Fast HEVC Intra Mode Decision Based on RDO Cost Prediction

Mohammadreza Jamali, *Student Member, IEEE*, and Stéphane Coulombe, *Senior Member, IEEE*

***Abstract*—High efficiency video coding (HEVC) increases the number of intra coding modes to 35 to provide higher coding efficiency than previous video coding standards. This results in an increased encoder complexity, since there are more modes to be processed by the high resource-demanding rate-distortion optimization (RDO). In this paper, we propose a novel method to reduce the HEVC intra mode decision computational complexity and encoding time. This method is based on the prediction of the RDO cost of intra modes from a low-complexity sum of absolute transformed differences-based cost. By predicting the RDO cost, we are able to exclude non-promising modes from further processing and thereby save substantial computations. Also, a gradient-based method, using the Prewitt operator, is proposed to eliminate the non-relevant directional modes from the list of candidates. For even more complexity reduction, a mode classification is proposed to adaptively reduce chroma intra modes based on block texture. Experimental results show that we achieve a 47.3% encoding time reduction on average with a negligible quality loss of 0.062 dB for the Bjøntegaard delta peak signal-to-noise ratio when we compare our method to the HEVC test model 15.0.**

***Index Terms*—High efficiency video coding (HEVC), H.265, video compression, intra video coding, mode decision.**

I. Introduction

HIGH Efficiency Video Coding (HEVC) [1] was developed by the Joint Collaborative Team on Video Coding (JCT-VC) to improve the coding efficiency as compared to previous video coding standards. It provides a significant bitrate reduction compared to H.264/AVC [2], which could be up to 50% for the same subjective quality. This is achieved through a very flexible coding structure which uses many new advanced tools. For intra prediction, compared to the 9 modes used in H.264/AVC, the number of modes are increased to 35, including dc, planar and the 33 directional modes shown in Fig. 1. The increased intra modes allow the HEVC encoder to perform predictions more precisely, and to consequently exploit the spatial correlation efficiently. To find the best candidate among these 35 modes, which results in the most efficient coding performance, a highly complex rate-distortion optimization (RDO) needs to be applied to each mode. Because of this process, the computational complexity and encoding time are increased significantly, making real-time coding challenging.

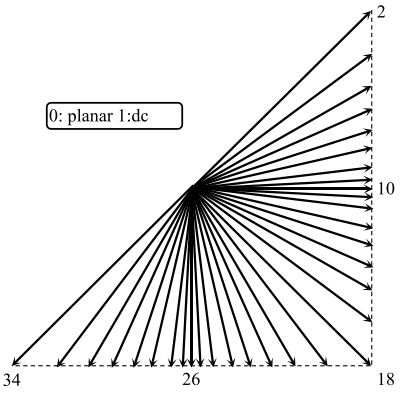

Fig. 1. HEVC intra prediction modes.

To lower this high complexity, in this paper, we propose an RDO cost prediction method, based on statistical models, to select only promising candidates for RDO. Based on this modeling, very few candidates are processed by RDO, which therefore results in a very low-complexity and high quality mode decision. Also, a fast chroma mode decision, which results in extra complexity reduction, is proposed to reduce the number of RDO candidates by applying a mode classification. Compared to our previous works, [3], [4], the contributions of this work are as follows:

- We develop an RDO cost statistical modeling (justified by goodness of fit test) with a new method for excluding non-promising candidates from further processing by rate-distortion optimization.
- We perform candidates selection based on bivariate normal distributions for RDO costs and considering correlation among these distributions.
- We carry out a chroma mode classification for fast chroma mode decision and conduct a performance analysis of different proposed fast chroma mode decision approaches.
- We conduct a performance analysis of different gradient kernels.

Compared to other works in this area, our method achieves the best trade-off between complexity reduction and coding

Manuscript received January 20, 2018; revised May 9, 2018; accepted May 10, 2018. This work was supported in part by the Vantrix Corporation, and in part by the Natural Sciences and Engineering Research Council of Canada through the Collaborative Research and Development Program under Grant NSERC-CRD 428942-11. *(Corresponding author: Mohammadreza Jamali.)*

The authors are with the Department of Software Engineering and Information Technology, École de technologie supérieure, Université du Québec, Montreal, QC H3C 1K3, Canada (e-mail: mohammadreza.jamali.1@ens.etsmtl.ca; stephane.coulombe@etsmtl.ca).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TBC.2018.2847464

Fig. 2. Frame splitting into CUs (red) and PUs (green) based on the quadtree structure, BlowingBubbles.

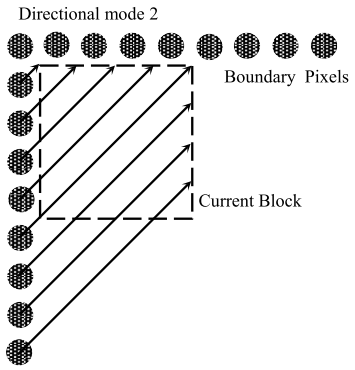

Fig. 3. Intra prediction based on directional mode 2.

efficiency by excluding candidates with no potential to be the best mode, and which would thus waste computational resources.

The rest of the paper is organized as follows. In Section II, a brief overview of the HEVC intra coding is presented. Section III discusses some related works in this area. Sections IV–VI present our proposed method for intra mode decision complexity reduction. Experimental results are shown in Section VII. Finally, Section VIII concludes the paper.

II. HEVC Intra Coding

HEVC intra coding follows the same structure as previous hybrid video codecs [5]. It is mainly based on spatial prediction and transform coding. However, it carries some new features, such as an increased number of prediction modes and a quadtree-based block splitting structure, which contribute to improved coding efficiency. This quadtree structure is based on the coding tree unit (CTU) as the root of the tree with maximum size of 64×64. A CTU is split recursively into coding units (CUs), and each coding unit (CU) is partitioned into prediction units (PUs) and transform units (TUs). The CU is a block which is coded by inter or intra prediction, and contains one or more non-overlapping PUs, as well as a quadtree of TUs. PUs inside an intra CU can be predicted by different intra modes. Fig. 2 shows how a frame is split into CUs and PUs based on the quadtree structure.

HEVC intra coding supports 35 prediction modes for the luma component for all PU sizes. There are 33 directional modes, allowing an efficient prediction of different directional video contents, a dc mode for homogeneous regions, and a planar mode to predict the smooth surfaces. Using the dc and planar modes let HEVC intra predict areas of the image which do not follow an edge model. Fig. 3 shows an example of intra prediction for a 4×4 block using directional mode 2. For chroma components, HEVC provides five intra modes, regardless of the block size, which include planar, dc, vertical, horizontal and the best luma mode. If the best luma mode is one of the first four modes, a diagonal mode (34) is used as the fifth mode. Using this concept of derived mode results in efficient signaling when chroma content follows the same structure as the luma, and thus, the same intra mode, which is a very likely scenario.

TABLE I
Number of Intra Mode Candidates, Selected by the HM, for Different PU Sizes

PU size	64×64	32×32	16×16	8×8	4×4
N	3	3	3	8	8

Although this increased number of modes results in an improved coding efficiency, it renders the encoding process much more complex due to the increased number of RDO computations required. To select the best luma mode, HEVC test model (HM) [6] selects N candidates among 35 modes based on the rough mode decision (RMD) cost and computes the RDO cost only for these candidates to find the best mode. Table I shows N for different PU sizes. The RMD cost (C_{RMD}) is computed as:

$$C_{RMD} = SATD_l + \lambda_p \times R_p \tag{1}$$

where $SATD_l$, as a measure of distortion, is the sum of absolute transformed differences (SATD) between the original luma component and its prediction block, R_p is the number of bits used for mode coding, and λ_p is the Lagrange multiplier, which is determined based on the quantization parameter (QP) [7]. The RDO cost (C_{RDO}) for the N selected candidates is computed using the following cost function:

$$C_{RDO} = (SSE_l + \omega_c \times SSE_c) + \lambda_m \times R_m \tag{2}$$

where SSE_l and SSE_c are the sum of squared errors (SSE) between the original luma and chroma components and their reconstructed blocks, ω_c is the chroma weight (depending on the QP), R_m is the number of bits for PU encoding and λ_m is the Lagrange multiplier. It should be mentioned that before RDO computations, a maximum of three most probable modes (MPMs) are added to the candidates list based on the best modes of the neighboring blocks. The final prediction mode is the candidate with the minimum RDO cost. There is no rough mode decision for chroma intra prediction, and to select the best mode among the five candidates, HM evaluates the RDO cost for all of them, which makes the chroma component a complex module as well. In the following section, we review some works which propose methods to reduce the encoding complexity by decreasing the number of modes to be processed by RDO.

III. Related Work

Many works have proposed different methods for reducing the complexity of HEVC and H.264 intra coding. These methods propose decreasing the number of modes processed by RDO (mode decision) [8]–[17], avoiding exhaustive search to find the optimal CU sizes inside a CTU (fast CU size decision) [18]–[23], pruning the transform tree to speed up the encoding process (transform tree optimization), or a combination of these techniques [24]–[33]. Since we are proposing a mode decision approach, in this review section, we focus on this category in order to include major state-of-the-art works which aim at decreasing the number of modes to be processed by RDO. Each method in this category can be viewed as an independent mode decision module and can be combined with other works to design a faster encoder.

In [8], a novel algorithm is presented based on the orientation detection which uses the local directional variance along some predefined lines. Based on this analysis, orientations with the lowest directional variances are considered as dominant orientations, and the number of intra mode candidates to be further processed by RDO is reduced accordingly.

Marzuki *et al.* [9] propose a context-adaptive fast intra coding based on the best modes of the upper CU and neighboring PUs. Further, their method includes early termination approaches for RDO, as well as a residual quadtree process (RQT) using adaptive thresholding.

Pakdaman *et al.* [10] propose a novel method based on dual-tree complex wavelet transform (CWT) to effectively determine edges leading to the selection of the most appropriate candidates for intra mode decision. This approach considers only few adjacent candidates for the best estimated mode selected by the edge analysis. CWT decomposes an image into various frequency sub-bands. The proposed algorithm uses the energy distribution in these sub-bands to exploit texture information and obtain the dominant direction for each block.

Tariq *et al.* [11] propose a method for intra coding complexity reduction based on a quadratic relation between the RDO cost and the sum of absolute differences (SAD). Using this method, it is possible to avoid entropy coding, Hadamard transform, and distortion computations. To this end, they formulate the distortion and rate as functions of SAD and QP in order to avoid performing the RDO process.

A method is presented in [12] for fast intra HEVC coding based on three optimized candidate sets. These sets include 1, 19 and 35 modes. Using the neighboring reference samples, the encoder selects the optimal set for each prediction unit (PU). This results in an accelerated encoder due to the reduced number of modes needed to be processed by RDO for the first two sets. Further, the number of bits needed for mode signaling is reduced.

In [13], the PU is converted from the pixel domain to the transform domain, and the main directions are determined in the latter. Based on these directions, a short list of candidate modes is selected for the RDO process.

Zhang *et al.* [24] propose a low-complexity HEVC intra coding method based on both fast mode decision and early CU size determination. For the fast mode decision, a gradient-based approach is used to reduce the number of candidate modes. For the fast CU size decision, texture homogeneity along with a support vector machine (SVM)-based approach are used to make early CU splitting and early CU termination decisions. The SVM uses the depth differences, Hadamard cost and rate-distortion (RD) cost as features.

In [25], a low-complexity intra HEVC encoder is proposed which is based on a fast PU mode decision (FPUMD) and a fast coding unit size decision (FCUSD). The FPUMD reduces the number of intra mode candidates based on the correlation of the PU mode and RD cost of the different depth levels. The FCUSD for its part excludes unnecessary CU sizes from further processing, using the depth of neighboring CUs and a RD cost threshold which is determined from previous frames. To update the coding parameters, an online method is used, and leads to an accurate decision for various sequences types.

In [26], a method is proposed for fast intra coding, and achieves significant time reduction by using the Hadamard cost-based progressive rough mode search (pRMS) and early CU split termination. Using pRMS, the algorithm selectively checks the potential modes instead of traversing all candidates. This allows fewer modes to enter the RDO process. Further, the early CU split termination excludes the lower depths if the estimated RD cost (aggregated cost of the sub-CUs) is already larger than the RD cost of the current CU.

In [31], a gradient-based pre-processing step is proposed to reduce the number of intra mode candidates. Moreover, they propose a gradient-based approach, based on texture complexity, to make an early decision for CU splitting.

A method is presented in [34] for inter CU size decision based on SAD estimation. At the first step, a new motion estimation (ME) algorithm is proposed which calculates the SAD costs for both upper CU and its sub-CUs. At the next step, a motion compensation rate-distortion cost is defined and is modeled based on the SAD cost using an exponential model. This modeling is used to decide on the CU size and make a fast CU size decision. In contrast to this work, we propose a stochastic model for intra RDO cost based on normal distributions which uses SATD for fast intra mode decision.

These works propose notable and novel ideas regarding intra mode decision complexity reduction. However, they exhibit some shortcomings, which could be addressed in a bid to design even faster HEVC intra encoders. In gradient-based approaches, gradients could be computed at the CTU level, instead of at the PU level, and then reused for every block inside the CTU. This makes it possible to use the same edge information at each depth and to avoid repetitive calculations for each block. Also, since the picture edges are not perfectly aligned with intra modes, works in this area need to propose a solid strategy to relate the gradients at each pixel to directional intra modes. Relying only on the modes of neighboring blocks as is proposed by some works, may result in a domino effect, where a wrong decision can propagate to other blocks and affect coding efficiency negatively. To avoid this drawback, these kinds of algorithms need to be accompanied by other approaches. Other works aimed at reducing the number

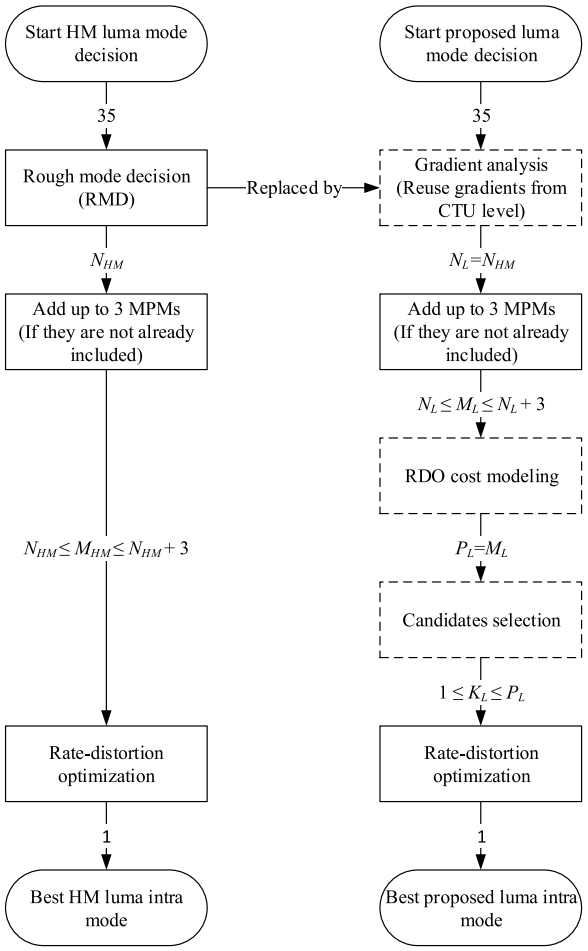

Fig. 4. Block diagram of luma mode decision at PU level for HM (left) and proposed method (right). Arrows show the number of modes in each step.

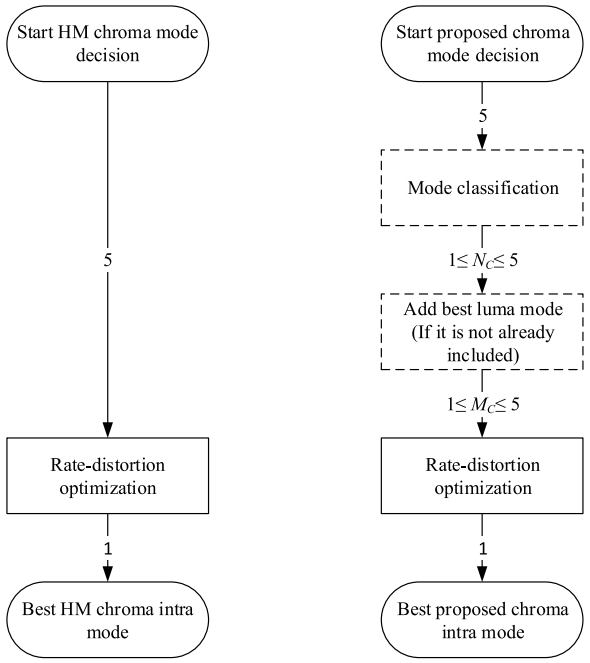

Fig. 5. Block diagram of chroma mode decision at PU level for HM (left) and proposed method (right). Arrows show the number of modes in each step.

of modes for RDO process based on low-complexity measures keep N best modes as a candidates list. Using a fixed number of candidates, they either waste computations by including modes with no potential to be the best mode or reduce the coding efficiency by excluding the best mode. Moreover, those methods, which select an adaptive number of candidates, end up with limited time reduction or high quality loss due to imperfect RDO candidates selection. In the next sections, we propose our method for intra mode decision complexity reduction to improve other approaches in this area by addressing the shortcomings noted.

IV. General Framework of the Proposed Method

To reduce the intra mode decision complexity, we add new modules to the HM. Fig. 4 shows the luma mode decision of the proposed method compared to HM at the PU level, where it illustrates the mode decision steps. In this diagram, N_{HM} and N_L are same as N in Table I. RMD is replaced by a gradient analysis step to select N_L luma promising candidates. Since we apply gradient computations for each CTU once and use the same information for all PUs inside the CTU, the process is much faster than the original RMD. Then, to exploit the spatial correlation, the same approach as HM is used to add three MPMs from neighboring blocks.

In the next step, a statistical model is proposed allowing the RDO cost to be predicted without the need to perform the RDO process. Next, based on this model, a reduced and adaptive number of candidates are selected to be processed by the high-complexity and high-demanding RDO. The entire algorithm keeps K_L promising candidates, which based on their RDO cost prediction, have a high potential to be the best intra luma mode for the current PU. For the chroma mode decision, we propose a mode classification which categorizes five modes into two groups of promising and non-promising candidates, and excludes non-promising ones from RDO. However, we always add the best luma mode to the group of promising candidates if it is not already included in the list. Fig. 5 shows the chroma mode decision of the proposed method compared to HM. The following presents the proposed method in detail.

V. Fast Luma Intra Mode Decision

A. Gradient Analysis

Although RMD is less complex than RDO, it still consumes a considerable amount of time, and needs to be replaced by a less complex process. To this end, we apply a gradient-based approach to exclude irrelevant angular modes from further processing. Sobel, Scharr, Prewitt and Roberts cross are computationally light and popular operators used to compute the approximation of the gradient. Their kernels are as follows, respectively:

$$G_x = \begin{bmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{bmatrix} \quad G_y = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix}. \quad (3)$$

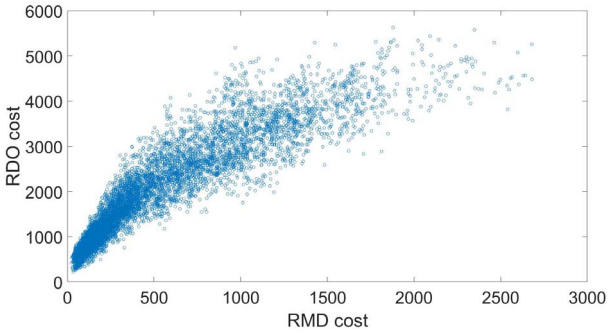

Fig. 6. Scatter plot for RMD cost and RDO cost, RaceHorses, 4×4 blocks, QP=32.

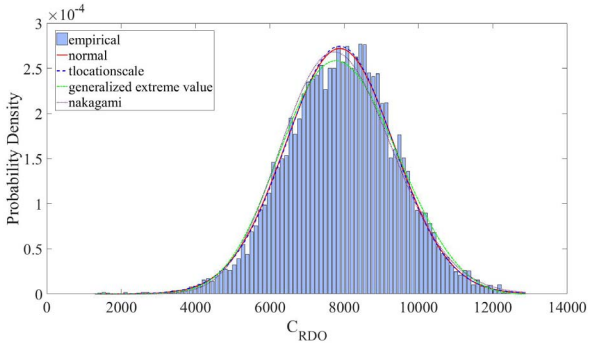

Fig. 7. Best fitted distributions for C_{RDO}, RaceHorses, 8×8 blocks, QP=32, 2903<C_{RMD}<2945.

$$G_x = \begin{bmatrix} -3 & 0 & 3 \\ -10 & 0 & 10 \\ -3 & 0 & 3 \end{bmatrix} \quad G_y = \begin{bmatrix} -3 & -10 & -3 \\ 0 & 0 & 0 \\ 3 & 10 & 3 \end{bmatrix}. \tag{4}$$

$$G_x = \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix} \quad G_y = \begin{bmatrix} -1 & -1 & -1 \\ 0 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix}. \tag{5}$$

$$G_x = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \quad G_y = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}. \tag{6}$$

For all operators, and at each pixel, the gradient magnitude is estimated by $|G_x| + |G_y|$ and the gradient direction is calculated by $\text{atan}(G_y/G_x)$. The gradient magnitude and direction are computed for each pixel at the CTU level, and this data is passed to all PUs inside the CTU. At the PU level, depending on the gradient direction, a mode is associated with each pixel. Finally, N_L most dominant modes are selected for each PU, considering the gradient magnitudes as weights.

To choose the most suitable operator, for our application, we look for the related features such as low complexity and accuracy. It needs to be faster, as compared to the RMD, and it needs to exclude only those directional modes which are unlikely to be the best mode. To this end, in Section VII, we compare these operators based on their accuracy in order to include the best mode in the candidates list; we also carry out a comparison based on the time reduction the operators provide. To the best of our knowledge, no study has carried out a performance analysis of different gradient operators in the context of HEVC intra coding.

B. RDO Cost Modeling

1) Model Selection: A predictive model for the RDO cost (C_{RDO}, Eq. (2)) is developed in this section based on a low-complexity measure to avoid the RDO process for non-promising candidates. Examples of such low-complexity measures include, but are not limited to, the RMD cost (C_{RMD}, Eq. (1)) which we use in this paper. Fig. 6 shows the correlation between the RMD cost and the RDO cost for a sample sequence, block size and QP. For this example, the Pearson correlation coefficient and the Spearman's rank correlation coefficient are 0.90 and 0.96, respectively. Similar results are observed for other sequences, block sizes and QPs. The Spearman's coefficient is more relevant to our RDO cost prediction goal since we are not looking for a linear correlation between two variables (as Pearson coefficient does), but rather, for rank correlation. That means if the relationship between the RMD cost and the RDO cost could be described by a monotonic function (Spearman's coefficient $= 1$), then the mode with the lowest RMD cost would be the same one with the lowest RDO cost, and in that case, the entire RDO process could be omitted.

Since there is no monotonic function describing the relationship between the RMD cost and the RDO cost, we propose a statistical model allowing the possible values of the RDO cost to be predicted from the RMD cost. Based on our observations, for a PU with a given C_{RMD}, or more specifically, with a given small range of C_{RMD} values, the C_{RDO} can be modeled by some well-known probability distributions. We compare some distributions based on the Bayesian information criterion [35] (BIC) to determine which of them can best describe empirical data. They are: Beta, Birnbaum-Saunders, Exponential, Extreme value, Gamma, Generalized extreme value, Generalized Pareto, Inverse Gaussian, Logistic, Log-logistic, Lognormal, Nakagami, Normal, Rayleigh, Rician, t location-scale and Weibull.

Thus, although we cannot predict a specific value for C_{RDO} based on C_{RMD}, we can predict the probability distribution function (PDF) of its values, i.e., the likelihood of each RDO cost value. Fig. 7 shows the top four models best fitted to the empirical C_{RDO} for a small range of C_{RMD} values for the RaceHorses sequence. We observe similar results for other video sequences with different QP values and block sizes with varying averages and variances. Having these models for all PU candidate modes allow us to omit non-promising candidates based on their C_{RDO} distribution, and to select an adaptive number of modes to go through the RDO process. In most cases, the best fitted model is normal, and we provide numeric results to support this claim.

2) Goodness of Fit Test: To show that the RDO costs come from normal distributions, a goodness of fit of a model is measured. To this end, tests such as Kolmogorov-Smirnov (KS), Anderson-Darling (AD) and Chi-Square (CS) can be used [36]. Table II shows sample results for these tests for the normal model. In this table, the p-value is the probability of finding the observed, or more extreme, values when the null hypothesis

TABLE II
GOODNESS OF FIT TEST RESULTS FOR THE NORMAL MODEL, RACEHORSES, 8 × 8 BLOCKS, QP=32

	CS	AD	KS
Significance level	0.05	0.05	0.05
h-value	0	0	0
p-value	0.49	0.07	0.33

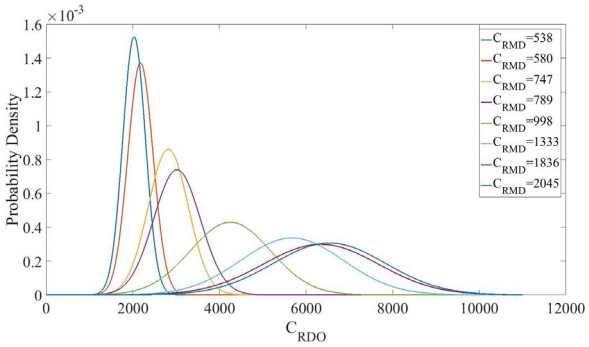

Fig. 8. Normal distributions for C_{RDO} from a sample PU of size 8 × 8, RaceHorses.

(i.e., data comes from a normal distribution) is true. If the *p*-value is greater than a significance level, the test decision or *h*-value is zero, otherwise it is 1. A recommended significance level is 0.05, which works well [36]. The test decision in our case is zero, which means that the observed data is not statistically different from what is expected from a normal distribution. Similar results are observed for different sequences with various block sizes and QPs.

C. Candidates Selection

From the previous sub-section, we consider a normal model for the RDO cost of each candidate mode. Fig. 8 shows these models for a sample PU, where $P_L = 8$. The RDO cost of each candidate mode $m_i\,(1 \leqslant i \leqslant P_L)$, with RMD cost of C_{RMD_i}, is represented by a random variable X_i which follows a normal distribution with the following parameters:

$$\begin{aligned}\mu_i &= f_\mu\big(C_{RMD_i}\big)\\ \sigma_i &= f_\sigma\big(C_{RMD_i}\big),\end{aligned} \tag{7}$$

where μ_i and σ_i are piecewise constant functions of RMD cost and obtained during the training phase. In the main phase (coding phase) for any candidate mode with a given RMD cost, these parameters are obtained based on the available functions and so the RDO cost distribution for the candidate mode is known. By using these distributions, we exclude those candidates with a low chance of being the best PU mode. To select the K_L promising candidates, we consider the probability that a particular candidate is the one with the lowest RDO cost. The probability that mode m_i is the best RDO mode is $P(X_i < Z_i)$, where

$$Z_i = \min_{\substack{1\leqslant j\leqslant P_L\\ j\neq i}} X_j. \tag{8}$$

Z_i is a random variable representing the minimum of the other $P_L - 1$ random variables. Then the following inequality is checked for all P_L candidates:

$$P(X_i < Z_i) > CL, i = 1 \ldots P_L, \tag{9}$$

where CL is a confidence level. Those candidates which satisfy the condition are selected as promising candidates, and the RDO process is run only for them. The above probability can be written as follows:

$$P(X_i < Z_i) = \int_{-\infty}^{+\infty} P(X_i = t)P(Z_i > t)dt, \tag{10}$$

where

$$P(Z_i > t) = P\left(\wedge_{\substack{j=1\\ j\neq i}}^{P_L} X_j > t\right), \tag{11}$$

where $\wedge$ is the 'logical and' symbol. Assuming the normal distributions are independent, we have:

$$\begin{aligned}P(Z_i > t) &= \prod_{\substack{j=1\\ j\neq i}}^{P_L} P\big(X_j > t\big)\\ &= \prod_{\substack{j=1\\ j\neq i}}^{P_L}\left(\frac{1}{2}\,\mathrm{erfc}\left(\frac{t-\mu_j}{\sigma_j\sqrt{2}}\right)\right).\end{aligned} \tag{12}$$

The last equality is based on the complementary cumulative distribution functions for normal distributions, where erfc is the following complementary error function:

$$\begin{aligned}\mathrm{erfc}(t) &= 1 - \mathrm{erf}(t)\\ &= \frac{2}{\sqrt{\pi}}\int_t^{\infty} e^{-s^2} ds.\end{aligned} \tag{13}$$

This gives us a mathematical formula that can be further expanded to have an analytical answer to the problem. However, implementing such a case, even with the independency assumption, is almost impractical.

To have a more practical and implementable solution, we break the problem into sub-problems consisting of comparisons of two normal distributions. In this approach, the best RMD mode (lowest C_{RMD}) and any other mode which is likely to be the best RDO mode (lowest C_{RDO}) constitute the set of K_L promising candidates. Considering the P_L random variables $X_i \sim N(\mu_i, \sigma_i^2)$ associated with P_L candidates, mode m_i is selected for the RDO process if:

$$P(X_i < X_1) > CL, \tag{14}$$

where, without loss of generality, we assume the μ_is are sorted in increasing order ($\mu_i \leqslant \mu_j,\ \forall\ i < j$) and m_i is the mode associated with X_i. X_1 has the lowest mean, and represents the best RMD mode. Without considering dependency between these variables, we can evaluate this probability as follows:

$$\begin{aligned}Y_i &= X_i - X_1\\ Y_i &\sim N\Big(\mu_i - \mu_1, \sigma_i^2 + \sigma_1^2\Big)\\ P(X_i < X_1) &= P(Y_i < 0),\end{aligned} \tag{15}$$

Algorithm 1 RDO Best Mode Selection

Input: $X_i \sim N(\mu_i, \sigma_i^2)$, $\mu_i = f_\mu(C_{RMD_i})$ and $\sigma_i = f_\sigma(C_{RMD_i})$, $1 \leqslant i \leqslant P_L$, $\mu_i \leqslant \mu_j \ \ \forall\ i < j$
Output: K_L, m_{best}
1: compute C_{RDO_1}
2: $C_{RDO_{best}} = C_{RDO_1}$
3: $m_{best} = m_1$
4: $K_L = 1$
5: **for** $i = 2$ to P_L **do**
6: **if** $P(X_i < C_{RDO_{best}}) > CL$ **then**
7: compute C_{RDO_i}
8: **if** $C_{RDO_i} < C_{RDO_{best}}$ **then**
9: $C_{RDO_{best}} = C_{RDO_i}$
10: $m_{best} = m_i$
11: **end if**
12: $K_L = i$
13: **else**
14: **break**
15: **end if**
16: **end for**
17: **return** K_L, m_{best}

where $P(Y_i < 0)$ is a familiar normal random variable probability.

To take into consideration the correlation between the normal random variables and to use the already computed RDO costs, we adopt an iterative approach which combines the two steps of candidates selection and rate-distortion optimization. Based on this approach, we carry out a successive evaluation of the probability that a certain variable is less than the current best RDO cost. Algorithm 1 shows the procedure for this. Using this algorithm, we obtain the number of candidate modes, K_L, and the best mode (m_{best}), which is the one with the lowest RDO cost. Based on this algorithm, as soon as one random variable does not satisfy the condition on line 6, we avoid checking this condition for other random variables with larger means because, based on our experiments, computing the extra RDO costs reduces the speedup, and does not improve the quality.

To compute the probability on line 6 of Algorithm 1, we show that the normal distributions related to candidate modes are, two by two, bivariate normal and then we use the properties of this joint distribution. To show a bivariate normal distribution for two random variables, many analytical methods are proposed in the literature. However, no best method exists since the results obtained by each of them are different under certain conditions. The most popular methods used to assess multivariate and bivariate normality are Mardia's, Henze-Zirkler's and Royston's, as well as graphical approaches such as Chi-Square Q-Q, perspective and contour plots [37]. Based on a comprehensive simulation study, presented in [38], on 13 statistical methods for testing the multivariate normal distribution (MVN) with a Monte Carlo study, the authors suggested using Henze-Zirkler's and Royston's tests because they show better results in terms of error control and power. We therefore use these two tests to show that the RDO costs of the candidate modes, two by two, follow a bivariate normal model. To this end, we apply MATLAB functions for these tests [39], [40] to vectors of RDO costs related to blocks with the RMD cost of C_{RMD_i} and C_{RMD_j}; hence, we have two vectors of RDO costs and it is a straightforward procedure to show they are bivariate normal using the referenced functions. Sample results are presented in Table III for these two tests. The significance level, the p-value and the test decision could be interpreted here similarly to what appears in Table II. Similar results are observed for different sequences, with various block sizes and QPs.

TABLE III
JOINT NORMALITY TEST RESULTS, RACEHORSES, 8×8 BLOCKS, QP=32

	Henze-Zirkler	Royston
Significance level	0.05	0.05
Test result	MVN verified	MVN verified
p-value	0.051	0.068

TABLE IV
CHROMA INTRA MODES

	Luma intra mode				
	0	26	10	1	X
Chroma mode (0)	34	0	0	0	0
Chroma mode (1)	26	34	26	26	26
Chroma mode (2)	10	10	34	10	10
Chroma mode (3)	1	1	1	34	1
Chroma mode (4)	0	26	10	1	X

Now, we consider the special property of a bivariate normal distribution which states that the observed value of one variable leads to a conditional distribution for the other unobserved one. If X_j and X_i are two random variables, among P_L variables, associated with two modes m_j and m_i ($j \neq i$), the mean and variance for the conditional variable $X_j|X_i = x_i$ are as follows:

$$\begin{aligned} \mu_{j|X_i=x_i} &= \mu_j + \sigma_j \rho (x_i - \mu_i)/\sigma_i \\ \sigma^2_{j|X_i=x_i} &= \sigma_j^2 \Big(1 - \rho^2\Big), \end{aligned} \tag{16}$$

where ρ is the correlation coefficient, and is obtained from training. As can be seen, the conditional mean depends linearly on the observed value, while the conditional variance does not depend on the observation. As a result of this analysis, conditional densities are used for computing the probability on line 6 of Algorithm 1, which once again, is a familiar normal random variable probability.

VI. FAST CHROMA INTRA MODE DECISION

HEVC follows a 4:2:0 chroma sampling format, which means each CTU of size $N \times N$ includes one luma coding tree block (CTB) of size $N \times N$ and two chroma CTBs of size $N/2 \times N/2$. Similarly, CUs, PUs and TUs include one luma block (coding block (CB), prediction block (PB), transform block (TB)) of size $N \times N$ and two chroma blocks of size $N/2 \times N/2$. As we mentioned in Section II, for both chroma components, HEVC presents five prediction modes, including the corresponding luma mode. Table IV shows these five modes. In this Table, X indicates luma modes other than planar

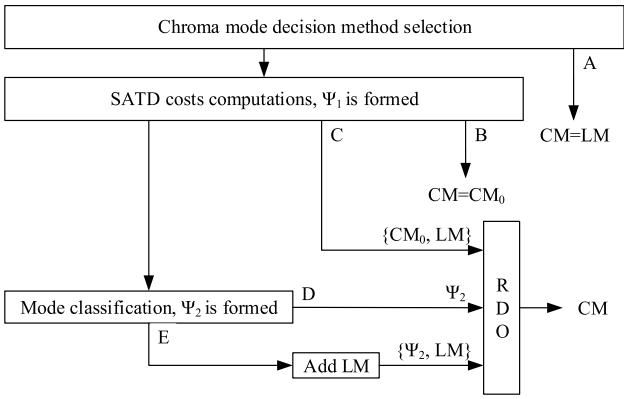

Fig. 9. Chroma mode decision methods.

(0), dc (1), vertical (26) or horizontal (10). Also, 34 indicates a diagonal mode.

To reduce chroma mode decision complexity, we propose a fast mode selection. Five different selection methods are examined, which reduce the number of modes going through the RDO process, and are compared to one another. These methods are as follows:

A) Using only the best luma mode
B) Using only the mode with the lowest SATD cost (SATD mode)
C) Using the mode with the lowest SATD cost and adding the best luma mode (SATD mode + luma mode)
D) Selecting a short list of candidate modes if their SATD costs are clearly separated by a gap from other candidates' costs (mode classification)
E) Selecting a short list of candidate modes by finding a gap and adding the best luma mode if it is not already included (mode classification + luma mode)

Fig. 9 shows how these five methods select the chroma mode (CM), among five candidates, which is the mode the encoder uses to obtain the residual for the current chroma PB. Method A simply chooses the same mode as the selected lume mode (LM). For other methods, first SATD costs of the five candidates are computed to form the set $\Psi_1 = \{\mathrm{CM}_0, \mathrm{CM}_1, \mathrm{CM}_2, \mathrm{CM}_3, \mathrm{CM}_4\}$ which includes the candidates in increasing order of their SATD cost. SATD cost is obtained based on the transformed distortion between the current PB and its predicted PB by applying the associated candidate mode. By using method B, CM_0 which is the candidate with the least SATD cost is selected as the final chroma mode. In method C, the encoder selects the mode with the least RDO cost between CM_0 and LM. For methods D and E, a mode classification step is proposed to select a subset of candidates for RDO process. Based on experimental results, this idea proved to be a very good strategy for chroma components. In this approach, we look for a manifest distance between SATD costs of two consecutive candidates from set Ψ_1. Then, based on this distance the candidates are categorized into two groups of promising and non-promising and we exclude the non-promising ones from further processing. The distance is defined as:

$$d = \alpha \times (Cost_{max} - Cost_{min}), \tag{17}$$

TABLE V
NUMBER OF SELECTED CHROMA MODES

Chroma method	M_C
A	1
B	1
C	2
D	$1 \leqslant M_C \leqslant 5$
E	$1 \leqslant M_C \leqslant 5$

Algorithm 2 Chroma Mode Classification

Input: $\Psi_1 = \{\mathrm{CM}_0, \mathrm{CM}_1, \mathrm{CM}_2, \mathrm{CM}_3, \mathrm{CM}_4\}$
Output: Ψ_2

```
1:  Ψ2 = Ψ1
2:  for i = 0 to 3 do
3:      if (Cost(CM_{i+1}) − Cost(CM_i)) > d then
4:          for j = i + 1 to 4 do
5:              Ψ2 = Ψ2\{CM_j}
6:          end for
7:          break
8:      end if
9:  end for
10: return Ψ2
```

where α is an adjustable parameter, trading off computational complexity and visual quality and, for N candidates ($N > 2$), it is constrained to the following condition:

$$\frac{1}{N-1} < \alpha < 1. \tag{18}$$

Algorithm 2 shows the procedure for mode classification which forms set Ψ_2 from set Ψ_1. Line 5 of the algorithm removes candidate CM_j from Ψ_2. Also, Table V shows the number of chroma modes, M_C, which are used in the RDO process to find the best mode. M_C is adaptive for the last two methods since when we are looking for a gap, the number of modes selected for the RDO process depends on the block context. Experimental results for these five different methods are provided in Section VII.

VII. EXPERIMENTAL RESULTS

To implement the proposed methods, we use the HEVC test model HM 15.0 and a PC equipped with an Intel Core i7-4790 CPU @ 3.60 GHz and 32 GB of RAM. The configuration and profile are set to *all-intra* and *Main profile*, respectively. We report the results, using the sequences recommended in [41], compared to HM based on time reduction (TR), Bjøntegaard delta rate [42] (BD-Rate) and Bjøntegaard delta peak signal-to-noise ratio [42] (BD-PSNR). The sequences are in different classes (A (2560×1600), B (1920×1080), C (832×480), D (416×240) and E (1280×720)) to cover various applications and resolutions. The results are averaged over four QPs: 22, 27, 32 and 37; they therefore cover low and high bitrate scenarios. We use first hundred frames of the recommended video sequences to achieve preliminary results for each sub-algorithm of our method. This number of frames makes it possible to have fair comparison with reference papers [3], [4], [14] which also use one hundred frames.

TABLE VI
ERROR RATE OF GRADIENT OPERATORS, CONFIG. A AND CONFIG. B

		Error (%)	
Video sequence	Kernel	Config A	Config B
RaceHorses	Sobel	24.7	11.9
	Prewitt	24.9	11.8
	Scharr	24.8	12.1
	Roberts cross	27.1	13.1
BasketballPass	Sobel	26.6	7.9
	Prewitt	27.3	8.2
	Scharr	26.9	8.1
	Roberts cross	29.3	8.8
BlowingBubbles	Sobel	23.7	11.5
	Prewitt	23.9	11.7
	Scharr	24.2	11.8
	Roberts cross	27.5	13.6

TABLE VII
EXPERIMENTAL RESULTS OF THE PROPOSED METHOD, IMPLEMENTING DIFFERENT GRADIENT OPERATORS, COMPARED TO HM

Video sequence	Kernel	TR (%)	BD-Rate (%)
RaceHorses	Sobel	-44.3	1.25
	Prewitt	-46.0	1.20
	Scharr	-42.2	1.21
	Roberts cross	-46.1	1.29
BasketballPass	Sobel	-47.1	1.70
	Prewitt	**-47.4**	**1.66**
	Scharr	-47.4	1.78
	Roberts cross	-46.2	1.71
BlowingBubbles	Sobel	-43.1	1.06
	Prewitt	**-43.7**	**1.05**
	Scharr	-43.2	1.05
	Roberts cross	-43.2	1.15

TABLE VIII
AVERAGE RESULTS (OVER ALL RECOMMENDED SEQUENCES) OF THE PROPOSED METHOD COMPARED TO HM USING DIFFERENT GRADIENT OPERATORS

Kernel	TR (%)	BD-Rate (%)	BD-PSNR (dB)
Prewitt	-47.2	1.36	-0.062
Sobel	-47.1	1.38	-0.063
Scharr	-47.0	1.40	0.064
Roberts cross	-47.4	1.60	-0.073

We also provide final results for all frames to fairly compare our work to reference papers [11], [24], [31] and to also present the results under common test conditions (CTC) recommended in [41]. Comparison conditions are explained in greater detail in Section VII-D. To report the results, time reduction is defined as:

$$TR(\%) = \frac{T_{Proposed} - T_{HM}}{T_{HM}} \times 100\%, \tag{19}$$

where $T_{Proposed}$ and T_{HM} are the total encoding time of the proposed encoder and HM encoder, respectively. To calculate BD-Rate and BD-PSNR, we apply Bjøntegaard algorithm by using the following combined PSNR [7]:

$$\mathrm{PSNR} = \frac{6 \cdot \mathrm{PSNR_Y} + \mathrm{PSNR_U} + \mathrm{PSNR_V}}{8}. \tag{20}$$

This assures us that we measure the impact of the proposed algorithms on the total quality. In the following, we present experimental results for each part of the method starting by gradient analysis.

A. Experimental Results for Gradient Analysis

In this section, we compare the four gradient operators, mentioned in Section V-A, based on accuracy and complexity to select the best one for our algorithm. Table VI shows the error rate of each operator (i.e., the rate at which the best mode is excluded from the candidates list) for the following two configurations:

A) Test model where RMD is replaced by a gradient operator and MPMs are not added;
B) Test model where RMD is replaced by a gradient operator (MPMs are added).

Also, Table VII shows the results of the entire algorithm, including the proposed luma and chroma mode decision approaches while different gradient kernels are implemented for three sequences with different video textures. Table VIII presents the results averaged over all sequences. Based on these results, Prewitt provides the best trade-off between time reduction and bit-rate increment, and is selected as our operator for the gradient analysis section. However, considering the time reduction table and error table, the difference between theses operators is not very significant, mostly because we apply the gradient analysis at the CTU level, and not at the PU level. This allows us to compute the gradient just once for each pixel, and to use this information for each PU which contains that pixel. To demonstrate the contribution of the gradient analysis in the overall algorithm, results for all test sequences while only gradient analysis is implemented are shown in Table IX. We observe that gradient analysis provides TR of 11.4% with a BD-Rate increase of 0.62%. In addition, Table X shows hit ratio of gradient analysis. It also shows the hit ratios of RDO cost modeling and chroma mode decision along with the combination of luma algorithms (gradient analysis and RDO cost modeling). The hit ratio is defined as the percentage of times that the best mode selected by the proposed algorithm is same as the mode selected by the HM.

B. Experimental Results for Statistical RDO Cost Modeling

In this section, we present the results for RDO cost modeling to show the contribution of this statistical method in the overall algorithm. Table IX shows the results while only RDO cost modeling and candidates selection are implemented (see Fig. 4). We observe that RDO cost modeling, one of our main contributions, leads to a time reduction of nearly 30% with a BD-Rate increase of 0.8%; a trade-off which is quite appealing. To derive the models (normal distributions), the RaceHorses sequence is used as the training sequence, and hence, the average results are presented with and without considering this sequence. This approach of presenting the results when a training sequence is selected among the common test sequences was adopted in [24]. Based on this training, the confidence level (CL) on line 6 of the Algorithm 1 is set to 0.2, which represents a very good trade-off between complexity reduction and coding efficiency. Depending on the application, CL can be changed to have a faster or higher quality encoder. Also, the correlation coefficient ρ is computed based on the training data and is averagely 0.48. To obtain ρ, for blocks of a specific RMD cost (e.g., C_{RMD_i}), we associate a

TABLE IX
EXPERIMENTAL RESULTS WITH DIFFERENT ALGORITHMS SEPARATELY IMPLEMENTED COMPARED TO HM

Class	Video sequence	Gradient analysis (Prewitt)			RDO cost modeling			Chroma mode decision (method E)		
		TR (%)	BD-Rate (%)	BD-PSNR (dB)	TR (%)	BD-Rate (%)	BD-PSNR (dB)	TR (%)	BD-Rate (%)	BD-PSNR (dB)
A	Traffic	-12.0	0.46	-0.021	-29.2	1.16	-0.055	-7.1	0.04	-0.002
	PeopleOnStreet	-11.5	0.64	-0.031	-32.4	1.23	-0.060	-6.6	0.04	-0.002
	Nebuta	-10.9	0.36	-0.022	-31.7	0.37	-0.023	-7.1	0.01	-0.001
	SteamLocomotive	-13.9	0.28	-0.011	-30.9	0.39	-0.015	-5.2	0.02	-0.001
B	Cactus	-11.9	0.78	-0.026	-29.6	0.74	-0.024	-5.9	0.07	-0.002
	Kimono	-14.1	0.45	-0.015	-30.2	1.49	-0.049	-5.9	0.06	-0.002
	ParkScene	-11.4	0.40	-0.016	-29.1	0.81	-0.031	-5.5	0.06	-0.002
	BasketballDrive	-11.9	1.47	-0.037	-31.1	0.92	-0.024	-7.6	0.11	-0.003
	BQTerrace	-10.3	0.44	-0.021	-30.0	0.45	-0.023	-5.2	0.05	-0.003
C	BQMall	-8.8	0.56	-0.029	-30.1	0.72	-0.038	-5.8	0.07	-0.004
	PartyScene	-9.7	0.51	-0.034	-24.7	0.48	-0.032	-5.8	0.09	-0.006
	RaceHorsesC	-11.1	0.35	-0.019	-26.8	0.48	-0.027	-5.7	0.05	-0.003
	BasketballDrill	-11.9	0.36	-0.016	-30.9	0.53	-0.024	-6.8	0.11	-0.005
D	RaceHorses*	-10.4	0.49	-0.028	-28.4	0.77	-0.044	-6.9	0.11	-0.006
	BasketballPass	-10.8	0.94	-0.051	-29.9	0.77	-0.041	-6.2	0.03	-0.002
	BlowingBubbles	-10.0	0.43	-0.023	-27.0	0.57	-0.030	-6.2	0.05	-0.002
	BQSquare	-9.1	0.77	-0.056	-24.1	0.54	-0.039	-5.8	0.08	-0.005
E	FourPeople	-11.9	0.70	-0.036	-31.6	1.19	-0.061	-5.3	0.06	-0.003
	Johnny	-12.6	0.93	-0.036	-31.3	1.30	-0.049	-4.5	0.09	-0.004
	KristenAndSara	-12.9	1.06	-0.049	-30.0	1.28	-0.060	-5.3	0.12	-0.005
Average (all sequences)		**-11.4**	**0.62**	**-0.029**	**-29.4**	**0.80**	**-0.038**	**-6.0**	**0.07**	**-0.003**
Average (without training sequence)					**-29.5**	**0.81**	**-0.037**			

*indicates the training sequence

TABLE X
HIT RATIO (%) FOR DIFFERENT ALGORITHMS, RACEHORSES (RH), BASKETBALLPASS (BP), BLOWINGBUBBLES (BB)

		Hit ratio (%)				
	QP	22	27	32	37	Average
RH	Gradient analysis (Prewitt)	71	73	75	77	74
	RDO cost modeling	74	78	79	82	78
	Chroma (method E)	73	79	83	87	80
	Luma algorithms combined	60	64	72	74	67
BP	Gradient analysis (Prewitt)	67	69	70	72	69
	RDO cost modeling	75	79	81	83	79
	Chroma (method E)	78	82	88	90	84
	Luma algorithms combined	57	60	69	70	64
BB	Gradient analysis (Prewitt)	72	74	77	79	76
	RDO cost modeling	77	81	84	86	82
	Chroma (method E)	77	82	87	90	84
	Luma algorithms combined	62	67	74	77	70

TABLE XI
EXPERIMENTAL RESULTS FOR THE PROPOSED CHROMA MODE DECISION METHODS, MODIFIED HM (CHROMA METHODS IMPLEMENTED) COMPARED TO STANDARD HM

Video sequence	Chroma method	TR (%)	BD-Rate (%)	BD-PSNR (dB)
RaceHorses	A	-12.6	1.75	-0.101
	B	-10.0	4.24	-0.246
	C	-7.6	0.21	-0.013
	D ($\alpha = 1/2$)	-4.5	0.22	-0.013
	E ($\alpha = 5/16$)	-6.9	0.11	-0.006
BasketballPass	A	-12.0	2.15	-0.112
	B	-10.0	5.63	-0.290
	C	-10.0	0.19	-0.010
	D ($\alpha = 1/2$)	-4.7	0.12	-0.006
	E ($\alpha = 5/16$)	-6.2	0.03	-0.002
BlowingBubbles	A	-9.9	1.36	-0.070
	B	-10.1	3.18	-0.165
	C	-8.0	0.21	-0.010
	D ($\alpha = 1/2$)	-3.9	0.18	-0.009
	E ($\alpha = 5/16$)	-6.2	0.05	-0.002

vector of RDO costs with distributions X_i and similarly we have another vector of RDO costs with distributions X_j associated to blocks with RMD cost of C_{RMD_j}. ρ_{ij} of these two distributions X_i and X_j is obtained by computing the Pearson correlation coefficient between these two vectors.

C. Experimental Results for Fast Chroma Mode Decision

In this section, we compare the five chroma mode decision approaches, presented in Section VI, based on quality and complexity reduction. Table XI shows the results in terms of time reduction, BD-Rate and BD-PSNR for the entire encoder compared to the HM, while implementing the proposed chroma methods instead of HM standard chroma implementation. Based on these experiments, and considering the trade-off between complexity reduction and coding efficiency, we believe the best results are achieved by using a combination of the mode classification and the best mode of the luma component (method E), since this results in significant speedup without noticeably affecting the rate-distortion performance. The results show that there is almost no quality loss using this proposed chroma method. However, a different method may be selected to achieve faster processing, with a penalty in terms of the quality. To show the contribution of chroma mode decision in the overall algorithm, Table IX presents the results for all sequences using method E. These experiments also show that chroma intra prediction is a complex component of the entire encoder, and that more research is required in the area of intra chroma complexity reduction.

D. Overall Results

Table XII shows the overall results of the proposed method for the first hundred frames of the recommended video

TABLE XII
EXPERIMENTAL RESULTS OF THE OVERALL PROPOSED METHOD COMPARED TO HM (FIRST HUNDRED FRAMES)

Class	Video sequence	TR (%)	BD-Rate (%)	BD-PSNR (dB)
A	Traffic	-49.1	1.44	-0.067
	PeopleOnStreet	-49.6	1.69	-0.082
	Nebuta	-48.0	0.68	-0.042
	SteamLocomotive	-49.1	0.60	-0.024
B	Cactus	-47.5	1.46	-0.048
	Kimono	-49.6	1.55	-0.050
	ParkScene	-47.3	1.00	-0.038
	BasketballDrive	-49.7	2.36	-0.060
	BQTerrace	-47.9	0.88	-0.042
C	BQMall	-46.6	1.26	-0.065
	PartyScene	-40.2	1.03	-0.068
	RaceHorsesC	-45.1	0.78	-0.044
	BasketballDrill	-48.2	0.88	-0.039
D	RaceHorses*	-46.0	1.20	-0.069
	BasketballPass	-47.4	1.66	-0.088
	BlowingBubbles	-43.7	1.05	-0.055
	BQSquare	-39.6	1.33	-0.096
E	FourPeople	-49.4	1.76	-0.090
	Johnny	-49.7	2.23	-0.085
	KristenAndSara	-49.3	2.21	-0.103
Average (with training sequence)		**-47.1**	**1.35**	**-0.063**
Average (without training sequence)		**-47.2**	**1.36**	**-0.062**

*indicates the training sequence

TABLE XIII
EXPERIMENTAL RESULTS OF THE OVERALL PROPOSED METHOD COMPARED TO HM (CTC CONDITIONS)

Class	Video sequence	TR (%)	BD-Rate (%)	BD-PSNR (dB)
A	Traffic	-48.8	1.46	-0.068
	PeopleOnStreet	-49.4	1.71	-0.084
	Nebuta	-49.2	0.68	-0.041
	SteamLocomotive	-48.9	0.70	-0.026
B	Cactus	-47.7	1.46	-0.048
	Kimono	-49.5	1.54	-0.051
	ParkScene	-47.4	1.02	-0.040
	BasketballDrive	-49.1	2.37	-0.061
	BQTerrace	-46.7	0.82	-0.035
C	BQMall	-47.0	1.48	-0.073
	PartyScene	-41.1	1.02	-0.068
	RaceHorsesC	-44.6	0.65	-0.035
	BasketballDrill	-48.7	0.85	-0.039
D	RaceHorses*	-46.5	1.22	-0.065
	BasketballPass	-46.8	1.71	-0.094
	BlowingBubbles	-44.2	1.03	-0.061
	BQSquare	-41.0	1.29	-0.085
E	FourPeople	-48.9	1.78	-0.091
	Johnny	-49.9	2.22	-0.085
	KristenAndSara	-49.5	2.21	-0.102
Average (with training sequence)		**-47.2**	**1.36**	**-0.063**
Average (without training sequence)		**-47.3**	**1.37**	**-0.062**

*indicates the training sequence

sequences. Also Table XIII shows the overall results for all frames of the sequences to present the results under CTC conditions. We provide the results for all recommended sequences, using Prewitt as our selected gradient operator and method E for fast chroma mode decision. Overall, the proposed method provides a 47.3% time reduction with a very low BD-Rate increase of 1.37%. As mentioned, parameters can be tuned to obtain a different trade-off between time reduction and BD-Rate. Figures 10 and 11 show the rate-distortion curves of the proposed method versus HM for the BasketballPass and PartyScene sequences. Based on these figures, the two curves

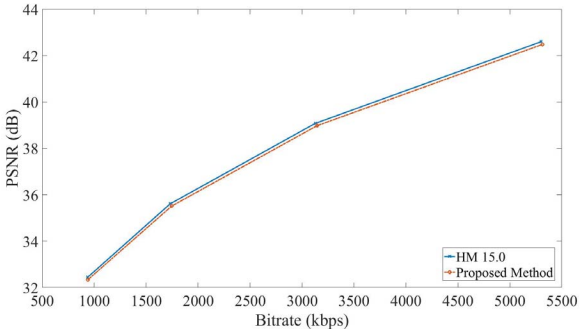

Fig. 10. RD curves of the proposed method and HM, BasketballPass.

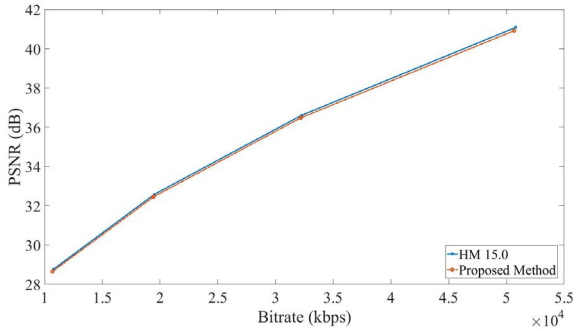

Fig. 11. RD curves of the proposed method and HM, PartyScene.

are very close, which shows that the proposed method leads to negligible quality loss, and is effective for different bitrates. Similar curves are achieved for other sequences with different video textures.

As we mentioned in Section III, some works exist on fast HEVC intra coding, based on fast CU size decision. Since our work covers the fast mode decision, in Table XIV, we compare the results of this work to other methods on this area. To have fair comparisons with the reference papers, Table XIV includes averages of different sets of sequences. Thus, based on these different sets of sequences we have different test conditions as follows:

A) Test condition includes sequences shown in Table IX other than ParkScene, BQMall, BasketballPass and Johnny, which are training sequences in [24] and are not used to compute the average. Also, we exclude RaceHorses as our training sequence. Thus, this test condition includes 15 sequences. All frames are considered for this test condition.

B) Test condition includes sequences shown in Table IX other than Nebuta, SteamLocomotive, BasketballDrive, BQTerrace, RaceHorsesC and BlowingBubbles and also Vidyo3 is used instead of KristenAndSara to have the same sequences as [14]. In addition, we exclude RaceHorses as our training sequence. Thus, this test condition includes 13 sequences. The first hundred frames are considered for this test condition.

c) Test condition includes sequences shown in Table IX (20 common test sequences as those considered in [31]). It should be noted that because this reference does not provide separate results for different test sequences for

TABLE XIV
COMPARISON OF THE PROPOSED METHOD WITH OTHER MODE DECISION (MD) METHODS

Test Condition	A		B		C		D		E		F	
	Proposed method	[24](MD)	Proposed method	[14]	Proposed method	[31](MD)	Proposed method	[11]	Proposed method	[4]	Proposed method	[3]
TR (%)	-47.2	-15.2	-47.3	-27.2	-47.2	-31.8	-47.2	-38.0	-47.0	-41.8	-47.2	-35.6
BD-Rate (%)	1.30	0.17	1.44	1.02	1.36	0.9	1.27	1.37	1.39	1.25	1.31	1.08
BD-PSNR (dB)	-0.060	-	-0.068	-0.056	-0.063	-	-0.060	-0.082	-0.063	-0.058	-0.060	-0.058

TABLE XV
EXPERIMENTAL RESULTS OF THE OVERALL PROPOSED METHOD COMPARED TO HM FOR CLASS E SEQUENCES WHICH ARE NOT PART OF CTC

Class	Video sequence	TR (%)	BD-Rate (%)	BD-PSNR (dB)
E	Vidyo1	-49.6	1.98	-0.087
	Vidyo3	-50.2	1.49	-0.071
	Vidyo4	-48.1	1.74	-0.070

their mode decision approach and it is not possible to remove RaceHorses for comparison, we compare our results including this sequence to their results to have exactly the same sequences as they use for computing the average. Considering the last two rows of Table XIII, the difference between the results including or excluding RaceHorses is negligible. All frames are considered for this test condition.

D) Test condition includes sequences shown in Table IX other than BasketballDrive, BQMall and BasketballPass, which are not used in [11]. Similar to test condition VII-D and since they do not provide separate results for different sequences (in their Top 1 Mode method), we include RaceHorses for comparison. Thus, this test condition includes 17 sequences. All frames are considered for this test condition.

E) Test condition includes sequences shown in Table IX other than Nebuta and SteamLocomotive. In addition, Vidyo1, Vidyo3 and Vidyo4 are used instead of FourPeople, Johnny and KristenAndSara to have the same sequences as [4]. Also, RaceHorses is removed as the training sequence. Thus, this test condition includes 17 sequences. The first hundred frames are considered for this test condition.

F) Test condition includes sequences shown in Table IX except for class E. For this class Vidyo1, Vidyo3 and Vidyo4 are used instead of FourPeople, Johnny and KristenAndSara to have the same sequences as [3]. Also, RaceHorses is removed as the training sequence. Thus, this test condition includes 19 sequences. The first hundred frames are considered for this test condition.

Table XV shows the results of the proposed method for extra sequences of class E which are not part of the common test conditions recommended in [41] but are used to achieve some results in some reference papers. They are presented here to permit a thorough comparison with other works as presented in Table XIV. Based on the comparison table, the proposed method offers a significantly higher time reduction compared to other state-of-the-art methods. It is nearly 10% faster than method [11] with an improved BD-Rate. Methods [31], [14] offer less than 0.5% BD-Rate improvement but are 15% and 20% slower respectively. Although method [24] offers a good BD-Rate, its time reduction is too small to be of much use in real-time systems. As it is increasingly difficult to maintain good BD-Rate as we increase the time reduction, the BD-Rate offered by the proposed method is quite impressive considering the time reuction of nearly 50%. In comparison to [14], our method does not use the neighboring blocks as a reference for mode decision since such an approach could result in a domino effect and lead to degradation of the quality. Also, they use a method for reduction of RDO candidates which is based on a heuristic idea of dominant directions and have a limited time reduction. In contrast, our method is a solid algorithm based on various tests and statistical modeling. In [11], the RDO cost is formulated as a function of SAD and QP. The function is quadratic and lets the encoder avoid performing RDO process by estimating RDO cost. While they describe the RDO cost as a deterministic function of SAD, we consider RDO cost as a stochastic model which is described in the form of probability distributions. Zhang *et al.* [24] use a gradient-based approach for mode decision. In the same way, a gradient-based mode decision is proposed in [31]. To improve this approach, they also propose an optimal mode selection based on the number of occurrences of each mode in the PU. In comparison to these two works, our method adds two significant extra steps of RDO cost prediction and chroma mode classification to achieve higher time reduction. In comparison to our previous works, [3], [4], we expand our RDO cost modeling and present goodness of fit tests and more analytical works. In addition, we present bivariate normal distributions for RDO costs and consider correlation among these distributions which are new compared to our previous works. Also, we add performance analysis and comparison of different gradient kernels. We also propose novel methods for chroma intra coding and conduct a performance analysis of these different fast chroma mode decision approaches. Overall, we have better performance compared to our previous works. It should be noted that there are other proposed methods in [31] and [24] which are related to fast CU size decision and are not related to our work on fast mode decision. It is important to note that the methods proposed in this work can be combined with fast CU size decision methods in order to achieve even higher complexity reduction.

VIII. CONCLUSION

In this paper, a method is proposed for accelerating the HEVC intra mode decision. At the first step, the rough mode decision of the HEVC test model (HM) is replaced by gradient analysis, based on the Prewitt operator, in order to

decrease the number of candidates by excluding non-relevant directional modes. Then, to reduce the number of candidates as much as possible, RDO cost prediction prevents high-demanding rate-distortion optimization for modes with no chance of being the best intra mode of the prediction unit. This is based on statistical modeling of the RDO cost using the low-complexity RMD cost. For fast intra chroma mode decision, a mode classification is proposed to avoid the RDO process for non-promising candidates. Experimental results show that the proposed method provides a 47.3% time reduction on average for all-intra profile with a 1.37% BD-Rate increase. The concepts presented in this paper decrease the number of candidates and provide a fast mode decision. They could be integrated with fast CU size decision methods and transform tree optimization algorithms to achieve a very fast HEVC intra encoder.

ACKNOWLEDGMENT

The authors would like to thank Jean-François Franche for providing the code to compute SATD for chroma modes.

REFERENCES

[1] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, "Overview of the high efficiency video coding (HEVC) standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[2] T. Wiegand, G. J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 7, pp. 560–576, Jul. 2003.

[3] M. Jamali, S. Coulombe, and F. Caron, "Fast HEVC intra mode decision based on edge detection and SATD costs classification," in *Proc. Data Compression Conf.*, Apr. 2015, pp. 43–52.

[4] M. Jamali and S. Coulombe, "RDO cost modeling for low-complexity HEVC intra coding," in *Proc. IEEE Can. Conf. Elect. Comput. Eng. (CCECE)*, Vancouver, BC, Canada, May 2016, pp. 1–5.

[5] J. Lainema, F. Bossen, W.-J. Han, J. Min, and K. Ugur, "Intra coding of the HEVC standard," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 22, no. 12, pp. 1792–1801, Dec. 2012.

[6] K. Il-Koo, *High Efficiency Video Coding (HEVC) Test Model 15 (HM15) Encoder Description*, document JCTVC-Q1002, Joint Collaborative Team Video Coding (JCT-VC) ITU-T SG16 WP3 ISO/IEC JTC1/SC29/WG11, Valencia, Spain, Apr. 2014.

[7] V. Sze, M. Budagavi, and G. J. Sullivan, "High efficiency video coding (HEVC)," in *Integrated Circuit and Systems, Algorithms and Architectures*. Cham, Switzerland: Springer, 2014, pp. 1–375.

[8] D. Ruiz, G. Fernández-Escribano, J. L. Martínez, and P. Cuenca, "Fast intra mode decision algorithm based on texture orientation detection in HEVC," *Signal Process. Image Commun.*, vol. 44, pp. 12–28, May 2016.

[9] I. Marzuki, J. Ma, Y.-J. Ahn, and D. Sim, "A context-adaptive fast intra coding algorithm of high-efficiency video coding (HEVC)," *J. Real Time Image Process.*, vol. 11, pp. 1–17, Mar. 2016.

[10] F. Pakdaman, M.-R. Hashemi, and M. Ghanbari, "Fast and efficient intra mode decision for HEVC, based on dual-tree complex wavelet," *Multimedia Tools Appl.*, vol. 76, no. 7, pp. 9891–9906, 2017.

[11] J. Tariq, S. Kwong, and H. Yuan, "HEVC intra mode selection based on rate distortion (RD) cost and sum of absolute difference (SAD)," *J. Vis. Commun. Image Represent.*, vol. 35, pp. 112–119, Feb. 2016.

[12] L.-L. Wang and W.-C. Siu, "Novel adaptive algorithm for intra prediction with compromised modes skipping and signaling processes in HEVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 10, pp. 1686–1694, Oct. 2013.

[13] M. Park and J. Jeong, "Fast HEVC intra prediction algorithm with enhanced intra mode grouping based on edge detection in transform domain," *J. Adv. Comput. Netw.*, vol. 3, no. 2, pp. 1686–1694, 2015.

[14] L. Gao, S. Dong, W. Wang, R. Wang, and W. Gao, "Fast intra mode decision algorithm based on refinement in HEVC," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Lisbon, Portugal, 2015, pp. 517–520.

[15] Q. Zhang, X. Huang, X. Wang, and W. Zhang, "A fast intra mode decision algorithm for HEVC using sobel operator in edge detection," *Int. J. Multimedia Ubiquitous Eng.*, vol. 10, no. 9, pp. 80–90, 2015.

[16] F. Pan *et al.*, "Fast mode decision algorithm for intraprediction in H.264/AVC video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 15, no. 7, pp. 813–822, Jul. 2005.

[17] S. Li, S. Chen, J. Wang, and L. Yu, "Second order prediction on H.264/AVC," in *Proc. Picture Coding Symp.*, Chicago, IL, USA, May 2009, pp. 1–4.

[18] S. Cho and M. Kim, "Fast CU splitting and pruning for suboptimal CU partitioning in HEVC intra coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 23, no. 9, pp. 1555–1564, Sep. 2013.

[19] M. Jamali and S. Coulombe, "Coding unit splitting early termination for fast HEVC intra coding based on global and directional gradients," in *Proc. IEEE Workshop Multimedia Signal Process. (MMSP)*, Montreal, QC, Canada, Sep. 2016, pp. 1–5.

[20] K. Saurty, P. C. Catherine, and K. M. S. Soyjaudah, "Terminating CU splitting in HEVC intra prediction using the Hadamard absolute difference (HAD) cost," in *Proc. SAI Intell. Syst. Conf. (IntelliSys)*, London, U.K., Nov. 2015, pp. 836–841.

[21] L. Shen, Z. Zhang, and Z. Liu, "Effective CU size decision for HEVC intracoding," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4232–4241, Oct. 2014.

[22] X. Fang, X. Zhu, L. Yu, and X. Shen, "Fast HEVC intra coding unit size decision based on an improved Bayesian classification framework," in *Proc. IEEE Picture Coding Symp. (PCS)*, San Jose, CA, USA, 2013, pp. 273–276.

[23] B. Min and R. C. C. Cheung, "A fast CU size decision algorithm for the HEVC intra encoder," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 25, no. 5, pp. 892–896, May 2015.

[24] T. Zhang, M. T. Sun, D. Zhao, and W. Gao, "Fast intra-mode and CU size decision for HEVC," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 8, pp. 1714–1726, Aug. 2017.

[25] X. Shang, G. Wang, T. Fan, Y. Li, and Y. Zuo, "Low-complexity intra-coding scheme for HEVC," *Circuits Syst. Signal Process.*, vol. 35, no. 12, pp. 4331–4349, 2016.

[26] H. Zhang and Z. Ma, "Fast intra mode decision for high efficiency video coding (HEVC)," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 4, pp. 660–668, Apr. 2014.

[27] Y. Song, Y. Zeng, X. Li, B. Cai, and G. Yang, "Fast CU size decision and mode decision algorithm for intra prediction in HEVC," *Multimedia Tools Appl.*, vol. 76, no. 2, pp. 2001–2017, 2017.

[28] L. Zhao, X. Fan, S. Ma, and D. Zhao, "Fast intra-encoding algorithm for high efficiency video coding," *Signal Process. Image Commun.*, vol. 29, no. 9, pp. 935–944, 2014.

[29] W. Zhao, L. Shen, Z. Cao, and Z. Zhang, "Texture and correlation based fast intra prediction algorithm for HEVC," in *Advances on Digital Television and Wireless Multimedia Communications*. Heidelberg, Germany: Springer, 2012, pp. 284–291.

[30] H. Zhang and Z. Ma, "Early termination schemes for fast intra mode decision in high efficiency video coding," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Beijing, China, 2013, pp. 45–48.

[31] A. BenHajyoussef, T. Ezzedine, and A. Bouallègue, "Gradient-based pre-processing for intra prediction in high efficiency video coding," *EURASIP J. Image Video Process.*, vol. 2017, no. 1, p. 9, 2017.

[32] X. Liu, Y. Liu, P. Wang, C.-F. Lai, and H.-C. Chao, "An adaptive mode decision algorithm based on video texture characteristics for HEVC intra prediction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 8, pp. 1737–1748, Aug. 2016.

[33] W. Zhu, Y. Yi, H. Zhang, P. Chen, and H. Zhang, "Fast mode decision algorithm for HEVC intra coding based on texture partition and direction," *J. Real Time Image Process.*, vol. 14, pp. 1–18, Apr. 2018, doi: 10.1007/s11554-018-0766-z.

[34] J. Xiong, H. Li, F. Meng, Q. Wu, and K. N. Ngan, "Fast HEVC inter CU decision based on latent SAD estimation," *IEEE Trans. Multimedia*, vol. 17, no. 12, pp. 2147–2159, Dec. 2015.

[35] H. S. Bhat and N. Kumar, *On the Derivation of the Bayesian Information Criterion*, School Nat. Sci., Univ. California, Berkeley, CA, USA, 2010.

[36] B. Yazici and S. Yolacan, "A comparison of various tests of normality," *J. Stat. Comput. Simulat.*, vol. 77, no. 2, pp. 175–183, 2007.

[37] S. Korkmaz, D. Goksuluk, and G. Zararsiz, "MVN: An R package for assessing multivariate normality," *R J.*, vol. 6, no. 2, pp. 151–162, 2014.

[38] C. J. Mecklin and D. J. Mundfrom, "A Monte Carlo comparison of the type I and type II error rates of tests of multivariate normality," *J. Stat. Comput. Simulat.*, vol. 75, no. 2, pp. 93–107, 2005.

[39] A. Trujillo-Ortiz, R. Hernandez-Walls, K. Barba-Rojo, and L. Cupul-Magana. (2007). *Hzmvntest:Henze-Zirkler's Multivariate Normality Test, A MATLAB file*. [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=17931

[40] A. Trujillo-Ortiz, R. Hernandez-Walls, K. Barba-Rojo, and L. Cupul-Magana. (2007). *Roystest:Royston's Multivariate Normality Test, A MATLAB File*. [Online]. Available: http://www.mathworks.com/matlabcentral/fileexchange/loadFile.do?objectId=17811

[41] F. Bossen, "Common test conditions and software reference configurations," document JCTVC-L1100, Joint Collaborative Team Video Coding (JCT-VC), Geneva, Switzerland, 2013.

[42] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," document VCEG-M33, Video Coding Experts Group (VCEG) ITU-T, VCEG-M33, Austin, TX, USA, Apr. 2001.

Mohammadreza Jamali (S'09) received the B.Sc. and M.Sc. degrees in electrical engineering from Khaje Nasir Toosi University, Tehran, Iran, in 2007 and 2010, respectively. He is currently pursuing the Ph.D. degree with the Department of Software Engineering and Information Technology, École de technologie supérieure, Université du Québec, Montreal, Canada, where he has been a Research Associate in Vantrix Industrial Research Chair in Video Optimization since 2013 and has been researching on video coding complexity reduction. His current research interests include the areas of image and video coding, virtual reality, and reinforcement learning.

Stéphane Coulombe (S'90–M'98–SM'01) received the B.Eng. degree in electrical engineering from École Polytechnique de Montréal, Canada, in 1991, and the Ph.D. degree in telecommunications (image processing) from INRS-Telecommunications, Montreal, in 1996. He is currently a Professor with the Software and IT Engineering Department, École de technologie supérieure, Université du Québec. From 1997 to 1999, he was with Nortel Wireless Network Group, Montreal, and from 1999 to 2004, he worked with the Nokia Research Center, Dallas, TX, USA, as a Senior Research Engineer and as a Program Manager in the Audiovisual Systems Laboratory. He joined ÉTS in 2004, where he currently carries out research and development on video processing and systems, compression, and transcoding. Since 2009, he has held the Vantrix Industrial Research Chair in Video Optimization.