \sum_{i=1}^{B_2}\sum_{j=1}^{B_2}\mathbf{P}_{B_2}(i,j)\mathbf{X}_T(i,j)\mathbf{H}_1(m,i)\mathbf{H}_2(j,n)\Big]^2 \end{aligned} \quad (7)$$

This is equivalent to the least square estimation problem of minimizing $||\mathbf{A}\mathbf{p}_{B_2} - \mathbf{b}||^2$, where $\mathbf{p}_{B_2}$ is the vector form (of size $B_2^2 \times 1$) of $\mathbf{P}_{B_2}$, $\mathbf{A}$ and $\mathbf{b}$ (of size of $B_1^2 \times B_2^2$ and $B_1^2 \times 1$) are quantities derived from the training data as,

$$\mathbf{A}(k,l) = \mathbf{X}_T(i,j)\mathbf{H}_1(m,i)\mathbf{H}_2(j,n), \quad (8)$$

$$\mathbf{b}(k) = \mathbf{Y}(m,n), \quad (9)$$

where, $k = mB_1 + n$ $(m,n = 0 \ldots B_1 - 1)$, and $l = iB_2 + j$ $(i,j = 0 \ldots B_2 - 1)$. The optimal solution for prediction coefficients is given as,

$$\mathbf{p}_{B_2} = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b}, \quad (10)$$

$$\mathbf{P}_{B_2}(i,j) = \rho_{i,j} = \mathbf{p}_{B_2}(l). \quad (11)$$

Estimating these coefficients in a conventional off-line closed-loop technique, suffers from the design instability problem, which is illustrated in Fig. 4. A set of prediction coefficients, $\mathbf{P}_{B_2}$, are designed to match the statistics of given reconstructed reference blocks, which includes the red block in frame 1 and the blue block in frame 2 (which are references for the red block in frame 2 and the blue block in frame 3, respectively). However, as we employ $\mathbf{P}_{B_2}$ in closed-loop, the reconstructed reference frames are updated, as seen in the shaded region of the blue block in frame 2, which implies the blue block in frame 3 is predicted from a reference for which the optimal predictor $\mathbf{P}'_{B_2}$ differs from the designed predictor $\mathbf{P}_{B_2}$. This mismatch in statistics between design and operation grows over time as the data is fed through the prediction loop, leading to substantial ineffectiveness of the designed prediction parameters.

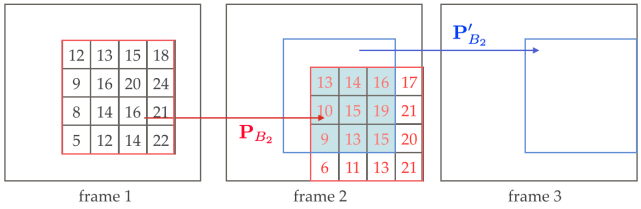


**Fig. 4**. The instability problem in closed-loop predictor design

Therefore, to overcome this design instability, we employ a two loop asymptotic closed-loop (ACL) approach similar to [4], where, in the inner loop we fix the encoder decisions, and optimize the predictor $\mathbf{P}_{B_2}$ to minimize *prediction error* $J$ via the ACL approach, and in the outer loop encoder decisions are updated with $\mathbf{P}_{B_2}$ fixed to optimize the *rate-distortion (RD) cost* in closed-loop (for a more formal and detailed description of the ACL approach see [4]). The mismatch in optimization criteria of the two loops eliminates convergence guarantee of the outer loop. Thus, in a deviation from [4] to achieve better overall RD performance, we compare the RD cost achieved at every iteration of the outer loop and select the prediction parameters of the outer loop iteration, which resulted in the minimum RD cost.

Note that $\rho_{i,j}$ in our framework is not strictly the temporal correlation coefficient per frequency any more, although

we observed it to be similar to temporal correlation in the low frequencies. These coefficients differ from the coefficients of basic TDTP to effectively account for the spatial correlation with pixels neighboring the reference block boundary. This also differentiates our framework from simply employing a larger inter-prediction block at the encoder. Although the TDTP and the interpolation filter interfere with each other, they can not be completely replaced by each other, since element-wise multiplication of TDTP is not equivalent to matrix multiplication of interpolation. However, the proposed framework paves a path for further performance improvement via joint design of TDTP and interpolation, which will be one of our future research directions.

## 4. EXPERIMENTAL RESULTS

To evaluate the performance, the proposed EB-TDTP is implemented in HM 14.0, and is compared to standard HEVC and HEVC with basic TDTP. To simplify the experiments, all the sequences are coded in IPPP format, only previous frame is allowed as reference, motion search is allowed only up to half-pixel precision, both prediction and transform block sizes are restricted to $8 \times 8$, and the sample adaptive offset function option is disabled. We evaluate the full potential of EB-TDTP by designing a specific set of coefficients for each sequence using the training method described above. Each sequence is tested at various bitrates with QP ranging from 22 to 37. Preliminary results in BD-rate improvement [11] of employing EB-TDTP is provided in Table 2, with up to $2.4\%$ extra average bitrate reduction over basic TDTP, and up to $12.9\%$ bitrate reduction over standard HEVC. The RD curve for sequence *coastguard* is shown in Fig. 5, with up to 0.75 dB PSNR improvements over standard HEVC. These results already demonstrate the utility of the proposed technique with consistent and reasonable gains over regular TDTP, and substantial performance improvements over standard HEVC, with reasonable complexity increase due to the extra DCT and inverse DCT. Moreover, the results also show the potential for substantial performance improvement that can be achieved with joint optimization of EB-TDTP and interpolation within the proposed framework.

## 5. CONCLUSION

This paper substantially extends the transform domain temporal prediction approach for video coding to *fully* account for the interference with sub-pixel interpolation filter, and the spatial correlations outside the reference block boundary, by completely disentangling the spatial and temporal correlations. Our proposed framework also paves a path for joint optimization of TDTP and interpolation. Experimental results demonstrate the effectiveness of the proposed technique with substantial gains over standard HEVC.

| Sequence | Bit rate reduction (%) (TDTP) | Bit rate reduction (%) (EB-TDTP) |
|---|---|---|
| *Coastguard* (CIF) | 8.69 | **10.61** |
| *Mobile* (CIF) | 11.91 | **12.91** |
| *Highway* (CIF) | 3.05 | **5.44** |
| *Bus* (CIF) | 5.95 | **6.70** |
| *Waterfall* (CIF) | 9.88 | **10.16** |
| *Tempete* (CIF) | 5.78 | **6.19** |
| *RaceHorse* (416x240) | 3.48 | **4.32** |

**Table 2**. Reduction in bitrate over reference encoder by employing TDTP and EB-TDTP

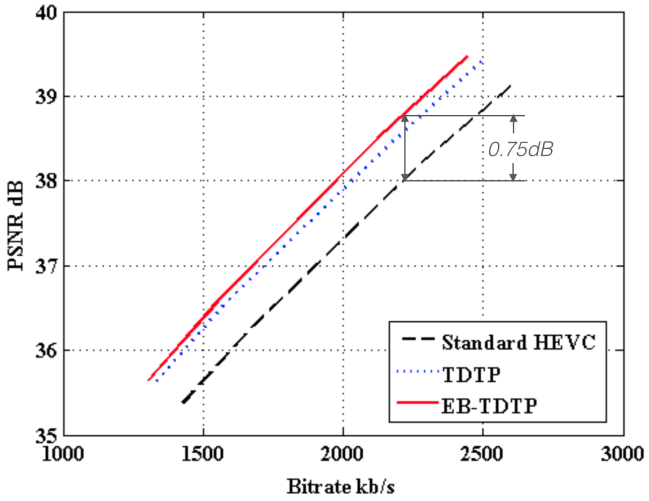


**Fig. 5**. Coding performance comparison for sequence *coastguard* at *CIF* resolution

## 6. REFERENCES


[1] 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.

[2] J. Han, V. Melkote, and K. Rose, "Transform-domain temporal prediction in video coding: exploiting correlation variation across coefficients," in *IEEE International Conference on Image Processing (ICIP)*, 2010, pp. 953–956.

[3] J. Han, V. Melkote, and K. Rose, "Transform-domain temporal prediction in video coding with spatially adaptive spectral correlations," in *IEEE International Workshop on Multimedia Signal Processing (MMSP)*, 2011, pp. 1–6.

[4] S. Li, T. Nanjundaswamy, Y. Chen, and K. Rose,

"Asymptotic closed-loop design for transform domain temporal prediction," in *IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 4907–4911.

[5] J. Kim and J. W. Woods, "Spatio-temporal adaptive 3-d kalman filter for video," *IEEE Transactions on Image Processing*, vol. 6, no. 3, pp. 414–424, 1997.

[6] T. Wedi, "Adaptive interpolation filter for motion and aliasing compensated prediction," in *Electronic Imaging 2002*. International Society for Optics and Photonics, 2002, pp. 415–422.

[7] S. Li, O. G Guleryuz, and S. Yea, "Reduced-rank condensed filter dictionaries for inter-picture prediction," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 1428–1432.

[8] S.-J. Choi and J. W. Woods, "Motion-compensated 3-d subband coding of video," *IEEE Transactions on Image Processing*, vol. 8, no. 2, pp. 155–167, Feb 1999.

[9] J.-R. Ohm, "Three-dimensional subband coding with motion compensation," *IEEE Transactions on Image Processing*, vol. 3, no. 5, pp. 559–571, Sep 1994.

[10] U.-V. Koc and K. J. R Liu, "Subpixel motion estimation in dct domain," in *Visual Communications and Image Processing'96*. International Society for Optics and Photonics, 1996, pp. 332–343.

[11] G. Bjontegard, "Calculation of average psnr differences between RD-curves," *Doc. VCEG-M33 (ITU-T SG16 Q.6), Austin, Texas, USA*, April 2001.