Joint Asymptotic Closed-Loop Design of Secondary Transform and Scan Order for Inter Coding in AV1

Kruthika Koratti Sivakumar, Bharath Vishwanath, Kenneth Rose
Department of Electrical and Computer Engineering
University of California, Santa Barbara
Santa Barbara, CA, 93106
{ kruthika, bharathvishwanath, kenrose } @ucsb.edu

***Abstract*—Most video coding systems employ separable transforms due to their low computational complexity and storage requirements, and despite their sub-optimal decorrelating capabilities. To achieve better decorrelation, recent codecs further apply a non-separable secondary transform to low-frequency primary transform coefficients of the intra-prediction residual. This paper focuses on effective design of non-separable secondary transforms for the *inter*-prediction residual. As the combination of primary and secondary transforms yields a set of ultimate transform coefficients for which the default zig-zag scan order is sub-optimal, we complement the secondary transform design with the design of corresponding coefficient scanning order modes that facilitate effective entropy coding. A critical stability challenge in this joint design, due to error propagation through the codec's prediction loop, is circumvented by leveraging the asymptotic closed loop (ACL) design paradigm. ACL operates in open-loop in each iteration to ensure stability, but in a manner that ultimately converges to parameter optimization for closed-loop operation. Experimental results show an average gains in BD rate of 0.76% for CIF and 0.41% for HD sequences, over the AV1 codec (libaom).**

***Index Terms*—AV1, inter-prediction, non-separable secondary transforms, coefficient scan order**

I. Introduction

Transform coding is a central component of state-of-the-art video coding. The intra or inter prediction residual is projected to the transform domain to achieve spatial decorrelation and gain energy compaction. Given the signal statistics, the optimality of the Karhunen-Loève transform (KLT) is well established (see, e.g., [1]). However, due to its computational complexity and dependence on signal statistics, KLT is often replaced with the discrete cosine transform (DCT), a fixed transform with fast implementation, which is known to closely approximate KLT in terms of its energy packing efficiency for certain Gauss-Markov processes [2]. Over the years, in addition to DCT, variants of the discrete sine transform (DST) have also been widely adopted to cater to blocks that exhibit one-sided smoothness [3], a scenario often encountered in intra-predicted blocks. One such example is [4], where asymmetric DST is analytically derived and shown to be optimal for the intra-prediction residual under certain Markov assumptions. Recently, there has also been a growing interest in switched transforms for the inter-predicted residual, which adapt to variations in signal statistics. Specifically, the AV1 codec [5] performs separable 2-D transformation by using combinations of 4 different 1-D kernels: DCT, asymmetric DST, flipped asymmetric DST and identity transform, resulting in a total of 16 2-D primary transform kernels. However, the optimality of the above trigonometric transforms is only ensured under certain model assumptions, the validity of which was justified for intra-prediction but is questionable for inter-prediction. Moreover, for complexity reasons, these trigonometric transforms are commonly used in a separable setting, and hence provide spatial decorrelation only in the horizontal and vertical directions.

Significant contributions have been made with the introduction of non-separable secondary transforms, which are performed on a subset of primary (separable) transform coefficients to achieve better spatial decorrelation. Typically, a non-separable secondary transform is performed on the top-left, i.e., low frequency primary transform coefficients, since they contain most of the residual energy. This helps reduce the required computational and storage complexities as opposed to applying a non-separable KLT on the entire block. In [6]-[10] various secondary transform schemes were proposed. In [6], trigonometric transforms are applied to the intra and inter-prediction residuals, and non-separable secondary transforms are designed to further transform the intra-residual. Further reduction in computational complexity is achieved by the low-frequency non-separable secondary transform (LFNST) scheme of [7], by zeroing out higher frequency coefficients to reduce the required transform kernel dimension, which has been included as a coding tool in the versatile video coding (VVC) standard [8]. In [9], a a non-separable unified secondary transform is proposed, which uses a single kernel size for different block sizes, while [10] exploits the correlation between the transform kernel choice of a block with the variance of its reconstructed neighbors, to avoid the cost of signaling. The aforementioned approaches focus on the intra-prediction residual since it often exhibits directionality that separable transforms fail to exploit. Conversely, secondary transform design for inter-prediction residual has not attracted much attention, mostly owing to the lack of distinct directionality features. However, given that a vast subset of the video blocks is predicted temporally, it is important to investigate possible

This work was supported in part by Google, Inc.

improvements in this area.

This paper is motivated by two main shortcomings of existing approaches to secondary transform design. First, they suffer from the inherent instability of iterative closed-loop design, which is due to error propagation through the prediction loop. More specific to the transform design, in a single run of the encoder, a newly designed set of transforms generates new reconstructions, wherein each reconstructed block affects the prediction of subsequent blocks and, in turn, modifies the residual statistics. Thus, the transforms are applied to a different set of residue statistic than the one they were originally designed for. The resulting error from the statistical mismatch propagates in the prediction loop, grows as the encoder proceeds further down the sequence, and causes design instability that may become catastrophic. Although this issue is less critical in intra-coding, where instability is limited to a single frame, it is a major hurdle in optimizing transforms for inter-prediction residual. We overcome this design instability by using the asymptotic closed-loop (ACL) algorithm that was originally proposed in the context of predictor and quantizer design [11]. In ACL, reconstructions are updated in an open-loop setting by applying transforms on the same residue statistic that they were designed for, thereby ensuring design stability. However, the algorithm asymptotically converges to closed-loop operation making the designed transforms optimal for the closed-loop operation of the codec.

The second major drawback of existing approaches is that they ignore the fact that coefficients produced by the secondary transform are inherently different from the remaining primary transform coefficients within which they are embedded, which raises doubts regarding the use the original scanning order. In [12], a method is proposed in which secondary transforms are performed on coefficients chosen in descending zigzag scan order (the same as the ultimate coefficient scan order for entropy coding) rather than based on raster scan order of the block, which achieves better decorrelation of neighboring coefficients. This approach ensures that the *secondary transform coefficients* are scanned in decreasing order of variance, though it relies on scanning order techniques designed for separable transform coefficients. Nevertheless, optimization of the scan order is required to realize the full potential of the secondary transforms.

The contributions of the method proposed herein include a technique for secondary transform design that circumvents the closed-loop design instability problem and whose scan order exploits the actual energy distribution across the ultimate transform coefficients. A related earlier work from our lab [13] proposed an ACL-based method for the design of separable transform modes for the inter-prediction residual. It focused on primary transforms to *replace* the existing trigonometric transforms used by VP9. Thus the contributions of the current paper can complement the prior work with a technique to design secondary transforms and scan order for inter-prediction residuals. The included experimental results will focus exclusively on the gains achieved by these contributions, although further gains are expected if and when the two approaches are combined.

The rest of the paper is organized as follows. Section II provides background on non-separable KLT transforms. Section III discusses the need for the design of scanning orders. Section IV describes the pitfalls of closed-loop design and how ACL design circumvents its inherent instability. Section V presents the overall algorithm description and pseudo-code. Section VI covers the experimental results and gains over the libaom codec, followed by conclusions in section VII.

II. Background

Given the signal statistics, KLT is the optimal orthogonal transform that removes spatial redundancy and achieves maximum energy compaction. The transform kernel is defined by the eigenvectors of the signal's covariance matrix.

Let $b^{(i)}$ be a data block of size $m \times n$ and let there be N such blocks in a given transform mode, i.e., $i = 1, 2, ..., N$. Please note that in our context, these blocks will be blocks of residual data after temporal or inter-prediction. For non-separable KLT, we first rearrange each block as a long vector $x^{(i)}$ of size $mn \times 1$. The (sample) mean vector is $\mu_x = \frac{1}{N}\sum_{i=1}^{N} x^{(i)}$. Then the (sample) covariance matrix of the data is defined as,

$$C_x = \frac{1}{N}\sum_{i=1}^{N}(x^{(i)} - \mu_x)(x^{(i)} - \mu_x)^T$$

The eigenvectors of C_x constitute the rows of KLT's transform matrix T, and KLT is performed as,

$$y^{(i)} = Tx^{(i)}, \quad i = 1, 2, ..., N.$$

The covariance matrix of the transformed data C_y is a diagonal matrix (implying that the transformed data is uncorrelated) with the eigenvalues of C_x on its diagonal. KLT maximizes coding gain by minimizing the geometric mean of the transform coefficient variances. This minimum equals the geometric mean of the eigenvalues of C_x.

Even though non-separable KLT designed on training signal statistics may achieve ideal coding gains, because of its computational and storage complexities, it is seldom applied directly to residual block. Instead, after a primary transform, the top-left low-frequency primary transform coefficients are picked as candidates for the non-separable secondary transform, since they often capture most of the energy in the block. At this stage, by only applying non-separable KLT to a small subset of consequential primary transform coefficients, we reduce the size of the non-separable transform significantly, and hence reduce complexity to an acceptable level.

III. Coefficient Scan Order

After separable primary transformation of a block of residual samples, the transform coefficients are expected to be in decreasing variance order in the top-down and left-right directions. For example, consider a 4×4 block of primary transform coefficients p :

$$p = \begin{bmatrix} p_{11} & p_{12} & p_{13} & p_{14} \\ p_{21} & p_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{bmatrix}$$

In general,

$$Var[p_{i1}] > Var[p_{i2}] > Var[p_{i3}] > Var[p_{i4}], \quad i = 1, 2, 3, 4.$$

Similarly,

$$Var[p_{1j}] > Var[p_{2j}] > Var[p_{3j}] > Var[p_{4j}], \quad j = 1, 2, 3, 4.$$

For the combination of primary trigonometric transforms used by AV1, although there is no deterministic relationship between the variance of p_{ij} and the variance of its right-diagonal neighbors, the zigzag scan order is used as a general rule (approximating decreasing order of variance) for all transform modes that do not use the identity transform in one of the two directions.

After non-separable secondary KLT is applied, a set of secondary transform coefficients replaces the primary coefficients in a top-left quadrant, say, $(p_{11}, p_{12}, p_{21}, p_{22})$ in the matrix p. Let s_{ij} denote a secondary transform coefficient for $i = 1, 2$ and $j = 1, 2$. The matrix containing the final (post secondary) transform coefficients is:

$$\begin{bmatrix} s_{11} & s_{12} & p_{13} & p_{14} \\ s_{21} & s_{22} & p_{23} & p_{24} \\ p_{31} & p_{32} & p_{33} & p_{34} \\ p_{41} & p_{42} & p_{43} & p_{44} \end{bmatrix}$$

It should be noted that the secondary transform is non-separable and is designed on the top-left quadrant of the primary coefficients as viewed in *raster scan order*. Hence,

$$Var[s_{11}] > Var[s_{12}] > Var[s_{21}] > Var[s_{22}]$$

We now have a coefficient matrix with its top-left quadrant of coefficients in decreasing variance in raster scan order. In addition to this, the remaining primary transform coefficients are still in decreasing variance order along the top-down and left-right directions. Again, we cannot define a deterministic relationship between the variances of the secondary and primary transform coefficients. Clearly, the zigzag scan order for the coefficients after secondary transform is, in general, sub-optimal. This motivates our design of different coefficient scan orders for the various transform modes.

In our offline training algorithm, after designing secondary transforms, we rearrange the final coefficients in decreasing order of variance, obtained from the effective transforms applied to the inter-prediction residual blocks. This gives us a final coefficient scan order for each mode, which is optimal for the collected residual statistics.

IV. Design Paradigm: Closed Loop vs. Asymptotic Closed Loop

The offline design of secondary transforms and scan orders based on signal statistics collected from encoding a training set of sequences, is an iterative approach. Detailed block diagrams of the standard closed-loop (CL) design technique and our asymptotic closed loop approach for an iteration i are shown in Fig. 1. In these block diagrams, we use x_n to denote a block of pixels in the nth frame and $\hat{x}_n^i$ to denote the reconstruction of the block in the ith iteration of the design algorithm. Similarly, e_n^i and p_n^i denote the respective ith iteration residue and the primary transform coefficients for block x_n. To emphasize the differences between the two approaches, the primary transform coefficients used in the design of T^i for CL design and the sample reconstructions used in the design of T^i in ACL, are marked in blue and red, respectively.

We will first consider a standard closed-loop design technique (see e.g., [14]). Using the primary transform coefficient statistics that are collected from iteration $i-1$, secondary transforms (T^{i-1}) and scan orders are designed. In iteration i, T^{i-1} is applied (along with the designed scan orders) to the primary transform coefficients p_n^i, p_{n+1}^i and so on. These coefficients are different from the primary transform coefficients from iteration $i-1$ for which T^{i-1} was designed. This drift occurs because of the new secondary transforms that are employed in every iteration which give rise to modified reconstructed residual blocks and sample reconstructions. It can be seen that the reconstructions used in iteration i for the prediction of blocks x_n and x_{n+1} are $\hat{x}_{n-1}^i$ and $\hat{x}_n^i$ which are different from the reconstructions used for prediction in iteration $i-1$ because iteration $i-1$ employs the transform $T^{i-2} \neq T^{i-1}$. Consequentially, the residual blocks (e_n^i, e_{n+1}^i, ...) are modified and hence, (p_n^i, p_{n+1}^i, ...) has different statistics than (p_n^{i-1}, p_{n+1}^{i-1}, ...) which was used in the design of T^{i-1}. This issue that stems from the inherent inter-dependency between predictions and reconstructions in a video coding system makes closed-loop design highly unstable.

In order to overcome this drawback of closed-loop design, we propose to make use of the asymptotic closed loop design paradigm. In ACL, the instability issue is addressed by eliminating the statistical mismatch that leads to error propagation in the prediction loop. This is done by providing reconstructions ($\hat{x}_{n-1}^{i-1}$, $\hat{x}_n^{i-1}$, ...) generated in iteration $i-1$ as open-loop reconstructions to the encoder in iteration i. This ensures that the encoder runs in an open-loop fashion making the design algorithm stable and generates residuals (e_n^{i-1}, e_{n+1}^{i-1}, ...) and hence, primary transform coefficients (p_n^{i-1}, p_{n+1}^{i-1}, ...) which are the same coefficients used for the design of T^{i-1}. Therefore, our approach is also immune to error propagation due to mismatch in reconstructions, which ensures increasingly better reconstructions over the iterations. This guarantees convergence, which upon occurrence leaves the reconstructions between iterations unchanged. In other words, after ACL converges, $(\hat{x}_{n-1}^{i-1}, \hat{x}_n^{i-1} \ldots) = (\hat{x}_{n-1}^i, \hat{x}_n^i, \ldots)$, which is equivalent to closed-loop operation.

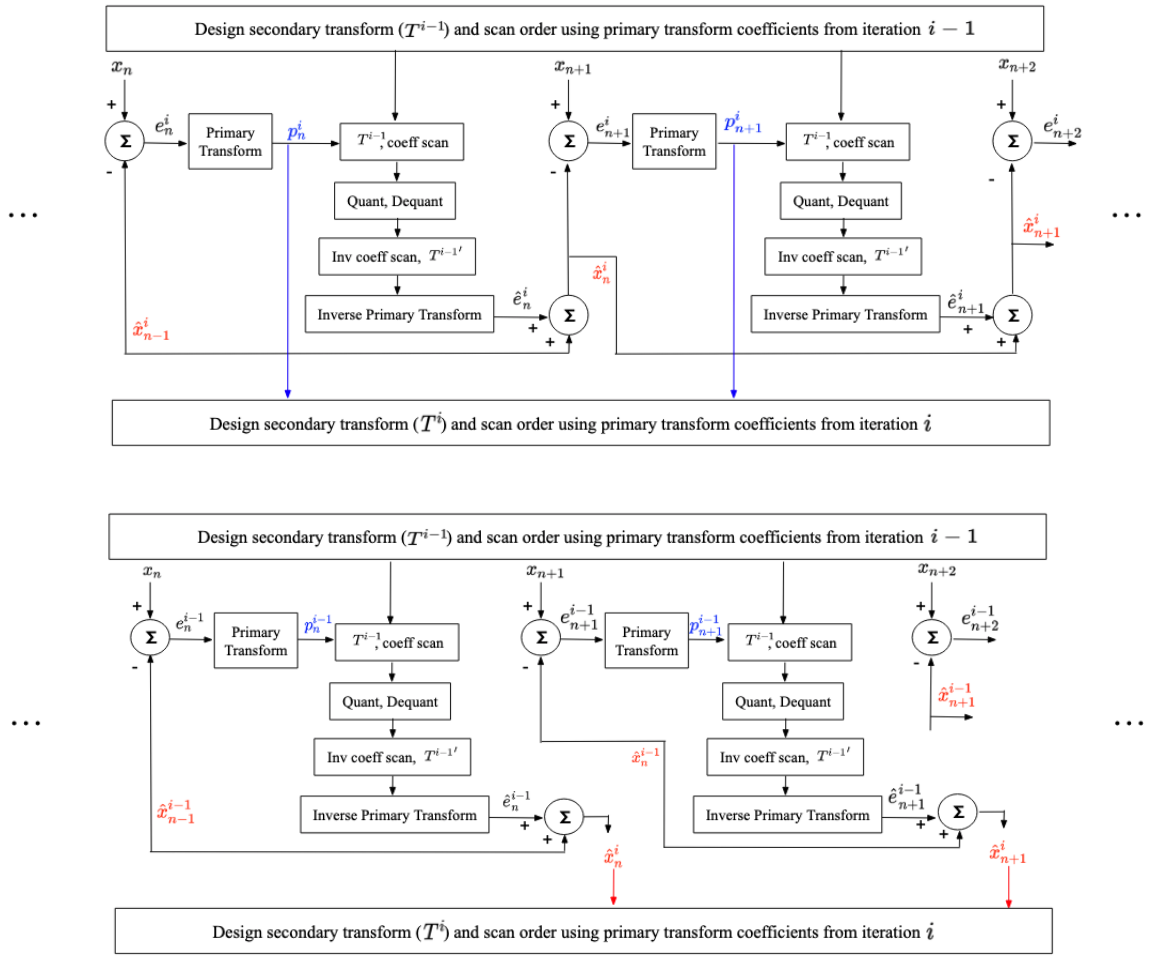

Fig. 1: Illustrations of the CL design algorithm (top) and ACL design algorithm (bottom) for two frames x_n and x_{n+1}, used to design secondary transforms (T^i) and scan orders in a given iteration i

V. Overall Training Algorithm and Pseudocode

To design non-separable secondary transforms and scan orders for a total of 16 transform modes used by AV1, we use an approach in the same vein as K-means clustering which we dub "K-modes clustering". The block subsets assigned to the 16 transform modes are equivalent to clusters and the transforms designed are comparable to the centroids of the clustering problem. Specifically, given the training set, the encoder assigns each block to one of the 16 modes (by RD-optimization), which constitutes our "nearest-neighbor" partition, and using the statistics of each cluster (of blocks), a KLT transform is designed for the low frequency (top-left quadrant) coefficients, which constitutes our "centroid" step. The "K-modes" algorithm is performed for every set of reconstructions generated by one run of the encoder on the training set. In other words, ACL's reconstruction update is implemented in an outer loop while "K-modes" forms the inner loop.

The overall design algorithm is illustrated in Algorithm 1. We initialize the secondary transforms and scanning orders to identity (effective initial scan order is the default zigzag order), collect primary transform coefficient data and encoder decision statistics for KLT and reconstructed sequences for ACL encoder runs.

Next, at each ACL iteration, we run "K-modes" to design transforms and scan orders based on available statistics, followed by mode reassignment to blocks by open-loop runs of the encoder. Upon k-modes convergence, we update reconstructions in ACL fashion and collect new statistics. The next ACL iteration updates the transforms and scan orders for these reconstructions, and so on until convergence, to yield the ultimate 16 secondary transforms and coefficient scan orders for each block size.

VI. Experimental Results

We designed non-separable secondary transforms and scan orders for a training set of sequences and compared the performance of our approach to the standard libaom (AV1) codec run at constant bit-rates 100, 150, 200 and 250 Kbps for CIF sequences and 400, 450, 500, 550 Kbps for HD sequences. We use 30 frames of each CIF sequence and 20 frames of each HD sequence in all of our experiments. Since each bit-

Algorithm 1: Overall Algorithm for Offline Design

Initialize:

Secondary transforms and coefficient scan orders (identity)

Collect data for training:

From closed loop runs of the encoder

- Primary transform coefficient statistics
- Mode assignments for each block
- Reconstructed sequences of the training set

while $ACL\_iter < max\_ACL\_iter$ **do**

  while $kmeans\_iter < max\_kmeans\_iter$ **do**

    1. Design secondary transforms and scan orders using collected statistics for each block size

    2. Update optimal mode assignments for each block using the newly designed transforms

    /* encoder runs in open-loop fashion since the reconstructions collected in the previous iteration (or at the initialization step) are being reused */

  end

  Update encoder decisions with new transforms

  Update reconstructions in ACL fashion

end

rate configuration has different encoder statistics, we design different transforms for different rates. As mentioned before, the AV1 codec uses 4 different 1-D trigonometric transform kernels to form a total of 16 2-D separable transforms. Taking four different rates, three block sizes ($4 \times 4, 8 \times 8, 16 \times 16$) and 16 transforms modes into consideration, we designed secondary transforms for the coefficients in the top-left quadrant of each block and scanning orders for all of the final transform coefficients. The training set consisted of 9 and 5 sequences for the transforms designed for CIF and HD resolutions, respectively.

BD bit-rate reduction over the baseline is calculated as per [15]. Results for the CIF and HD test sequences are shown in tables I and II, respectively. Fig. 2 also shows the rate-distortion curves for the *Meerkat* HD sequence. Despite occasional losses for a few sequences like soccer (CIF) and factory (HD), our approach yields an overall average gain of 0.76% for CIF and 0.41% for HD sequences, over the AV1 codec. The average encoding time for all the CIF test sequences is 117%. This can be lowered by using the designed transforms to derive equivalent integer transforms. Also, since we apply secondary transforms on the top-left quadrant of blocks of size 4×4, 8×8 and 16×16, the largest secondary transform applied on a set of 64 coefficients has comparable complexity to a separable 64×64 transform by AV1.

From our experiments during training, we observed that for inter-prediction residual, significant gains can be obtained only when secondary transforms and scan order are jointly designed and employed. In a setting where secondary transforms are employed without scan order optimization (default zigzag scan order is used), the potential gains do not manifest on account of the subpar coefficient scanning order. On the other hand,

CIF Sequence	BD-rate gains (%)
hall	-1.32
stefan	-1.03
football	0.47
harbour	0.84
husky	-0.54
soccer	0.45
paris	-0.78
ice	-0.76
akiyo	-1.94
deadline	-1.49
mother-daughter	-0.61
news	-1.05
sign-irene	-2.08
Average	**-0.76**

TABLE I: Bit-rate savings (%) for CIF test sequences over AV1

HD Sequence	BD-rate gains (%)
rush hour	-0.67
station2	-1.57
sunflower	-0.67
factory	1.27
in to tree	-0.11
meerkat	-1.49
old town cross	0.14
walking	-0.20
Average	**-0.41**

TABLE II: Bit-rate savings (%) for HD test sequences over AV1

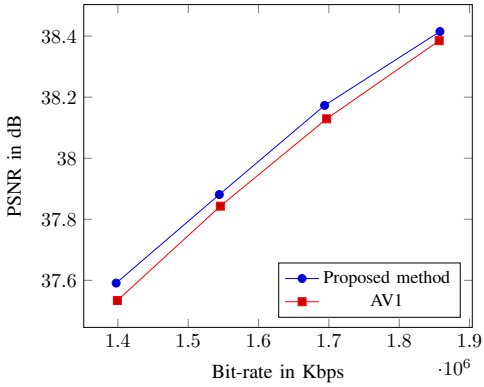

Fig. 2: RD curves for the *Meerkat* HD sequence

when secondary transforms are not employed, the requirement for scan order design does not arise since the default zigazag scan order has shown to work well for primary separable transform coefficients. Based on the results for the joint design of secondary transforms and scan order, we conclude that an overall improvement in coding performance is obtained only when they are allowed to work together.

Note that in an earlier work from our lab [13], the average BD-rate gain for ACL-based design of *primary separable* transforms for the luma component over VP9 (which only uses DCT) is reported as 6% for a test set of CIF sequences. The focus of the current paper is exclusively on gains due to non-separable secondary transforms, and future work will combine

both approaches for a joint design of primary and secondary transforms for AV1.

VII. Conclusions

This paper proposes an efficient offline design procedure for the joint design of multi-modal non-separable secondary transforms and coefficient scanning orders for inter-prediction residual. The design algorithm overcomes the instability issue in standard closed-loop design techniques by using the asymptotic closed loop approach. Experimental results on top of *libaom* showcase gains obtained over the AV1 standard. The proposed approach could be adapted to any state-of-the-art video codec.

VIII. Acknowledgment

We thank Bohan Li of Google for useful discussions, as well as for help with the libaom codebase.

Use was made of computational facilities purchased with funds from the National Science Foundation (OAC-1925717) and administered by the Center for Scientific Computing (CSC). The CSC is supported by the California NanoSystems Institute and the Materials Research Science and Engineering Center (MRSEC; NSF DMR 1720256) at UC Santa Barbara.

References

[1] R Dony et al., "Karhunen-loeve transform," *The transform and data compression handbook*, vol. 1, no. 1-34, pp. 29, 2001.

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

[3] K. Rose, A. Heiman, and I. Dinstein, "DCT/DST alternate-transform image coding," *IEEE Transactions on Communications*, vol. 38, no. 1, pp. 94–101, 1990.

[4] 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.

[5] J. Han *et al*., "A technical overview of AV1," *Proceedings of the IEEE*, vol. 109, no. 9, pp. 1435–1462, 2021.

[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] M. Koo, M. Salehifar, J. Lim, and S. H. Kim, "Low frequency non-separable transform (LFNST)," in *2019 Picture Coding Symposium (PCS)*, 2019, pp. 1–5.

[8] B. Bross *et al*., "Overview of the versatile video coding (VVC) standard and its applications," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 10, pp. 3736–3764, 2021.

[9] X. Zhao and S. Liu, "Unified secondary transform for intra coding beyond AV1," in *IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 3393–3397.

[10] S. Kahu, M. P. Krishnan, X. Zhao, and S. Liu, "Context-adaptive secondary transform for video coding," in *IEEE International Conference on Image Processing (ICIP)*, 2021, pp. 2039–2043.

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

[12] X. Zhao *et al*., "Study on coding tools beyond AV1," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*, 2021, pp. 1–6.

[13] B. Vishwanath, S. Li, and K. Rose, "Asymptotic closed-loop design of transform modes for the inter-prediction residual in video coding," in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 3403–3407.

[14] 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.

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