A JOINT SPATIO-TEMPORAL FILTERING APPROACH TO EFFICIENT PREDICTION IN VIDEO COMPRESSION

Yue Chen, Jingning Han†, Tejaswi Nanjundaswamy, and Kenneth Rose

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

ABSTRACT

A novel filtering approach that naturally combines information from both intra-frame and motion compensated referencing for efficient prediction is proposed to fully exploit the spatio-temporal correlations of video signals, thereby achieving superior compression performance. Inspiration was drawn from our recent work on extrapolation filter based intra prediction, which views the spatial signal as a non-separable first-order Markov process and employs a 3-tap recursive filter to effectively capture the statistical characteristics. This work significantly extends the scope to further incorporate motion compensated reference in a filtering framework, whose coefficients were optimized via a "k-modes"-like iteration that accounts for various factors in the compression process including variation in statistics in the prediction loop, to minimize the *rate-distortion* cost. Experiments validate the efficacy of the proposed spatio-temporal approach, which translates into consistent coding performance gains.

Index Terms— Spatio-temporal prediction, extrapolation filter, rate-distortion optimization, video coding

1. INTRODUCTION

Modern video codecs exploit temporal and spatial redundancies in the format of inter and intra predictions, respectively. Inter prediction employs motion compensation to predict from previously coded frames, while Intra prediction generates the prediction from previously reconstructed boundary pixels in the same frame along a given angle to imitate the directionality of the texture content [1][2].

An inter-frame coded block usually has access to multiple information sources, namely, reconstructed top and left boundaries in the same frame, and motion compensated reference block in the prior frames. Current video coders, however, choose amongst the two prediction modes separately, and hence rendering the prediction sub-optimal due to the fact that such ad hoc switch can not fully utilize all the available information.

This work was supported by Google, Inc.

† Jingning Han is now with Google Inc., 1600 Amphitheatre Parkway, Mountain View, CA 94043.

A recent approach to intra prediction building on 2-D separable first order Markov models appeared in [3]. Our own recent work on intra frame coding was premised on the realization of the importance of non-separable models to capture the true spatial correlations [4]. Besides spatial correlations one has to account for temporal correlations, which are typically modeled as a first order (motion compensated) Markov process. This motivates our proposed spatio-temporal filtering approach that efficiently combines information from the available boundaries of the same frame and from the motion compensated reference of previous frames for optimal prediction and hence compression.

Prior work on jointly exploiting spatial and temporal redundancies includes [5], where it predicts a block as a simple linear combination of the inter-frame reference block and the intra predicted block. It largely ignores the variation in statistics across the block and is unable to fully exploit the interaction of spatial and temporal correlations therein. Recent algorithms that exploit all neighboring information to design adaptive optimal predictors inevitably incurs overly expensive computational complexity [6, 7, 8]. In [9], a prediction performance study of higher order spatio-temporal filtering which subsumes sub-pixel motion compensation is presented without recourse to the important step of rate-distortion optimized integration within a video coder.

Built upon our prior work on recursive extrapolation approach to intra prediction [4], we propose a joint spatio-temporal 4-tap prediction filter approach (where 3 tap are for spatial information and 1 tap captures the temporal correlation). It recursively predicts block content from the boundary, and is naturally capable of capturing the variations in spatial correlations in both the current and the motion compensated reference blocks, thereby exploiting all the available information for superior coding performance. The filters are optimized via a "K-modes" like iterative training modified to account for various factors in the predictive coding loop. In particular, it consists three major phases: (1) optimal linear filter estimation (with simplifying Markov model assumptions) to obtain a good initialization for the parameters; (2) direct gradient descent adjustments that do not depend on any model assumptions; and (3) optimization to incorporate rate-distortion optimization process. Experimental results validate

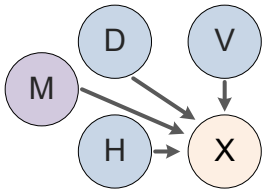

Fig. 1. Spatio-Temporal 3-D Markov model

that the proposed scheme provides substantial coding performance gains over the conventional approach.

2. THE RECURSIVE SPATIO-TEMPORAL PREDICTOR

In [4], we proposed a recursive extrapolation filter to tackle the underutilization of available boundary information in conventional 'pixel copying' based intra prediction. The image signal were modeled by a 2-D non-separable first-order Markov process whose evolution recursion can be written as:

$$X = c_v V + c_h H + c_d D + \epsilon \,, \tag{1}$$

where V, H, and D are neighbors of X, and ϵ denotes the innovation. The coefficients c_v, c_h, and c_d effectively capture the correlation gradients in the 2-D space, or 'directionality' of the image signal. This intra prediction scheme achieved significant coding performance gains.

We extend this framework to further incorporate the temporal correlation:

$$X = c_v V + c_h H + c_d D + c_m M + \epsilon \,, \tag{2}$$

where, M is the motion compensated prediction of X drawn from either one previous frame or as a filtered output of several references, and the coefficient c_m is the weight of temporal reference in the spatio-temporal model. An example illustration of the proposed 3-D model is shown in Fig. 1.

In a medium to high bit-rate setting, the reference pixels are well approximated by their reconstructions, thus we propose to use a linear spatio-temporal predictor:

$$\tilde{X} = c_v \hat{V} + c_h \hat{H} + c_d \hat{D} + c_m \hat{M} \,. \tag{3}$$

An illustration of the proposed prediction paradigm is shown in Fig. 2. One problem we can see from this figure is that in a block-based video coder, only the top-left pixel has previously reconstructed spatial neighbors available. We overcome this limitation by predicting target pixels inside the block from predictions of its neighbors. Specifically, prediction is started from the pixel adjacent to the boundary and then continued recursively in a raster scanning order. This structure, well characterizes the decaying correlation with the

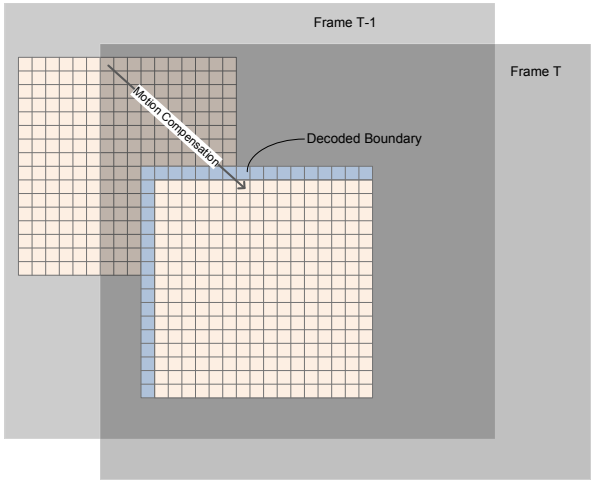

Fig. 2. Spatio-Temporal Prediction Paradigm

spatial boundary across the block and a consistent correlation with temporal reference in the entire block. The choice of filter coefficients controls directionality of spatial prediction and dependency on motion compensation. Note that the proposed filter subsumes conventional intra prediction and motion compensation, e.g., $c_v = 1$, $c_d = 0$, $c_h = 0$ and $c_m = 0$ corresponds to vertical 'pixel copying' intra mode, and $c_v = 0$, $c_h = 0$, $c_d = 0$ and $c_m = 1$ corresponds to pure motion compensation.

We integrate this prediction scheme into a video coder by introducing a set of spatio-temporal prediction modes, tailored for varying texture directionalities and dependency on motion compensation. The prediction filters are designed off-line using training data and embedded in the coder. Note that the additional side information rate to indicate the mode selected is very low when compared to directly transmitting the filter coefficients.

3. FILTER DESIGN

In this section, an off-line design of K candidate spatio-temporal prediction filters for blocks of size $B \times B$ is described. First, frames from a diverse set of video sequences are divided into $B \times B$ blocks to form our training set. The temporal references (M) are produced by employing the regular inter coder to estimate optimal motion compensation with quarter-pixel precision, at a high bit-rate. A variant of K-means clustering is then applied to iteratively partition the training set into clusters (or equivalently, modes), and an optimal spatio-temporal filter is redesigned per cluster/mode, i.e, a "K-modes" iterative approach is employed. We first minimize the mean squared prediction error over the training data and then extend the optimization to account for the overall rate-distortion criterion.

3.1. K-Modes Iterative Clustering to Obtain LMS Filters

Consider the non-separable spatio-temporal Markov model of (2) with known correlation statistics. Let $\underline{c} = [c_v, c_h, c_d, c_m]$. The predictor coefficients minimizing overall squared prediction error are given by

$$\underline{c}_{opt} = \begin{bmatrix} R_{XV} \\ R_{XH} \\ R_{XD} \\ R_{XM} \end{bmatrix}^T \begin{bmatrix} R_{VV} & R_{VH} & R_{VD} & R_{VM} \\ R_{VH} & R_{HH} & R_{HD} & R_{HM} \\ R_{VD} & R_{HD} & R_{DD} & R_{DM} \\ R_{VM} & R_{HM} & R_{DM} & R_{MM} \end{bmatrix}^{-1}, \quad (4)$$

where R_{AB} denotes the cross correlation between A and B.

We initialize the algorithm by clustering the training data only based on the conventional intra prediction modes, corresponding to commonly occurring textures. Then, 2-step iterations involving re-design of filters and re-partitioning described below are executed until convergence. As the overall mean squared prediction error is monotonically non-increasing in every step, convergence is guaranteed.

Re-design of prediction filters: Given a partition of the training set, LMS filters are derived from the statistics of each subgroup. The directional cross-correlations in equation (4) are estimated using the original pixel values, $x_{i,j}$, and the motion compensation reference, $x^M_{i,j}$. For instance, R_{XV} and R_{HM} is estimated as

$$\begin{aligned} R_{XV} &= \sum (x_{i,j} - \overline{x})(x_{i+1,j} - \overline{x}), \\ R_{VM} &= \sum (x_{i+1,j} - \overline{x})(x^M_{i,j} - \overline{x}^M), \end{aligned} \quad (5)$$

where, $\overline{x}$ is the block mean of original pixels values, and $\overline{x}^M$ is the block mean of motion compensated reference. This ensures reduction in overall squared prediction error as each of the K filters best serve the subset they represent.

Re-partitioning of the training set: Each training block is now assigned to the mode minimizing squared prediction error. This again ensures reduction in overall squared prediction error due to the reduction of error for each block.

After convergence we have LMS solution for a set of K spatio-temporal filters.

3.2. Gradient Decent Approach to Minimize Actual Prediction Error

A second phase of design is motivated by the recognitions that, there is possible mismatches between the Markov model and real signals, and in block-wise prediction, some pixels are predicted from boundary and others are predicted from predictions of pixels, which is ignored in the first phase. Starting from the optimal filter in Section 3.1, a gradient descent approach is used to minimize the prediction error resulting from applying the 4-tap recursive filter to the entire block. The "K-modes" technique is again employed to iterate between assigning modes to each block and re-optimizing the filters by gradient descent approach in each mode. To minimize the squared prediction error $J = \sum_{\forall \text{ blocks}} \sum_{i,j} (\tilde{x}_{i,j} - x_{i,j})^2$, we do line search along the negative of the gradient of squared prediction error for each mode. For a mode k, the gradient is given as,

$$\nabla_k = \begin{bmatrix} \sum_{\text{blocks} \in \text{ mode } k} \sum_{i,j} \frac{\partial (\tilde{x}_{i,j} - x_{i,j})^2}{\partial c_{v,k}} \\ \sum_{\text{blocks} \in \text{ mode } k} \sum_{i,j} \frac{\partial (\tilde{x}_{i,j} - x_{i,j})^2}{\partial c_{h,k}} \\ \sum_{\text{blocks} \in \text{ mode } k} \sum_{i,j} \frac{\partial (\tilde{x}_{i,j} - x_{i,j})^2}{\partial c_{d,k}} \\ \sum_{\text{blocks} \in \text{ mode } k} \sum_{i,j} \frac{\partial (\tilde{x}_{i,j} - x_{i,j})^2}{\partial c_{m,k}} \end{bmatrix}^T, \quad (6)$$

where the vector elements are partial derivatives with respect to the filter coefficients. The partial derivative with respect to the vertical coefficient is,

$$\begin{aligned} \frac{\partial J}{\partial c_{v,k}} &= \sum_{\text{blocks} \in \text{ mode } k} \sum_{i,j} \frac{\partial (\tilde{x}_{i,j} - x_{i,j})^2}{\partial c_{v,k}} \\ &= \sum_{\text{blocks} \in \text{ mode } k} \sum_{i,j} 2\,(\tilde{x}_{i,j} - x_{i,j})\,\frac{\partial \tilde{x}_{i,j}}{\partial c_{v,k}}. \end{aligned} \quad (7)$$

where $\frac{\partial \tilde{x}_{i,j}}{\partial c_{v,k}}$ can be derived using (3), as

$$\frac{\partial \tilde{x}_{i,j}}{\partial c_{v,k}} = \tilde{x}_{i-1,j} + c_{v,k}\frac{\partial \tilde{x}_{i-1,j}}{\partial c_{v,k}} + c_{h,k}\frac{\partial \tilde{x}_{i,j-1}}{\partial c_{v,k}} + c_{d,k}\frac{\partial \tilde{x}_{i,j-1}}{\partial c_{v,k}}.$$

Similarly, the other partial derivatives can be derivated through

$$\begin{aligned} \frac{\partial \tilde{x}_{i,j}}{\partial c_{h,k}} &= \tilde{x}_{i,j-1} + c_{v,k}\frac{\partial \tilde{x}_{i-1,j}}{\partial c_{h,k}} + c_{h,k}\frac{\partial \tilde{x}_{i,j-1}}{\partial c_{h,k}} + c_{d,k}\frac{\partial \tilde{x}_{i,j-1}}{\partial c_{h,k}}, \\ \frac{\partial \tilde{x}_{i,j}}{\partial c_{d,k}} &= \tilde{x}_{i-1,j-1} + c_{v,k}\frac{\partial \tilde{x}_{i-1,j}}{\partial c_{d,k}} + c_{h,k}\frac{\partial \tilde{x}_{i,j-1}}{\partial c_{d,k}} + c_{d,k}\frac{\partial \tilde{x}_{i,j-1}}{\partial c_{d,k}}, \\ \frac{\partial \tilde{x}_{i,j}}{\partial c_{m,k}} &= x^M_{i,j} + c_{v,k}\frac{\partial \tilde{x}_{i-1,j}}{\partial c_{m,k}} + c_{h,k}\frac{\partial \tilde{x}_{i,j-1}}{\partial c_{m,k}} + c_{d,k}\frac{\partial \tilde{x}_{i,j-1}}{\partial c_{m,k}}. \end{aligned}$$

Even in this design phase the overall squared prediction error is monotonic non-increasing in every step, guaranteeing convergence.

3.3. Overall Rate-Distortion Optimization

Until here the filter design minimized the squared prediction error, however, the performance of a video codec is evaluated according to the rate-distortion cost. This cost has complex dependency on both the quantization error energy, and bit-rate required to encode the mode indices and the quantized transformed prediction residuals. In such a complex system, simply minimizing the prediction residue need not improve the codec performance. Thus the ultimate rate-distortion criteria of a video coder is taken into account in the third filter design phase by building on the training results of Section 3.2.

Let $c_i,\ 1 \le i \le 4K$, denote all the filter coefficients. First, the prediction filters obtained from Section 3.2 are

incorporated into the encoder, and the initial rate-distortion cost, L_{opt}, is calculated by running the encoder. Then, each filter coefficient is fine tuned by running the following iterations until convergence:

1. Update $c_i = c_i + \Delta$, calculate the new rate-distortion cost L.
2. If $L < L_{opt}$, update $L_{opt} = L$, repeat Step 1. If not, update $c_i = c_i - \Delta$, continue to Step 3.
3. Update $c_i = c_i - \Delta$, calculate the new rate-distortion cost L.
4. If $L < L_{opt}$, update $L_{opt} = L$, repeat Step 3. If not, update $c_i = c_i + \Delta$, continue to Step 1.

4. EXPERIMENTAL RESULTS

The above spatio-temporal prediction approach was implemented in the VP9 reference framework as a new coding option in addition to the inter and intra prediction modes available for the inter frames. A preliminary experiment that applied $K = 7$ filters to prediction of blocks using a single frame as motion compensation reference and ranging from 8×8 to 32×32 was included to validate the potential of the proposed scheme. The filter coefficients were optimized using the iterative "K-modes" approach of Sec. 3. The test sequences (all apart from those used as training set to obtain the filter coefficients) were coded in $IPPP$ format and the performance gains, in terms of BD-rate[10], over the reference VP9 codec are presented in Table 1. Our spatio-temporal filtering scheme achieves consistent gains over the conventional separate inter/intra coding, even under limited and and preliminary settings. Future directions include enabling the proposed approach for prediction of blocks using compound-frames as motion compensation reference and also for the remaining block sizes, hence fully exploiting its efficacy.

Table 1. BD-rate reduction due to the spatio-temporal prediction approach relative to the VP9 Inter coder.

Test Sequence	Resolution	Bit Savings (%)
$foreman$	CIF	0.274
$football$	CIF	0.820
$crew$	CIF	1.072
ice	CIF	0.539
$city$	720p	0.615

5. CONCLUSION

A novel recursive filtering approach was proposed to jointly exploit spatio-temporal redundancy of video signal for optimal prediction. It builds on a non-separable first-order Markov model, which well approximates the underlying statistical characteristics. The requisite filter coefficients, which effectively capture the spatial and temporal correlations, were optimized using a "k-modes" like framework that iteratively minimizes the overall rate-distortion cost. It was experimentally demonstrated that the proposed scheme achieved consistent performance gains.

6. REFERENCES

[1] J. Bankoski, R.S. Bultje, A. Grange, Q. Gu, J. Han, J. Koleszar, D. Mukherjee, P. Wilkins, and Y. Xu, "Towards a next generation open-source video codec," *IS&T/SPIE Electronic Imaging*, 2013.

[2] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the h.264/avc video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 560–576, 2003.

[3] F. Kamisli, "Intra prediction based on markov process modeling of images," *IEEE Transactions on Image Processing*, vol. 22, pp. 3916–3925, Oct. 2013.

[4] Y. Chen, J. Han, and K. Rose, "A recursive extrapolation approach to intra prediction in video coding," *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, 2013.

[5] J. Xin, K. N. Ngan, and G. Zhu, "Combined inter-intra prediction for high definition video coding," *Picture Coding Symposium*, 2007.

[6] J. Seiler and A. Kaup, "Spatio-temporal prediction in video coding by spatially refined motion compensation," *IEEE International Conference in Image Processing (ICIP)*, pp. 2788–2791, 2008.

[7] J. Seiler, H. Lakshman, and A. Kaup, "Spatio-temporal prediction in video coding by best approximation," *Picture Coding Symposium*, pp. 1–4, 2009.

[8] J. Seiler, T. Richter, and A. Kaup, "Spatio-temporal prediction in video coding by non-local means refined motion compensation," *Picture Coding Symposium*, pp. 318–321, 2010.

[9] I. Matsuda, K. Unno, H. Aomori, and S. Itoh, "Block-based spatio-temporal prediction for video coding," *Proc. European Signal Processing Conference*, pp. 2052–2056, 2010.

[10] G. Bjontegaard, "Calculation of average psnr differences between rd curves," *ITU-T SC16/Q.6 VCEG-M33*, Apr. 2001.