April 09, 2019

A Temporal Dependency Model for Rate-Distortion Optimization in Video Coding

James Bankowski

Paul Wilkins

Yaowu Xu

Jingning Han

A Temporal Dependency Model for Rate-Distortion Optimization in Video Coding

ABSTRACT

Many video codecs use motion compensated prediction to achieve compression efficiency. Motion compensated prediction may produce temporal dependency across frames. For example, quantization distortion in a block may propagate through motion compensated prediction and affect the coding efficiency of blocks in subsequent frames. Identifying temporal dependencies may improve the rate-distortion optimization and produce coding performance gains. Block-based motion trajectories and correlations between source pixel blocks along a motion trajectory may be used estimate a distortion propagation model, which may represent the correlation between the distortion propagation and the effect of quantization. A temporal dependency model that accounts for both block correlation and the quantization effect may provide compression gains over the use of a distortion propagation model.

DETAILED DESCRIPTION

Video compression techniques exploit the temporal correlations in a video signal, such as in the form of motion compensated prediction, to achieve coding efficiency. Motion compensated prediction produces a dependency between a coding block and a corresponding reference block. The reconstruction quality of a block can influence the compression efficiency of blocks in subsequent frames. For a reference block that is an accurate prediction of a current block, an adjustment to the bit allocation for the reference block to improve the reference block reconstruction quality may reduce the overall distortion and may maintain the rate cost. For a reference block that is a less accurate prediction of the current block, adjusting bit allocations for the blocks individually may improve the rate-quality balance.

A rate allocation approach to optimizing the performance of a hierarchical coding structure may be to use lower quantization parameters (QP) for frames at lower temporal layers, which serve as the reference frames for later higher temporal layer frames. A trellis-based rate allocation optimization, wherein each node corresponds to a frame encode at a given QP, may include finding the path that minimizes the overall rate-distortion cost as the optimal frame QP combinations for a sequence. To achieve optimality for frame level QP selection, the encoding complexity increases significantly.

Rate allocation optimization may utilize inter frame dependency. A linear model, which may be trained offline, may capture frame level distortion propagation. The inter frame dependency model may be updated according to the coding statistics from previously coded frames at the same temporal layer. The frame QP may be adapted according to the derived dependency models. A block based temporal dependency model may include conducting a forward search over subsequent coding frames to measure the impact of the reconstruction distortion of a current block, based on which the Lagrangian multiplier of each coding tree unit in a current frame may be adjusted. The model may be simplified for single pass encoding. For example, a motion trajectory on-grid alignment constraint, wherein a coding block may be constrained to be on grid and a corresponding reference block may have unconstrained grid alignment, may be relaxed. In another example, inter-mode coding may be used and the use of intra-prediction may be omitted.

A macroblock-tree (MB-tree) scheme may implemented that tracks the temporal dependency through block level motion trajectories, such as in a two-pass encoding approach. The first pass may include rate-distortion optimization based mode decision that omits accounting for the distortion impact on the subsequent frames. The second pass may use the motion vectors and inter/intra mode decisions available from the first pass to build motion trajectories over the source frames (i.e., uncompressed frames). To estimate each block's impact on the subsequent blocks in the same motion trajectory, the MB-tree scheme uses a linear model of the intra- and inter-prediction errors. The correlation between the two blocks is estimated by the difference error divided by the intra-prediction error. The correlations are recursively propagated through the motion trajectories, which form a temporal dependency model. Based on the temporal dependency model, the encoder adjusts the rate allocation to improve reconstruction quality of blocks that have higher impact on corresponding subsequent blocks in the motion trajectories. The MB-tree scheme may provide significant compression gains over some other frame type dependent constant quantization parameter coding schemes. The MB-tree scheme may build a model based on the inter- and intra-prediction errors over the source signals, which may omit using the quantization effect on the temporal dependency.

The distortion propagation may relate to the relative energy value between the innovation term and the quantization error. For high resolution quantization, the quantization error may be substantially smaller than the innovation, and the distortion propagation between frames may be minimal. In some cases, such as in the medium bit-rate range, the quantization error may be similar to or greater than the innovation and the distortion propagation may be related to the correlation between blocks in respective motion trajectories.

A temporal dependency model may account for the quantization effect on the distortion propagation through the motion trajectory, which may provide considerable compression efficiency improvement compared to the MB-tree scheme.

The MB-tree scheme may estimate the amount of information each block contributes to the prediction of subsequent frames, which may be used to weight the rate-distortion trade-off for each block based on the respective contribution. The MB-tree scheme may include reverse frame processing order over the source frames, propagating information from future frames back to the current frame.

For each frame, a propagation determination may be performed for each block. The propagation determination may include estimating the intra prediction cost (*intra_cost*) in terms of sum of absolute Hadamard transform difference (SATD). The propagation determination may include obtaining the motion information available from the first-pass encoding and may estimate the inter prediction cost (*inter_cost*). For hybrid inter/intra prediction mode encoding, the inter cost value (*inter_cost*) may be upper bounded by intra cost (*intra_cost*). A propagation cost variable (*propagation_cost*) may be used to collect the information flowed back from future processing frames. The propagation cost variable (*propagation_cost*) may be initialized to zero (0) for the blocks in the last processing frame in a sequence, such as a group of pictures (GOP).

The propagation determination may include estimating the fraction of information from a current block to be propagated towards its reference block, which may reflect percentage of prediction error reduction associated with the motion compensated reference, and which may be expressed as the following:

$$propagation\_fraction = (1 - inter\_cost / intra\_cost).$$

(Equation 1)

The propagation determination may include estimating the amount of information the current block contributes to the sequence based on a sum of the intra prediction cost (*intra_cost*) and the propagation cost variable (*propagation_cost*), which may be expressed as the following:

$$MB\_cost = intra\_cost + propagation\_cost.$$

(Equation 2)

The information that is propagated toward the reference blocks (*propagation_amount*) may be obtained, which may be expressed as the following:

$$propagation\_amount = (intra\_cost + propagation\_cost) * propagation\_fraction.$$

(Equation 3)

The reference block may be unaligned with the grid of blocks. The amount of information propagated toward the reference blocks (*propagation_amount*) may be allocated, or dispensed, to the blocks that overlap (*overlap_area*) with the reference block (*MB_area*). The corresponding block in the reference frame may accumulate a respective propagation cost (*propagation_cost*), which may represent backwards propagation, and which may be expressed as the following:

$$propagation\_cost \mathrel{+}= (overlap\_area / MB\_area) * propagation\_amount.$$

(Equation 4)

The distortion propagation factor (*DPF*) of a block may be evaluated, wherein the intra prediction cost (*intra_cost*) represents the impact of the block on subsequent frames in the sequence, which may be expressed as the following:

$$DPF = 1 + propagation\_cost \,/\, intra\_cost.$$

(Equation 5)

The rate allocation may be adjusted according to the distortion model such that blocks with a higher distortion propagation factor are allocated a higher rate allocation and blocks with a lower distortion propagation factor are allocated a lower rate allocation.

The temporal dependency model may include a distortion propagation model that differs from the distortion propagation model of the MB-tree scheme, which may account for the quantization effect.

In an example, for a frame index (k), a source pixel block (M_k), and a reconstructed reference block (M_{k-1}), the second moment of inter prediction error may be expressed as the following:

$$\sigma_k^2 = E\{\|M_k - \widehat{M}_{k-1}\|^2\} .$$

(Equation 6)

The innovation term between M_k and M_{k-1} may be largely uncorrelated with the quantization noise at $\widehat{M}_{k-1}$, the prediction error σ_k^2 may be approximately decomposed into the innovation term σ_o^2 and the quantization distortion in the reference block D_{k-1}, which may be expressed as the following:

$$\sigma_k^2 = E\{\|M_k - M_{k-1} + M_{k-1} - \widehat{M}_{k-1}\|^2\}$$

$$\approx E\{\|M_k - M_{k-1}\|^2\} + E\{\|M_{k-1} - \widehat{M}_{k-1}\|^2\}$$

$$= \sigma_o^2 + D_{k-1}.$$

(Equation 7)

The expected distortion of a block (M_k) may be expressed, based on a quantization and dequantization process (Q), as the following:

$$D_k = Q(\sigma_k^2) = Q(\sigma_o^2 + D_{k-1}).$$

(Equation 8)

For high resolution the quantization distortion may be linear with the input signal energy, wherein the effective linear relationship α is based on the bit-rate and the probability distribution of the input signal, which may be expressed as the following:

$$D_k = \alpha(\sigma_o^2 + D_{k-1}).$$

(Equation 9)

Equation 9 may apply in other bit-rate ranges. In some implementations, the value of the effective linear relationship α may be a defined value, such as 0.94 or 1.0. In some implementations, the effective linear relationship α may be directly estimated per block.

Building the temporal dependency model may include the encoder accessing the source blocks M_k M_{k-1}, and the difference between the source blocks may be expressed as the following:

$$R_k = M_k - M_{k-1}.$$

(Equation 10)

Quantized transform coefficients may be obtained by applying a Hadamard transform to the difference R_k, which may be expressed as the following:

$$\widehat{R}_k = T^{-1}Q(T(R_k)).$$

(Equation 11)

A discrete cosine transform may be used to improve compression performance. The Hadamard transform approximates the Discrete Cosine Transform and has simple and fast implementation to reduce the encoder complexity increase. Applying the transform may be used to determine the transform coding gains for evaluating the quantization noise. Estimating the distortion on the innovation term may be expressed as the following:

$$D_{k0} \approx E\{\|R_k - \widehat{R}_k\|^2\}.$$

(Equation 12)

Estimating the prediction error may be expressed as the following:

$$\sigma_o^2 = E\{\|R_k\|^2\}.$$

(Equation 13)

A frame may have a larger or similar quantization parameter as compared to its reference frames, which indicate that $D_k \geq D_{k-1}$. The quantization noise may be bounded by the innovation term, such as $\sigma_o^2 \geq D_k$. The linear relationship expressed in Equation 9 may indicate that $\sigma_o^2 \geq D_{k-1}$. The quantization effect on the innovation term may capture the linear relationship expressed in Equation 9, which may be expressed as the following:

$$\alpha \approx \frac{D_{k0}}{\sigma_o^2}.$$

(Equation 14)

The approximate contribution to block M_k of the distortion in the reference block $\widehat{M}_{k-1}$ may be expressed as the following:

$$\alpha D_{k-1} = \frac{D_{k0}}{\sigma_o^2} D_{k-1}.$$

(Equation 15)

The term D_{k0} may express the quantization effect, the term $(1 - \frac{inter\_cost}{intra\_cost})$ may express the mutual information between the reference and the current blocks, and the distortion propagation model expressed in Equation 1 may be expressed as the following:

$$propagation\_fraction = \frac{D_{k0}}{\sigma_o^2} \cdot (1 - \frac{inter\_cost}{intra\_cost}).$$

(Equation 16)

The quantization noise may be significantly smaller than the innovation energy and the inter frame distortion propagation may be close to 0, which correspond with omitting accounting for the distortion impact on future frames in the construction of the rate-distortion optimization for coding a current frame. The quantization noise may be comparable to the innovation process and the impact of a current block on subsequent blocks in the motion trajectory may depend on the respective correlations, which may be quantified as the percentage of intra prediction error reduction due to inter prediction.

FIG. 1 shows an example flowchart for building the temporal dependency model.

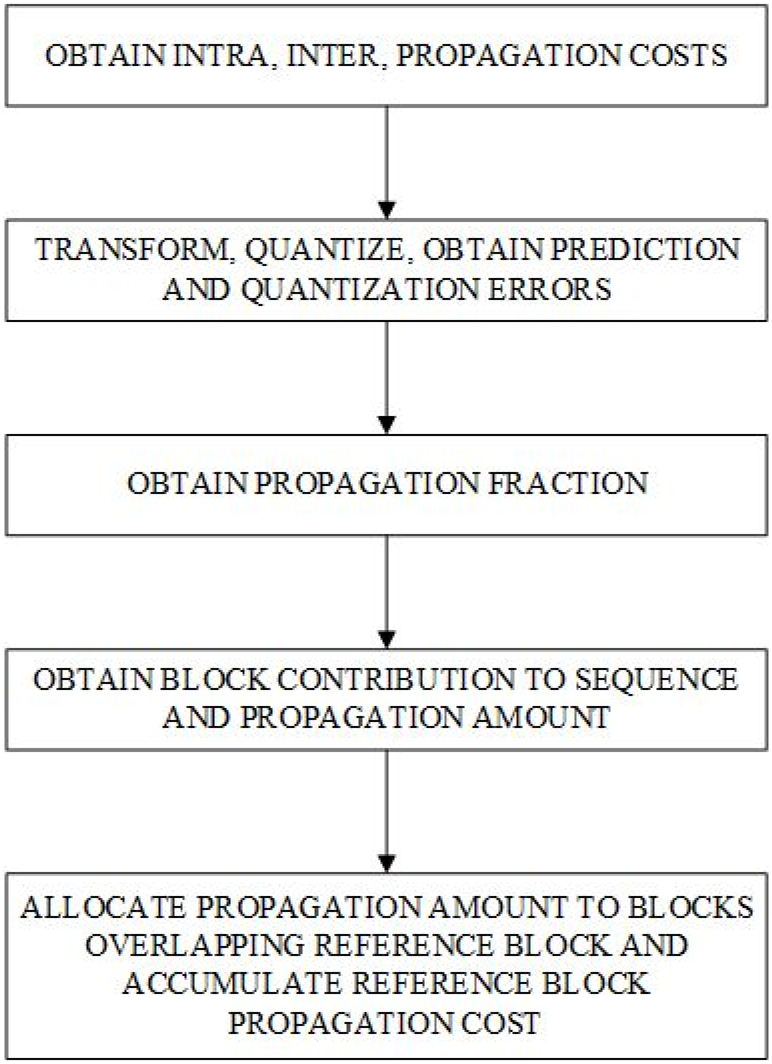

FIG. 1

As shown in FIG. 1, building the temporal dependency model may include obtaining the *intra_cost*, the *inter_cost*, and the *propagation_cost*. Building the temporal dependency model may include applying a Hadamard transform or a discrete cosine transform to the inter prediction residuals, and, subsequently, quantization. Building the temporal dependency model may include obtaining the prediction error σ_o^2 and the quantization error D_{k0}.

Building the temporal dependency model may include estimating a fraction of information from a current block to be propagated towards its reference block as shown in Equation 16.

Building the temporal dependency model may include estimating the amount of information the current block contributes to the sequence as a sum of the *intra_cost* and the *propagation_cost*. The information that the current block propagates towards the corresponding reference block may be determined as shown in Equation 3.

Building the temporal dependency model may include allocating, or dispensing, the *propagation_amount* to the blocks that overlap with the reference block. The corresponding block in the reference frame may accumulate a respective propagation cost (*propagation_cost*), which may represent backwards propagation, and which may be expressed as the following:

$$propagation\_cost \mathrel{+}= (overlap\_area / block\_area) * propagation\_amount. \quad \text{(Equation 17)}$$

Building the temporal dependency model may include obtaining the distortion propagation factor of a block may as shown in Equation 5.

The performance of the temporal dependency model (TPL) may be evaluated with reference to the performance of the MB-tree model, such as in a two-pass encoding framework, wherein the first pass gathers inter frame statistics to optimize the frame level rate control in the second pass. The efficacy of the temporal dependency model may be evaluated using the temporal dependency model may and the MB-tree model respectively to adapt the Lagrangian multiplier at 64×64 coding block level.

For each 64×64 block in a frame, the respective distortion propagation factor may be obtained. For example, for the block index i, the distortion propagation factor may be expressed as the following:

$$dist\_prop[i] = 1 + \frac{propagation\_cost[i]}{intra\_cost[i]}.$$

(Equation 18)

Determining the frame level distortion propagation factor may be expressed as the following:

$$dist\_prop = 1 + \frac{\Sigma_i\, propagation\_cost[i]}{\Sigma_i\, intra\_cost[i]}.$$

(Equation 19)

Equation 19 may be used to normalize the block distortion propagation factor in Equation 18 and the Lagrangian multiplier at 64×64 may be adapted with a multiplier λ_0 associated with the frame level quantization parameter, which may be expressed as the following:

$$\lambda[i] = \lambda_0 * \frac{dist\_prop}{dist\_prop[i]}.$$

(Equation 20)

A block with a higher relative distortion propagation factor may have a smaller Lagrangian multiplier, which biases the rate-distortion optimization to reduce the reconstruction distortion. Other algorithms may be used to optimize the rate allocation based on the temporal dependency information.

Evaluation encoding parameters may include configuring the encoder to use a maximum complexity mode for optimal compression efficiency. The evaluation may be performed using CIF and HD resolution sequences. The operating bit-rates may be set to cover a $35dB$ to $45dB$ range for each sequence. The compression performance of the Lagrangian multiplier optimization using the MB-tree model and the TPL model respectively relative to a baseline encoder in terms of BD-rate reduction, wherein negative values indicate an improvement in compression efficiency are shown in Table 1.

	MB-tree		TPL	
	PSNR	SSIM	PSNR	SSIM
basketballpass_240p	-1.72%	-3.71%	-1.63%	-4.65%
keiba_240p	-0.75%	-1.22%	-1.00%	-2.00%
football_cif	-0.18%	-0.86%	-0.20%	-1.40%
ice_4cif	-1.79%	-3.09%	-2.08%	-5.42%
RaceHorses_480p	-1.25%	-1.63%	-1.22%	-2.10%
soccer_4cif	-1.27%	-1.81%	-1.44%	-3.05%
harbour_4cif	-1.18%	-1.31%	-1.16%	-1.34%
BalloonFestival_720p	-0.59%	-3.70%	-0.75%	-4.52%
Market3_720p	-0.65%	-2.43%	-0.87%	-3.52%
parkjoy_1080p	-2.38%	-4.19%	-2.68%	-5.58%
factory_1080p	-0.54%	-0.38%	-0.69%	-0.93%
tennis_1080p	-0.82%	-0.59%	-0.90%	-1.39%
pedestrian_1080p	-2.48%	-2.95%	-2.71%	-4.07%
parkscene_1080p	-1.70%	-1.40%	-1.54%	-2.26%
ducks_take_off_1080p	-0.46%	-0.25%	-0.66%	-0.65%
cyclists_720p	-0.11%	0.204%	-0.79%	-2.70%

Table 1

The MB-tree provides fairly consistent compression gains over the baseline, where the frame level QP is optimized according to the first pass encode statistics. The TPL model further outperforms MB-tree when the innovation to quantization noise ratio varies significantly across 64x64 blocks within a frame, e.g., ice_4cif and pedestrian_1080p. When the innovation process is largely uniform across the frame, e.g., harbour_4cif, the quantization effect, as expressed in Equation 14, uniformly applies to the 64×64 blocks. The quantization effect may be substantially cancelled out by the normalization step in the above Lagrangian multiplier adaptation scheme.