ASYMPTOTIC CLOSED-LOOP DESIGN FOR TRANSFORM DOMAIN TEMPORAL PREDICTION

Shunyao Li, Tejaswi Nanjundaswamy, Yue Chen, Kenneth Rose

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

ABSTRACT

Current video coders exploit temporal dependencies via prediction that consists of motion-compensated pixel copying operations. Such per-pixel temporal prediction ignores important underlying spatial correlations, as well as considerable variations in temporal correlation across frequency components. In the transform domain, however, spatial decorrelation is first achieved, allowing for the true temporal correlation at each frequency to emerge and be properly accounted for, with particular impact at high frequencies, whose lower correlation is otherwise masked by the dominant low frequencies. This paper focuses on effective design of transform domain temporal prediction that: *i)* fully accounts for the effects of sub-pixel interpolation filters, and *ii)* circumvents the challenge of catastrophic design instability due to quantization error propagation through the prediction loop. We design predictors conditioned on frequency and sub-pixel position, employing an iterative open-loop (hence stable) design procedure that, on convergence, approximates closed-loop operation. Experimental results validate the effectiveness of both the asymptotic closed-loop design procedure and the transform-domain temporal prediction paradigm, with significant and consistent performance gains over the standard.

Index Terms— Temporal prediction, motion compensation, sub-pixel interpolation, DCT, video coding

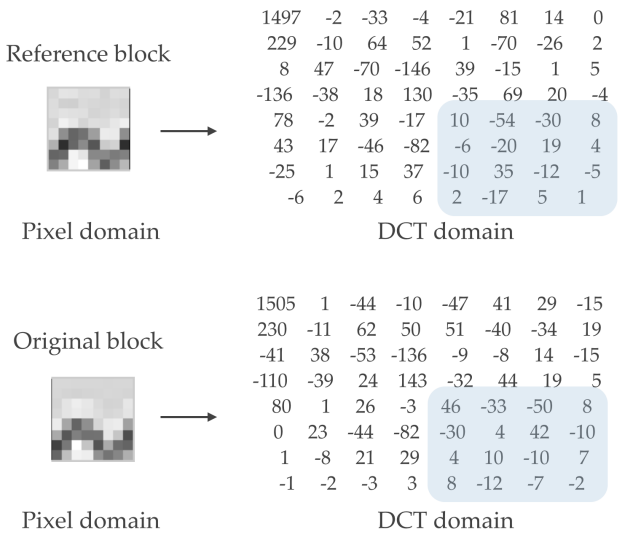
1. INTRODUCTION

Modern video coding standards, such as HEVC, exploit the inherent temporal dependencies in a video sequence via inter prediction [1]. Instead of directly encoding the raw pixel values for each block, the encoder predicts them from a similar reference block in previously reconstructed frames through pixel domain block matching. The prediction error is then transformed, typically by the discrete cosine transform (DCT), and the transform coefficients are quantized and coded. While a considerable amount of research has been focused on the accuracy of motion compensation, including motion vector estimation (e.g., [2, 3, 4]), and variable block sizes (e.g., [5, 6]), very few questions have been raised about the limitations of pixel domain block matching and copying.

In most video sequences, each pixel is highly correlated with its neighbors, thus a given pixel in the current block is correlated with a group of pixels in the reference block. This implies that the conventional one-to-one pixel-copying approach is suboptimal. Some multi-tap filtering approaches [7, 8, 9] and motion-compensated three-dimensional subband coding approaches [10, 11] have been proposed to account for such spatial correlations. However, a more effective approach to model complex spatio-temporal correlation is via DCT-domain temporal prediction, where spatial decorrelation is (largely) achieved first and allows for optimality of subsequent one-to-one transform coefficient prediction. Moreover, while the pixel domain correlation coefficient ρ is typically close to 1, transform coefficients exhibit temporal correlation that varies considerably with frequency, as illustrated in Fig. 1 for an example block. In Fig. 1(a) we observe that the reference block and the original block look very similar in the pixel domain, and this is roughly true for the lower frequency DCT coefficients, yielding ρ values close to 1. However, the similarity tends to break down for higher frequency coefficients with consequently decreasing ρ values, as illustrated in Fig. 1(b). Clearly, this variation in correlation across frequencies is masked in the pixel domain by the dominant low frequencies, and the resulting $\rho \approx 1$ led to the prevalence of block matching and copying techniques in current coders. Thus, the advantages of transform domain temporal prediction (TDTP) can be viewed from two perspectives: *i)* an effective paradigm to disentangle spatial and temporal correlations allowing for optimal prediction, and *ii)* a means to make explicit, and hence properly account for, the variation in temporal correlation across frequency, which is otherwise hidden in the pixel domain.

The significant potential of TDTP was recognized in an earlier paper from our group [12] where a TDTP approach was proposed in conjunction with full-pixel motion, with coefficients trained from original video sequences, which yielded substantial coding gains. It was then extended in [13] to update coefficients along the motion trajectory using a backward spatially adaptive approach. In this paper, we return to forward adaptive prediction with focus on the two-fold challenge of effective offline predictor design for TDTP at sub-pixel motion compensation. First, sub-pixel interpolation employs low-pass filters, which implies that high frequencies are scaled down, as suggested by the filter's magnitude response. It should be emphasized that the preceding statement applies to the Fourier transform domain and we are concerned with the effect on DCT coefficients. Nevertheless, it is clear that the interpolation filter interferes with TDTP, and thus needs to be properly accounted for during the predictor design. Since every sub-pixel location incurs a different but fixed combination of vertical and horizontal sub-pixel interpolation filters, we propose to train predictors conditioned on the sub-pixel location. The second, and critical challenge, is instability of the predictor design due to quantization error propagation in closed-loop operation. If the prediction parameters are modified during the design, to match the statistics observed in a pass through a sequence with its corresponding reconstructed reference frames, they will then be employed in the next pass. But each consecutive frame is now predicted from a differently reconstructed reference, resulting in prediction statistics incompatible with the designed parameters and, moreover, with such deviation in statistics potentially growing in magnitude as the coder advances through the sequence, as the quality of actual prediction impacts the quality of reconstruction and thereby the next

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



(a) Reference block and original block in pixel and DCT domain

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

(b) Transform prediction coefficients for 8x8 DCT coefficients for *mobile* sequence at QP=22

Fig. 1. An illustration of difference in correlations between pixel domain and DCT domain

frame's prediction, and so on. We address this issue by employing an iterative open-loop design technique, leveraging inspiration from an early paper from our lab on the design of vector quantizers in a predictive coding setting [14]. Here, a complete sequence of reconstructed frames from the previous iteration provides reference frames to predict a sequence of frames in the current iteration, with prediction parameters designed for exactly this sequence of reference frames. The coder then produces a sequence of reconstructions which is now fixed as sequence of reference frames for the next iteration, and the prediction parameters are updated. This ensures there is no incompatibility between design and deployment. As the reconstructed sequences converge, the above open-loop prediction effectively becomes stable closed-loop prediction. Hence, we call this the asymptotic closed-loop (ACL) design technique. Simulation results provide evidence of significant coding gains over standard HEVC.

2. PREDICTION MODEL

Conventional motion-compensated prediction assumes that pixels along a motion trajectory form a temporal first-order AR process, neglecting all the spatial correlation. Instead we operate in the DCT domain, where spatial decorrelation has been achieved, and assume that frequency coefficients of blocks along a motion trajectory form a first-order AR process per frequency. Let's denote by x_n a DCT coefficient at a particular frequency of an inter-coded block in frame n, and by x_{n-1} the corresponding DCT coefficient of its motion compensated reference block in frame $(n-1)$, then the AR process is given as,

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

where ρ is the transform domain correlation coefficient, which captures the temporal dependency of transform coefficients at a given frequency along the motion trajectory, and z_n is the innovation. Without loss of generality, we assume that the motion compensated reference block is in the immediately previous frame. We need closed-loop prediction so that the decoder can exactly mimic the encoder, so we use the *reconstructed* DCT coefficient, $\hat{x}_{n-1}$, as a reference. Thus the optimal prediction for each frequency coefficient is

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

with ρ now the corresponding correlation coefficient. We note that the conventional pixel domain block matching and copying is equivalent to employing $\rho = 1$ at all frequencies. We estimate ρ to minimize the mean square prediction error,

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

The optimal prediction coefficient ρ is

$$\rho = \frac{E(x_n \hat{x}_{n-1})}{E(\hat{x}_{n-1}^2)}, \tag{4}$$

which forms the basis of the off-line design technique described below.

3. OFFLINE ASYMPTOTIC CLOSED-LOOP DESIGN

The closed-loop operation causes instability of the predictor design due to quantization error propagation. Let's denote a sequence of DCT coefficients at a certain frequency for blocks along the motion trajectory as, $x_1, x_2, \ldots, x_N$. The first frame is intra coded so we have the first reconstructed coefficient $\hat{x}_1$. As discussed earlier the current temporal prediction is equivalent to predicting in transform domain with $\rho = 1$ at all frequencies. Thus the predicted coefficient in frame 2 is, $\tilde{x}_2 = \hat{x}_1$, the prediction error, $e_2 = x_2 - \tilde{x}_2$, is quantized to generate $\hat{e}_2$, and the final reconstructed coefficient is $\hat{x}_2 = \hat{e}_2 + \tilde{x}_2$. Similarly, the reconstruction of frame 3 is generated using $\hat{x}_2$, and so forth. This closed-loop encoder system can be summarized as

$$\hat{x}_n = \hat{x}_{n-1} + \hat{e}_n \tag{5}$$

Using these samples as a reference, we can get a first estimate of the prediction coefficient as,

$$\rho_1 = \frac{E(x_n \hat{x}_{n-1})}{E(\hat{x}_{n-1}^2)}, \tag{6}$$

which works well for this set of reconstructed samples as reference. On using this prediction coefficient in the coder, first reconstructed coefficient, $\hat{x}_1$ is unaltered as it is intra coded. However, for the coefficient in frame 2, we generate a new prediction, $\tilde{x}'_2 = \rho_1 \hat{x}_1$, and the corresponding new prediction error, $e'_2 = x_2 - \tilde{x}'_2$, is quantized to generate $\hat{e}'_2$, which is used to generate the new reconstruction $\hat{x}'_2 = \hat{e}'_2 + \tilde{x}'_2$. In a closed loop operation, mimicking (5), frame 3 is generated using this new reconstruction $\hat{x}'_2$. Clearly, the correlation between $\hat{x}'_2$ and x_3 is different from what the prediction coefficient ρ_1 was designed for (i.e., correlation between $\hat{x}_2$ and x_3), leading

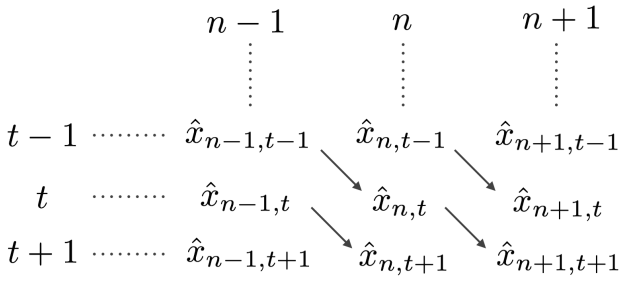


Fig. 2. Asymptotic Closed-Loop (ACL) training approach

to ineffective prediction. Moreover, this may result in "build up" of deviation in statistics, as we proceed to future frames. This instability problem is particularly serious at low bitrates, where the residual is not well encoded and the reconstruction is more dependent on the prediction quality.

To address this instability problem, we propose an iterative open-loop design technique that asymptotically optimizes the system for closed-loop operation, similar to the asymptotic closed-loop (ACL) approach previously proposed by our lab [14] for an instability problem in predictive vector quantizer design. We use double subscripts, e.g., $x_{n,t}$ to indicate variables from frame n and iteration t. The basic idea of ACL is to employ open-loop prediction to avoid the instability problem of closed-loop prediction, while updating the prediction parameters in each iteration. Once the parameters converge, it becomes equivalent to closed-loop operation. Given a set of reconstructed coefficients along a motion trajectory for a frequency at iteration $t-1$, $\hat{x}_{1,t-1}, \hat{x}_{2,t-1}, \ldots, \hat{x}_{N,t-1}$, we estimate the prediction coefficient for iteration t as,

$$\rho_t = \frac{E(x_n \hat{x}_{n-1,t-1})}{E(\hat{x}^2_{n-1,t-1})}. \tag{7}$$

This ρ_t is employed in open-loop to predict coefficients in frame n of iteration t, $\tilde{x}_{n,t} = \rho_t \hat{x}_{n-1,t-1}$. So instead of (5) where all the samples are of the same iteration, the reconstruction in this open-loop scheme is given as,

$$\hat{x}_{n,t} = \rho_t \hat{x}_{n-1,t-1} + \hat{e}_{n,t}, \tag{8}$$

where $\hat{e}_{n,t}$ is the quantized prediction error, $e_{n,t} = x_n - \tilde{x}_{n,t}$. Clearly, there is no instability problem here, as ρ_t is directly optimized for the correlation between $\hat{x}_{n-1,t-1}$ and x_n. Since the prediction coefficients are specifically designed to be optimal for the statistics they are applied to at each iteration, the prediction is guaranteed to improve. Better prediction (usually) leads to better reconstruction, and vice versa. The reconstruction error is generally decreasing and would approach convergence. On convergence, the reconstruction remains the same, i.e., $\hat{x}_{n-1,t-1} = \hat{x}_{n-1,t}$, which makes it equivalent to the closed-loop system, i.e.,

$$\hat{x}_{n,t} = \rho_t \hat{x}_{n-1,t} + \hat{e}_{n,t}, \tag{9}$$

and the prediction coefficients converge as well, i.e., $\rho_{t+1} = \rho_t$. An illustration of ACL is provided in Fig. 2.

If the motion vectors and other encoder decisions for quantization and entropy coding are fixed, then we can employ the ACL scheme described above to estimate optimal prediction coefficients. However, motion vectors, quantization, and entropy coding decisions are dependent on the prediction coefficients. Thus, we propose a two-loop design scheme for prediction coefficients. In the inner loop, we estimate prediction coefficients via ACL while fixing the motion vectors and other encoder decisions for quantization and entropy coding such as prediction type and merge/skip flag. In practical implementation, to keep the design complexity in check, instead of waiting for full ACL convergence, we stop the inner loop when the prediction error energy, $E(e^2_{n,t})$, no longer decreases significantly. The encoder decisions are updated in the outer loop, while using the prediction coefficients estimated in the inner loop. The outer-loop decisions are updated in the closed-loop operation of the encoder to minimize the rate-distortion (RD) cost. The outer-loop is stopped when inner-loop converges quickly, as we observed that further outer-loop iterations did not help.

Since prediction coefficients are trained using reconstructed reference samples, their statistics change based on encoding quality. Thus we design and employ different set of prediction coefficients for different quantization parameters (QP) of the encoder. The training is done via the two-loop method described above, while operating the encoder in a constant QP mode. Also as discussed earlier, the interpolation filter interferes with the prediction coefficients, thus to account for different interpolation effect at each sub-pixel location, we design and employ different prediction coefficients for each sub-pixel location, e.g., at half pixel precision, we have 4 sets of coefficients, 1 for full-pixel location and 3 others for sub-pixel location. Further, at low bitrates, the encoder opts to encode vast majority of blocks via skip mode, where prediction error of all DCT coefficients is quantized to 0. For these blocks, the overall reconstruction error is same as the prediction error. However, for regular blocks, when a fixed quantizer is used (i.e., under constant QP operation) the overall reconstruction error does not vary much. Clearly, the stability of reconstructed data varies between these two modes during the ACL design. Thus, to exploit this difference effectively, we design and employ different set of prediction coefficients for these two modes. For further effectiveness of the prediction, we updated the motion estimation criteria to minimize the transform domain prediction error, which is in sync with the prediction design criteria.

The basic assumption for convergence that better prediction and better reconstruction are mutually supportive is not always guaranteed. For real world sequences we observe that the prediction error cost decreases initially and then hits a limit cycle. Thus we simply stop the inner loop iterations when this cost stops decreasing, and still achieve significant performance improvements. Moreover, while the inner-loop iterations minimize the mean squared prediction error, the encoder decisions in the outer loop are updated to minimize the RD cost. This mismatch in the optimization criteria does not ensure full convergence in the outer loop, and hence we stop the outer loop when the inner loop converges quickly. Resolving this mismatch in optimization criteria and optimizing the prediction coefficients for the overall RD cost in both loops will be one of our future research directions.

4. EXPERIMENTAL RESULTS

The proposed TDTP is implemented in HM 14.0, and compared with standard HEVC. To simplify the experiments, all sequences are coded in format IPPP (although the method is applicable to bidirectional prediction), sample adaptive offset (SAO) function option is disabled, both prediction size and transform size are restricted to 8x8, and the motion search is at half-pixel precision. As mentioned earlier the motion search criterion is the mean squared transform domain prediction error instead of 1-norm (SAD).

Two experiments were conducted to validate the efficacy of TDTP for different application scenarios. In experiment 1, we target

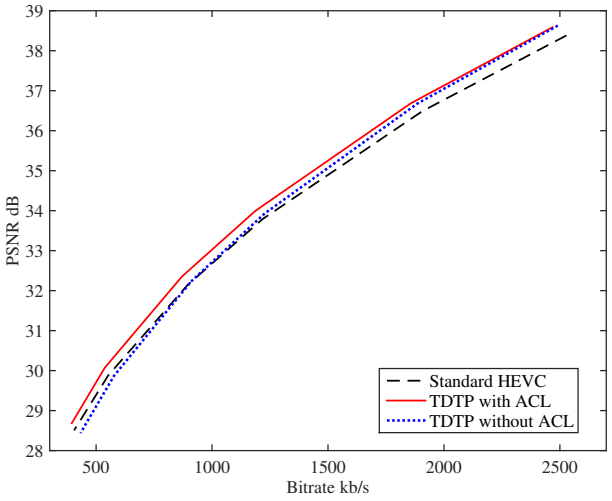


Fig. 3. Coding performance comparison for sequence *bus* at *CIF* resolution

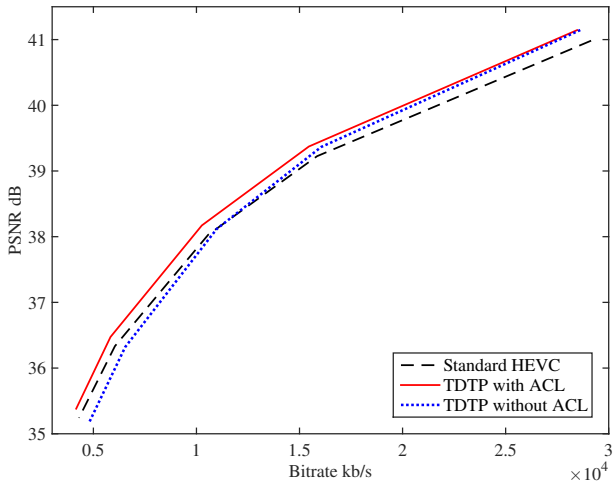


Fig. 4. Coding performance comparison for sequence *ParkScene* at *1080p* resolution

Sequence	Bit rate reduction (%) (Experiment 1)	Bit rate reduction (%) (Experiment 2)
BQTerrace	12.88	10.23
BasketballDrive	6.21	5.36
Kimono	7.88	4.53
ParkScene	7.41	7.00
Keiba	5.38	5.17
RaceHorse	3.44	3.03
Waterfall	9.29	4.38
Vidyo1	5.75	4.56
Bus	5.59	5.37
Tennis	2.23	1.50
Tempete	5.56	5.30
FourPeople	6.81	3.04
Average	6.53	4.96

Table 1. Reduction in bitrate over reference encoder by employing TDTP.

video storage applications where encoding is performed off-line, and design a specific set of coefficients for each sequence using the training method described above, to exploit the full potential of TDTP. The overhead of storing a set of coefficients per sequence is negligible. The percentage bitrate reduction (calculated as per [15]) achieved by TDTP over standard HEVC encoder is a significant 6.53% on the average. Individual bitrate reductions for 12 sequences is presented in the first column of Table 1. In experiment 2, we evaluate performance outside the training set, by providing a choice of fixed 8 sets of prediction coefficients (including all 1 coefficients, i.e. no transform prediction) at the encoder. These choices are provided to cover varying statistics in video content and are obtained using a small set of training sequences different from the test set. The overhead here is a mere 3 bits per sequence. The performance gains for experiment 2 is presented in the second column of Table 1. The difference in gains between column 1 and column 2 suggests further scope for improving sequence and frame wise adaptivity and will be one of our future research directions. The RD curves for sequence *bus* and *ParkScene* are shown in Fig. 3 and Fig. 4 with performance comparison to employing prediction coefficients trained without the ACL technique. Note that at low bitrates, training without ACL suffers greatly from the instability problem, resulting in worse performance than the standard, while the training with ACL has no such issues, clearly providing substantial evidence for the effectiveness of our proposed technique.

5. CONCLUSION

This paper substantially extends the transform domain motion compensated prediction approach for video coding, so as to account for spatial correlations during temporal prediction, and the true temporal correlations in the AR process at different spatial frequencies along motion trajectories. A novel design scheme for prediction coefficients is proposed to account for the sub-pixel interpolation filter and the effect of quantization error propagation in the prediction loop. A two-loop training method is applied, wherein an iterative open-loop design technique is employed to address a major instability problem of closed-loop design. The overall system provides substantial gains compared to standard HEVC, providing evidence for the effectiveness of the proposed technique.

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] L. K. Liu and E. Feig, "A block-based gradient descent search algorithm for block motion estimation in video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 6, no. 4, pp. 419–422, 1996.

[3] A. M. Tourapis, O. C. Au, and M. L. Liou, "Predictive motion vector field adaptive search technique (pmvfast)-enhancing

block based motion estimation," in *Proceedings of SPIE*, 2001, vol. 4310, pp. 883–892.

[4] J. Konrad and E. Dubois, "Bayesian estimation of motion vector fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 9, pp. 910–927, 1992.

[5] M. Chan, Y. Yu, and A. Constantinides, "Variable size block matching motion compensation with applications to video coding," *IEE Proceedings I (Communications, Speech and Vision)*, vol. 137, no. 4, pp. 205–212, 1990.

[6] J. Lee, "Optimal quadtree for variable block size motion estimation," in *IEEE International Conference on Image Processing (ICIP)*. IEEE, 1995, vol. 3, pp. 480–483.

[7] 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.

[8] 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.

[9] S. Li, O. 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, p. to appear.

[10] 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.

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

[12] 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)*. IEEE, 2010, pp. 953–956.

[13] 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)*. IEEE, 2011, pp. 1–6.

[14] H. Khalil, K. Rose, and S. L. Regunathan, "The asymptotic closed-loop approach to predictive vector quantizer design with application in video coding," *IEEE Transactions on Image Processing*, vol. 10, no. 1, pp. 15–23, 2001.

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