# Deterministic Annealing Based Transform Domain Temporal Predictor Design for Adaptive Video Coding

Bharath Vishwanath, Tejaswi Nanjundaswamy, and Kenneth Rose

Department of Electrical Engineering
University of California, Santa Barbara, CA 93106
{*bharathv*, *tejaswi*, *rose*}@*ece.ucsb.edu*

## Abstract

Current video coders employ motion compensated pixel-to-pixel prediction, which largely ignores significant spatial correlations and the fact that true temporal correlations vary with spatial frequency. Earlier work from our lab proposed to first spatially decorrelate the block of pixels by performing temporal prediction in the transform domain, and to effectively account for both spatial and temporal correlations. To adapt to variations in video signal statistics, the encoder switches between a set of appropriately designed prediction modes. This setting critically depends on efficient offline learning of transform domain temporal prediction modes. Significant challenges include: *i)* issues of instability and mismatched statistics inherent to closed loop design; and *ii)* severe non-convexity of the cost function trapping the system in poor local minima. Statistics mismatch is tackled by an appropriate paradigm for system design in a stable open loop fashion, but which asymptotically mimics closed loop operation. The non-convexity is handled by deterministic annealing, a powerful non-convex optimization tool whose probabilistic formulation allows for direct optimization of the cost function with respect to the discrete set of prediction modes, and whose annealing schedule avoids poor local minima. Experimental results validate the method's efficacy.

## Introduction

Modern video coders exploit temporal correlations by employing motion compensated prediction [1, 2]. Simple pixel copying of the best (motion-compensated and possibly interpolated) block from the reference frame is used to obtain the prediction signal. The resulting prediction error is then decorrelated by the discrete cosine transform (DCT) and the transform coefficients are quantized and sent to the decoder. Such pixel-to-pixel temporal prediction is suboptimal in that it ignores significant spatial correlations in the video signal. Several approaches that account for spatial correlations include multi-tap filtering [3, 4] and three-dimensional subband coding [5, 6], which incur high encoder complexity. An earlier work from our lab [7] proposed an effective way to account for complex spatio-temporal correlations by first spatially decorrelating a block via DCT and subsequently preforming temporal prediction of the resulting uncorrelated transform coefficients. Temporal evolution of each DCT coefficient in a block, along a motion trajectory, is modelled as a first order auto-regressive process. Moreover, transform domain temporal prediction (TDTP) perfectly captures and exploits how the temporal correlation varies across frequencies, which is otherwise masked in the pixel domain.

Video signals exhibit considerable variations in local statistics and hence significant diversity in temporal correlations of transform coefficients. In an earlier variant,

2375-0359/19/$31.00 ©2019 IEEE
DOI 10.1109/DCC.2019.00027

we proposed optimizing the TDTP parameters for each sequence being compressed, a multi-pass approach, for applications involving offline video compression [8]. Real-time encoding, however, requires adaptation to signal statistics, and a reasonable approach involves a set of prediction modes for the encoder to switch between, which were designed offline from a training set of diverse video sequences. Early experimentation has shown that realizing the full potential of the approach depends critically on effective design of the prediction modes. This is the main motivation for the work described in this paper. The non-linear operation of TDTP mode assignment makes it a challenging design problem, since the cost derivatives with respect to the TDTP mode decisions vanish almost everywhere. Having encountered a similar difficulty in the context of intra-mode design, we proposed to recourse to an iterative "K-means" clustering [9] style design in [10], wherein the input is partitioned into clusters based on the modes that minimize their prediction residual, followed by an update of the mode representing each cluster. However, this greedy approach does not guarantee a globally optimal solution and is, in fact, often seen to converge to poor local minima.

TDTP modes are optimized to enhance prediction performance, but prediction is applied to the reconstructed samples, which are modified by update to the TDTP modes. This provides a first glimpse of the closed loop conundrum discussed below. Thus, we need an iterative design procedure, wherein we design TDTP modes for a given reconstructed sequence, and then update these reconstructions. In this context, various design techniques have been proposed. In the standard closed loop design [11], the predictor from the previous design iteration is used to update reconstructions in a closed-loop manner. However, this leads to mismatch in statistics, which often yields catastrophic error buildup at low rates, due to error propagation through the prediction loop. Specifically, the predictor is optimized for the reconstructions in the previous iteration, while it operates on the reconstructions of the current iteration. To tackle this, asymptotic closed-loop (ACL) design was proposed in [12]. ACL operates in open loop, in that reconstructed samples from the previous iteration are used as the prediction reference, providing a stable platform for design. Moreover, as the system approaches convergence, the reconstructed sequence remains largely unchanged from iteration to the next, effectively implementing closed-loop operation.

Although ACL provides a stable design platform, experiments show that it is nevertheless highly sensitive to initialization, indicating severe non-convexity of the cost surface. Each TDTP mode can be viewed as a high dimensional vector of parameters, and the mode design is essentially vector quantizer design, a notoriously non-convex optimization problem. This difficulty is further exacerbated by the design instabilities. In this paper, we overcome these shortcomings by embedding the design within a deterministic annealing (DA) framework. Inspired by principles of statistical physics and information theory, DA was developed and shown to be a powerful non-convex optimization tool [13]. The inherent probabilistic nature of DA allows us to deterministically optimize the effective cost function, an appropriate expectation function that efficiently accounts for and replaces the stochastic wandering on the cost surface of the classical method of simulated annealing [14]. With a careful annealing schedule DA is shown to avoid poor local minima. The optimization benefits of DA are complemented by the design stability of ACL. The resulting DA-ACL framework effectively

overcomes all the design challenges and provides an efficient way to design TDTP modes. Experimental results demonstrate the efficacy of the proposed approach.

## Problem Statement

Let us consider an input training set which is partitioned into segments, each of which is called a group of pictures (GOP). Let the set of frames in a GOP be denoted by $N_g$. Let us consider a block $b$ in a frame $n$ that is inter-predicted by a block in reference frame $n-1$. We spatially decorrelate the block and perform prediction in the DCT domain. The encoder switches between prediction modes $\{\alpha_k\}$, where each $\alpha_k$ is a block of prediction coefficients. Let the prediction mode chosen for a GOP $g$ be $\hat{\alpha}_g$. The temporal evolution of each DCT coefficient $x_{n,b,p}$ in the block is modelled as a first-order auto-regressive process:

$$x_{n,b,p} = \hat{\alpha}_{g,p}\hat{x}_{n-1,b,p} + e_{n,b,p} \tag{1}$$

where $\hat{x}_{n-1,b,p}$ is the corresponding DCT coefficient of the reference block and $e_{n,b,p}$ is the innovation. The problem at hand is to design the prediction modes to minimize the overall prediction error of the training set. Thus the cost function is,

$$E = \sum_{g}\sum_{n\in N_g}\sum_{b}\sum_{p\in b}(x_{n,b,p} - \hat{\alpha}_{g,p}\hat{x}_{n-1,b,p})^2 \tag{2}$$

## Background

*Iterative K-mode predictor design*

As mentioned earlier, the non-linear TDTP mode assignment makes it a challenging problem to optimize the overall prediction error w.r.t TDTP modes. We can thus design TDTP modes in an iterative manner that are optimized for a given set of reconstructions $\hat{\mathbf{x}}_{b,n}$ at the encoder. With some initialization of the TDTP modes, the algorithms iterates between the following steps:

- Mode assignment: For a given GOP $g$, assign the best mode from the set of TDTP modes which minimizes the prediction error for the GOP.
- TDTP modes update: Let $N_k$ be the set of frames that share the same TDTP mode. The $p^{th}$ component of optimal prediction mode $\alpha_k$ for this cluster is given by,

$$\alpha_{k,p} = \frac{\sum_{n\in N_k}\sum_b x_{n,b,p}\hat{x}_{n-1,b,p}}{\sum_{n\in N_k}\sum_b \hat{x}^2_{n-1,b,p}} \tag{3}$$

With the new set of prediction modes, the reconstructed samples at the encoder are updated and these steps are repeated until convergence. The reconstructions can be updated in different ways, leading to the following design paradigms.

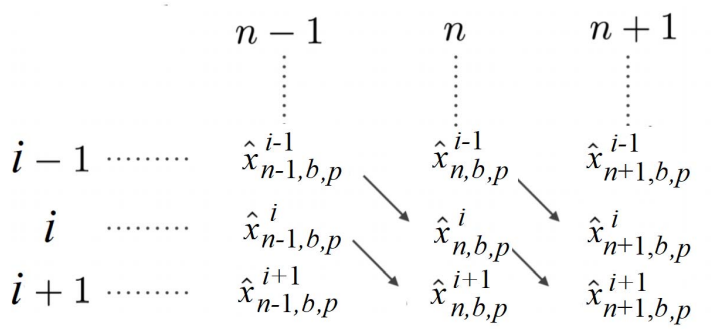


Figure 1: Asymptotic closed loop design

*Open-loop, closed-loop and asymptotic closed-loop design*

Several techniques have been proposed for the design of predictors and quantizers. Since most modern video codecs employ a fixed standard quantizer, the current paper focuses on the predictor design problem. Prediction design in open-loop approach is based on the original sample statistics (see, e.g., [15]). A major shortcoming of this approach is that the predictor must operate in closed loop as the decoder does not have access to original samples. In closed loop design [11], predictors are designed iteratively. Let $\hat{\alpha}_g^{i-1}$ denote the prediction coefficients for GOP $g$ at iteration $i-1$. The reconstructed transform coefficients for the corresponding GOP in iteration $i$ is updated as,

$$\hat{x}^i_{n,b,p} = \hat{\alpha}^{i-1}_{g,p}\hat{x}^i_{n-1,b,p} + \hat{e}^i_{n,b,p} \tag{4}$$

where $\hat{e}^i_{n,b,p}$ is the quantized prediction error $e_{n,b,p} = x_{n,b,p} - \hat{\alpha}^{i-1}_{g,p}\hat{x}^i_{n-1,b,p}$. Predictor $\hat{\alpha}^{i-1}_{g,p}$ was designed for $\hat{x}^{i-1}_{n,b,p}$. However, it is used with the reconstructed samples $\hat{x}^i_{n,b,p}$ of the next iteration. Errors due to mismatch propagate through the prediction loop, causing design instability, which often proves catastrophic at low rates. To address this issue, ACL was proposed in [12]. ACL operates in an open loop fashion, while eventually optimizing the system for closed-loop operation. The reconstructed transform coefficients for iteration $i$ is updated as,

$$\hat{x}^i_{n,b,p} = \hat{\alpha}^{i-1}_{g,p}\hat{x}^{i-1}_{n-1,b,p} + \hat{e}^i_{n,b,p} \tag{5}$$

where $\hat{e}^i_{n,b,p}$ is the quantized prediction error $e_{n,b,p} = x_{n,b,p} - \hat{\alpha}^{i-1}_{g,p}\hat{x}^{i-1}_{n-1,b,p}$. The predictor $\hat{\alpha}^{i-1}_{g,p}$ is applied to reconstructions $\hat{x}^{i-1}_{b,n,p}$, the same set of reconstructions that it was optimized for, thereby eliminating statistical mismatch. The new set of reconstructions are then used to design predictor $\hat{\alpha}^i_k$. Upon convergence, the reconstructed samples remain largely the same from one iteration to the next. Thus, predicting from $\hat{x}^{i-1}_{b,n,p}$ is the same as predicting from $\hat{x}^i_{b,n,p}$, which is essentially closed loop operation. Fig. 1 illustrates the ACL design paradigm.

## Proposed Method

The proposed approach is inspired by, and builds on deterministic annealing (DA) [13]. DA is derived from principles of information theory and is motivated by intuition gained from the annealing process in physical chemistry, wherein certain systems are driven to their low energy states by gradually cooling the system. Analogously, we randomize the TDTP mode assignment and deterministically minimize the overall prediction error subject to the specified level of randomness. The amount of randomness is measured by Shannon's entropy and is essentially characterized by the "temperature" of the system. In contrast with the hard TDTP mode assignment in (2), the prediction mode assignment in DA is probabilistic, and is differentiable everywhere (with respect to the assignment probabilities), paving the way to effective overall optimization of the system. The process starts at high temperature, where all TDTP modes are coincident, effectively converging to a single distinct globally optimal mode, regardless of initialization. The system is then gradually cooled, and the minimum is tracked. During annealing, the reconstructions are updated within the ACL paradigm, providing a stable design platform. At ACL iteration $i$, let $P^i_{k|g}$ denote the probability of assigning TDTP mode $\alpha^i_k$ to GOP $g$. The prediction error to be minimized is given by the expectation,

$$J = \sum_g \sum_k \sum_{n \in N_g} \sum_b \sum_p P_g P^i_{k|g} (x_{n,b,p} - \alpha^i_{k,p} \hat{x}^i_{n-1,b,p})^2 \tag{6}$$

where $P_g$ denotes the probability of the input GOPs (assumed uniform). The degree of randomness is measured by the Shannon entropy,

$$H = -\sum_g \sum_k P^i_{gk} \log(P^i_{gk}), \tag{7}$$

where $P^i_{gk} = P_g P^i_{k|g}$ is the joint distribution over TDTP modes and input GOPs. The problem can be posed as the minimization of the Lagrangian cost function, directly analogous to Helmholtz free energy of statistical physics:

$$F = J - TH \tag{8}$$

The degree of randomness is controlled by $T$, which characterizes the temperature. As the temperature decreases, the system trades entropy for distortion. At the limit of zero randomness, the process directly minimizes the overall prediction error. Assuming uniform distribution over the training set, it is straightforward to show that the association probabilities that minimize the Lagrangian cost (subject to the obvious additional constraint $\sum_k P^i_{k|g} = 1$), are given by the Gibbs distribution:

$$P^i_{k|g} = \frac{e^{\frac{-\sum_{n \in N_g} \sum_b \sum_p (x_{n,b} - \alpha^i_{k,p} \hat{x}^i_{n-1,b,p})^2}{T}}}{\sum_j e^{\frac{-\sum_{n \in N_g} \sum_b \sum_p (x_{n,b,p} - \alpha^i_{j,p} \hat{x}^i_{n-1,b,p})^2}{T}}} \tag{9}$$

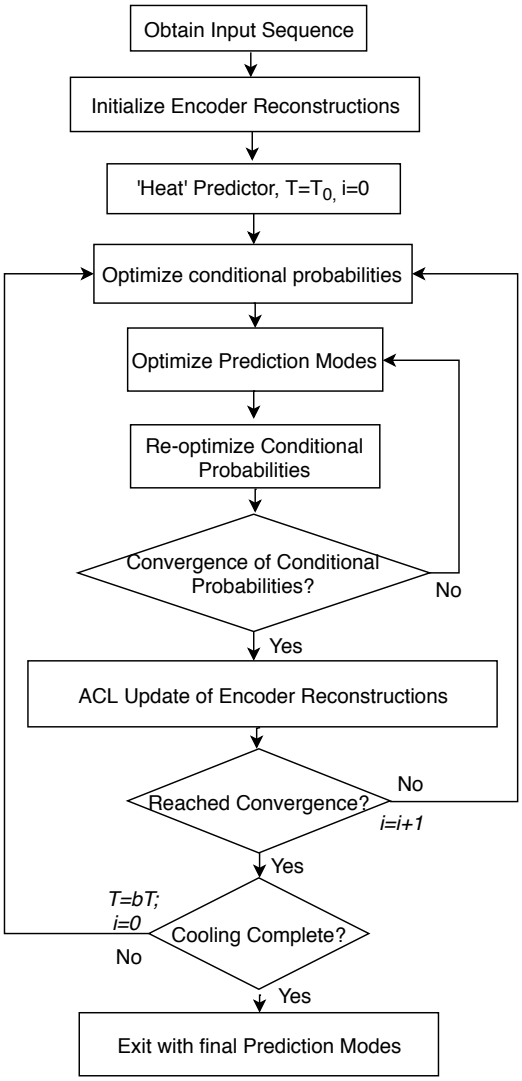


Figure 2: Flow chart of the proposed algorithm

Note that, at high temperatures, the association probabilities are uniform, and hence the system achieves maximum entropy. As the temperature is lowered, more probability is assigned to nearby TDTP modes, making the system more discriminatory.

To obtain the optimal prediction mode we set the partial derivative to zero:

$$\begin{aligned} \frac{\partial J}{\partial \alpha_{k,p}^{i}} &= \sum_{g} \sum_{n \in N_g} \sum_{b} 2 P_g P_{k|g}^{i} (x_{b,n,p} - \alpha_{k,p}^{i} \hat{x}_{b,n-1,p}^{i})(-\hat{x}_{b,n-1,p}^{i}) \\ &= 0 \end{aligned} \tag{10}$$

Thus, the $p^{th}$ component of the optimal prediction modes is given by,

$$\alpha_{k,p}^{i} = \frac{\sum_g \sum_{n \in N_g} \sum_b P_{k|g}^{i} x_{b,n,p} \hat{x}_{b,n-1,p}^{i}}{\sum_g \sum_{n \in N_g} \sum_b P_{k|g}^{i} (\hat{x}_{b,n-1,p}^{i})^2} \tag{11}$$

At high temperatures, as the association probabilities are uniform, it follows from (11) that all the prediction modes are coincident, i.e., and essentially implement a single optimal prediction mode for the entire training set *regardless of initialization*. In other words, the method is invariant to initialization as it achieves the global optimum at high temperature. As the temperature is lowered, the system becomes more deterministic with the emergence of more prediction modes through a sequence of phase transitions.

The reconstructed samples in GOP $g$ are now updated via ACL as,

$$\hat{x}_{n,b,p}^{i+1} = \sum_k P_{k|g}^{i} (\alpha_{k,p}^{i} \hat{x}_{n-1,b,p}^{i} + \hat{e}_{k,n,b,p}^{i}) \tag{12}$$

where, $\hat{e}_{k,n,b,p}^{i}$ is the quantized prediction error $e_{k,n,b,p}^{i} = x_{n,b,p} - \alpha_{k,p}^{i} \hat{x}_{n-1,b,p}^{i}$. The overall design procedure is illustrated in Fig. 2.

## Experimental Results

The proposed method is implemented in HM 14.0. To simplify the experiments, all the sequences are coded in IPPP format, only the previous frame is allowed as reference, both prediction and transform block sizes are restricted to $8 \times 8$, and the sample adaptive offset function option is disabled. A set of six TDTP modes were designed by each of the following design methods: *i)* The proposed method referred as DA-ACL *ii)* K-mode design with "plain ACL"referred to as ACL iii) Our previous "crude" method of providing a set of TDTP modes, each chosen from a distinct training sequence, referred to as TDTP. The average bit-rate reduction is calculated as per [16]. The training set sequences are listed in Table 1. The bit-rate savings over standard HEVC for the test set sequences are tabulated in Table 2. The consistent performance gains and significant bit-rate reduction over the test set provide clear evidence for the utility of the proposed approach.

## Conclusions

This paper presents a novel near-optimal procedure for designing transform domain temporal prediction modes. It effectively resolves significant shortcomings due to statistical mismatch and design instability of standard approaches. The deterministic annealing-based framework enables direct optimization of the overall cost with respect to prediction mode decisions, and avoids many poor local minima that trap its competitors. Substantial and consistent gains over HEVC, in terms of bit-rate reduction over the test sequences, demonstrates the efficacy of the proposed approach.

| Sequence |
|---|
| BasketballDrive (1080p) |
| BQTerrace (1080p) |
| ParkScene (1080p) |
| BQSquare (240p) |
| RaceHorses (240p) |
| City (cif) |
| Stefan (cif) |
| Container (cif) |
| Tempete (cif) |
| Waterfall (cif) |
| Highway (cif) |
| Coastguard (cif) |
| Bus (cif) |
| Mobile (cif) |

Table 1: The training set of video sequences

| Sequence | TDTP | ACL | DA-ACL |
|---|---|---|---|
| Cactus (1080p) | 8.4 | 8.7 | 8.8 |
| KristenAndSara (720p) | 0.9 | 1.4 | 4.4 |
| Johnny (720p) | 2.7 | 3.3 | 2.4 |
| vidyo3 (720p) | 3.5 | 4.8 | 6.1 |
| vidyo1 (720p) | 4.2 | 2.1 | 5.1 |
| FourPeople (720p) | 0.4 | 0.3 | 2.9 |
| RaceHorses (480p) | 2.1 | 2.6 | 3.9 |
| Mobisode2 (480p) | 1.0 | 1.8 | 3.4 |
| Kimono1 (480p) | 10.0 | 10.2 | 12.1 |
| Keiba (480p) | 3.6 | 3.9 | 4.1 |
| BlowingBubbles (240p) | 0.2 | 1.0 | 1.0 |
| Mother-daughter (cif) | 0.2 | 0.2 | 2.0 |
| Suzie (qcif) | 1.7 | 1.0 | 2.1 |
| **Average** | **3.0** | **3.2** | **4.4** |

Table 2: Performance over the test set: bit-rate savings over HEVC ( in % ) for the Y component

## References


[1] H. Schwarz, D. Marpe, and T. Wiegand, "Overview of the scalable video coding extension of the h. 264/avc standard," *IEEE Transactions on circuits and systems for video technology*, vol. 17, no. 9, pp. 1103–1120, 2007.

[2] G. J. Sullivan, J.-R. Ohm, W.-J. Han, T. Wiegand, et al., "Overview of the high efficiency video coding(hevc) standard," *IEEE Transactions on circuits and systems for video technology*, vol. 22, no. 12, pp. 1649–1668, 2012.

[3] J. Kim and J. W. Woods, "Spatiotemporal adaptive 3-d kalman filter for video," *IEEE Transactions on Image Processing*, vol. 6, no. 3, pp. 414–424, 1997.

[4] T. Wedi, "Adaptive interpolation filter for motion and aliasing compensated prediction," in *Visual Communications and Image Processing 2002*. International Society for Optics and Photonics, 2002, vol. 4671, pp. 415–423.

[5] S.-J. Choi and J. W. Woods, "Motion-compensated 3-d subband coding of video," *IEEE Transactions on image processing*, vol. 8, no. 2, pp. 155–167, 1999.

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

[7] J. Han, V. Melkote, and K. Rose, "Transform-domain temporal prediction in video coding: exploiting correlation variation across coefficients," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*. IEEE, 2010, pp. 953–956.

[8] S. Li, T. Nanjundaswamy, Y. Chen, and K. Rose, "Asymptotic closed-loop design for transform domain temporal prediction," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4907–4911.

[9] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A k-means clustering algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.

[10] S. Li, Y. Chen, J. Han, T. Nanjundaswamy, and K. Rose, "Rate-distortion optimization and adaptation of intra prediction filter parameters," in *Image Processing (ICIP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 3146–3150.

[11] 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.

[12] 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.

[13] K. Rose, "Deterministic annealing for clustering, compression, classification, regression, and related optimization problems," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2210–2239, 1998.

[14] P. JM. Laarhoven and E. HL. Aarts, "Simulated annealing," in *Simulated annealing: Theory and applications*, pp. 7–15. Springer, 1987.

[15] V. Cuperman and A. Gersho, "Vector predictive coding of speech at 16 kbits/s," *IEEE Transactions on Communications*, vol. 33, no. 7, pp. 685–696, 1985.

[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.