ASYMPTOTIC CLOSED-LOOP DESIGN OF TRANSFORM MODES FOR THE INTER-PREDICTION RESIDUAL IN VIDEO CODING

Bharath Vishwanath, Shunyao Li, and Kenneth Rose

Department of Electrical and Computer Engineering
University of California, Santa Barbara, CA 93106
$\{bharathv, shunyao\_li, rose\}$@$ece.ucsb.edu$

ABSTRACT

Transform coding is a key component of video coders, tasked with spatial decorrelation of the prediction residual. There is growing interest in adapting the transform to local statistics of the inter-prediction residual, going beyond a few standard trigonometric transforms. However, the joint design of multiple transform modes is highly challenging due to critical stability problems inherent to feedback through the codec's prediction loop, wherein training updates inadvertently impact the signal statistics the transform ultimately operates on, and are often counter-productive (and sometimes catastrophic). It is the premise of this work that a truly effective switched transform design procedure must account for and circumvent this shortcoming. We introduce a data-driven approach to design optimal transform modes for adaptive switching by the encoder. Most importantly, to overcome the critical stability issues, the approach is derived within an asymptotic closed loop (ACL) design framework, wherein each iteration operates in an effective open loop, and is thus inherently stable, but with a subterfuge that ensures that, asymptotically, the design approaches closed loop operation, as required for the ultimate coder operation. Experimental results demonstrate the efficacy of the proposed optimization paradigm which yields significant performance gains over the state-of-the-art.

Index Terms— inter-prediction, multi-modal transforms, asymptotic closed-loop, spatial transform

1. INTRODUCTION

Transform coding is an essential component in image and video compression, wherein it is applied to the prediction residual after a block of pixels has undergone intra or inter-prediction. Here, the objective of transform coding is to eliminate spatial correlations in the prediction residual, and hence achieve energy compaction in the transform domain. Given known and stationary signal statistics, it is well known that the Karhunen- Loève transform (KLT) is the optimal decorrelating transform. However, its dependency on signal statistics and its high computational complexity compromise its practicality. Instead, the discrete cosine transform (DCT) has been the most widely adopted transform due to its fast implementation and good energy compaction property, as well as the theoretical justification provided to its ability to approximate performance of KLT on certain Gauss-Markov processes [1].

Recently, there has been growing interest in switched transforms that adapt to variations in signal statistics. Much of the work on such transform design focused on the intra-prediction residual, including the derivation of asymmetric trigonometric transforms to leverage the directionality of intra-prediction, which were further shown to approach KLT optimality under mild Markovian assumption [2] as well as several other approaches to mode-dependent transforms (e.g., [3], [4]). The design of transforms for inter-prediction residuals, however, attracted significantly less attention, perhaps due to the fact such transforms do not exhibit as "obvious" properties such as the directionality inherent to intra prediction modes, and are hence more challenging to design. It is nevertheless important to note that the vast majority of video blocks are predicted temporally, which implies that progress here is likely to have more impact on the overall performance. This motivates the focus of this paper, namely, transforms optimization for inter-prediction residual statistics. The latest open source codec AV1 [5] allows switching within a set of known trigonometric transforms such as DCT and ADST in order to capture some additional gains. The authors in [6] also propose to use known trigonometric transforms for inter-prediction residuals which was later adopted in JEM codec [7]. However, to realize the full potential of multi-modal transforms, it is necessary to look beyond the known trigonometric transforms, and employ a data-driven approach that statistically learns the optimal set of transforms. Few recent contributions (e.g., [8]) propose online learning of transforms. As such an approach significantly increases the computational complexity of the encoder, we will focus on the practical alternative of an offline design paradigm. For a recent approach in this vein see [9], where residue statistics are collected from a training set and the resulting KLT is given as an option during encoding. Other

This work was supported by Google Inc.

relevant approaches include the 1-D transforms developed in [10], directional DCTs in [11], row column transforms in [12] and layered-Givens transforms in [13]. All these approaches largely ignore what is a critical difficulty (a variant on the proverbial "chicken and egg" problem) in closed-loop iterative design of modules of a predictive coding system. Specifically, in the case of transforms, an updated transform changes the reconstructions, which in turn modify the prediction residual statistics on which the transform update was premised. This fundamental difficulty strongly motivates the work in this paper, and is further discussed below.

A major challenge in the joint design of multiple transform modes is due to the instability inherent to the closed-loop design of the coder. Updated transforms are applied to prediction residuals to obtain new reconstructions, which in turn affect the prediction residual statistics. This complex interplay between the transforms and reconstructions makes effective transform design quite elusive. Standard closed-loop design often suffers from significant (and sometimes catastrophic) design instability due to error propagation through the prediction loop. An effective remedy, called asymptotic closed-loop (ACL) design, was proposed in [14] in the context of predictor and quantizer design. In this paper we extend the ACL paradigm to effective transform design. Specifically, transforms are designed iteratively, in an open loop that ensures design stability, but with a subterfuge that guarantees that upon convergence, the transforms are optimal for closed loop operation. Thus, in this paper, we use ACL as a stable hence effective platform for designing multi-mode transforms. Note that, while the focus is on the design of separable transforms which are preferred due to their lower complexity, the proposed design paradigm is general and applicable to non-separable transforms.

2. BACKGROUND

2.1. Separable KLT

Let $\mathbf{e}$ be a random vector of (say, prediction residual) samples, whose covariance matrix is $\mathbf{C_e}$. Let $\mathbf{T}$ be the transform matrix. The transform-domain signal vector $\mathbf{y}$ is given by

$$\mathbf{y} = \mathbf{Te}, \tag{1}$$

and its covariance matrix $\mathbf{C_y}$ is

$$\mathbf{C_y} = \mathbf{TC_eT}' \tag{2}$$

The optimal transform that diagonalizes $\mathbf{C_y}$, i.e., decorrelates the components of $\mathbf{y}$ is precisely KLT, whose basis vectors are the eigenvectors of $\mathbf{C_e}$.

In the context of video coding, let $\mathbf{E}$ be the random prediction residual block. Let $\mathbf{T_r}$ and $\mathbf{T_c}$ be the respective KLTs for the row covariance $\mathbf{C_r}$ and column covariance $\mathbf{C_c}$, of $\mathbf{E}$. The transform-domain signal can be written as,

$$\mathbf{Y} = \mathbf{T_cET_r}' \tag{3}$$

KLTs are optimal for given statistics of the prediction residual signal. But updating the transforms changes the reconstructions and hence also the residual signal statistics, requiring a new KLT calculation. Thus, transform design requires an iterative procedure. Next we summarize the standard iterative approach.

2.2. Closed-Loop Design

Standard closed-loop techniques (see e.g., [15]), when applied to transform design, employ transforms trained on the residual sequence of iteration i to transform the residue in the next iteration $i+1$, i.e,

$$\mathbf{y}^{i+1} = \mathbf{T}^i\mathbf{e}^{i+1} \tag{4}$$

where the residual $\mathbf{e}^{i+1} = \mathbf{x}_n - \hat{\mathbf{x}}_{n-1}^{i+1}$ is the prediction error in iteration $i+1$ (assuming prediction coefficient of one, as is common practice in video coding). Thus transform $\mathbf{T}^i$, optimal for the previous residual sequence $\{\mathbf{e}^i\}$, is in fact applied to a potentially very different residual sequence in iteration $i+1$. This results in statistical mismatch which tends to grow as errors propagate in the prediction loop, and the resulting instability may prove catastrophic at low rates. Fig. 1 illustrates closed-loop design.

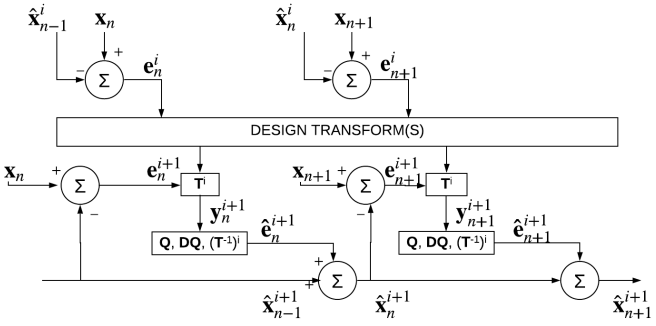

Fig. 1. Closed-loop design

3. PROPOSED METHOD

In this section, we propose a stable design paradigm for learning the transforms. Before delving into the design, we note that the inter-prediction residual exhibits significant variations in statistics. To cover a wide spectrum of statistics, we propose to design "super-modes" of transforms, wherein each super-mode is a collection of M transform modes. The adaptivity is such that the encoder can switch between super-modes at the group-of-pictures (GOP) level, and can further switch between the transforms in a super-mode at the block level. Thus, the problem at hand is to design a set of S

super-modes $\{\mathbf{T_s}\}, s = 1, 2..S$, wherein each super-mode s consists of M pairs of row and column transform modes denoted $\{\{\mathbf{T_{s,m,r}}, \mathbf{T_{s,m,c}}\}\}, m = 1, 2..M$. An iterative design technique is needed to optimize transforms and update reconstructions. A clustering based framework is presented first, to enable super-modes design, given a training set of residual sequences.

3.1. Clustering

Let $\{\mathbf{E}^i_{b,n,g}\}$ be the training sequence of prediction residual where $\mathbf{E}^i_{b,n,g}$ is block b in frame n of GOP g, obtained by subtracting from source block $\mathbf{X}_{b,n,g}$ its motion-compensated prediction $\hat{\mathbf{X}}^i_{b^{mv},n-1,g}$

$$\mathbf{E}^i_{b,n,g} = \mathbf{X}_{b,n,g} - \hat{\mathbf{X}}^i_{b^{mv},n-1,g} \tag{5}$$

To design the super-modes, we employ an algorithm in the spirit of "K-means clustering", which iterates between assigning to each GOP the best super-mode super-mode ("nearest neighbor" step) and then optimizing the super-modes to match their GOP clusters ("centroid" step), which specifically means designing M transform modes that optimally match the statistics of all GOPs that share the super-mode. These M modes are again designed in a "K-means clustering" fashion, where blocks are assigned to M modes followed by optimal row and column transforms design for each mode. Note that the mode assignment decisions are RD-optimal and take into account the total cost of coding the transform coefficients and signaling these modes to the decoder. This constitutes the re-estimation of the super-modes. With the designed super-modes, the GOPs are re-clustered and the process is repeated until convergence.

We next consider how to embed within the approach an ACL paradigm for transform design so as to avoid the notorious instability of closed-loop design.

3.2. Asymptotic Closed Loop Design

As discussed in 2.2, the main shortcoming of the closed-loop approach is the design instability due to error propagation in the prediction loop. ACL design effectively resolves the stability issue by updating the reconstructions in an open-loop fashion as illustrated in Fig. 2. The updated transforms are used with the same set of residual sequences for which they were designed. This ensures increasingly better reconstructions over the iterations. However, on convergence, the reconstructed sequence remains essentially unchanged. Therefore, predicting from the previous iteration's reconstructions approaches equivalence with predicting from the current iteration, i.e., it effectively operates in closed-loop. Thus, ACL asymptotically optimizes transforms for closed-loop operation. For the problem at hand, given optimal super-modes from a design iteration i, the transform signal is obtained as,

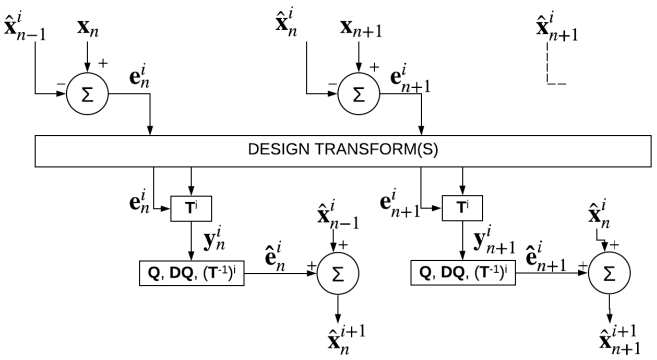

Fig. 2. Asymptotic closed-loop design

$$\mathbf{Y}^i_{b,n,g} = \mathbf{T}^i_{\mathbf{c},\text{best}} \mathbf{E}^i_{b,n,g} \mathbf{T}^{i\prime}_{\mathbf{r},\text{best}} \tag{6}$$

where $\mathbf{T}^i_{\mathbf{c},\text{best}}, \mathbf{T}^i_{\mathbf{r},\text{best}}$ are the best row and column transforms chosen by the encoder from the transforms designed in iteration i. This is followed by, quantization, de-quantization and inverse transform to obtain the block $\hat{\mathbf{E}}^i_{b,n,g}$. The reconstructions are updated as,

$$\hat{\mathbf{X}}^{i+1}_{b,n,g} = \hat{\mathbf{X}}^i_{b^{mv},n-1,g} + \hat{\mathbf{E}}^i_{b,n,g} \tag{7}$$

The overall design procedure has been illustrated in Algorithm 1. First, a closed-loop initialization is performed to obtain a reconstructions sequence and random assignment of GOPs to super-modes. Standard trigonometric transforms are used as initialization for the M transform modes in each super-mode. The algorithm then iterates between super-modes design for a given residual statistics and the reconstruction update in ACL fashion. Note that, after a reconstruction update, we update the encoder decisions including the motion vectors, ensuring optimal encoder decisions for the new reconstructions. We use these decisions to generate prediction residual statistics for the next iteration. Upon convergence, both the reconstructions and the encoder decisions remain the same, and hence the system effectively operates in closed-loop.

4. EXPERIMENTAL RESULTS

4.1. Main Results for VP9

We designed transforms for the VP9 codec, whose experimental features included a set of sixteen separable transform modes, namely, the four transforms { DCT, ADST, FlipADST, IDTX } are available as row and column transforms. This experimental feature of VP9 is now a part of AV1. We considered the nine separable trigonometric transform modes obtained as row-column transform combinations of { DCT, ADST, FlipADST }, since the identity transform (IDTX) is mostly intended for screen content sequences. To

initialize: reconstructed sequence from closed loop encoder, super-mode assignment, transform modes;
while *ACL_iter* $<$ *max_ACL_iter* **do**
  Generate residue statistics;
  while *Super_mode_iter*$<$ *Max_super_mode_iter* **do**
    (a) Update super-modes:
    while *M_mode_iter*$<$ *Max_M_mode_iter* **do**
      (i) Assign best transform-mode to each block ;
      (ii) Design KLTs for each mode ;
      break on convergence;
    end
    (b) Assign best super-mode to each GOP ;
    break on convergence;
  end
  Update reconstructions in ACL fashion ;
  Update encoder decisions with new transforms;
  break on convergence ;
end

Algorithm 1: Overall design approach

simplify experiments, the block size was fixed at 8x8. The training set consisted of nine cif sequences { bridge-far, mobile, bride-close, highway, foreman, tempete, flower, city, bus }. We trained four super-modes where each super-mode consisted of nine separable transform modes. The training proved to be sensitive to initialization, and we employed multiple initializations for the super-mode assignment. The training was done at constant bit-rate configuration of VP9. The target bit-rates were chosen to be 200, 300, 500 and 800 Kbps. Since the residue statistic changes with bit-rate, we design four super-modes for each target bit-rate. The base-line codec uses only DCT as the transform for inter-prediction residual. Bit-rate reduction over the baseline is calculated as per [16]. The results for the test set sequences (cif resolution) are shown in Table 1. The default trigonometric transforms now used by AV1 yield on average 2.5% bit-rate reduction. Designing super-modes with closed-loop (CL) design yields 4.6% bit-rate reduction over base-line codec. The proposed method with ACL design paradigm gains 3.5% and 1.4 % over the trigonometric transforms and closed-loop design respectively and yields significant gains of 6% over the base-line codec.

4.2. Preliminary Results with AV1

To obtain very preliminary results for AV1, we replaced its trigonometric transforms with the transforms we had designed for VP9. The target bit-rate is reduced here to operate in the same range of PSNRs as in the previous experiment. Bit-rate reduction over AV1 for the test set is presented in Table. 2. Note that, despite the fact that the transforms were not designed directly on the residual statistics of AV1, they already offer half the gains, namely, 1.7% bit-rate reduction on average. Experimental work to train the transforms within the AV1 framework, is underway.

Test Sequence	Trigonometric Transforms	CL Design	Proposed ACL Design
silent	2.0	2.9	3.5
soccer	2.7	3.8	5.4
akiyo	3.1	6.0	8.1
bowing	-0.18	8.3	9.3
hall	1.2	3.3	3.7
mother-daughter	3.2	4.4	6.5
paris	2.3	3.0	3.8
coastguard	4.0	4.8	6.4
stefan	3.6	4.4	4.7
ice	4.1	5.6	9.0
Average	**2.6**	**4.6**	**6.0**

Table 1. VP9 experiment: % bit-rate savings on test set, for Y component over base-line VP9 (uses DCT only)

Test Sequence	Bit-rate Savings over AV1
silent	1.1
soccer	1.0
akiyo	5.2
bowing	0.5
hall	0.6
mother-daughter	0.5
paris	0.8
coastguard	2.0
stefan	0.7
ice	3.8
Average	**1.7**

Table 2. Preliminary AV1 experiment (with transforms designed for VP9): % bit-rate savings on test set for Y component over AV1 (uses trigonometric transforms).

5. CONCLUSIONS

This paper presents an efficient offline-design procedure to learn transforms for inter-prediction residuals. Critical design instability was circumvented by deriving the method within the asymptotic-closed loop framework. Significant bit-rate reduction substantiates the potential of this data-driven approach to effectively learn transforms and outperform standard trigonometric transforms.

6. REFERENCES

[1] K. R. Rao and P. Yip, *Discrete cosine transform: algorithms, advantages, applications*, Academic press, 2014.

[2] J. Han, A. Saxena, V. Melkote, and K. Rose, "Jointly optimized spatial prediction and block transform for video and image coding," *IEEE Transactions on Image Processing*, vol. 21, no. 4, pp. 1874–1884, 2011.

[3] C. Yeo, Y. H. Tan, and Z. Li, "Low-complexity mode-dependent klt for block-based intra coding," in *2011 18th IEEE International Conference on Image Processing*. IEEE, 2011, pp. 3685–3688.

[4] A. Arrufat, P. Philippe, and O. Déforges, "Non-separable mode dependent transforms for intra coding in hevc," in *2014 IEEE Visual Communications and Image Processing Conference*. IEEE, 2014, pp. 61–64.

[5] Y. Chen, D. Murherjee, J. Han, A. Grange, Y. Xu, Z. Liu, S. Parker, C. Chen, H. Su, U. Joshi, et al., "An overview of core coding tools in the AV1 video codec," in *2018 Picture Coding Symposium (PCS)*. IEEE, 2018, pp. 41–45.

[6] X. Zhao, J. Chen, M. Karczewicz, A. Said, and V. Seregin, "Joint separable and non-separable transforms for next-generation video coding," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2514–2525, 2018.

[7] J. Chen, M. Karczewicz, Y. Huang, K. Choi, J-R. Ohm, and G. J. Sullivan, "The joint exploration model (JEM) for video compression with capability beyond hevc," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[8] C. Lan, J. Xu, W. Zeng, G. Shi, and F. Wu, "Variable block-sized signal-dependent transform for video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 8, pp. 1920–1933, 2017.

[9] K. Fan, R. Wang, W. Lin, L. Y. Duan, and W. Gao, "Signal-independent separable KLT by offline training for video coding," *IEEE Access*, vol. 7, pp. 33087–33093, 2019.

[10] F. Kamisli and J. Lim, "1-D transforms for the motion compensation residual," *IEEE Transactions on Image Processing*, vol. 20, no. 4, pp. 1036–1046, 2010.

[11] B. Zeng and J. Fu, "Directional discrete cosine transforms—A new framework for image coding," *IEEE transactions on circuits and systems for video technology*, vol. 18, no. 3, pp. 305–313, 2008.

[12] H. E. Egilmez, O. G. Guleryuz, J. Ehmann, and S. Yea, "Row-column transforms: Low-complexity approximation of optimal non-separable transforms," in *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016, pp. 2385–2389.

[13] B. Li, O. G. Guleryuz, J. Ehmann, and A. Vosoughi, "Layered-Givens transforms: Tunable complexity, high-performance approximation of optimal non-separable transforms," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 1687–1691.

[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] P-C. Chang and R. Gray, "Gradient algorithms for designing predictive vector quantizers," *IEEE transactions on acoustics, speech, and signal processing*, vol. 34, no. 4, pp. 679–690, 1986.

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