R-D Spatio-Temporal Adaptive Quantization based on Temporal Distortion Backpropagation in HEVC

Michael Ropert, Julien Le Tanou, Maxime Bichon, Médéric Blestel

R-D Spatio-Temporal Adaptive Quantization based on Temporal Distortion Backpropagation in HEVC

Michael Ropert, Julien Le Tanou, Maxime Bichon and Médéric Blestel
Ericsson R&D, TV & Media
35136 Saint Jacques de la Lande, France
firstname.lastname@ericsson.com

***Abstract*—Nowadays, Rate-Distortion Optimization (RDO) is commonly used in hybrid video coding to maximize coding efficiency. Usually, the rate distortion tradeoff is explicitly computed in offline encoder implementations whereas R(D) model are used in live encoders to select the best decisions at a lower computational cost. For sake of simplicity, this (mathematical) modelling is often performed for each coding unit (CU) individually and independently, obliterating the spatial or temporal dependency between CUs. In this paper, we provide a new spatio-temporal algorithm to compute local quantizers, based on a theoretical framework able to describe the temporal distortion propagation from an R-D standpoint. In particular, we model the temporal distortion propagation making possible the retro accumulations of any (spatial) psycho-visually weighted distortion onto reference images. Using the R(D) Shannon bound, its high bitrate approximation, and a Lagrange optimization, analytical solutions are obtained for the local quantizers and the Lagrange multiplier. The proposed algorithm shows -4.4% BD-BR SSIM gains in average over state-of-the art algorithm in HEVC, using the same SSIM-based psycho-visual function.**

***Keywords*—** *Temporal Dependency, Rate-Distortion Optimization, Local Quantization, Distortion Propagation, HEVC*

I. INTRODUCTION

Rate-distortion optimization (RDO) [1] aims to minimize the distortion D subject to a rate constraint R. Lagrange multiplier method is usually used to remove the constraint on R [1], with λ the Lagrange multiplier trading between D and R. RDO ultimately minimizes the R-D cost function J defined as:

$$J = D + \lambda \times R \quad (1)$$

Video encoders based on MPEG video compression standards, such as HEVC [2], are block-based video coding systems which involves several processes (prediction, transformation, quantization, entropy coding) applied sequentially for each block of pixels within a sequence of images. These block-wise processes rely on the selection of a set of coding parameters per block. The set of coding parameters $\overrightarrow{p_i}$ for a given block i is usually optimized to minimize J, the R-D cost function, where D measures the pixel-wise distance (e.g. L2 norm) between source and compressed signals. R is the amount of bits from quantized residues and various syntax data entropy coded. For sake of simplicity RDO is commonly performed for each block or CU individually and independently, obliterating the spatial or temporal dependency between CUs. Inter-CU dependencies, such as temporal distortion propagation, are often not considered.

However, several studies have tried to model coding dependencies in order to improve the global coding efficiency. Many of these studies [3,4] focused on Rate-Control (RC) algorithms. In [3], authors approximate an Inter-frame dependent R-D model in order to optimize frame quantizers for RC purpose. Frame quantizers and model parameters are updated on-the-fly using a sliding-window at the coding stage. Fiengo et al. [4] express distortion as a convex function of all frame rates. Using multiple-pass encoding to estimate model parameters and solving a convex optimization problem, the solution ends up close to the global optimum, but is intractable for real-time application.

In [5], authors assume that distortion of one CU induces a bitrate increase when this unit is used as reference, due to a motion-compensated prediction error increase. By modelling this bitrate increment, authors propose a new RDO formulation, which results in scaling the λ parameter. However, only the direct dependency between the current CU and its reference (depth 1) is considered. In [6], a similar approach is extended for multiple-reference motion prediction and compensation, but again it only considers a "one-depth" dependency between blocks. A computationally complex proposal was made by Winken et al. [7]. This solution describes the dependencies between all coefficients levels after DCT/DST transform, for a given set of frames. It relies onto a two passes encoding process. Coefficients levels dependencies result into a multi-frame transform coefficient optimization problem, solved iteratively in order to minimize the global R-D cost over all frames. In intra coding, benefits of inter-block consideration for intra prediction mode optimization are experimented in [8].

Adaptive quantization based on CUs importance within a GOP is introduced and widely discussed in [9]. The base for the propagation of CUs inter dependencies with a non-limited (temporal) depth is laid. A well felt intuitive propagation mechanism is described based on pragmatic considerations. Consequently, the activation of the so-called "*MB-Tree*" algorithm results in an improvement in compression efficiency. However, the intuitive essence of the *MB-Tree* (or equivalently *CU-Tree*) algorithm makes complex any modifications for further improvements and some weaknesses can be pointed out:

- The nature of propagated information is debatable.
- Rate control considerations are nested with the design making R-D gains evaluations difficult.

- Only the "temporal local quantization" is based on the propagation technique: spatial and temporal local quantization contributions are combined as a weighted sum to get a final result.
- The strength of delta quantization, which is a scale factor, is a parameter to be defined by the user.

In this paper, we provide an improved algorithm to compute local quantizers, based on a theoretical framework able to describe the distortion propagation from an R-D standpoint, separated from rate control considerations and not requiring any ad-hoc spatial-temporal mix. The final goal is to get an analytical solution for the local quantizers that can be used in real time. In particular, we model the temporal distortion propagation making possible the retro accumulations of any psycho-visually weighted distortion onto reference images. Using the R(D) Shannon bound, its high bitrate approximation, and a Lagrange optimization, analytical solutions are obtained for local quantizers and the Lagrange multiplier. The proposed algorithm shows -4.4% BD-BR SSIM gains in average over state-of-the art algorithm in HEVC, using the same SSIM-based psycho-visual function.

The paper is organized as follows. In section II, the R-D local quantization optimization problem is stated, along with the particular distortion and rate constraint considered. Section III gives a description of the temporal dependencies between CUs and leads to the introduction of the proposed temporal distortion propagation model. Formulas of the forward temporal distortion and the backward recursive propagation of the distortion derivative are exhibited. The analytical solution to the global minimization is then reported in section IV. Experimental results, demonstrating the benefits of the proposed model and its theoretical framework are exhibited in section V. Finally, conclusions are given in section VI.

II. LOCAL QUANTIZATION OPTIMIZATION PROBLEM

In this paper, the only considered variables to optimize are the local quantizers. In order to introduce a psycho-visual importance to each distortion (typically based on L2 norm), a psycho-visual weighting factor ψ is introduced. The ψ function can be any psycho-visual function that modeled HVS, for instance spatial masking based on local variance, average luminance or contrast [10, 11, 12]. The objective is to find the set of local quantizers q_{i_t} of a group of pictures (GOP) able to minimize the psycho-visually weighted distortion $D_{Tot} = \sum_t \sum_{i_t} \psi_{i_t} D_{i_t}$, where t is the temporal index, and i the block or coding unit (CU) index in the image. N is the total number of CUs within an image, while T is the length of the GOP. A constraint over the global rate $R_{Tot} = \sum_t \sum_{i_t} R_{i_t}$ is added to avoid getting a trivial solution with all the quantizers q_{i_t} set to the minimum value.

The minimization problem is set as:

$$\begin{gathered} \left\{q_{k_\tau}\right\}_{k_\tau \in Idx} = ARGMIN(D_{Tot}) \\ s.t.\ \ \sum_t \sum_{i_t} R_{i_t} = R_{Tot} \end{gathered} \quad (2)$$

where Idx is the complete set of indexes in the GOP sequence: $Idx = \left\{k_\tau \setminus k \in \{1, N\}, \tau \in \{0, T-1\}\right\}$.

Thanks to the Lagrange multiplier method, this problem is re-written as follow:

$$\left\{q_{k_\tau}\right\}_{k_\tau \in Idx} \cup \{\lambda^*\} = ARGMIN\left(\underbrace{D_{Tot} + \lambda\left(\sum_t \sum_{i_t} R_{i_t} - R_{Tot}\right)}_{J_{Tot}}\right) \quad (3)$$

To solve this problem, the ideal situation would have been to dispose of motion prediction information linking blocks to its references and mode decisions used at the encoding stage, in order to build the tree of dependencies. It would require a complex two passes approach. Besides, motion estimation is an uncertain process, so prediction errors are also uncertain, as well as decided intra/inter/skip blocks and encoding modes. Consequently, instead of a two passes algorithm, a coarse probabilistic model is proposed to emulate the encoder behavior similar to [9].

III. TEMPORAL DISTORTION PROPAGATION MODEL

At a CU level, the encoding can be modeled as the addition of a distortion onto the input signal, as the reconstruction process is not lossless. Of course, a part of this distortion is propagated on spatial CU neighbors and even if not considered in this document for brevity reasons, it can be treated by the same propagation approach. For the temporal propagation, two weights are of major interest:

- p the INTER probability of a current CU. It is the proportion of the past distortions from the reference areas used for motion compensation and captured by the current CU; INTRA CUs are not temporally impacted by the past distortions.
- r the proportion of absorbed distortion considered as proportional to the surface ratio of CU copied by the motion compensation.

Several inter prediction cases (predictive or bi-predictive picture, with or without multiple references) are gathered with a unique propagation formula (r_{j_{t-1}, i_t} and $Ref(i_t)$) are adjusted to cover the various possible frame types):

$$\eta_{i_t} = p_{i_t} \sum_{j_{t_{ref}} \in Ref(i_t)} r_{j_{t_{ref}}, i_t} \cdot D_{j_{t_{ref}}} \quad (4)$$

where

- t is the temporal index (frame number), i_t is the CU index number i in the frame numbered t.
- $Ref(i_t) = \left\{j_{t_{ref1}}, \ldots j_{t_{refk}}\right\}$ is the set of reference CU covered by the pixels used by the motion compensation.
- $j_{t_{ref}}$ is one particular index belonging to $Ref(i_t)$.
- p_{i_t} is the INTER probability of the CU numbered i_t.
- $r_{j_{t_{ref}}, i_t}$ is the pixel surface ratio involved in the motion compensation to go from pixels of $j_{t_{ref}}$ to i_t.
- D_{i_t} is the CU distortion number i in the frame numbered t.
- η_{i_t} is the projected distortion (from the reference CUs) on the i^{th} CU in the frame numbered t.

The distortion of the i^{th} CU in the frame numbered t is then the summation of its own distortion and the projected one:

$$D_{i_t} = \eta_{i_t} + d_{i_t} \tag{5}$$

Where d_{i_t} is the intrinsic distortion of the i^{th} CU in the frame numbered t. d_{i_t} depends on its quantizer q_{i_t}. The past distortions do not depend on q_{i_t} due to the referencing and the propagation mechanisms. q_{i_t} only impacts future distortions made on next coded CUs. Consequently, in terms of optimization, only the impact on future distortions has to be taken into account, and without loss of generality, we can start from current image and try to minimize the impact of the distortion over a stack of T images.

A. Forward temporal distortion

One "next" (in coding order) CU is partially impacted (according to its motion prediction information and its INTER probability) by the distortion produced by the "current" CU. Generalizing the propagation to several images leads to successively accumulate distortions along the GOP, such the total distortion formula D_{Tot} is:

$$\sum_{t=0}^{T-1}\left(\sum_{i_t}\psi_{i_t}\left(p_{i_t}\sum_{i_{t-1}\,\epsilon\,Ref(i_t)} r_{i_{t-1},i_t}\left(p_{i_{t-1}}\sum_{i_{t-2}\,\epsilon\,Ref(i_{t-1})} r_{i_{t-2},i_{t-1}}\dots\right.\right.\right.$$
$$\left.\left.\left.\dots\left(\dots p_{i_1}\sum_{i_0\,\epsilon\,Ref(i_1)} r_{i_0,i_1}d_{i_0}+d_{i_1}\right)+\dots\right)+d_{i_t}\right)\right) \tag{6}$$

We then introduce the derivative of the total distortion with the hypothesis that the distortion attached to a spatial position only depends on its local quantizer and that the bits produced (rate) only depend on its local quantizer. After some mathematical development, we get:

$$\frac{\partial D_{Tot}}{\partial q_{k_\tau}} = \frac{\partial d_{k_\tau}}{\partial q_{k_\tau}}\left(\underbrace{\psi_{k_\tau}+\sum_{t=\tau+1}^{T-1}\sum_{i_t}\sum_{i_{t-1}\,\epsilon\,Ref(i_t)}\dots\sum_{i_{\tau+1}\,\epsilon\,Ref(i_{\tau+2})} P}_{V(k_\tau,\ T)}\right) \tag{7}$$

with $P = \psi_{i_t} p_{i_t}\, r_{i_{t-1},i_t} p_{i_{t-1}}\, r_{i_{t-2},i_{t-1}} \dots\, r_{i_{\tau+1},i_{\tau+2}} p_{i_{t+1}} r_{k_\tau,i_{t+1}}$.

B. Backward temporal distortion derivative

The distortion propagation can be described by a particular multi-layer perceptron [13], where each image is a layer and each CU is a linear neuron. Probabilities and pixel surface ratios are the weights. It can also be shown that the value of $V(k_0,\ T)$ can be obtained directly by the backpropagation [14] of the ψ_{k_τ} values along the GOP onto the first image 0. The recursion to apply is defined by:

$$U_{k_{\tau-1}} = \sum_{i_\tau} p_{i_\tau}\,\rho_{i_{\tau-1},i_\tau}\,U_{i_\tau} + \psi_{k_{\tau-1}} \text{ and } U_{n_{T-1}} = \psi_{n_{T-1}} \tag{8}$$

$$\text{with } \rho_{j_{t-1},i_t} = \begin{cases} 0 & if \quad j_t \notin Ref(i_t) \\ r_{j_{t-1},i_t} & if \quad j_{t-1} \in Ref(i_t) \end{cases}$$

More generally, all along the propagation:

$$U_{k_\tau} = V(k_\tau, T) \tag{9}$$

U_{k_τ} is an accumulation factor that is dependent on neither the distortion nor the rate. It is just a particular value weighting the distortion derivative.

$$\frac{\partial D_{Tot}}{\partial q_{k_\tau}} = \frac{\partial d_{k_\tau}}{\partial q_{k_\tau}}\,U_{k_\tau} \tag{10}$$

IV. Analytical soluton

A. Computation of the Lagrangian

The problem described by (3) consists in finding the minimum of expression J_{Tot}. The necessary condition to find the minimum of J_{Tot} is determined by the condition of all the derivatives equal to zero $\forall k \in \{1, \dots, N\}$, $\forall \tau \in \{0, \dots, T-1\}$:

$$\frac{\partial J_{Tot}}{\partial q_{k_\tau}} = \frac{\partial D_{Tot}}{\partial q_{k_\tau}} + \lambda \frac{\partial}{\partial q_{k_\tau}} \sum_{t=0}^{T-1} \sum_{i_t} R_{i_t} = 0 \tag{11}$$

We assume the simplified hypothesis of independence of rates:

$$\frac{\partial R_{ij}}{\partial q_{i'j'}} = \frac{\partial R_{ij}}{\partial q_{ij}}\,\delta_{(i-i',j-j')} \tag{12}$$

We consequently obtain the total rate derivative:

$$\frac{\partial}{\partial q_{k_\tau}} R_{Tot} = \frac{\partial R_{k_\tau}}{\partial q_{k_\tau}} \tag{13}$$

And

$$\frac{\partial d_{k_\tau}}{\partial q_{k_\tau}} U_{k_\tau} + \lambda \frac{\partial R_{k_\tau}}{\partial q_{k_\tau}} = 0 \tag{14}$$

Then simplifying (14) and combining with (5):

$$\lambda^* = -U_{k_\tau}\left(\frac{\partial R_{k_\tau}}{\partial d_{k_\tau}}\right)^{-1} = -U_{k_\tau}\left(\frac{\partial R_{k_\tau}}{\partial D_{k_\tau}}\right)^{-1} \tag{15}$$

Introducing the R-D Shannon bound, with $\sigma_{k_\tau}^2$ the variance of the residual signal for the block k_τ, c a constant depending on the statistical properties of the transformed coefficients [15]:

$$R_{k_\tau} = -\frac{1}{2} log_2\left(\frac{D_{k_\tau}}{c \cdot \sigma_{k_\tau}^2}\right) \tag{16}$$

We deduce the optimal Lagrange multiplier:

$$\lambda^* = 2 \cdot ln(2) \cdot U_{k_\tau} \cdot D_{k_\tau} \tag{17}$$

B. Computation of $\boldsymbol{R_{k_\tau}}$

To simplify equation writings we define $\lambda' = \lambda/(2 \cdot ln(2))$

$$\lambda' = U_{k_\tau} \cdot D_{k_\tau} \Rightarrow log_2(\lambda') = log_2\left(U_{k_\tau} \cdot D_{k_\tau}\right) \tag{18}$$

Summing on both side over all the CU of the GOP:

$$log_2(\lambda') = \frac{\sum_{t=0}^{T-1}\sum_{i_t} log_2(U_{i_t}\cdot D_{i_t})}{T\cdot N} \quad (19)$$

Combining (18) and (19) and separating the log of products:

$$\frac{\sum_{t=0}^{T-1}\sum_{i_t} log_2(U_{i_t}\cdot D_{i_t})}{T\cdot N} = log_2\left(U_{k_\tau}\cdot D_{k_\tau}\right) \quad (20)$$

At the other side, we compute the $\frac{2\,R_{Tot}}{t\cdot N}$ based on the R-D Shannon bound and mix it with the previous equality:

$$\begin{aligned}\frac{2\,R_{Tot}}{T\cdot N} &= \frac{2}{T\cdot N}\sum_{t=0}^{T-1}\sum_{i_t} R_{i_t} \\ &= \underbrace{-\frac{\sum_{t=0}^{T-1}\sum_{i_t} log_2\left(D_{i_t}\right)}{T\cdot N} + \frac{\sum_{t=0}^{T-1}\sum_{i_t} log_2\left(c\cdot\sigma_{i_t}^2\right)}{T\cdot N}}_{Shannon\ bound\ utilization} \\ &= \underbrace{-log_2\left(U_{k_\tau}\cdot D_{k_\tau}\right) + \frac{\sum_{t=0}^{T-1}\sum_{i_t} log_2\left(U_{i_t}\right)}{T\cdot N}}_{replacing\ -\frac{1}{T\cdot N}\sum_{t=0}^{T-1}\sum_{i_t} log_2\left(D_{i_t}\right)} \\ &\quad + \frac{\sum_{t=0}^{T-1}\sum_{i_t} log_2\left(c\cdot\sigma_{i_t}^2\right)}{T\cdot N}\end{aligned} \quad (21)$$

Adding a null contribution $log_2\left(c\cdot\sigma_{k_\tau}^2\right) - log_2\left(c\cdot\sigma_{k_\tau}^2\right)$ to (21), we can exhibit R_{k_τ} :

$$\begin{aligned}R_{k_\tau} = \frac{R_{Tot}}{T\cdot N} + \frac{1}{2}\Big(&log_2\left(U_{k_\tau}\right) + log_2\left(c\cdot\sigma_{k_\tau}^2\right) \\ &- \frac{\sum_{t=0}^{T-1}\sum_{i_t} log_2\left(U_{i_t}\right)}{T\cdot N} - \frac{\sum_{t=0}^{T-1}\sum_{i_t} log_2\left(c\cdot\sigma_{i_t}^2\right)}{T\cdot N}\Big)\end{aligned} \quad (22)$$

C. Computation of Δqp_{k_τ}

From (22), and re-introducing the R-D Shannon bound in R_{k_τ}:

$$\begin{aligned}-log_2\left(D_{k_\tau}\right) = \frac{2\,R_{Tot}}{T\cdot N} &+ log_2\left(U_{k_\tau}\right) \\ &- \frac{\sum_{t=0}^{T-1}\sum_{i_t} log_2\left(U_{i_t}\right)}{T\cdot N} - \frac{\sum_{t=0}^{T-1}\sum_{i_t} log_2\left(c\cdot\sigma_{i_t}^2\right)}{T\cdot N}\end{aligned} \quad (23)$$

Utilizing the high bitrate approximation: $D_{k_\tau} = \frac{\left(q_{k_\tau}\right)^2}{12}$; and the float quantizer to the quantification parameter relation $q_{k_\tau} = 2^{\frac{qp_{k_\tau}-4}{6}}$:

$$\begin{aligned}log_2\frac{\left(2^{\frac{qp_{k_\tau}-4}{6}}\right)^2}{12} &= \frac{qp_{k_\tau}-4}{3} - log_2(12) \\ &= -\frac{2\,R_{Tot}}{T\cdot N} + \frac{\sum_{t=0}^{T-1}\sum_{i_t} log_2\left(U_{i_t}\right)}{T\cdot N} \\ &\quad -log_2\left(U_{k_\tau}\right) + \frac{\sum_{t=0}^{T-1}\sum_{i_t} log_2\left(c\cdot\sigma_{i_t}^2\right)}{T\cdot N}\end{aligned} \quad (24)$$

We get:

$$\begin{aligned}qp_{k_\tau} = 3\Bigg[&-\frac{2\,R_{Tot}}{T\cdot N} - \Bigg(log_2\left(U_{k_\tau}\right) - \frac{\sum_{t=0}^{T-1}\sum_{i_t} log_2\left(U_{i_t}\right)}{T\cdot N} \\ &- \frac{\sum_{t=0}^{T-1}\sum_{i_t} log_2\left(c\cdot\sigma_{i_t}^2\right)}{T\cdot N}\Bigg) + log_2(12)\Bigg] + 4\end{aligned} \quad (25)$$

If we assume that the whole sequence is encoded with a unique quantizer ($q = 2^{\frac{qp-4}{6}}$) then the R_{Tot} expression becomes:

$$\begin{aligned}\frac{2\,R_{Tot}}{T\cdot N} &= \frac{2}{T\cdot N}\sum_{t=0}^{T-1}\sum_{i_t} R_{i_t} \\ &= -\frac{1}{T\cdot N}\sum_{t=0}^{T-1}\sum_{i_t} log_2\left(D_{i_t}\right) + \frac{1}{T\cdot N}\sum_{t=0}^{T-1}\sum_{i_t} log_2\left(c\cdot\sigma_{i_t}^2\right) \\ &= -\frac{1}{T\cdot N}\sum_{t=0}^{T-1}\sum_{i_t}\left(\frac{qp-4}{3} - log_2(12)\right) \\ &\quad + \frac{1}{T\cdot N}\sum_{t=0}^{T-1}\sum_{i_t} log_2\left(c\cdot\sigma_{i_t}^2\right)\end{aligned} \quad (26)$$

Combining (25) and (26), the final equation is then:

$$\Delta qp_{k_\tau} = qp_{k_\tau} - qp = -3\left(log_2\left(U_{k_\tau}\right) - \frac{\sum_{t=0}^{T-1}\sum_{i_t} log_2\left(U_{i_t}\right)}{T\cdot N}\right) \quad (27)$$

Equation (27) is a unique formulation for any spatial only (i.e. $T = 1$), temporal only (i.e. $\psi = 1$) or spatio-temporal optimized local quantizers.

V. Experimental Results

The x265 software [16] was used for experiments due to its low complexity but still R-D efficient encoding process. In particular, we took advantage of the availability of an efficient *lookahead* estimating inter and intra prediction costs, motion prediction information, etc. prior encoding. Estimations are typically computed on half the input sequence resolution, using 8x8 block size and SATD scores. Further detailed of x265/x264 *lookahead* design can be found in [9]. Besides, recent comparison of x265 performances against HEVC Reference Model software (HM) [17] has been published in [18]. The x265 configuration settings are presented in Table I below.

TABLE I. x265 Configuration

Version	x265 version 2.4+28-f850cdbe381c
Common Settings	--preset slower --psnr --ssim --ipratio 1.1 --bframes 3 --b-adapt 0 --no-open-gop --psy-rd 0 --psy-rdoq 0 --cutree --aq-mode 0

Test conditions follow the recommendations of the Joint Collaborative Team on Video Coding (JCT-VC) [19] in Random Access configurations. Coding efficiency is measured using Bjøntegaard BD-BR [20]. BD-BR reflects the percentage of bit savings to achieve equivalent YUV distortion, measured for both Peak Signal to Noise Ratio (PSNR) and Structural Similarity (SSIM) in this paper; with the advantage for SSIM [21] to better correlate the psycho-visual video quality. Since, BD-BR is the difference of areas between two R-D functions, we choose to add a fifth R-D point at QP = 42 in order to cover a larger bitrate range with the same metric.

The reference is the x265 without adaptive quantization algorithm. Four adaptive quantization methods are compared:

1. *CUTree*: native temporal model of x265 inherited from x264 [9]; thanks to the mathematical formalism introduced in sections III and IV, the final delta qp equation can be rewritten

in a more understandable form similar to (27), with $\forall i,t \;\; \psi_{i_t} = intraCost_{i_t}/intraCost_{k_\tau}$ in the recursion (8) and $intraCost_{i_t}$ the estimated intra prediction error for the block i_t

We inform the readers that in its original form *CUTree* delta qp equation does not consider the regularization term $\sum_{t=0}^{T-1}\sum_{i_t} log_2\left(U_{i_t}\right)/(T.N)$ in (27) resulting from the rate constraint. It has been added for fair comparison against proposed model and avoids damaging bitrate drift for *CUTree*.

2. *RDTQ*: proposed temporal distortion propagation model with retro-accumulation of the distortion without spatial psycho-visual weighting. The final delta qp is given by equation (27) with $\forall i,t \;\; \psi_{i_t} = 1$ in the recursion (8); with $\psi_{i_t} = 1$, by design, we optimize PSNR-based score.

3. *CUTree+AQmode*: native temporal model and spatial model of x265 both activated; the spatial model [10] (default --aq-mode 1) uses the local pixel variances for the spatial psycho-visual criteria, such: $(\Delta qp_{k_\tau})_{spa} = \log_2(\; MAX(1,\; e_{k_\tau}^2)) - C$, with C a constant, $e_{i_t}{}^2 = (\sigma_Y^2 + \sigma_U^2 + \sigma_V^2)_{i_t}$ the cumulative-sum of pixel variances on planes Y, U and V for the block i_t. As demonstrated in [10, 22], the use of the local pixel variance weighting the MSE specifically optimizes SSIM-based score. Finally, spatial and temporal contributions are combined as a simple weighted sum of spatial and temporal delta qps, such: $(\Delta qp_{k_\tau})_{tpl} = s_{spa}(\Delta qp_{k_\tau})_{spa} + s_{tpl}(\Delta qp_{k_\tau})_{tpl}$, with $(\Delta qp_{k_\tau})_{tpl}$ the *CUTree* delta qp detailed point 1.

4. *RDSTQ(PSY)*: proposed temporal distortion propagation model with retro-accumulation of the distortion with spatial psycho-visual weighting. The weighting is based on the same criteria than point 3, designed for optimizing SSIM. The final delta qp is given by equation (27) with $\forall i,t \;\; \psi_{i_t} = 1/MAX\left(1,\; e_{i_t}\right)$ in recursion (8).

We point out that all models share the same statistics estimated on source signal. Consequently, no complexity overhead is added by the proposed *RDTQ* or *RDSTQ(PSY)* models in comparison to *CUTree* or *CUTree+AQmode*, respectively.

BD-Rate results are presented in Table II. We observe almost systematic bitrate savings against no local quantization for all considered models. *BQTerrace* sequence suffers from R-D losses in almost all cases. One explanation is that inter probability estimations in the *lookahead* tend to be close to 0.5. Consequently, the temporal distortion propagation is very uncertain with multiple mismatches between *lookahead* estimations and final encoder decisions.

The adaptive quantization models *RDTQ* and *CUTree* only consider the temporal propagation of the L2 distortion (MSE) expecting to optimize PSNR-based scores. In average, one observes very close behavior between the two models with an average bitrate savings of about -10% based on PSNR (up to -18%) and of about -12.0% based on SSIM (up to -25%). *RDTQ* has the advantage of a simpler and analytical formulation.

When combining the *AQmode* with the *CUTree*, one observes further R-D improvements in term of both BD-BR PSNR (about -1.6%) and BD-BR SSIM (about -3%). While *CUTree* tends to tremendously improve quality of reference frames versus non-reference frames, the spatial-based quantization *AQMode* will efficiently balance this trade-off according to the local spatial pixel variance, further optimizing SSIM-based scores.

The last configuration *RDSTQ(PSY)* is the proposed spatio-temporal local quantization. It shows average savings against reference of -8.0% and -19.4% for PSNR and SSIM, respectively. The spatial propagated criteria being designed for optimizing SSIM, we observes, as expected, efficiency improvements based on SSIM against *RDTQ*. We also note that despite the average BD-BR PSNR is slightly decreased, the worst BD-BR PSNR result is improved from +3.3% to +0.7% against *RDTQ*. Besides, *RDSTQ(PSY)* outperforms *CUTree+AQMode* of about -4.4% BD-BR SSIM gain using the same SSIM-based spatial psycho-visual weighting. These results exhibit the relevance of the proposed distortion model that may optimize any psycho-visual function. Most notably, worst BD-BR results (for both PSNR and SSIM) are significantly improved against *CUTree+AQMode*.

VI. Conclusion

In this paper, we demonstrate the benefits of considering inter-block dependencies for adaptive quantization. We provide a new spatio-temporal algorithm to compute local quantizers, based on a theoretical framework able to describe the temporal distortion propagation from an R-D standpoint. In particular, we model the temporal distortion propagation making possible the (temporal) retro accumulations of any (spatial) psycho-visually weighted distortion onto reference images. Using the R(D) Shannon bound, its high bitrate approximation, and a Lagrange optimization, analytical solutions are obtained for the local quantizers and the Lagrange multiplier. The proposed *RDSTQ(PSY)* algorithm shows -4.4% BD-BR SSIM gains in average over state-of-the art algorithm *CUTree+AQMode* in HEVC/x265 encoder, using the same SSIM-based psycho-visual function. However, the proposed model can be easily improved in multiple aspects. First, in considering the SKIP mode probability: prediction mode that does not induce any quantization error. Then, most probably in using more accurate Inter probability and R(D) models.

TABLE II. CODING EFFICIENCY OF ADAPTIVE QUANTIZATION MODELS OVER NO LOCAL QUANTIZATION IN X265

Test sequences		BD-BR based on PSNR					BD-BR based on SSIM			
		CUTree	RDTQ	CUTree + AQmode	RDSTQ (PSY)		CUTree	RDTQ	CUTree + AQmode	RDSTQ (PSY)
Class B	BasketballDrive	-6.4%	-6.5%	-7.1%	-3.6%		-9.3%	-9.4%	-10.6%	-19.3%
	BQTerrace	+3.5%	+3.3%	+2.5%	+0.7%		+2.6%	+3.2%	+0.7%	-15.8%
	Cactus	-9.0%	-9.2%	-9.4%	-8.2%		-8.7%	-9.2%	-9.1%	-16.9%
	Kimono	-7.7%	-7.6%	-8.1%	-4.0%		-9.8%	-9.8%	-10.4%	-10.6%
	ParkScene	-11.7%	-11.6%	-12.3%	-10.4%		-15.2%	-15.0%	-17.5%	-21.1%
	Average	**-6.3%**	**-6.3%**	**-6.9%**	**-5.1%**		**-8.1%**	**-8.0%**	**-9.4%**	**-16.7%**
Class C	BasketballDrill	-17.9%	-18.1%	-19.9%	-16.7%		-21.7%	-21.7%	-28.6%	-30.8%
	BQMall	-12.4%	-12.1%	-13.9%	-11.4%		-15.4%	-14.3%	-19.4%	-20.6%
	PartyScene	-18.0%	-17.8%	-18.5%	-17.1%		-25.0%	-24.5%	-26.3%	-28.5%
	RaceHorses	-5.0%	-5.1%	-5.2%	-0.7%		-12.6%	-13.0%	-13.1%	-16.5%
	Average	**-13.3%**	**-13.3%**	**-14.4%**	**-11.5%**		**-18.7%**	**-18.4%**	**-21.9%**	**-24.1%**
Class D	BasketballPass	-10.6%	-10.6%	-9.2%	-7.4%		-22.1%	-21.6%	-22.5%	-29.1%
	BlowingBubbles	-14.9%	-14.8%	-14.7%	-14.2%		-21.7%	-21.2%	-22.1%	-25.0%
	BQSquare	-9.5%	-8.9%	-9.9%	-7.3%		-12.1%	-10.7%	-17.0%	-26.8%
	RaceHorses	-7.1%	-7.2%	-7.0%	-3.9%		-14.3%	-14.6%	-14.9%	-21.0%
	Average	**-10.5%**	**-10.4%**	**-10.2%**	**-8.2%**		**-17.5%**	**-17.0%**	**-19.1%**	**-25.5%**
Class E	FourPeople	-13.1%	-13.2%	-22.7%	-15.4%		-9.2%	-9.2%	-23.0%	-16.7%
	Johnny	-6.8%	-7.0%	-17.6%	-8.2%		-3.2%	-3.5%	-9.6%	-10.5%
	KristenAndSara	-14.1%	-14.2%	-22.3%	-12.1%		-5.3%	-5.5%	-12.5%	-11.3%
	Average	**-11.3%**	**-11.5%**	**-20.9%**	**-11.9%**		**-5.9%**	**-6.1%**	**-15.0%**	**-12.8%**
Class F	basketballDrillText	-16.5%	-16.5%	-18.4%	-15.0%		-19.1%	-18.7%	-26.2%	-29.4%
	chinaspeed	-13.9%	-14.1%	-14.0%	-5.4%		-6.2%	-6.2%	-7.2%	-14.8%
	slideediting	-0.9%	-1.2%	-1.5%	-0.8%		-1.2%	-1.5%	-1.9%	-4.7%
	slideshow	-9.6%	-11.1%	-4.5%	+0.3%		-10.3%	-12.2%	-9.6%	-18.6%
	Average	**-10.2%**	**-10.7%**	**-9.6%**	**-5.3%**		**-9.2%**	**-9.6%**	**-11.2%**	**-16.8%**
All	**Average**	**-10.1%**	**-10.2%**	**-11.7%**	**-8.0%**		**-12.0%**	**-11.9%**	**-15.0%**	**-19.4%**
	Best	**-18.0%**	**-18.1%**	**-22.7%**	**-17.1%**		**-25.0%**	**-24.5%**	**-28.6%**	**-30.8%**
	Worst	**+3.5%**	**+3.3%**	**+2.5%**	**+0.7%**		**+2.6%**	**+3.2%**	**+0.7%**	**-4.7%**

REFERENCES

[1] GJ Sullivan, T Wiegand, "Rate-distortion optimization for video compression", IEEE Signal Processing Mag., vol. 15, pp. 74-90, 1998.

[2] ISO/IEC 23008-2 HEVC (ITU-T Rec. H.265) "High Efficiency Video Coding," Final Draft International Standard and ITU-T, 2013.

[3] I. Li et al., "Inter-Dependent rate-distortion modelling for video coding and its application to rate control", IEEE Int. Conf. on Multimedia and Expo., 2014.

[4] A. Fiengo et al., "Rate Allocation in Predictive Video Coding Using a Convex Optimization Framework", IEEE Trans. Image Process., 2017.

[5] S. Li et al., "Lagrangian Multiplier Adaptation for Rate-Distortion Optimization with Inter-Frame Dependency", IEEE Transactions on Circuits and Systems for Video Technology, vol. 26, pp. 117-129, 2016.

[6] Y. Gao, C. Zhu, S. Li, "Hierarchical Temporal Dependent Rate-Distortion Optimization for Low-Delay Coding", IEEE Int. Symposium on Circuits and Systems, 2016.

[7] M.Winken, A. Roth, H. Schwarz and T. Wiegand, "Multi-Frame Optimized Quantization for High Efficiency Video Coding", Picture Coding Symposium, 2015.

[8] M. Bichon et al., "Inter-Block Dependencies Consideration for Intra Coding in H.264/AVC and HEVC standards", IEEE Int. Conf. on Acoustics, Speech and Signal Processing, 2017.

[9] J. Garrett-Glaser, "A novel macroblock-tree algorithm for high-performance optimization of dependent video coding in H264/AVC", [Online]. Available: http://x264.nl/developers/Dark_Shikari/MBtree paper.pdf , 2011.

[10] J. Garrett-Glaser, "Variance-based adaptive quantization", [Online]. Available: http://permalink.gmane.org/gmane.comp.video.x264.devel/8875, 2012.

[11] S. Winkler, "Digital Video Quality: Vision Models and Metrics" in Chichester:Wiley, pp. 35-149, 2005.

[12] S. Rimac-Drlje, D. Zagar, G. Martinovic, "Spatial Masking and Perceived Video Quality in Multimedia Applications", Int. Conf. on Systems, Signals and Image Processing, 2009.

[13] F. Rosenblatt. "Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms". Spartan Books, Washington DC, 1962.

[14] D. E. Ruineihart, & al, "Learning Internal Representation by Error Propagation", ICS Report 8506, 1985.

[15] T. Wiegand and H. Schwarz, "Source Coding: Part I of Fundamentals of Source and Video Coding," Foundations and Trends in Signal Processing, vol. 4, pp. 1–222, 2011.

[16] X265, [Online]. Available: https://bitbucket.org/multicoreware/x265.

[17] K. McCann et al., "High Efficiency Video Coding (HEVC) Test Model 16 (HM 16) Encoder Description", JCTVC-R1002, 2014.

[18] D. Grois, T. Nguyen, D. Marpe, "Coding Efficiency Comparison of AV1/VP9, H.265/MPEG-HEVC, and H.264/MPEG-AVC Encoders", Picture Coding Symposium, 2016.

[19] F. Bossen, "Common test conditions and software reference configurations", JCTVC-L1100, 2013

[20] G. Bjøntegaard, "Calculation of average PSNR differences between RD-curves", VCEG-M33, 2001

[21] Z. Wang et al., "Image quality assessment: From error visibility to structural similarity", IEEE Trans. Image Process., vol. 13, no. 4, pp. 600-612, 2004.

[22] C. Yeo, H. L. Tan, Y. H. Tan, "SSIM-based adaptive quantization in HEVC", IEEE Int. Conf. Acoustics, Speech and Signal Processing, 2013.