

# General Embedded Quantization for Wavelet-Based Lossy Image Coding

Francesc Aulí-Llinàs, *Member, IEEE*

**Abstract**—Embedded quantization is a mechanism employed by many lossy image codecs to progressively refine the distortion of a (transformed) image. Currently, the most common approach to do so in the context of wavelet-based image coding is to couple uniform scalar deadzone quantization (USDQ) with bitplane coding (BPC). USDQ+BPC is convenient for its practicality and has proved to achieve competitive coding performance. But the quantizer established by this scheme does not allow major variations. This paper introduces a multistage quantization scheme named general embedded quantization (GEQ) that provides more flexibility to the quantizer. GEQ schemes can be devised for specific decoding rates achieving optimal coding performance. Practical approaches of GEQ schemes achieve coding performance similar to that of USDQ+BPC while requiring fewer quantization stages. The performance achieved by GEQ is evaluated in this paper through experimental results carried out in the framework of modern image coding systems.

**Index Terms**—General embedded quantization, lossy image coding, JPEG 2000.

## I. INTRODUCTION

QUANTIZATION [1] is a signal processing technique that maps a large set of input values to a smaller set of output values called quantization indices. A quantizer is an algorithmic procedure, or a device, that converts the input values to the output indices. The dequantizer reverses this operation by reconstructing the original values using the quantization indices, which commonly implies a loss on the signal fidelity called quantization error [2].

Scalar and vector quantization are two classic techniques that have been widely employed in the field of lossy image coding. Scalar quantization maps one input sample to one quantization index, whereas vector quantization maps a set of image samples, called vector, to an index. Commonly, quantization is applied on the coefficients of an image that has undergone transformation—rather than on the original image samples—to gain coding efficiency. When the quantizer assigns a fixed rate for all indices, compression is achieved at the expense of larger quantization error by means of reducing the indices' rate. Compression can also be achieved by using indices of variable rate, though variable rate quantization commonly requires more elaborated methods of rate-distortion optimization.

Copyright (c) 2013 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending a request to [pubs-permissions@ieee.org](mailto:pubs-permissions@ieee.org).

Dr. Francesc Aulí-Llinàs is with the Department of Information and Communications Engineering, Universitat Autònoma de Barcelona, Spain (phone: +34 935813571; fax: +34 935814477; e-mail: [fauli@deic.uab.es](mailto:fauli@deic.uab.es)). This work has been partially supported by the Spanish Government (MINECO), by FEDER, and by the Catalan Government under Grants RYC-2010-05671, TIN2009-14426-C02-01, TIN2012-38102-C03-03, and 2009-SGR-1224.

Besides brute compression efficiency, most modern lossy image codecs provide an interesting feature called quality progressivity. Quality progressivity refers to the ability of the coding system to code an image from a coarse to a fine quality level in successive stages. From the point of view of quantization, quality progressivity is generally achieved by means of a quantizer that produces indices that can be split in short words. Each word is a suffix of the previous ones (if any), so that they can be transmitted separately and combined by the dequantizer to reconstruct the original value with more or less precision depending on the transmitted words. This technique is called embedded quantization, progressive quantization, or successive approximation quantization [3].

Embedded quantization enables the generation of a code-stream that can be truncated at different rates with neither penalizing coding performance nor requiring re-compression. This is of utility for image transmission, progressive decoding, or transcoding, for example, so embedded quantization has been studied thoroughly and has been adopted by many coding systems. Without aiming to be exhaustive, scalar quantization schemes that are adaptively adjusted as more data are transmitted are investigated in [4]–[6], the best size for the deadzone of uniform scalar quantizers is determined in [7], progressively refinable vector quantization schemes are studied in [8]–[12] and the popular SPIHT [13] algorithm is adapted to a multistage vector quantization scheme in [14], [15]. Embedded and multistage trellis coded quantization schemes [16] are explored in [17]–[22] and ordering strategies for wavelet data are examined in [23].

Currently, most wavelet-based image codecs carry out embedded quantization using uniform scalar deadzone quantization (USDQ) [24] together with bitplane coding (BPC) [25]. USDQ is a quantization scheme that partitions the range of input values into uniform intervals of the same width  $\Delta$ .  $\Delta$  is referred to as the quantization step size. The deadzone is the interval that contains zero (i.e.,  $(-\Delta, 0] \cup [0, \Delta)$ ), and is the only interval of width  $2\Delta$  because all coefficients in this interval are mapped to zero. Contrarily to coefficients in other intervals, no sign is transmitted for the coefficients in the deadzone. The BPC strategy splits the binary representation of the quantization indices into words of one bit. The same words or, otherwise stated, the bits at the same binary position, from all quantized coefficients form the so-called bitplane. The bits from all quantization indices are transmitted from bitplane  $M - 1$  to bitplane 0,  $M$  denoting a sufficient number of bits to represent all coefficients. The dequantizer interprets this procedure as a multistage quantization scheme that starts with a step size of  $\Delta 2^{M-1}$  that is then reduced in each stage by a power of two until lessened to  $\Delta 2^0 = \Delta$ . The sign of

the coefficient is transmitted just after the bit indicating that the coefficient is to be reconstructed outside the deadzone. Fig. 1(a) illustrates this scheme.

Embedded quantization emerges naturally in the scheme of USDQ+BPC. Its competitive coding performance and the convenient use of the binary representation for hardware architectures have made this scheme very popular. Nonetheless, the quantizer established by USDQ+BPC does not allow major variations. Work on quantization has shown that USDQ might be an appropriate quantizer for a variety of sources [26]–[28] but, to the best of our knowledge there is no study seeking embedded quantizers specifically designed to achieve optimal coding performance when they are introduced into modern coding engines based on the wavelet transform. The embedded quantizer is a key-piece of lossy image codecs, so the goal of this work is to devise a quantizer that minimizes quantization error at a range of selected decoding rates. To achieve this purpose, we investigate embedded quantizers that are not restricted to the scheme imposed by USDQ+BPC. The quantizer that arises from our analysis is then introduced in the core coding system of JPEG 2000 [29]. Experiments validate our observations during quantizer design: the scheme of USDQ+BPC is near-optimal from the point of view of compression efficiency, though sub-optimal from the point of view of number of quantization stages. This paper extends our previous works [30], [31] with detailed descriptions and comparisons of the quantization schemes proposed, comprehensive evaluations of the theory and practice behind them, and an extended set of experimental results.

The paper is structured as follows. Section II introduces a flexible scheme for embedded quantization and poses the optimization problem to achieve quantizers with optimal performance for a range of decoding rates. Section III describes a methodology to exhaustively explore the quantizers' efficiency and provides a low complexity algorithm to design quantizers that achieve near-optimal performance. The insights provided by this analysis are employed in Section IV to devise a practical quantizer that can be employed in modern coding systems based on the wavelet transform. Section V provides experimental results for different types of images when the proposed quantizer is utilized within the framework of JPEG 2000. The last section summarizes this work and provides concluding remarks.

## II. QUANTIZER DESIGN

### A. General embedded quantization

General embedded quantization (GEQ) is defined as a coding procedure that transmits the coefficients' magnitude of a transformed image through a multistage quantization scheme that uses arbitrary quantization intervals. To this end, let  $T_k$  denote the quantization threshold employed in stage  $k$ , with  $T_k \in (0, \mathcal{W})$ ,  $\mathcal{W}$  being the largest magnitude of all coefficients. Fig. 1(b) depicts the procedure carried out by the GEQ. The first quantization stage indicates whether the magnitude of each coefficient lies above or below  $T_1$ . In other words, it defines quantization intervals  $[0, T_1)$  and  $[T_1, \mathcal{W}]$ . The second quantization stage partitions one of these intervals

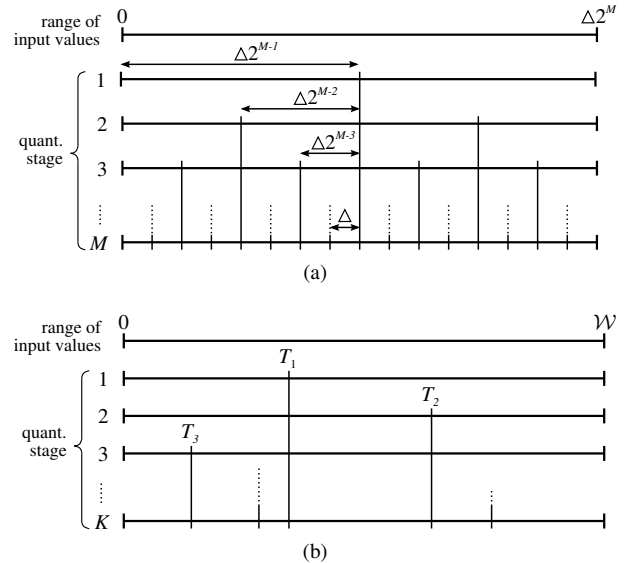


Fig. 1: Illustration of the quantization intervals produced by two multistage quantization schemes. The straight line on the top of each figure represents the range of the input values. Only the magnitude is depicted (omitting the sign) in all figures since symmetry about zero is assumed. The intervals produced in each quantization stage are depicted with a new line from the top to the bottom in each figure. (a) and (b) illustrate the USDQ+BPC and the GEQ scheme, respectively.

into two intervals. In the example of Fig. 1(b), the second stage operates only on coefficients whose magnitudes lie within the quantization interval  $[T_1, \mathcal{W}]$ . For this case,  $T_2$  is used to indicate whether the relevant coefficient magnitudes lie in  $[T_1, T_2)$  or in  $[T_2, \mathcal{W}]$ . Nothing is emitted for coefficients with magnitudes outside  $[T_1, \mathcal{W}]$  during the second quantization stage of this example. The third stage partitions one of the three intervals previously defined. In the example of Fig. 1(b), the third stage operates on coefficients within the deadzone, indicating whether coefficient magnitudes in  $[0, T_1)$  lie in  $[0, T_3)$  or  $[T_3, T_1)$ . The procedure continues in this fashion resulting in  $k + 1$  intervals at the end of the quantization stages. The coding of a stage that partitions the deadzone (first and third stages in the example of Fig. 1(b)) is referred to as significance coding, whereas the coding of other stages is referred to as refinement coding. The GEQ does not impose any restriction on the quantization thresholds employed in each stage, thus the order in which quantization intervals are partitioned is not fixed a priori. The possibilities for the quantizer design are hence immense.

Coefficients inside the deadzone are reconstructed by the dequantizer as zero. The sign of coefficients outside the deadzone is transmitted just after the symbol indicating that the coefficient is to be reconstructed as nonzero. The dequantization operation for such a coefficient, say  $\omega$ , consists of assigning a reconstruction value  $\hat{\omega}$  that lies somewhere in the indexed quantization interval. The dequantization operation is expressed as  $\hat{\omega} = \text{sign}(\omega) (T_l + \delta(T_h - T_l))$ , with  $[T_l, T_h)$  denoting the quantization interval of the reconstructed coefficient.  $\delta \in [0, 1)$  adjusts the reconstruction value in the interval.

The procedure carried out to select  $\delta$  in each interval is similar to that described in [32]. It is employed herein rather than the classic mid-point reconstruction (i.e., selecting  $\delta = 0.5$  for all intervals) to maximize distortion decreases in each quantization interval, which provides a robust framework to evaluate embedded quantizers [7]. As common, the distortion metric employed to evaluate the quantization error is mean squared error (MSE).

GEQ distinguishes from USDQ+BPC in that each quantization stage operates only on selected coefficients—rather than on all of them—, and in that the quantization thresholds are selected without any restriction. One quantization stage of GEQ is *not* equivalent to the coding of one bitplane as defined in USDQ+BPC. We note that, except for the first stage of the quantizer, the GEQ scheme requires the application of multiple stages to refine the magnitude of all coefficients.

GEQ requires that quantization thresholds  $\{T_k\}$  are known by coder and decoder to reproduce the same procedure. This section and the following one are aimed to explore designs for optimal quantizers, assuming that  $\{T_k\}$  are known. This aspect is approached in Section IV through a practical mechanism.

### B. Optimization problem

Our analysis adopts a coding engine that uses GEQ to apply successive stages of quantization until a pre-defined target rate, denoted as  $R$ , is attained. The rate and the distortion achieved at the end of quantization stage  $k$  are denoted as  $r_k$  and  $d_k$ , respectively. The goal of the coding engine is to produce a quality progressive codestream that, when decoded at any rate  $r \in (0, R]$ , results in the smallest possible distortion. To this end, let  $f_R(r)$  be the probability density function (pdf) indicating the probability of decoding the codestream at a specified rate  $r$  in the range  $(0, R]$ . As stated in [33],  $f_R(r)$  can be used to model a variable-rate channel, or to assign different weights to different decoding rates. For simplicity, this density is assumed to be uniform herein (i.e.,  $f_R(r) = 1/R$ ), though other distributions such as the exponential or the Laplacian might also be employed. The objective is then to find quantization thresholds  $\mathbf{t} = \{T_1, T_2, \dots, T_K\}$  that produce a set of rate/distortion points  $\{r_k, d_k\}, 1 \leq k \leq K$ , that minimize the expected multirate distortion in the rate range, i.e.,

$$\min_{\mathbf{t}} \int_0^R f_R(r) \cdot d(r) dr \quad (1)$$

with

$$T_k \in (0, W) \quad \forall \quad 1 \leq k \leq K. \quad (2)$$

The distortion achieved at rates between the end of quantization stages is assumed to decrease linearly with rate. That is,  $d(r)$  in (1) is determined according to

$$d(r) = d_k - (d_k - d_{k+1}) \cdot \frac{r - r_k}{r_{k+1} - r_k}, \quad r_k < r < r_{k+1}, \quad (3)$$

which corresponds to well-known models that use MSE as the distortion metric [29, Ch. 5.4], [34].

The wavelet transform [35], [36] is employed by many image coding systems [13], [25], [37] and standards [38], [39] to decorrelate image data. The following analysis and the subsequent practical approach investigate GEQ schemes for the coding of coefficients produced by a wavelet transform. Two approaches are employed in this and following sections to appraise the performance of each quantizer tested. The first approach utilizes the tier-1 coding procedure of the JPEG 2000 standard [29], [38]. JPEG 2000 is chosen due to its widespread use, excellent coding performance and context-based adaptive mechanisms, which are sound for many types of data and coding parameters. As originally formulated, tier-1 codes wavelet coefficients quantized by USDQ performing multiple coding passes per bitplane, which is a common practice of BPC strategies [37]. Herein, tier-1 is modified to allow the coding through the GEQ. To do so, one coding pass is employed at each stage of the GEQ scheme that operates on the relevant coefficients for that stage. The context-based probability model for the symbols emitted, as well as the arithmetic coder MQ, defined by JPEG 2000 are left unmodified in this approach.

The second approach employed by our analysis estimates the rate and the distortion achieved by the GEQ scheme using an i.i.d. generalized Gaussian distribution (GGD) that models wavelet data [40]–[42]. The magnitude of a wavelet coefficient is denoted as  $\nu$ , and its pdf is determined through a GGD denoted as  $f_V(\nu)$ . When the interval being partitioned in quantization stage  $k$  is the deadzone, the rate and the distortion achieved just after coding stage  $k$  are estimated according to

$$\begin{aligned} \hat{r}_k &= \hat{r}_{k-1} + \int_0^{T_H} f_V(\nu) d\nu \cdot \\ &[ \mathcal{H}(P(T_k \leq \nu < T_H \mid \nu < T_H)) + \\ &P(T_k \leq \nu < T_H \mid \nu < T_H) ] \end{aligned} \quad (4)$$

and

$$\hat{d}_k = \hat{d}_{k-1} - \int_{T_k}^{T_H} f_V(\nu) \cdot (\nu^2 - (\nu - \hat{\nu})^2) d\nu, \quad (5)$$

with  $T_H$  denoting the upper limit of the quantization interval (i.e.,  $[0, T_H)$  is the deadzone before quantizing stage  $k$ ), and  $\mathcal{H}(\cdot)$  denoting the binary entropy function.<sup>1</sup> The second term of Equation (4) represents the rate increase due to the quantization stage. The integral  $\int_0^{T_H} f_V(\nu) d\nu$  represents the fraction of coefficients coded in the stage, whereas the parenthetical portion of this term (i.e., second and third lines of (4)) corresponds to the rate increase per coefficient coded in

<sup>1</sup>The maximum probability of the entropy function in Equation (4) is set to 0.9999 by means of replacing  $\mathcal{H}(P(T_k \leq \nu < T_H \mid \nu < T_H))$  by  $\mathcal{H}(\min(P(T_k \leq \nu < T_H \mid \nu < T_H), 0.9999))$ . This approximately approaches the performance limit of real implementations of arithmetic coding.

the stage.  $P(T_k \leq \nu < T_H \mid \nu < T_H)$  refers to the probability of coefficients becoming significant, more precisely,

$$P(T_k \leq \nu < T_H \mid \nu < T_H) = \frac{\int_{T_k}^{T_H} f_V(\nu) d\nu}{\int_0^{T_H} f_V(\nu) d\nu}. \quad (6)$$

The quantity  $\mathcal{H}(P(T_k \leq \nu < T_H \mid \nu < T_H))$  accounts for coding the coefficient magnitudes, whereas  $P(T_k \leq \nu < T_H \mid \nu < T_H)$  accounts for coding the signs. The cost of coding the sign is assumed to be 1 bit in this approach since  $f_V(\nu)$  is symmetric about 0. Equation (5) determines the distortion achieved just after coding stage  $k$  as the square error before coding that stage (i.e.,  $\hat{d}_{k-1}$ ) minus the squared error decrease produced in that stage. Such a decrease is determined as the squared error before coding coefficient  $\nu$  (i.e.,  $\nu^2$ ) minus the squared error after coding coefficient  $\nu$  (i.e.,  $(\nu - \hat{\nu})^2$ ). The derivation of  $\hat{r}_k$  and  $\hat{d}_k$  for intervals other than the deadzone is determined similarly as in Equations (4) and (5). An extended description of rate and distortion models similar to those employed in this work can be found in [32], [37], [42].

The use of a codec like JPEG 2000 in the first approach provides actual coding performance. The rate and the distortion achieved after coding each quantization stage corresponds *exactly* with those achieved when using the sophisticated mechanisms [29], [43], [44] employed by modern engines such as context adaptive coding of coefficient magnitudes and signs, scanning order strategies aimed to reduce the number of emitted symbols, etc. Contrarily, the rate determined by the second approach via Equation (4) is an estimate that uses zero-order entropy, without taking into account the aforementioned mechanisms. Even so, the second approach is employed because the deployment of a real codec like JPEG 2000 in the exhaustive search below can be computationally expensive. Estimates of rate and distortion allow the extension of our analysis to quantizers containing a number of stages that would otherwise be too computationally complex to evaluate.

Each quantizer tested produces a codestream with a particular set of rate/distortion points  $\{r_k, d_k\}$ . These points are calculated from the compressed codestream in the first approach, and using Equations (4) and (5) in the second approach. The first approach employs the high vertical- low horizontal-frequency subband of the first decomposition level produced by the irreversible 9/7 wavelet transform applied on the ‘‘Portrait’’ image.<sup>2</sup> To simplify the performance analysis of the next section, all coefficients in the subband are coded together –rather than using small sets of coefficients as defined by JPEG 2000. Wavelet data in this subband have  $\mathcal{W} = 199$ . This value is measured for this particular image and wavelet subband. Reported results are similar for other subbands and images. The GGD employed by the second approach is set with mean, shape parameter, and standard deviation 0, 0.45

<sup>2</sup>The image belongs to the ISO 12640-1 corpus, and is  $2560 \times 2048$ , 8 bit, gray-scale.

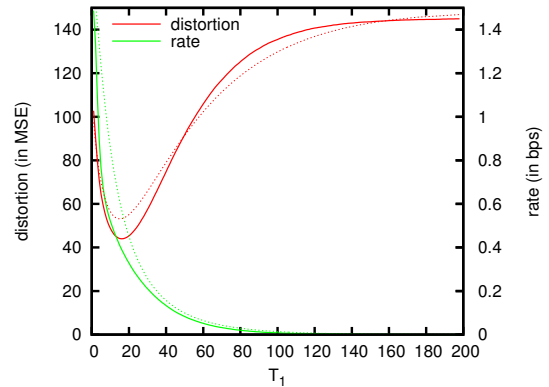


Fig. 2: Distortion and rate achieved by a GEQ scheme with  $K = 1$  and  $\Delta = 1.0$ . Results for the JPEG 2000- and estimation-based approach are depicted by the continuous and the dotted lines, respectively.

and 18.68, respectively. These parameters approximate the data of the wavelet subband used in the first approach.

Quantizers are designed by means of selecting a set of distinct quantization thresholds. We discretize the problem of threshold selection by assuming that all thresholds are a multiple of a given step size  $\Delta$ . Under this assumption, there are  $\Gamma = \lfloor \mathcal{W}/\Delta \rfloor$  possible thresholds. Assuming all thresholds are distinct, there are then  $\Gamma!/(\Gamma - K)!$  different quantizers that satisfy (2), with  $K$  denoting the number of quantization stages.

### III. QUANTIZER EVALUATION

#### A. Preliminary evaluation

First, the rate and the distortion achieved by a GEQ scheme with  $K = 1$  and thresholds resulting from  $\Delta = 1.0$  are examined to illustrate the differences achieved by the use of different quantization thresholds. Fig. 2 depicts the results achieved by the both approaches described before. The horizontal axis of the figure is the threshold  $T_1$ , whereas the vertical axes are the distortion (left axis) and rate (right axis) achieved after coding the first quantization stage using  $T_1$ . Rate is reported in bits per sample (bps). Results indicate that little rate is required when  $T_1$  is high. This is because most coefficients are inside the deadzone, which decreases the entropy and does not require the transmission of signs. When  $T_1$  decreases, more coefficients are dequantized outside the deadzone, which requires more rate due to higher entropy and transmission of more signs. The distortion has a different behavior. The higher the  $T_1$ , the larger the distortion because most coefficients are reconstructed as 0. Distortion decreases as lower is  $T_1$ , until  $T_1 = 16$  (this value is rounded off). For these data and quantization stage,  $T_1 = 16$  achieves the minimum quantization error, i.e.,  $\int_0^{T_1} \nu^2 d\nu + \int_{T_1}^{\mathcal{W}} (\nu - \hat{\nu})^2 d\nu$  is minimized when  $T_1 = 16$ . Distortion escalates again for  $T_1 < 16$ .

Different data achieve slightly different results than those depicted in Fig. 2, though rate and distortion behaves similarly depending on the quantization threshold employed. This evaluation is aimed to remark the differences in distortion and rate

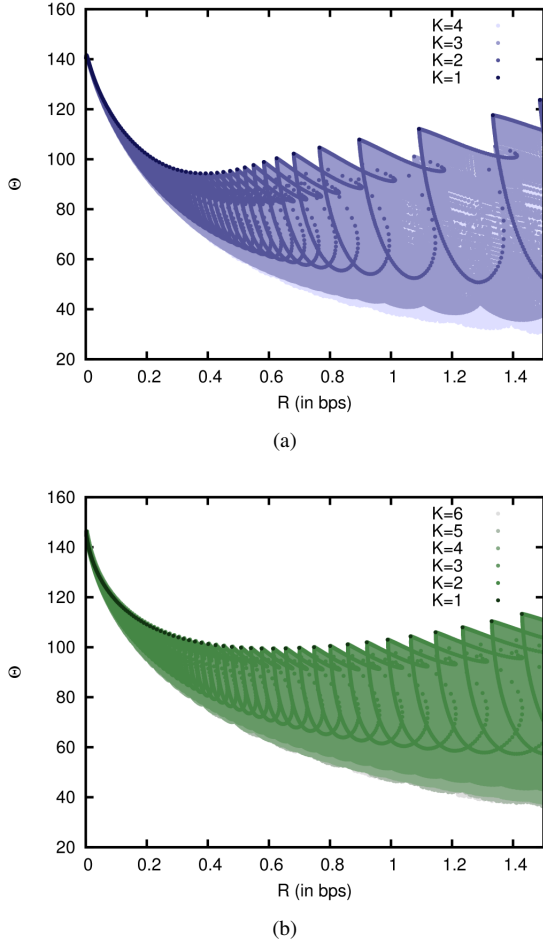


Fig. 3: Evaluation of the performance achieved with GEQ when up to  $K$  quantization stages are coded. (a) First approach (JPEG 2000-based) (b) Second approach (estimation-based).

produced by the use of different quantization thresholds. Differences are also noteworthy for quantizers with  $K > 1$  (not illustrated in Fig. 2). The rate and the distortion achieved at each quantization stage determine the quantizer’s performance via (1) and (3), hence, quantizers with distinct thresholds may achieve considerably different performance.

### B. Exhaustive search

The preliminary test reported in Fig. 2 evaluates a GEQ scheme with a single stage. Next, quantizers with  $K \geq 1$  are assessed. To appraise their performance according to (1), in what follows the expected multirate distortion is referred to as

$$\Theta = \int_0^R f_R(r) \cdot d(r) dr, \quad (7)$$

and serves as the performance metric to evaluate all GEQ schemes tested.  $d(r)$  in (7) is computed via (3) using the rate/distortion points achieved by the quantizer at each quantization stage.

Our analysis makes no assumptions in advance, so an exhaustive search is performed over all possible quantizers

in order to disclose those with best performance. We recall that the exhaustive search evaluates  $\Gamma! / (\Gamma - K)!$  quantizers for a scheme having  $K$  stages. Results achieved for the first approach with quantizers using 1, 2, 3, and 4 stages are depicted in Fig. 3(a). For  $K = 1$  and 2,  $\Delta = 1.0$ . To alleviate the computational load when  $K = 3$  and 4,  $\Delta$  is chosen as 3.0 and 8.0, respectively. The horizontal axis of the figure is the rate achieved for a given quantizer after the last quantization stage is coded (i.e.,  $R = r_K$ ). The vertical axis is the performance, expressed as  $\Theta$ . The performance of *every* quantizer tested is shown in this figure; *not* just the ones with best performance. As expected, there are multiple quantizers that yield the same rate, but different values of the performance metric  $\Theta$  (lower is better). Results suggest that at very low rates, approximately until 0.05 bps, the differences in performance among the different quantizers are very small. At higher rates, quantizers coding one stage (i.e., with  $K = 1$ ) achieve inferior performance to quantizers coding more stages.

Careful examination of Fig. 3(a) reveals some interesting patterns. Consider first a quantizer having one stage. Selection of the single threshold  $T_1$  then uniquely determines the rate and distortion achieved (at the end of the single stage). The results obtained are illustrated by the darkest points along the top of the shaded area in the figure. Each such point corresponds to a different value of  $T_1$ . For a given  $T_1$ , say  $T_1 = T$ , consider now all quantizers having 2 stages, with thresholds  $T_1 = T$  and  $T_2 = T'$ ,  $T' \in (0, \mathcal{W})$ . With respect to the single stage quantizer with threshold  $T$ , the addition of a second stage can only increase  $R$ . Consider now thresholds  $T' > T$ . When  $T'$  is close to the opposite end of its quantization interval  $[T, \mathcal{W})$  or, in other words, when  $T'$  is close to  $\mathcal{W}$ , the increase in  $R$  is negligible. This is because the density of coefficients between  $T'$  and  $\mathcal{W}$  is negligible, resulting in negligible entropy in the second stage. As  $T'$  moves away from  $\mathcal{W}$ , the increase in  $R$  will grow. When  $T'$  approaches  $T$  the entropy of the second quantization stage decreases, and the associated increase in  $R$  grows smaller. This can be seen in Fig. 3(a) as a “trace” leaving each darkest point (single stage quantizer) that then reverses its direction. Contrarily, for thresholds  $T' < T$  the rate always grows as  $T'$  moves away from  $T$  because the second quantization stage codes coefficients inside the deadzone, requiring the transmission of the sign for all significant coefficients. This is observed in Fig. 3(a) as a “trace” that leaves each darkest point, though this “trace” is more difficult to identify in the figure due to the high density of points. A similar behavior occurs for quantizers having 2 stages with  $T_1 = T''$ ,  $T'' \in (0, \mathcal{W})$  and  $T_2 = T$ . This produces four “traces” that leave each darkest point in Fig. 3(a).

Fig. 3(b) reports the results achieved by the second approach. The GGD support in this experiment is cut-off at  $\mathcal{W} = 199$ . Coefficient densities for  $\nu > 199$  are negligible for the selected parameters of the GGD. Quantizers with up to 6 stages are analyzed.  $\Delta$  is selected as 5.0, 8.0 and 14.0, respectively for  $K = 4, 5$ , and 6, and as  $\Delta = 1.0$  otherwise. The similarity between Figs. 3(a) and 3(b) seems to indicate that the two approaches are consistent to appraise quantizers. Similar patterns as those found in Fig. 3(a) are also observed



in this figure. These patterns can be more clearly observed in the video sequence at [45], which extends Fig. 3(b) showing the formation of selected quantizers in a sequential fashion. The sequence begins with a quantizer having one stage.  $T_1$  is selected as  $T_1 = \mathcal{W} - \Delta$  at the beginning of the sequence, and is decreased by  $\Delta$  until  $T_1 = \Delta$ . The threshold of the last quantizer tested is indicated with the label at the top-center of the figure, whereas its performance is indicated with a red cross. The performance of the previous quantizers tested are indicated with green points. After depicting the performance of the single stage quantizers, quantizers with  $K = 2$  and either  $T_1$  or  $T_2$  equal to 15 are showed. The patterns described in the previous paragraph can be observed in the video sequence. After depicting the performance of quantizers with two stages, selected quantizers with  $K = 3$  are showed. “Traces” with similar patterns as those described before are discerned, and similar observations also hold for quantizers with  $K > 3$  (not shown in the video). It also worth noting that the quantizers that achieve the best performance are those that partition the deadzone in all stages. This is also seen in the following test.

The performance and the overall design of many GEQ schemes are studied in the test of Fig. 3. The aim of the following test is to examine more in-depth the design of the best GEQ schemes depicted in that figure. To this end, the quantization thresholds of the best GEQ scheme achieved at a pre-defined target rate are analyzed. We note that the target rate  $R$  is now defined a priori instead of assuming that  $R$  coincides with  $r_K$  as in Fig. 3. For this purpose, all quantizers that achieve  $r_K \geq R$  should be evaluated via (3) and (7) and that with lowest  $\Theta$  kept. Unfortunately, the use of large  $\Delta$ s in the previous test, which is required to contain computational costs, does not produce sufficient finely rate-spaced GEQ schemes to approximate the target rate with precision. Therefore, an alternative procedure is devised whose goal is to approximate quantizers that attain  $R$  with more precision. The estimated performance of these quantizers may not be exact, but the proposed technique is computationally simple and provides the necessary insights to further the research.

The main idea is to linearly interpolate the distortion and the threshold of GEQ schemes having the last quantization thresholds that differ by  $\Delta$ . Let us define two GEQ schemes that each codes  $K$  stages, achieving rate/distortion points  $\{r_k, d_k\}$  and  $\{r'_k, d'_k\}$ , respectively, with  $1 \leq k \leq K$ . The quantization thresholds respectively employed by these quantizers are denoted by  $\{T_k\}$  and  $\{T'_k\}$  with  $T_k = T'_k \forall 1 \leq k < K$  and  $T'_K = T_K \pm \Delta$ . Of course, then  $r_k = r'_k$  and  $d_k = d'_k \forall 1 \leq k < K$  as well. When the target rate  $R$  lies between the final rate achieved by these two quantizers (i.e.,  $r_K < R < r'_K$  or  $r_K > R > r'_K$ ), a GEQ scheme having the same thresholds and rate/distortion points –except for the case of  $k = K$ – is fabricated. The final desired rate of this quantizer, say  $\tilde{r}_K$ , is  $R$  and linear interpolation is employed to estimate both the required threshold  $\tilde{T}_K$  and the achieved distortion  $\tilde{d}_K$  at the last quantization stage according to  $\tilde{T}_K = T_K + \alpha(T'_K - T_K)$  and  $\tilde{d}_K = d_K + \alpha(d'_K - d_K)$ , with  $\alpha = \frac{R - r_K}{r'_K - r_K}$ . This technique only modifies the last quantization stage and it does not affect previous stages.

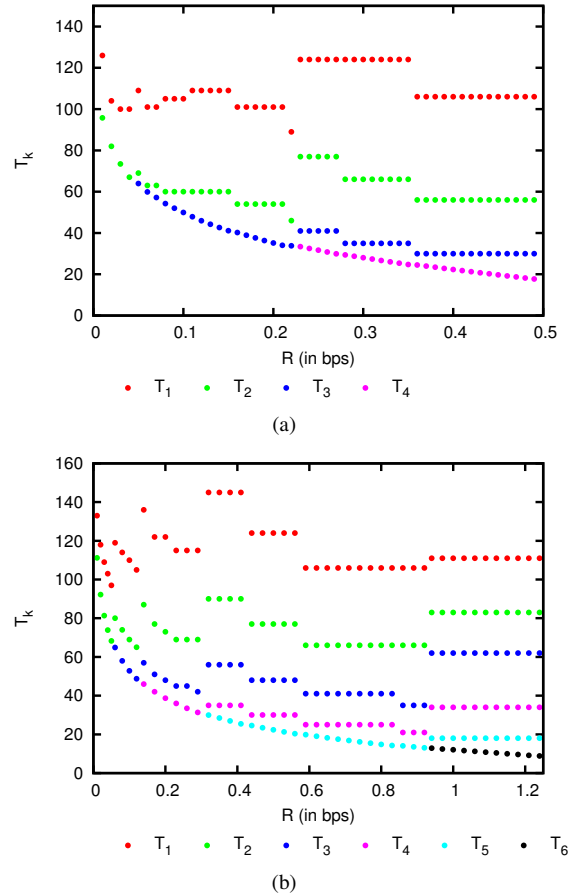


Fig. 4: Quantization thresholds achieved at different rates corresponding to GEQ schemes of Fig. 3 with best performance. (a) First approach (JPEG 2000-based) (b) Second approach (estimation-based).

The technique described before results in quantizers with a more regular structure than those achieved when the codestream produced by a quantizer with  $r_K \geq R$  is truncated at  $R$ . The reason is that codestream truncation does not alter the thresholds of the quantizer, which complicates the evaluation of quantization thresholds when one quantizer is found optimal at more than one target rate. This may happen often at high rates, since the exhaustive search using large  $\Delta$  obtains few near-optimal quantizers at these rates.

Fig. 4 reports the results obtained when the procedure described before is carried out over a range of  $R$ , for both approaches. When two or more quantizers with different number of quantization stages achieve same performance at the same target rate, that with the lowest number of stages is selected. The horizontal axis of the figure is the desired rate  $R$ , whereas the vertical axis is the thresholds employed by the quantizer. The results depicted in this figure are achieved under the same conditions as those of Fig. 3. Our first observation is that quantizers use more stages as the target rate increases, which coincides with our previous analysis. Quantization thresholds follow an exponentially decreasing pattern, approximately. This is most clear for the last quantization threshold employed, which has more variability due to the interpolation procedure

described before. Finally, we note again that all quantizers depicted in the figures carry out significance coding exclusively. This can be seen by noting that

$$T_{k+1} < T_k \quad \forall 1 \leq k < K. \quad (8)$$

### C. Low complexity algorithm

The previous analysis aids the development of an algorithm aimed to determine near-optimal GEQ schemes at a given target rate. The main insight behind the proposed algorithm is that quantization thresholds in consecutive quantization stages have strictly lower magnitudes, as expressed in (8). The algorithm is devised to determine the thresholds of an embedded quantizer with a fixed number of quantization stages. Roughly speaking, the algorithm begins by initializing the thresholds of a quantizer with the highest values that respect (8). Then the quantizer is iteratively modified to increase the target rate while maximizing performance. In each iteration, the rate-performance slope (decrease in  $\Theta$  divided by the increase in  $R$ ) is computed assuming a single threshold is provisionally decreased by  $\Delta$ . This slope is computed for all thresholds as permitted by (2) and (8). That threshold achieving the highest slope is then actually decreased. This process is iterated until  $R$  is achieved. Quantizers found by the algorithm differ only slightly in each iteration, so the target rate  $R$  can be achieved with high precision. The algorithm does not assure optimality, though the exponentially decreasing pattern of Fig. 4 together with the following experimental results seem to indicate that the quantizers found by this method achieve near-optimal performance. Despite using the rate-performance slope, the procedure carried out by this algorithm is *not* based on Lagrange optimization.

Alg. 1 summarizes the steps carried out by the proposed method. The function “evaluateGEQ( $\cdot$ )” (lines 4 and 12) returns the rate/distortion points achieved when the coding engine uses  $K$  quantization stages using the provided quantization thresholds. The function “evaluatePerformance( $\cdot$ )” (lines 5 and 13) computes the metric  $\Theta$  for the given set of points. The algorithm finishes execution when  $R$  is achieved or when a quantizer with that number of stages is not found for that  $R$ .

The computational complexity of Alg. 1 is much lower than that of the exhaustive search employed in the previous section. A maximum of  $K$  different quantizers are evaluated in each iteration of the algorithm, and each threshold can be decreased  $\lfloor W/\Delta \rfloor$  times, at most. The computational complexity of Alg. 1 is then  $O(K \lfloor \frac{W}{\Delta} \rfloor)$ .

Alg. 1 allows us to determine a near-optimal quantizer at a pre-defined rate without needing the interpolation technique described before. This facilitates an accurate analysis of optimally designed quantizers. To this end, Alg. 1 is executed to find quantizers with  $K = 1..10$  for values of  $R$  varying from a low to a medium rate. From the collection of all quantizers, that with the smallest  $\Theta$  at rate  $R$  is kept. Fig. 5(a) reports the performance achieved by the quantizers determined by the proposed algorithm. This figure reports results for

---

### Algorithm 1

---

```

1: for  $k \in [1, K]$  do
2:    $T_k^\bullet \leftarrow \lceil W/\Delta \rceil \Delta - k\Delta$ 
3: end for
4:  $\{r_k^\bullet, d_k^\bullet\} \leftarrow \text{evaluateGEQ}(\{T_k^\bullet\})$ 
5:  $\Theta^\bullet \leftarrow \text{evaluatePerformance}(\{r_k^\bullet, d_k^\bullet\})$ 
6: while  $r_K^\bullet < R$  do
7:    $\beta^\circ, \{T_k^\circ\}, \{r_k^\circ, d_k^\circ\}, \Theta^\circ \leftarrow \text{null}$ 
8:   for  $k \in [1, K]$  do
9:      $\{T_k^*\} \leftarrow \{T_k^\bullet\}$ 
10:    if  $(T_k^* > T_{k+1}^* + \Delta)$  or  $(k = K \text{ and } T_k^* > \Delta)$  then
11:       $T_k^* \leftarrow T_k^* - \Delta$ 
12:       $\{r_k^*, d_k^*\} \leftarrow \text{evaluateGEQ}(\{T_k^*\})$ 
13:       $\Theta^* \leftarrow \text{evaluatePerformance}(\{r_k^*, d_k^*\})$ 
14:       $\beta^* \leftarrow (\Theta^* - \Theta^\bullet) / (r_K^* - r_K^\bullet)$ 
15:      if  $(r_K^* > r_K^\bullet)$  and  $(\beta^\circ = \text{null or } \beta^* < \beta^\circ)$  then
16:         $\beta^\circ, \{T_k^\circ\}, \{r_k^\circ, d_k^\circ\}, \Theta^\circ \leftarrow \beta^*, \{T_k^*\}, \{r_k^*, d_k^*\}, \Theta^*$ 
17:      end if
18:    end if
19:  end for
20:  if  $\beta^\circ \neq \text{null}$  then
21:     $\{T_k^\bullet\}, \{r_k^\bullet, d_k^\bullet\}, \Theta^\bullet \leftarrow \{T_k^\circ\}, \{r_k^\circ, d_k^\circ\}, \Theta^\circ$ 
22:  else
23:    exit
24:  end if
25: end while
26: output $(\{T_k^\bullet\}, \{r_k^\bullet, d_k^\bullet\}, \Theta^\bullet)$ 

```

---

the second approach, though similar results hold for the first too. This figure is similar to Fig. 3(b), so to facilitate the comparison between the quantizers determined by Alg. 1 and those tested in the previous exhaustive search, all quantizers depicted in Fig. 3(b) are also depicted in Fig. 5(a) in gray. Results indicate that all quantizers determined by Alg. 1 always achieve equal or lower  $\Theta$  than those tested in the exhaustive search, which suggests that they are (close to) the optimal solution. Evidently, the quantizers achieving lower  $\Theta$  than those found in the exhaustive search were not evaluated in the exhaustive search due to the use of large  $\Delta$ .

The thresholds employed by the near-optimal quantizers determined by Alg. 1 are reported in Fig. 5(b). The exponentially decreasing pattern of the thresholds is evident in this figure. As in Fig. 4, the last threshold of all quantizers shows the clearest pattern, though the remaining thresholds behave similarly.

It is worth noting that quantizers determined by the proposed method can attain any target rate with more precision than that achieved by the classic USDQ+BPC scheme. To illustrate this point, Fig. 6 reports the rate/distortion points obtained by the optimal GEQ schemes depicted in Fig. 5 at target rates  $R = 0.05, 0.6$ , and  $1.4$  bps. The target rate  $R$  is represented as a dotted horizontal line in this figure. Fig. 6 also reports the performance achieved by the USDQ+BPC scheme,<sup>3</sup> which is identical in each subfigure. Evidently, the codestream generated by USDQ+BPC needs to be truncated at  $R$ , which rarely coincide with the end of quantization stages. Contrarily, there exist a high density of optimal GEQ schemes in any rate range. This is illustrated

<sup>3</sup>Rate-distortion points of the USDQ+BPC scheme correspond to a BPC strategy that carries out two coding passes per bitplane (significance and refinement coding), which is a common approach of coding engines [37].

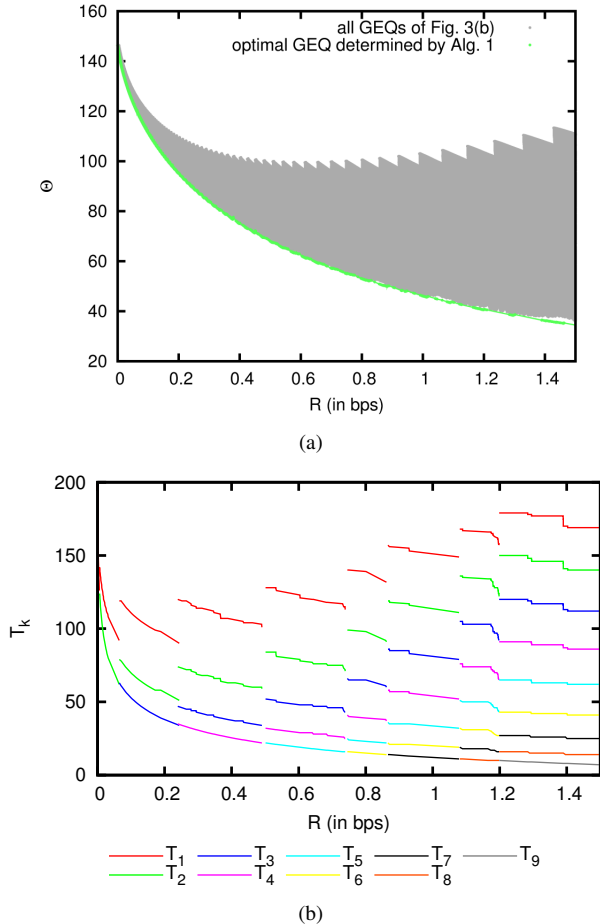


Fig. 5: Evaluation of the quantizers generated by the proposed algorithm. Data corresponds to the estimation-based approach. (a) Performance evaluation (b) Thresholds analysis.

in the video sequence at [46], which extends Fig. 6 showing all quantizers found by Alg. 1 between 0.001 to 1.4 bps in a sequential fashion. Since  $f_R(r)$  in (1) is chosen to be uniform, the quantizer’s performance for a given  $R$  can be conceptually seen as the area below the rate-distortion points up to rate  $R$ . The smaller the area (or, equivalently, the lower the distortion over the rate range  $(0, R]$ ), the better the performance. In general, the more precise attainment of the target rate achieved by the optimal GEQ scheme provides slightly better coding performance than that achieved by USDQ+BPC. This evaluation is carried out for low-to-medium rates since Algorithm 1 is devised employing the insights provided in the exhaustive search described before. The algorithm is not appropriate for high rates. The next section approaches this issue.

#### IV. PRACTICAL APPROACH

Sections II and III analyze the efficiency that can be achieved by the GEQ. Alg. 1 describes a low-complexity procedure to achieve optimal GEQ schemes for a target rate. Four issues need to be addressed before implementing the optimal GEQ scheme in practice:

- 1) *Target rate*: in order to find an optimal quantizer for a target rate, Alg. 1 needs to code the data more than

once. In practical terms, this means coding the image many times, which lays a computational burden on the codec that is not assumable in practice.

- 2) *Rate-distortion optimization procedures*: some modern coding systems enhance the scalability features of the codestream by means of coding small sets of wavelet coefficients (commonly called codeblocks) independently [38], [43], [44]. After coding the codeblocks, a rate-distortion optimization procedure truncates the codestream generated for each codeblock selecting the segments that are included in the final file. The difficulty of using the optimal GEQ scheme in such a coding system is that neither the rate for the codeblock’s codestream nor the distortion-rate slope threshold [29] selected to attain the rate of the final codestream is known a priori. As seen before, different target rates require different GEQ schemes, so the truncation of a codestream at an arbitrary rate would penalize performance unless data are coded again.
- 3) *Side information*: the GEQ scheme employs distinct quantization thresholds depending on the data coded. These thresholds are selected by the encoder and have to be known by the decoder. The rate needed to transmit the thresholds is negligible when the whole image, or when every wavelet subband, is coded together because only one set of thresholds, or one set of thresholds per subband, is used. But this side information may grow significantly when codeblocks are used since one distinct GEQ scheme may be necessary for each codeblock.
- 4) *High target rates*: the analysis of the previous section explores the design of optimal GEQ schemes only for low and medium rates. Real codecs require the coding of data for high rates as well.

To overcome these four issues we need to design a GEQ scheme that: 1) codes the data only once, 2) allows the truncation of the codestream at any point, 3) requires the transmission of none (or little) side information, and 4) is able to generate a codestream for low, medium, and high target rates. To fulfill these requirements, a GEQ scheme different from that produced by Alg. 1 is needed. Although a different GEQ scheme may penalize the quantizer’s performance, empirical evidence indicates that GEQ has enough flexibility to achieve a quantizer that fulfills the previous requirements without significantly penalizing its performance. Let us explain further. The exhaustive search of Section III-B discloses many existing quantizers at a fixed target rate. Note in Fig. 3(b), for instance, that at  $R = 0.1$  bps there exist many quantizers having performance  $\Theta$  between 111 to 122, approximately. Although the most interesting quantizers are those with lowest  $\Theta$ , the examination of quantizers that are close to the optimal reveals an interesting aspect regarding their design. Fig. 7 reports the thresholds employed by quantizers shown in Fig. 3(b) at  $R = 0.1$  bps,  $K = 3$  and  $\Theta \in [111.32, 111.9]$ . The horizontal axis of the figure is  $\Theta$ , whereas the vertical axis is the quantization thresholds employed by the quantizer. The quantizer depicted on the most left side of the figure achieves the best performance (i.e., is that in Fig. 3(b) achieving the



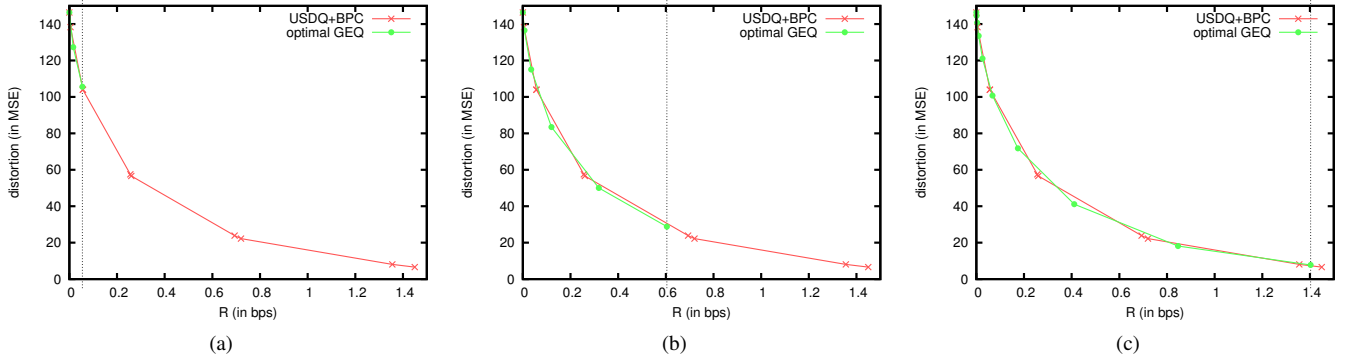


Fig. 6: Evaluation of the rate-distortion performance achieved by the USDQ+BPC scheme and the optimally designed GEQ schemes. Data corresponds to the estimation-based approach. The target rate  $R$  is indicated with an horizontal dotted line, being  $R = 0.05$  bps for (a),  $R = 0.6$  bps for (b) and  $R = 1.4$  bps for (c).

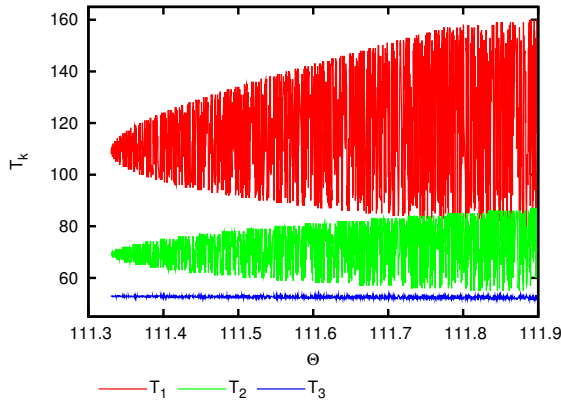


Fig. 7: Thresholds employed by quantizers with  $K = 3$  that achieve different performance at the target rate  $R = 0.1$  bps. Results for the estimation-based approach.

lowest  $\Theta$  at 0.1 bps). From left to right, the performance of quantizers decreases, though only slightly. Although all quantizers depicted in the figure achieve similar performance, their quantization thresholds differ significantly. In this figure, threshold  $T_1$  ranges from 90 to 160, whereas  $T_2$  ranges from 55 to 85, approximately. The last threshold exhibits less variability since small changes on  $T_3$  may produce significant variations on the rate due to the larger density of significant coefficients coded by the threshold. Similar results are achieved for other target rates, which seems to indicate that quantizers with slightly different thresholds to those used by the optimal GEQ scheme can also achieve competitive performance.

The flexibility on the selection of the quantization thresholds together with the insights provided in Section III guide the design of a GEQ scheme that fulfills the aforementioned requirements. Three main structural ideas are used in the conception of the practical GEQ scheme. The first is that the thresholds of the quantizer need not to vary depending on the target rate. This eliminates the need of coding the data many times, addressing the first and second issue described before. As seen in the next section, this reduces only slightly the efficiency of the quantizer due to the non-optimality of the thresholds employed at some rates and due to the target rate

may not coincide exactly with the end of a quantization stage.

The second structural idea behind the design of the practical GEQ is to establish the thresholds depending on the largest magnitude of the coefficients to be coded (i.e.,  $\mathcal{W}$ ). This entails negligible computational costs since  $\mathcal{W}$  is easy to compute for subbands or codeblocks, and allows the use of a distinct set of thresholds for each subband/codeblock requiring the transmission of little side information. Experimental evidence indicates that the rate increase due to the transmission of  $\mathcal{W}$  is negligible. Our experience indicates that only using  $\mathcal{W}$  to design the quantizer already achieves near-optimal performance, so the use of other parameters such as the variance or shape of the distribution can be avoided. The third aforementioned issue is thus resolved.

The third idea behind the practical GEQ is to devote the first stages of the quantizer to significance coding, as indicated by the analysis of the previous sections. After coding several significance stages, our experience seems to indicate that the most effective means of attaining higher rates is to interleave stages of refinement with stages of significance coding. This is intuitively explained by considering that every significance coding stage lessens the width of the deadzone interval, which reduces the density of coefficients in the deadzone and the rate produced in further significance stages. By carrying out first several significance coding stages and then interleaving significance with refinement, the rate can be grown until necessary without penalizing performance significantly. This strategy addresses the fourth issue mentioned before. We note that most USDQ+BPC strategies, including tier-1 of JPEG 2000, also interleave significance coding with refinement coding.

The design of the practical GEQ is depicted in Fig. 8(a). We note that, for some rates, the proposed design may have thresholds with variations with respect to the optimal one that are larger than those depicted in Fig. 7. The first three quantization stages carry out significance coding, with thresholds  $T_1 = \frac{2}{3}\mathcal{W}$ ,  $T_2 = \frac{1}{3}\mathcal{W}$ , and  $T_3 = \frac{1}{6}\mathcal{W}$ . These thresholds have been determined empirically to yield competitive performance for a variety of data found in wavelet subbands. Small variations on these thresholds do not affect the quantizer's performance significantly. As indicated in Fig. 8(a), the fourth stage of

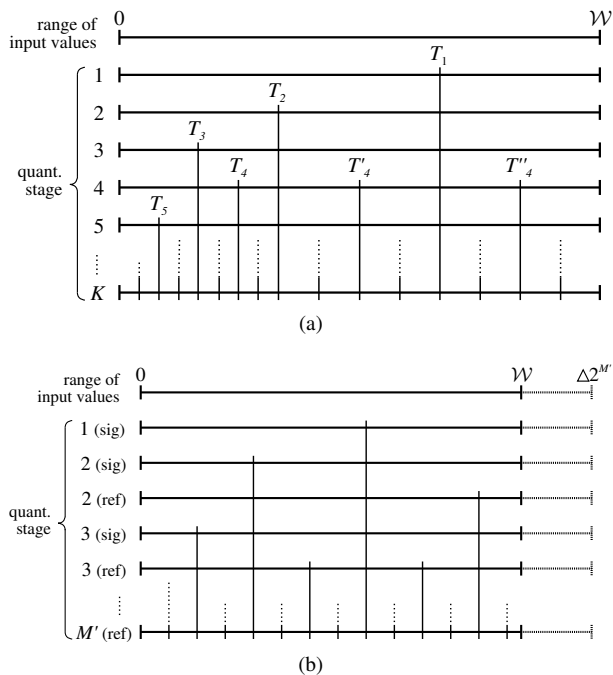


Fig. 8: Illustration of the quantization intervals produced by two multistage quantization schemes. The representation of the quantization stages is equivalent to that of Fig. 1. (a) and (b) illustrate the proposed GEQ scheme for practical implementations and an observation of the USDQ+BPC scheme, respectively.

the quantizer partitions all quantization intervals except the deadzone. This requires a small variation on the structure of the GEQ as is described in Section II-A because more than one quantization threshold is necessary in this stage. Such a variation allows the refinement of all coefficients outside the deadzone in a single stage. The refinement thresholds are in the middle of the quantization interval. Our experience indicates that to select different thresholds, or to split the refinement in more stages, does not affect performance. These thresholds are selected as  $T_4 = T_3 + \frac{1}{2}(T_2 - T_3)$ ,  $T_4' = T_2 + \frac{1}{2}(T_1 - T_2)$ , and  $T_4'' = T_1 + \frac{1}{2}(\mathcal{W} - T_1)$ .

The fifth stage of the practical GEQ carries out significance coding, partitioning the interval that contains zero, this is,  $T_5 = \frac{1}{2}T_3 = \frac{1}{12}\mathcal{W}$ . The sixth stage carries out refinement coding partitioning all intervals except the deadzone. The quantizer continues interleaving one significance stage with one refinement stage until the target rate is attained. This design achieves a multistage scheme with quantization intervals having two different widths depending on the interval laying on the left or right side of  $T_2$ . See in Fig. 8(a) that in quantization stage 5 all intervals on the left side of  $T_2$  have width  $w = T_2/4$ , whereas all intervals on the right side of  $T_2$  have width  $w' = (\mathcal{W} - T_2)/4$ . More generally, the interval width can be expressed as  $w = T_2/2^{(k-1)/2}$ ,  $k = 3, 5, 7, \dots$  and  $w' = (\mathcal{W} - T_2)/2^{k/2}$ ,  $k = 2, 4, 6, 8, \dots$  for intervals on the left and right side of  $T_2$ , respectively. The intervals on the right side of  $T_2$  have a width twice as large as that of the intervals on the left side of  $T_2$ .

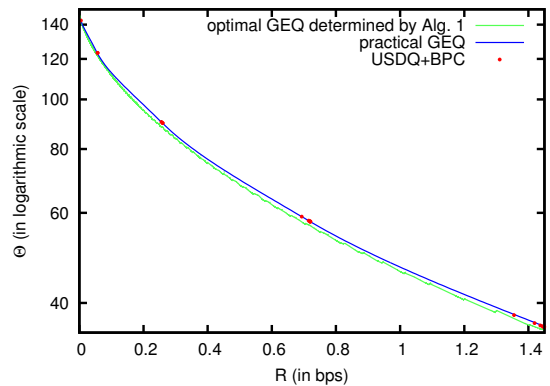


Fig. 9: Evaluation of the performance achieved by the optimal GEQ, the practical GEQ, and USDQ+BPC. Data corresponds to the estimation-based approach. For the sake of clarity, the vertical axis of the figure is in logarithmic scale.

The design of the practical GEQ is convenient since the density  $f_V(\nu)$  of coefficients on the right side of  $T_2$  (i.e., coefficients with large magnitudes) is very low. For the data used by the first and second approach of the previous section, for example, the density  $\int_{\frac{1}{3}\mathcal{W}}^{\mathcal{W}} f_V(\nu) d\nu$  is 0.5% and 0.6%, respectively. The larger intervals on the right side of  $T_2$  adjust coefficients  $\nu \geq T_2$  with less precision than coefficients  $\nu < T_2$ , but this does not affect performance—even when many quantization stages are applied—because less than 1% of coefficients are greater than  $T_2$ , in general. Fig. 9 reports the performance  $\Theta$  achieved by the practical GEQ as compared to the optimal GEQ produced by Alg. 1. As stated before, the quantizers produced by Alg. 1 can not be employed in practice, so, in this context, the practical GEQ and the quantizers produced by Alg. 1 can only be compared through the evaluation provided in this figure. The “optimal GEQ” plot reported in this figure is the same as that in Fig. 5(a). Results suggest that the performance achieved by the practical GEQ is near-optimal, commonly being less than 3% worse than the optimal. Fig. 9 also reports the performance achieved by USDQ+BPC, which is almost the same as that of the practical GEQ. We remark that the design of the practical GEQ restricts some features of GEQ. The loss in flexibility is inevitable to achieve practicality.

It is worth noting one observation from the structure of the practical GEQ when compared to the classic scheme of USDQ+BPC. Commonly, USDQ+BPC is described as a multistage quantization scheme that produces intervals of uniform width. Contrarily, the practical GEQ produces non-uniform quantization intervals due to the use of particular thresholds at the first three quantization stages. This may seem an structural difference between USDQ+BPC and the practical GEQ but, studied more in detail, this difference renders insignificant in practice. Let us recall the use of codeblocks, or similar conceptual partitions, to code the data of wavelet subbands. In general, one quantization step size  $\Delta$  is employed for the whole subband [29], and the number of magnitude bits necessary to code all coefficients in the codeblock is transmitted

to the decoder. This strategy is employed by JPEG 2000, for instance. Through this method, the available range for input values at the most significant bitplane of the codeblock, say  $M'$ , is utilized only partially. This is due to the largest magnitude of all coefficients (i.e.,  $\mathcal{W}$ ) does not commonly coincide with the largest magnitude available at that bitplane, which is  $2^{M'} - 1$ . Although this does not affect the coding performance achieved in practice due to the use of arithmetic coding, it conceptually approaches the quantization schemes of the practical GEQ and the USDQ+BPC.

Fig. 8(b) illustrates the scheme of USDQ+BPC taking into account the range of input values that may not be utilized in practice. The line at the top of figure represents the available range for the input values, with  $\mathcal{W}$  situated on the left side of the maximum value. This happens often in practice. The range  $(\mathcal{W}, \Delta 2^{M'})$  is depicted as a dotted line in the figure to indicate that there is none coefficient at this range. Since the interval partitioning is carried out through bitplane coding, the resulting quantization intervals achieve a similar structure as that of the practical GEQ, especially at the first stages. Two quantization stages per bitplane are depicted in the figure to differentiate significance coding from refinement coding. Although the practical GEQ and USDQ+BPC are devised using different points of view, this observation indicates that both achieve a similar structure. We stress this point since provides a theoretical justification behind the well-known competitive performance of USDQ+BPC.

## V. EXPERIMENTAL RESULTS

The practical GEQ is implemented in the tier-1 coding procedure of JPEG 2000. The resulting system is not JPEG 2000 compliant since the standard does not support GEQ-like quantizers, but provides an appropriate framework to evaluate the proposed method. Tier-1 defines three coding passes per bitplane called Significance Propagation Pass (SPP), Magnitude Refinement Pass (MRP), and Cleanup Pass (CP) [29]. SPP and CP carry out significance coding. The main difference between them is that SPP scans those coefficients that are more likely to become significant in the current bitplane. MRP refines the magnitude of coefficients that become significant in previous bitplanes. The practical GEQ is integrated in tier-1 *without* modifying these coding passes. Only the significance/refinement state of coefficients is computed differently when the practical GEQ is in use. The significance/refinement state is the function that determines whether a coefficient has to be reconstructed in the upper or lower quantization interval defined in that stage. In a compliant JPEG 2000 implementation these thresholds are imposed by the USDQ+BPC scheme, whereas the proposed method uses the thresholds defined by the practical GEQ. Also, the practical GEQ executes MRP only when defined by the quantizer. All the aspects of a conventional JPEG 2000 codec, including the use of codeblocks, rate-distortion optimization procedures, etc. need not to be modified when the practical GEQ is implemented in tier-1.

The images employed in the experiments are chosen from different corpora. “Portrait” ( $2048 \times 2560$ ) and “Flowers”

( $2731 \times 2048$ ) are natural images that belong to the ISO 12640-1 and ISO 12640-2 corpus, respectively. “Barcelona” ( $4096 \times 4096$ ) is an aerial image provided by the Cartographic Institute of Catalonia [47] that belongs to the remote sensing community. “Hip” ( $2048 \times 2495$ ) is a computer radiology provided by the UDIAT Centre Diagnostic [48] that belongs to the medical community.

The evaluation of the practical GEQ considers two aspects of the quantizer: coding performance and complexity. Coding performance is evaluated by compressing the image at different target rates and computing the Peak Signal to Noise Ratio (PSNR). This is a conventional test to assess the performance of image codecs that provides PSNR values for selected target rates. The quantizer’s complexity is assessed in terms of quantization stages or, equivalently, in terms of coding passes included in the codestream at a target rate. The quantizer’s complexity evaluation is relevant since, in general, a codec that codes fewer coding passes spends less computational resources as well [37]. The number of coding passes of the practical GEQ is computed as  $2k'$  when  $k' \leq 3$  (recall that each significance coding stage requires 2 coding passes), and as  $(3k' + 3)/2$  and  $(3k' + 2)/2$  when  $k'$  is odd and even, respectively, with  $k'$  denoting the number of quantization stages coded for that codeblock.

Fig. 10 reports the results obtained when coding the aforementioned images using codeblocks of size  $64 \times 64$ . The horizontal axis of each figure is the rate, whereas the vertical axes are the PSNR (left axis) and the number of coding passes (right axis). For clarity, the right axis is reversed (higher means fewer coding passes). Coding performance and coding passes are reported with crosses and points, respectively. Results suggest that, in this framework, the coding performance of the practical GEQ and the USDQ+BPC scheme are virtually the same. This is in agreement with the analysis of the previous sections, which indicated the reasons behind the near-optimal performance of USDQ+BPC and the practical GEQ. Results also suggest that the complexity of the practical GEQ is lower than that of USDQ+BPC. At some rates this difference is significant. See, for instance, that at 0.8 bps the practical GEQ codes approximately 30% fewer coding passes than USDQ+BPC, for the “Hip” image. Similar results hold for other coding parameters, and for the reversible 5/3 wavelet transform included in JPEG 2000. Although the differences between the coding performance and the computational complexity achieved by the practical GEQ and JPEG 2000 vary depending on the image, these results also hold for other images as well.

Table I reports the same results as those of Fig. 10 for a selected set of target rates. The column “JP2” reports the PSNR, number of coding passes, and  $\Theta$  achieved by JPEG 2000 at the target rate specified on the top of the column. The column “GEQ” reports the difference between the practical GEQ and JPEG 2000. For the “Portrait” image at 0.50 bps, for instance, the practical GEQ achieves a PSNR 0.04 dB lower than that achieved by JPEG 2000. These results indicate that the practical GEQ achieves a coding performance (in terms of PSNR and  $\Theta$ ) very similar to that of USDQ+BPC while significantly reducing the number of passes coded.

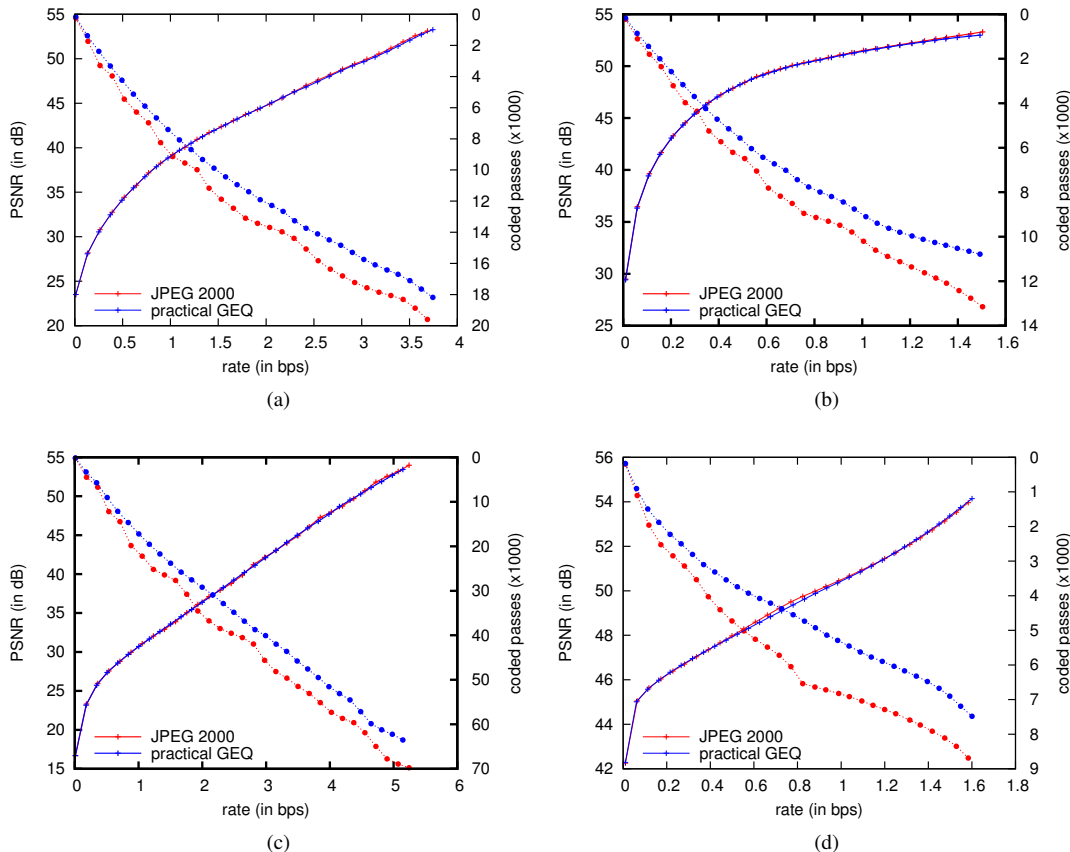


Fig. 10: Evaluation of the coding performance (plots with crosses) and the quantizer’s complexity (plots with dots) achieved by JPEG 2000 and the practical GEQ. Codeblocks of size  $64 \times 64$ . (a) “Portrait” (b) “Flowers” (c) “Barcelona” (d) “Hip”.

TABLE I: Evaluation of the coding performance, the quantizer’s complexity, and the multirate distortion measure  $\Theta$  achieved by JPEG 2000 and the practical GEQ. Codeblocks of size  $64 \times 64$ . The last three columns report the minimum, maximum, and average values depicted in the “GEQ” columns.

“Portrait”	0.50 bps		1.00 bps		1.50 bps		2.00 bps		2.50 bps		3.00 bps		3.50 bps		min, max	aver.
	JP2	GEQ	JP2	GEQ	JP2	GEQ	JP2	GEQ	JP2	GEQ	JP2	GEQ	JP2	GEQ		
	PSNR	34.24	-0.04	39.08	-0.03	42.20	-0.03	44.73	0.03	47.40	-0.17	49.72	-0.14	52.28		
cod. pass.	5317	-1016	9023	-1478	11742	-1674	13612	-1499	15614	-1613	17436	-1772	18493	-1405	-1772, -1016	-1494
$\Theta$	66.08	-0.03	40.15	0.01	28.66	0.01	22.21	0.02	18.11	0.01	15.23	0.03	13.15	0.01	-0.03, 0.03	0.01
“Flowers”	0.20 bps		0.40 bps		0.60 bps		0.80 bps		1.00 bps		1.20 bps		1.40 bps		min, max	aver.
	JP2	GEQ	JP2	GEQ	JP2	GEQ	JP2	GEQ	JP2	GEQ	JP2	GEQ	JP2	GEQ		
	PSNR	43.03	0.00	47.12	0.01	49.33	-0.09	50.54	-0.05	51.49	-0.07	52.20	-0.06	52.93		
cod. pass.	3103	-541	5642	-859	7743	-1229	9107	-1218	10167	-1178	11310	-1371	12390	-1864	-1864, -541	-1180
$\Theta$	11.62	0.02	6.81	0.00	4.86	0.01	3.81	0.00	3.15	0.01	2.70	0.00	2.36	0.01	0.00, 0.02	0.01
“Barcelona”	0.70 bps		1.40 bps		2.10 bps		2.80 bps		3.50 bps		4.20 bps		4.90 bps		min, max	aver.
	JP2	GEQ	JP2	GEQ	JP2	GEQ	JP2	GEQ	JP2	GEQ	JP2	GEQ	JP2	GEQ		
	PSNR	28.73	0.04	32.86	0.15	37.08	-0.13	41.18	-0.19	44.91	0.13	48.75	0.20	52.55		
cod. pass.	14250	-1667	26390	-3850	36724	-6317	41836	-3350	51487	-5586	58672	-5148	67773	-5993	-6317, -1667	-4559
$\Theta$	207.64	-0.34	130.95	-0.10	94.48	-0.14	73.03	-0.13	59.09	-0.09	49.47	-0.09	42.49	-0.08	-0.34, -0.08	-0.14
“Hip”	0.20 bps		0.40 bps		0.60 bps		0.80 bps		1.00 bps		1.20 bps		1.40 bps		min, max	aver.
	JP2	GEQ	JP2	GEQ	JP2	GEQ	JP2	GEQ	JP2	GEQ	JP2	GEQ	JP2	GEQ		
	PSNR	46.22	0.02	47.40	0.00	48.56	-0.12	49.65	-0.18	50.49	-0.06	51.44	0.00	52.62		
cod. pass.	2702	-588	4095	-846	5219	-1218	6342	-1726	6837	-1504	7288	-1359	7849	-1356	-1726, -588	-1228
$\Theta$	1.81	0.00	1.59	0.00	1.40	0.00	1.25	0.00	1.13	0.01	1.03	0.01	0.94	0.01	0.00, 0.01	0.00

Fig. 11 reports the results obtained when all coefficients of the subband are coded together (i.e., without using codeblocks). This is not supported in JPEG 2000, but it is reported herein to appraise the practical GEQ when codeblocks are not in use. Again, results suggest that the coding performance achieved by both USDQ+BPC and the practical GEQ is very

similar. Also, the practical GEQ codes fewer coding passes than USDQ+BPC, although in this test the differences are not as significant as when codeblocks are used. These results suggest that the practical GEQ adjusts with high precision the quantization thresholds especially when small portions of data (i.e., codeblocks) are coded independently.

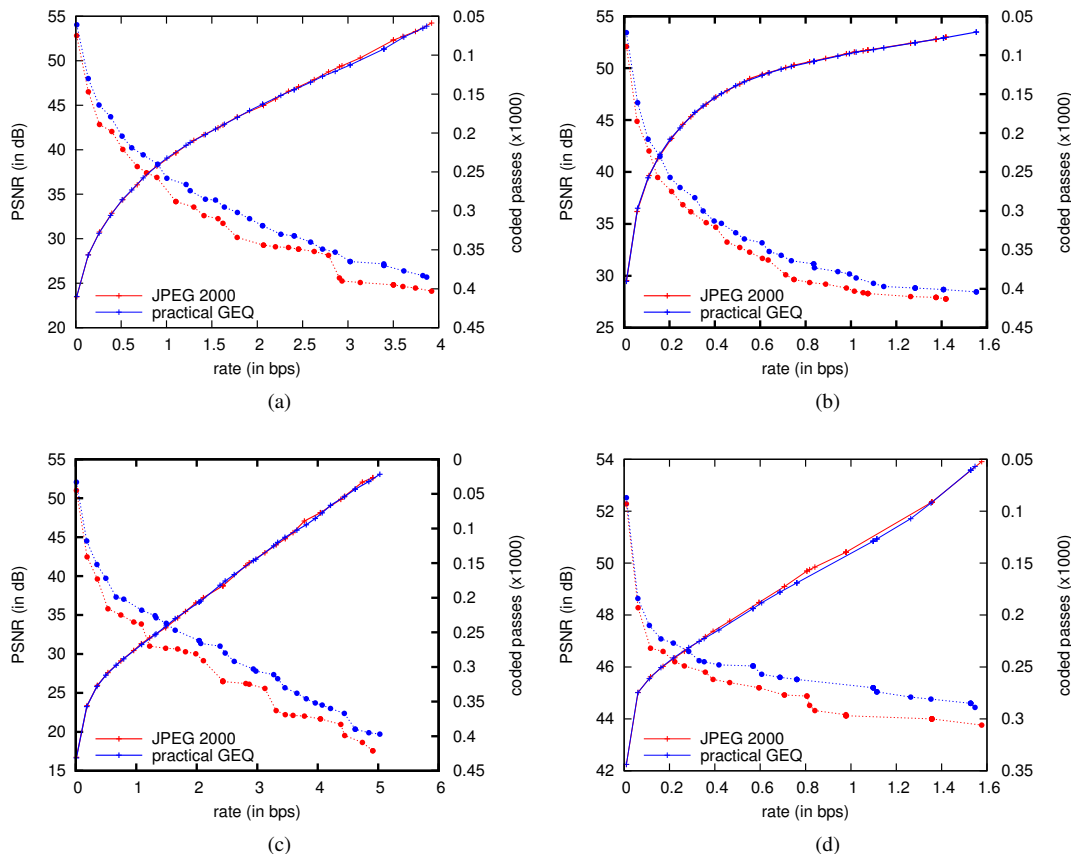


Fig. 11: Evaluation of the coding performance (plots with crosses) and the quantizer’s complexity (plots with dots) achieved by JPEG 2000 and the practical GEQ. All the coefficients of the subband are coded together. (a) “Portrait” (b) “Flowers” (c) “Barcelona” (d) “Hip”.

## VI. CONCLUSIONS

Embedded quantization is a fundamental mechanism employed by lossy image coding systems to generate a quality progressive codestream. This work explores embedded quantizers aimed to the wavelet-based lossy, or lossy-to-lossless, compression of images. First, general embedded quantization (GEQ) is introduced. GEQ is a multistage quantization scheme that codes arbitrary quantization intervals in each quantization stage. This provides a greater flexibility to the quantizer than that provided by other schemes. Second, the optimization problem posed to achieve optimal performance for a selected range of decoding rates is described. This specifies an appropriate metric to appraise the performance of quantizers tested. Third, an exhaustive search evaluates the performance of GEQ for two different approaches. The first approach uses a codec based on JPEG 2000, which precisely appraise the efficiency of GEQ when introduced into an image codec employing advanced coding mechanisms. The second approach uses estimates of rate and distortion, which permits the extension of the exhaustive search to quantizers that would otherwise be too complex to evaluate. Fourth, the design of the quantizers achieving best performance found in the exhaustive search is studied, and a low-complexity algorithm that produces near-optimal GEQ schemes is proposed. Fifth, the insights learned

during this analysis are employed to devise a practical GEQ scheme that achieves performance close to the optimal, and that can be implemented in an image codec. Sixth, the practical GEQ is integrated in the core coding system of JPEG 2000 and evaluated in terms of coding performance and quantizer’s complexity. Experimental results suggest that the coding performance achieved by the practical GEQ is very similar to other conventional quantization schemes, though it requires fewer quantization stages. This is convenient for codecs since, in general, fewer stages implies less computational resources. This seems to indicate that the implementation of the practical GEQ into codecs may reduce computational costs without sacrificing coding performance.

The development carried out in this work seeks quantizers designed to achieve optimal coding performance. Nonetheless, the flexibility provided by GEQ can also be employed to devise quantizers with other purposes, such as the reduction of the maximum distortion or the maximum rate in each stage, the accurate attainment of a target rate, or the lossless coding of images. Also, GEQ may be adopted in other disciplines such as speech or audio coding.

## ACKNOWLEDGMENT

The author thanks M. Marcellin, J. Serra-Sagrsta, Thomas Richter, and the reviewers of the IEEE Transactions on Signal Processing and the Data Compression Conference for their comments and remarks.

## REFERENCES

- [1] R. M. Gray and D. L. Neuhoff, "Quantization," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2325–2383, Oct. 1998.
- [2] J. Max, "Quantizing for minimum distortion," *IEEE Trans. Inf. Theory*, vol. 6, no. 1, pp. 7–12, Mar. 1960.
- [3] W. H. R. Equitz and T. M. Cover, "Successive refinement of information," *IEEE Trans. Inf. Theory*, vol. 37, no. 2, pp. 269–275, Mar. 1991.
- [4] P. W. Wong, "Progressively adaptive scalar quantization," in *Proc. IEEE International Conference on Image Processing*, vol. 1, Oct. 1996, pp. 357–360.
- [5] Z. Xiong, K. Ramchandran, and M. T. Orchard, "Space-frequency quantization for wavelet image coding," *IEEE Trans. Image Process.*, vol. 6, no. 5, pp. 677–693, May 1997.
- [6] A. Ortega and M. Vetterli, "Adaptive scalar quantization without side information," *IEEE Trans. Image Process.*, vol. 6, no. 5, pp. 665–676, May 1997.
- [7] G. J. Sullivan, "On embedded scalar quantization," in *Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 4, May 2004, pp. 605–608.
- [8] W.-Y. Chan, S. Gupta, and A. Gersho, "Enhanced multistage vector quantization by joint codebook design," *IEEE Trans. Commun.*, vol. 40, no. 11, pp. 1693–1697, Nov. 1992.
- [9] H. Jafarkhani and N. Farvardin, "A scalable wavelet image coding scheme using multi-stage pruned tree-structured vector quantization," in *Proc. IEEE International Conference on Image Processing*, vol. 3, Oct. 1995, pp. 81–84.
- [10] C. F. Barnes, S. A. Rizvi, and N. M. Nasrabadi, "Advances in residual vector quantization: A review," *IEEE Trans. Image Process.*, vol. 5, no. 2, pp. 226–262, Feb. 1996.
- [11] E. A. B. da Silva, D. G. Sampson, and M. Ghanbari, "A successive approximation vector quantizer for wavelet transform image coding," *IEEE Trans. Image Process.*, vol. 5, no. 2, pp. 299–310, Feb. 1996.
- [12] K. Bao and X.-G. Xia, "Image compression using a new discrete multiwavelet transform and a new embedded vector quantization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, no. 6, pp. 833–842, Sep. 2000.
- [13] A. Said and W. A. Pearlman, "A new, fast, and efficient image codec based on set partitioning in hierarchical trees," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, no. 3, pp. 243–250, Jun. 1996.
- [14] D. Mukherjee and S. K. Mitra, "Successive refinement lattice vector quantization," *IEEE Trans. Image Process.*, vol. 11, no. 12, pp. 1337–1348, Dec. 2002.
- [15] —, "Vector SPIHT for embedded wavelet video and image coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, no. 3, pp. 231–246, Mar. 2003.
- [16] M. W. Marcellin and T. R. Fischer, "Trellis coded quantization of memoryless and Gauss-Markov sources," *IEEE Trans. Commun.*, vol. 38, no. 1, pp. 82–93, Jan. 1990.
- [17] A. Aksu and M. Salehi, "Multistage trellis coded quantisation (MS-TCQ) design and performance," *IEE Proceedings- Communications*, vol. 144, no. 2, pp. 61–64, Apr. 1997.
- [18] H. Brunk and N. Farvardin, "Embedded trellis coded quantization," in *Proc. IEEE Data Compression Conference*, Apr. 1998, pp. 93–102.
- [19] H. Jafarkhani and V. Tarokh, "Successively refinable trellis coded quantization," in *Proc. IEEE Data Compression Conference*, Apr. 1998, pp. 83–92.
- [20] A. Bilgin, P. J. Sementilli, and M. W. Marcellin, "Progressive image coding using trellis coded quantization," *IEEE Trans. Image Process.*, vol. 8, no. 11, pp. 1638–1643, Nov. 1999.
- [21] P. Seigneurbieux and Z. Xiong, "Progressive trellis-coded space-frequency quantization for wavelet image coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 12, no. 7, pp. 587–591, Jul. 2002.
- [22] S. Steger and T. Richter, "Universal refinable trellis coded quantization," in *Proc. IEEE Data Compression Conference*, Mar. 2009, pp. 312–321.
- [23] M. D. Gaubatz and S. S. Hemami, "Ordering for embedded coding of wavelet image data based on arbitrary scalar quantization schemes," *IEEE Trans. Image Process.*, vol. 16, no. 4, pp. 982–996, Apr. 2007.
- [24] M. W. Marcellin, M. A. Lepley, A. Bilgin, T. J. Flohr, T. T. Chinen, and J. H. Kasner, "An overview of quantization in JPEG 2000," *ELSEVIER Signal Processing: Image Communication*, vol. 17, no. 1, pp. 73–84, Jan. 2002.
- [25] J. M. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans. Image Process.*, vol. 41, no. 12, pp. 3445–3462, Dec. 1993.
- [26] H. Gish and J. N. Pierce, "Asymptotically efficient quantizing," *IEEE Trans. Inf. Theory*, vol. 14, no. 5, pp. 676–683, 1968.
- [27] N. Farvardin and J. W. Modestino, "Optimum quantizer performance for a class of non-gaussian memoryless sources," *IEEE Trans. Inf. Theory*, vol. 30, no. 3, pp. 485–497, May 1984.
- [28] G. J. Sullivan, "Efficient scalar quantization of Exponential and Laplacian random variables," *IEEE Trans. Inf. Theory*, vol. 42, no. 5, pp. 1365–1374, Sep. 1996.
- [29] D. S. Taubman and M. W. Marcellin, *JPEG2000 Image compression fundamentals, standards and practice*. Norwell, Massachusetts 02061 USA: Kluwer Academic Publishers, 2002.
- [30] F. Auli-Llinas, M. W. Marcellin, L. Jimenez-Rodriguez, I. Blanes, and J. Serra-Sagrsta, "Embedded quantizer design for low rate lossy image coding," in *Proc. IEEE Data Compression Conference*, Apr. 2012, pp. 89–98.
- [31] F. Auli-Llinas, J. L. Monteagudo-Pereira, J. Serra-Sagrsta, and J. Bartrina-Rapesta, "Low-complexity lossy image coding through a near-optimal general embedded quantizer," in *Proc. IET Image Processing Conference*, Jul. 2012, pp. 1–6.
- [32] F. Auli-Llinas and M. W. Marcellin, "Distortion estimators for bitplane image coding," *IEEE Trans. Image Process.*, vol. 18, no. 8, pp. 1772–1781, Aug. 2009.
- [33] X. Wu, S. Dumitrescu, and N. Zhang, "On multirate optimality of JPEG2000 code stream," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2012–2023, Dec. 2005.
- [34] F. Auli-Llinas, A. Bilgin, and M. W. Marcellin, "FAST rate allocation through steepest descent for JPEG2000 video transmission," *IEEE Trans. Image Process.*, vol. 20, no. 4, pp. 1166–1173, Apr. 2011.
- [35] S. Mallat, "A theory of multiresolution signal decomposition: the wavelet representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, pp. 674–693, Jul. 1989.
- [36] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image coding using wavelet transform," *IEEE Trans. Image Process.*, vol. 1, no. 2, pp. 205–220, Apr. 1992.
- [37] F. Auli-Llinas and M. W. Marcellin, "Scanning order strategies for bitplane image coding," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1920–1933, Apr. 2012.
- [38] *Information technology - JPEG 2000 image coding system - Part 1: Core coding system*, ISO/IEC Std. 15444-1, Dec. 2000.
- [39] *Image Data Compression*, Consultative Committee for Space Data Systems Std. CCSDS 122.0-B-1, Nov. 2005.
- [40] R. W. Buccigrossi and E. P. Simoncelli, "Image compression via joint statistical characterization in the wavelet domain," *IEEE Trans. Image Process.*, vol. 8, no. 12, pp. 1688–1701, Dec. 1999.
- [41] Z. He and S. K. Mitra, "A unified rate-distortion analysis framework for transform coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 12, pp. 1221–1236, Dec. 2001.
- [42] F. Auli-Llinas, "Stationary probability model for bitplane image coding through local average of wavelet coefficients," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2153–2165, Aug. 2011.
- [43] D. Taubman, "High performance scalable image compression with EBCOT," *IEEE Trans. Image Process.*, vol. 9, no. 7, pp. 1158–1170, Jul. 2000.
- [44] W. A. Pearlman, A. Islam, N. Nagaraj, and A. Said, "Efficient, low-complexity image coding with a set-partitioning embedded block coder," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 14, no. 11, pp. 1219–1235, Nov. 2004.
- [45] F. Auli-Llinas. (2012) Formation of selected GEQ schemes. [Online]. Available: [http://www.deic.uab.es/~francesc/research/embedded\\_quantization/performance.avi](http://www.deic.uab.es/~francesc/research/embedded_quantization/performance.avi)
- [46] —. (2012) Near-optimal GEQ schemes in a sequential fashion. [Online]. Available: [http://www.deic.uab.es/~francesc/research/embedded\\_quantization/quantizers.avi](http://www.deic.uab.es/~francesc/research/embedded_quantization/quantizers.avi)
- [47] Institut Cartografic de Catalunya. (2007, May) ICC. Barcelona 08038 (Spain). [Online]. Available: <http://www.icc.cat>
- [48] Corporacio Parc Tauli. (2007, May) UDIAT centre diagnostic. Sabadell 08208 (Spain). [Online]. Available: <http://www.parctauli.es/webcspt/udiat>





**Francesc Aulí-Llinàs** (S'2006-M'2008) is a Fellow of the Ramón y Cajal program funded by the Spanish Government. He is currently in the Department of Information and Communications Engineering, at the Universitat Autònoma de Barcelona (Spain). He received the B.Sc. and B.E. degrees in Computer Management Engineering and Computer Engineering in 2000 and 2002, respectively, both from the Universitat Autònoma de Barcelona (Spain), and for which he was granted with two extraordinary awards of Bachelor (awards given to the first students of the promotion). In 2004 and 2006 he respectively received the M.S. degree and

the Ph.D. degree (with honors), both in Computer Science from the Universitat Autònoma de Barcelona. Since 2002 he has been consecutively awarded with doctoral and postdoctoral fellowships in competitive calls. From 2007 to 2009 he carried out two research stages of one year each with the group of David Taubman, at the University of New South Wales (Australia), and with the group of Michael Marcellin, at the University of Arizona (USA). He is the main developer of BOI, a JPEG2000 Part 1 implementation that was awarded with a free software mention from the Catalan Government. His research interests include a wide range of image coding topics, including highly scalable image and video coding systems, rate-distortion optimization, distortion estimation, and interactive transmission, among others.