Transform-domain Temporal Prediction in Video Coding with Spatially Adaptive Spectral Correlations

Jingning Han, Vinay Melkote*, and Kenneth Rose

Department of Electrical and Computer Engineering, University of California Santa Barbara
Santa Barbara, CA 93106, USA
{jingning,melkote,rose}@ece.ucsb.edu

***Abstract*—Temporal prediction in standard video coding is performed in the spatial domain, where each pixel block is predicted from a motion-compensated pixel block in a previously reconstructed frame. Such prediction treats each pixel independently and ignores underlying spatial correlations. In contrast, this paper proposes a paradigm for motion-compensated prediction in the transform domain, that eliminates much of the spatial correlation before individual frequency components along a motion trajectory are independently predicted. The proposed scheme exploits the true temporal correlations, that emerge only after signal decomposition, and vary considerably from low to high frequency. The scheme spatially and temporally adapts to the evolving source statistics via a recursive procedure to obtain the cross-correlation between transform coefficients on the same motion trajectory. This recursion involves already reconstructed data and precludes the need for any additional side-information in the bit-stream. Experiments demonstrate substantial performance gains in comparison with the standard codec that employs conventional pixel domain motion-compensated prediction.**

I. Introduction

Motion compensation is a central component in most standard video coding techniques, such as H.264, to exploit the inherent temporal redundancy in the video sequence [12]. A motion search that involves pixel domain block matching associates each on-grid block in the current frame with a reference block in previously coded frames, which is subtracted from the former to generate a residual block of pixels. This is then subjected to a spatial transformation (typically DCT), and the transform coefficients quantized and coded. A considerable volume of prior research is devoted to accurate motion compensation with focus on various issues, including the use of a long-term buffer [13], overlapping estimation [9], variable size partition [2], fractional sample interpolation [11], etc.

The efficacy of motion compensated prediction builds on the assumption that pixel blocks along a motion trajectory form an autoregressive (AR) sequence. The reason for the direct subtraction of the reference pixel block from the current block is that the temporal correlation coefficient, as appears between pixel blocks along the same motion trajectory, is typically close to 1. An alternative viewpoint that transform coefficients of the blocks in a motion trajectory form a scalar AR process, at each spatial frequency, was inspired by [4] where we proposed an estimation-theoretic (ET) approach to delayed video decoding, and [10] where an ET approach for predictive scalable coding was proposed. In both [4] and [10], such a viewpoint was necessary to explicitly account for the quantization interval information exploited by the ET framework, which is available only in the transform domain. The innovation of each scalar AR process (per frequency) was modeled as Laplacian [1], and the temporal correlation coefficient of each AR process at different spatial frequencies was assumed to be unity (as is the common practice in pixel domain). Note that in this case the transform domain model is congruent with the pixel domain AR model due to the unitarity of the spatial transform. An analysis of the actual correlations (see Sec. II and Sec. V) at each frequency reveals that this assumption is valid only for the case of DC spatial frequency, while the temporal correlation varies noticeably at higher frequencies. Further, the correlation at any particular spatial frequency exhibits substantial variation across motion trajectories, and evolves over time. The concentration of energy in the DC component results in its high (near unity) temporal correlation dominating any pixel domain evaluation of the correlation coefficient.

The above observation motivates the proposed transform-domain motion-compensated prediction (TDMCP) scheme that first performs a spatial transformation (DCT) on both the reference and original blocks to eliminate spatial correlations, and then optimally predicts individual transform coefficients by exploiting the true temporal correlations underlying the video signal. Since these temporal correlations markedly vary across motion trajectories, and evolve over time, the proposed approach employs a recursive procedure to update running estimates of the first and second moments of transform coefficients along motion trajectories that enter each frame, which are in turn employed to evaluate the temporal correlations along these trajectories. The overall system effectively adapts to spatial and temporal variations in the statistics of the video signal, and provides substantial coding performance

*Vinay Melkote is now with Dolby Laboratories Inc., 100 Potrero Avenue, San Francisco, CA 94103.

This work was supported in part by Qualcomm, Inc.

978-1-4577-1434-4/11/$26.00 ©2011 IEEE

gains compared to the standard pixel-domain motion compensated prediction employed in H.264. Some peripherally relevant work includes motion-compensated three-dimensional subband coding, see e.g., [3], [8], where blocks in consecutive frames are realigned and coded via 3D DCT/DWT. We note that the approach proposed herein optimally exploits the true temporal correlation in a DPCM (in the format of inter-frame prediction) framework.

Some preliminary results with a simplified transform domain prediction scheme were presented in [6], where we assumed a spatio-temporally stationary model for the video signal, and employed fixed values for the temporal correlation coefficient at different spatial frequencies (which are conveyed to the decoder as side information), while ignoring statistical variations across motion trajectories and time. The purpose of [6] was to preliminarily test and validate the potential benefit to performing prediction in the transform domain. The proposed approach in this paper eschews such limitations via adaptive estimation of the correlation coefficient. The estimation procedure utilizes prior reconstructed data, and does not require the transmission of these coefficients in the bitstream. Moreover, the approach is extended to incorporate the commonly used fractional sample interpolation methods.

II. Transform Domain Prediction: Model and Motivation

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

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

where the innovation samples z_n are zero-mean, independent and identically distributed.

To substantiate the motivation for transform domain prediction, we ran the 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 4×4. The temporal correlation coefficient ρ 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. I is the matrix of these 16 temporal correlation coefficients of sequence $coastguard$ at $QCIF$ resolution. Note that the correlation is close to 1 for DC, but quite different otherwise. Table. II 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 dominates any evaluation of block-wise temporal correlation in the pixel domain. Such characteristics are also exhibited by other video sequences. More detailed analysis (see Sec. V) demonstrates that correlation variation also exists in spatial and temporal directions, in addition to frequency.

TABLE I
Matrix of temporal correlation coefficients for the 16 DCT coefficients in $coastguard$ at $QCIF$ resolution.

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 II
Matrix of variances of the 16 DCT coefficients in $coastguard$ at $QCIF$ resolution.

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

We emphasize that the conventional AR model inherently assumes pixels of the blocks form *independent* scalar temporal AR process, i.e., the model completely ignores *inter-pixel (spatial) correlation* within each block when exploiting the *temporal correlation*. In contrast, the proposed model first removes the spatial correlation via DCT, and then the resulting transform coefficients (which are almost uncorrelated) are individually modeled by a temporal scalar AR process per frequency.

As an aside, it should be noted that the zero-mean innovations in (1) 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 ρ). It was indeed observed in the experiment that the mean of DCT coefficients at any AC frequency was always nearly zero. The DC coefficient in general is not zero-mean since 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.

In the context of inter frame prediction, the optimal prediction for each frequency coefficient x_n is

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

where $\hat{x}_{n-1}$ is the corresponding frequency coefficient of the motion compensation reference block, in the previously reconstructed frame $n-1$, and ρ is the correlation coefficient appropriate to that frequency and motion trajectory. We note that conventional pixel domain prediction is equivalent to employing $\rho = 1$ at all times. Clearly ρ captures the temporal dependency of transform coefficients at the same frequency along the motion trajectory, and thus is potentially adaptive to the spatial location and time instance. Estimation of temporal correlation at each frequency *along the motion trajectory for every on-grid block of the current frame*, at both encoder and decoder, would ideally require a motion search spanning multiple past frames, so that an adequate number of transform coefficient samples along these trajectories can be collected. This formidable and computationally intensive task is circumvented by instead employing a scheme that recursively computes the statistical information (up to second

order) of the transform coefficients of on-grid blocks, which completely captures the underlying temporal correlations.

III. RSPICE: Recursive Spectral per-Coefficient Interframe Correlation Estimate

We develop the recursions to compute temporal correlations along the motion trajectories. Ideally statistical model evaluation is performed with full access to the original signal. However, since the decoder does not have such access, and sending side information to indicate temporal correlation per transform coefficient will inevitably incur significant overheads, we rely on the reconstructed samples and assume they approximate the original signal for this purpose. The efficacy of this approximation will be experimentally verified in Sec.V.

Consider encoding, X_n^k, the on-grid block k in frame n. The motion trajectory associated with X_n^k, is composed of (potentially off-grid) blocks $U_{n,n-1}^k$ in frame $n-1$, and $U_{n,n-2}^k$ in frame $n-2$, connected by motion vectors as depicted in Fig 1. In general, the notation $U_{n,r}^k$, $r < n$, denotes a block in frame r that lies in the same motion trajectory as the on-grid block X_n^k in frame n. Performing spatial transform (typically DCT) on the blocks, results in a set of scalar sequences, one per frequency, which are largely decorrelated of each other. Let $x_n^{k,m}$ denote the unquantized value of transform coefficient m in X_n^k, and $u_{n,r}^{k,m}$ be the transform coefficient at the same frequency index in $U_{n,r}^k$. Correspondingly their reconstructions are denoted by $\hat{x}_n^{k,m}$ and $\hat{u}_{n,r}^{k,m}$. In viewing $\{\cdots, u_{n,n-2}^{k,m}, u_{n,n-1}^{k,m}, x_n^{k,m}\}$ as part of a first order AR process, the optimal prediction for $x_n^{k,m}$ is given by (2):

$$\tilde{x}_n^{k,m} = \rho_{n-1}\hat{u}_{n,n-1}^{k,m}. \tag{3}$$

Here ρ_{n-1} is the correlation coefficient of the AR process, estimated from reconstructed samples up to frame $n-1$, and is computed as

$$\rho_{n-1} = \frac{S_{n-1}(\hat{u}_{n,p}^{k,m}\hat{u}_{n,p-1}^{k,m})}{S_{n-1}((\hat{u}_{n,p}^{k,m})^2)}. \tag{4}$$

The notation $S_n(z_p)$ denotes in general the moving average of the sequence $\{z_p\}_{p\le n}$. Thus, (4) is the estimate of the correlation coefficient obtained by averaging pair-wise correlations at a particular spatial frequency m along the motion trajectory containing the block X_n^k.

Ideally, evaluation of (4) requires construction of a motion trajectory ending in X_n^k and tracking back into the past, so that the sequence $\{\cdots, u_{n,n-2}^{k,m}, u_{n,n-1}^{k,m}, x_n^{k,m}\}$ can be specified. The intractable complexity of such a motion trajectory construction is circumvented by employing the proposed RSPICE scheme. Define

$$u_{n,n}^{k,m} = x_n^{k,m} \tag{5}$$

so that the sequence $\{\cdots, u_{n,n-2}^{k,m}, u_{n,n-1}^{k,m}, x_n^{k,m}\}$ can be compactly denoted as $\{u_{n,p}^{k,m}\}_{p\le n}$. Assume availability of the following averages: $S_{n-1}(\hat{u}_{n-1,p}^{k,m})$, $S_{n-1}((\hat{u}_{n-1,p}^{k,m})^2)$, and $S_{n-1}(\hat{u}_{n-1,p}^{k,m}\hat{u}_{n-1,p-1}^{k,m})$. Note that by the definition (5), $\hat{u}_{n-1,n-1}^{k,m} = x_{n-1}^{k,m}$. Therefore, these averages involve sequences that end in a transform coefficient of an on-grid block in frame $n-1$. We now describe evaluation of the moments $S_{n-1}(\hat{u}_{n,p}^{k,m}\hat{u}_{n,p-1}^{k,m})$ and $S_{n-1}((\hat{u}_{n,p}^{k,m})^2)$ required in (4), which involve sequences that potentially end in off-grid blocks of frame $n-1$, from moment estimates for on-grid blocks in the same frame, that have for now been assumed available. These on-grid moments are themselves recursively updated once frame n is coded.

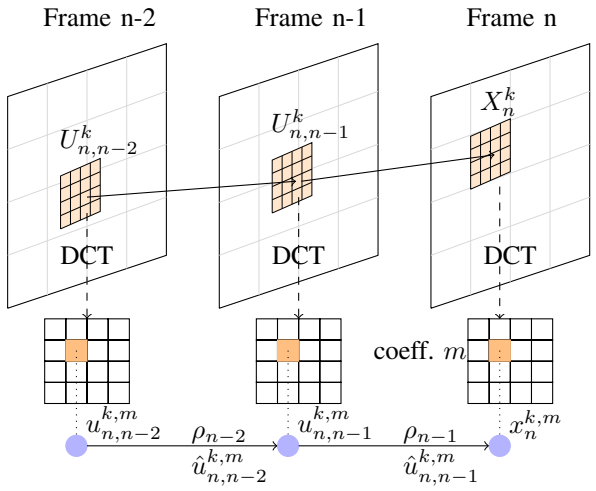


Fig. 1. Temporal prediction scheme per transform coefficient: each on-grid block in the current frame n (e.g., block k) is associated with a motion trajectory constructed by a sequence of reference blocks (potentially off-grid) connected by the motion vectors. Applying spatial transformation to each individual block in the sequence, the resulting transform coefficients at the same frequency are viewed as an AR process in temporal direction.

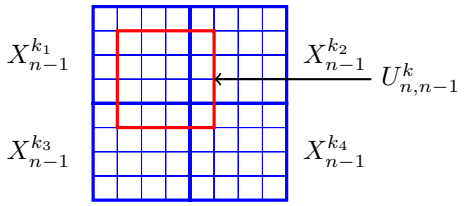


Fig. 2. An off-grid block overlaps 4 on-grid blocks. The blue blocks are on-grid, and the red off-grid block is employed as motion compensation reference for the block in subsequent frame.

Any off-grid block overlaps at most four on-grid blocks. The block $U_{n,n-1}^k$ shown in Fig. 2 is the motion compensation reference for X_n^k. It overlaps with on-grid blocks $\{X_{n-1}^{k_i}\}$ in frame $n-1$. The block $U_{n,n-1}^k$ consists of coefficients $u_{n,n-1}^{k,m}$, and $X_{n-1}^{k_i}$ consists of $x_{n-1}^{k_i,m}$. Due to linearity of the transform, there exists a set of constants $a_{i,m}$ named *construction constants*, such that

$$u_{n,n-1}^{k,m} = \sum_{i=1}^{4}\sum_{m=0}^{15} a_{i,m}x_{n-1}^{k_i,m}\ . \tag{6}$$

The construction constants only depend on the relative position of $U_{n,n-1}^k$ in this four-block grid. Inspired by the above linear

combination, and employing the redefinition of $x_{n-1}^{k,m}$ via (5), we make the following approximation:

$$S_{n-1}(\hat{u}_{n,p}^{k,m}) \approx \sum_{i=1}^{4} \sum_{m=0}^{15} a_{i,m} S_{n-1}(\hat{u}_{n-1,p}^{k_i,m}) \tag{7}$$

The approximation is based on the assumption that the motion trajectory containing the block sequence $\{U_{n,p}^{k,m}\}_{p<n}$ that culminates in the potentially off-grid block $U_{n,n-1}^{k}$, is very similar to the motion trajectories of the sequences $\{U_{n-1,p}^{k_i,m}\}_{p\leq n-1}$, $1 \leq i \leq 4$, that contain the 4 neighboring on-grid blocks $X_{n-1}^{k_i}$, i.e., the motion field possesses spatially local stationarity. Thus given on-grid averages $S_{n-1}(\hat{u}_{n-1,p}^{k,m})$, the off-grid average $S_{n-1}(\hat{u}_{n,p}^{k,m})$ can be calculated via (7).

The denominator of (4) requires moving average $S_{n-1}((\hat{u}_{n,p}^{k,m})^2)$ as an estimate of the marginal second moment. Approximating as before, we have:

$$\begin{aligned} &S_{n-1}((\hat{u}_{n,p}^{k,m})^2) \\ &\quad \approx \sum_{i=1}^{4} \sum_{j=1}^{4} \sum_{m=0}^{15} \sum_{l=0}^{15} a_{i,m} a_{j,l} S_{n-1}(\hat{u}_{n-1,p}^{k_i,m} \hat{u}_{n-1,p}^{k_j,l}). \end{aligned} \tag{8}$$

Note that in the above summation, terms with $i = j$ and $l = m$ correspond to second moments, $S_{n-1}((\hat{u}_{n-1,p}^{k,m})^2)$, of sequences that end in on-grid blocks of frame $n-1$. Exact evaluation of (8), however, also mandates storage of cross-correlations between transform coefficients of different on-grid blocks/frequency in each frame. A major advantage of operating in the transform domain is that such intensive computation/buffer requirements can be circumvented if we assume a largely *decorrelating* transformation as is indeed sought in compression applications, such as DCT in video coding. Specifically, the following approximation of "uncorrelatedness" holds well in the DCT domain[1]:

$$\begin{aligned} S_{n-1}(\hat{u}_{n-1,p}^{k_i,m} \hat{u}_{n-1,p}^{k_j,l}) \;&\approx S_{n-1}(\hat{u}_{n-1,p}^{k_i,m}) S_{n-1}(\hat{u}_{n-1,p}^{k_j,l}), \\ &\text{when } j \neq i \text{ or } l \neq m. \end{aligned} \tag{9}$$

With the above approximations, the marginal first and second moments of on-grid blocks in frame $n-1$ are sufficient to evaluate the term $S_{n-1}((\hat{u}_{n,p}^{k,m})^2)$ required in (4).

Now consider the term $S_{n-1}(\hat{u}_{n,p}^{k,m} \hat{u}_{n,p-1}^{k,m})$ in the numerator of (4). Again, this is approximated as:

$$\begin{aligned} &S_{n-1}(\hat{u}_{n,p}^{k,m} \hat{u}_{n,p-1}^{k,m}) \\ &\quad \approx \sum_{i=1}^{4} \sum_{j=1}^{4} \sum_{m=0}^{15} \sum_{l=0}^{15} a_{i,m} a_{j,l} S_{n-1}(\hat{u}_{n-1,p}^{k_i,m} \hat{u}_{n-1,p-1}^{k_j,l}) \; . \end{aligned} \tag{10}$$

In the above summation terms with $i = j$ and $l = m$ are of the form $S_{n-1}(\hat{u}_{n-1,p}^{k,m} \hat{u}_{n-1,p-1}^{k,m})$. Note that these terms correspond to the moving average estimating temporal cross-correlation at the *same spatial frequency* along motion trajectories that enter on-grid blocks in frame $n-1$, and are assumed available as discussed earlier. The remaining terms in (10) are again approximated:

$$\begin{aligned} S_{n-1}(\hat{u}_{n-1,p}^{k_i,m} \hat{u}_{n-1,p-1}^{k_j,l}) \;&\approx S_{n-1}(\hat{u}_{n-1,p}^{k_i,m}) S_{n-2}(\hat{u}_{n-1,p}^{k_j,l}), \\ &\text{when } j \neq i \text{ or } l \neq m. \end{aligned} \tag{11}$$

Note that in (11) above, the term $S_{n-2}(\hat{u}_{n-1,p}^{k_j,l})$ is a first moment of the motion compensated reference of on-grid block $X_{n-1}^{k_j}$ in frame $n-1$. This moment would have already been evaluated via (7) at time $n-1$, and can be available in a buffer. This, in addition to the first moment $S_{n-1}(\hat{u}_{n-1,p}^{k,m})$, and cross-correlation $S_{n-1}(\hat{u}_{n-1,p}^{k,m} \hat{u}_{n-1,p-1}^{k,m})$, corresponding to on-grid blocks in frame $n-1$ are therefore sufficient to evaluate the numerator in (4).

Once ρ_{n-1} has been determined by the above method, $x_n^{k,m}$ is predicted as (3), and the prediction residual is encoded. The reconstruction $\hat{x}_n^{k,m}$ is then employed to update the on-grid averages as follows:

$$\begin{aligned} &S_n(\hat{u}_{n,p}^{k,m}) = \alpha S_{n-1}(\hat{u}_{n,p}^{k,m}) + (1-\alpha)\hat{x}_n^{k,m}, \\ &S_n((\hat{u}_{n,p}^{k,m})^2) = \alpha S_{n-1}((\hat{u}_{n,p-1}^{k,m})^2) + (1-\alpha)(\hat{x}_n^{k,m})^2, \\ &S_n(\hat{u}_{n,p}^{k,m} \hat{u}_{n,p-1}^{k,m}) = \alpha S_{n-1}(\hat{u}_{n,p}^{k,m} \hat{u}_{n,p-1}^{k,m}) \\ &\qquad\qquad + (1-\alpha)\hat{x}_n^{k,m} \hat{u}_{n,n-1}^{k,m}. \end{aligned} \tag{12}$$

The above equations to update moving averages employ the forgetting factor α to give more weight to recent samples, also referred to as "exponential smoothing". This provides adaptation to the slowly time-varying statistical characteristics of the AR processes in natural video sequences. The update equations in (12) for on-grid moments along with (7), (8) and (10), thus describe a recursion.

To complete the recursion, the intra-coded blocks, which are often interpreted as containing emerging objects in the frame, are used as initialization for the process:

$$\begin{aligned} S_n(\hat{u}_{n,p}^{k,m}) &= \hat{x}_n^{k,m} \\ S_n(\hat{u}_{n,p}^{k,m} \hat{u}_{n,p-1}^{k,m}) &= S_n((\hat{u}_{n,p}^{k,m})^2) = (\hat{x}_n^{k,m})^2. \end{aligned} \tag{13}$$

We now summarize the recursive update procedure of RSPICE.

RSPICE recursion at time n

Given the moving averages of the transform coefficients of on-grid blocks in frame $n-1$ (and an auxiliary buffer containing that of their motion compensated references):

1) Identify the motion-compensated reference block $U_{n,n-1}^{k}$ in frame $n-1$ for each on-grid block k in frame n.
2) Compute the moving averages of $U_{n,n-1}^{k}$, i.e., $S_{n-1}(\hat{u}_{n,p}^{k,m})$, $S_{n-1}((\hat{u}_{n,p}^{k,m})^2)$, and $S_{n-1}(\hat{u}_{n,p}^{k,m} \hat{u}_{n,p-1}^{k,m})$ via (7), (10), and (8), respectively. Update the auxiliary buffer with $S_{n-1}(\hat{u}_{n,p}^{k,m})$.
3) Perform transform-domain prediction for on-grid blocks in frame n using (3), and code the residuals.
4) Update the transform-domain moving averages of on-grid blocks in frame n, i.e., $S_n(\hat{u}_{n,p}^{k,m})$, $S_n((\hat{u}_{n,p}^{k,m})^2)$, and $S_n(\hat{u}_{n,p}^{k,m} \hat{u}_{n,p-1}^{k,m})$, via (12).

[1]We note that a similar philosophy of using the uncorrelatedness property to overcome the off-grid motion compensation issue was adopted in a completely different context in [5], [7], where we developed an optimal end-to-end distortion estimation for error resilience over lossy networks.

IV. Extension of RSPICE for Sub-Pixel Motion Compensated Coding

The previous section inherently asssumed full-pixel motion compensation when discussing the construction of the reference block $U^k_{n,n-1}$. Sub-pixel motion compensated coding typically adopts a 6-tap FIR filter to produce samples at fractional positions [12]. Therefore, generating a reference block $U^k_{n,n-1}$ located on the sub-pixel grid could potentially involve as many as 9 on-grid blocks (12x12 pixels) of frame $n-1$. Since interpolation and DCT are both linear operations, there exist a new set of construction constants $b_{i,m}$, such that,

$$u^{k,m}_{n,n-1} = \sum_{i=1}^{9}\sum_{m=0}^{15} b_{i,m} x^{k_i,m}_{n-1} \ . \tag{14}$$

In the case of sub-pixel motion compensated reference, the required moving averages are computed with $b_{i,m}$ akin to (7), (8), and (10).

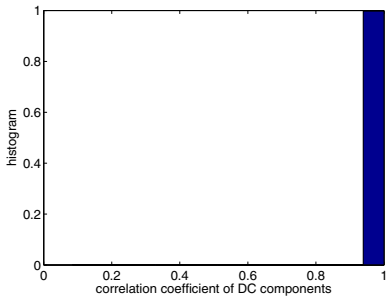
V. Simulation Results

The first part considers the statistical characteristics of natural video sequences, and demonstrates the variations of temporal correlations across transform coefficient frequencies. The second part provides experimental evidence that RSPICE captures the variations in temporal correlation, and ultimately that TDMCP exploits these to improve the coding performance. To simplify the implementation and focus the comparisons, all the sequences are coded in format $IPPP$, although the method is applicable to bidirectional prediction. Also, the deblocking function option is disabled in the experiments, and the motion search is at half-pixel resolution. We note that the proposed approach is extensible to other linear operations.

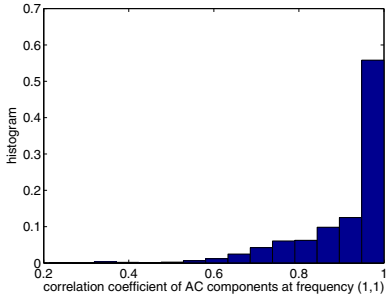
A. Spectral and Temporal Variations in Interframe Correlation

The sequence $foreman$ at $QCIF$ resolution and frame rate $30f/s$ is encoded. The correlation coefficient at each spatial frequency is calculated along the motion trajectory of every on-grid block in frame 10 by employing the RSPICE approach in Sec. III. The distribution of correlation coefficient (i.e., in the range $[0,1]$) for the DC frequency is provided in Fig. 3(a).[2] Note that this distribution is practically an impulse at the the value 1, indicating that there is very little spatial variation in the temporal correlation at DC frequency. Similarly, Fig 3(b) and Fig. 3(c) indicate the distribution of the temporal correlation coefficients at AC spatial frequencies $(1,1)$ and $(2,1)$. In contrast to the DC case, the temporal AR processes at AC frequencies have substantial variation in the correlation coefficient value, i.e., for the same AC frequency the temporal correlation along different motion trajectories that enter the frame can be quite different. Fig. 3(d) again plots the distribution of the correlation at AC frequency $(2,1)$ but by running the RSPICE scheme up to frame 45. Contrasting Fig. 3(c) with Fig. 3(d) provides an illustration of the typical temporal variations in the distribution of the correlation coefficient.

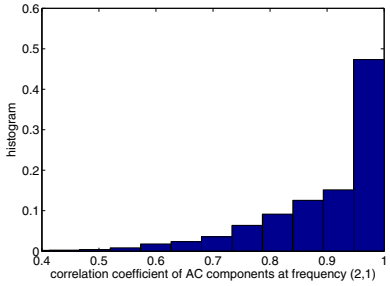
[2]To better visualize the distribution, the following histograms normalize the number of elements in each bin by the total number of blocks in a frame.



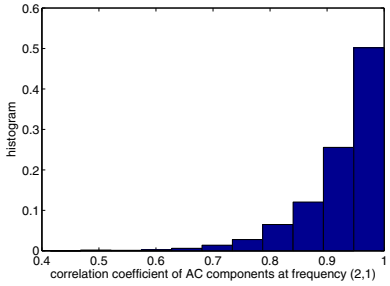
(a) DC Frame 10



(b) AC (1,1) Frame 10



(c) AC (2,1) Frame 10



(d) AC (2,1) Frame 45

Fig. 3. Histogram of correlation coefficient distribution of transform coefficients captured at different time instances. The sequence under test is $foreman$ at $QCIF$ resolution.

B. Coding Performance

The proposed TDMCP approach with the RSPICE scheme for adaptive correlation calculation is implemented in the H.264 framework, and compared against the reference implementation that employs pixel domain subtraction in motion-compensated prediction. Either encoder employs regular pixel domain search for motion vectors at half-pixel resolution. Blocks at fractional pixel positions are generated using the 6-tap FIR filter. Figs. 4-7 compare the the coding performance of the two codecs, standard H.264 and the proposed motion

trajectory adaptive TDMCP, for several typical test sequences at $QCIF$ or CIF resolutions in terms of PSNR at different bit-rates. The proposed approach generally provides gains of 0.5-1dB at different bit rates.

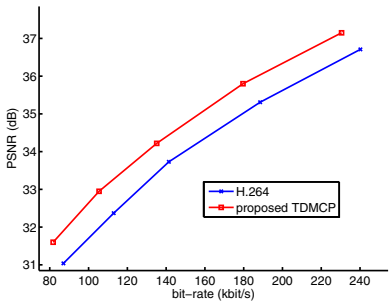


Fig. 4. Coding performance for sequence $foreman$ at $QCIF$ resolution. Both the H.264 and the proposed TDMCP schemes employ the 6-tap FIR filter to generate fractional pixels.

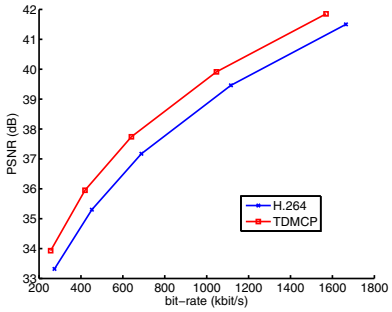


Fig. 5. Coding performance for sequence $foreman$ at CIF resolution. Both the H.264 and the proposed TDMCP schemes employ the 6-tap FIR filter to generate fractional pixels.

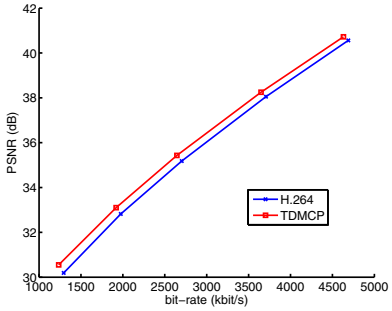


Fig. 6. Coding performance for sequence $mobile$ at CIF resolution. Both the H.264 and the proposed TDMCP schemes employ the 6-tap FIR filter to generate fractional pixels.

VI. Conclusion

We propose a transform-domain motion compensated prediction (TDMCP) approach for video coding that accounts for the true temporal correlations in the AR process at different spatial frequencies along motion trajectories in the video signal. Rather than simply subtracting the motion compensated reference from the current block, as is common in standard pixel domain techniques, the proposed approach first performs

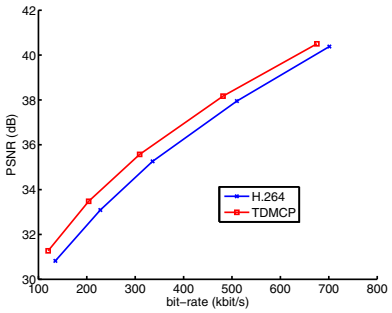


Fig. 7. Coding performance for sequence $coastguard$ at $QCIF$ resolution. Both the H.264 and the proposed TDMCP schemes employ the 6-tap FIR filter to generate fractional pixels.

a spatial transformation on both blocks, and then individually predicts the transform coefficients while optimally exploiting the temporal correlation at each spatial frequency. Such nuances of temporal correlations are hidden from the perspective of standard approaches that employ pixel domain prediction, due to the fact that the high variance of the DC component inundates pixel values. In order to adapt to the temporal and spatial evolution of source statistics, the proposed TDMCP scheme is complemented with a recursive method to adaptively evaluate temporal correlations along motion trajectories that enter each frame. The overall system provides substantial coding gains compared to a standard H.264 codec, while requiring transmission of no additional side-information to the decoder.

References

[1] 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.

[2] 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.

[3] S. J. Choi and J. W. Woods, "Motion-compensated 3D subband coding of video," *IEEE Trans. on Image Processing*, vol. 8, pp. 155–167, Feb 1999.

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

[5] ——, "A recursive optimal spectral estimate of end-to-end distortion in video communications," *Proc. Packet Video*, Dec 2010.

[6] ——, "Transform-domain temporal prediction in video coding: exploiting correlation variation across coefficients," *Proc. IEEE ICIP*, Sep 2010.

[7] ——, "A spectral approach to recursive end-to-end distortion estimation for sub-pixel motion-compensated video coding," *Proc. IEEE ICASSP*, May 2011.

[8] J. R. Ohm, "Three-dimensional subband coding with motion compensation," *IEEE Trans. on Image Processing*, vol. 3, pp. 559–571, Sep 1994.

[9] 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.

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

[11] T. Wedi and H. G. Musmann, "Motion- and aliasing-compensated prediction for hybrid video coding," *IEEE Trans. Circ. Sys. Video Tech.*, vol. 13, pp. 577–586, July 2003.

[12] 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.

[13] 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.