ON EMBEDDED SCALAR QUANTIZATION

Gary J. Sullivan

Microsoft Corporation, One Microsoft Way, Redmond, WA 98052 USA
garysull@microsoft.com

ABSTRACT

This paper studies the rate-distortion performance of symmetric scalar quantizers having a large (effectively infinite) number of steps and using the same step size for all steps except the one containing the zero input value. Quantizers of this form have been shown to have good performance for a variety of sources, and are precisely optimal for the Laplacian source. Performance is particularly investigated for embedded quantization, in which the representation of a source quantity is refined successively by forming finer quantizers from further segmenting the steps of coarser quantizer constructions. Although the use of a double-wide dead-zone has dominated prior embedded quantization practice, it is shown that any rational number can be maintained as a stable dead-zone ratio. Two forms are investigated in more depth – quantizers with dead-zone ratios of 1 and 2, and a ratio of 1 is shown to often provide a significant performance advantage (up to 1 dB). Performance is explored primarily in the context of the generalized Gaussian pdf using the squared-error distortion measure, but should also apply in other contexts.

1. INTRODUCTION

This paper studies the rate-distortion performance of symmetric scalar quantizers having a large (effectively infinite) number of steps and using the same step size for all steps except the one containing the zero input value. Quantizers of this form have been shown to have good performance for a variety of sources, and are precisely optimal for the Laplacian source. Performance is particularly investigated for applications using embedded quantization, in which the representation of a source quantity is refined successively by forming finer quantizers from further segmenting the steps of coarser quantizer constructions.

While embedded quantization (e.g., as used in the JPEG standards) has been typically formed by using a double-wide dead-zone and dividing the non-zero steps of such a quantizer by two in each embedded refinement stage, this paper describes and analyzes a more general form of such embedded quantizer constructions. It is shown herein that any rational number can be maintained as a stable ratio between the size of the embedded quantizer dead-zone and the size of the remaining steps. Two forms are investigated in more depth – quantizers with dead-zone ratios of 1 and 2, and a ratio of 1 is shown to often provide a significant performance advantage. Quantizers using optimal, single-offset, and mid-point reconstruction rules are investigated. Performance is explored primarily in the context of the generalized Gaussian source pdf using the squared-error distortion measure, but should also apply in other contexts.

2. SCALAR QUANTIZATION ANALYSIS

2.1. Scalar quantization

Define a *scalar quantizer* as an approximating functional mapping $x \rightarrow Q[x]$ that can be decomposed into two distinct stages, the first being a classifier functional mapping $x \rightarrow A[x]$ that maps a real-valued input variable x to an integer-valued quantization index $A[x]$, and the second being a reconstructor functional mapping $k \rightarrow \beta[k]$ that maps each quantization index k to a real-valued reconstruction value $\beta[k]$, so that $Q[x] = \beta[A[x]]$. Define the distortion introduced by such a quantizer using a difference-based distortion measure $d(x–Q[x])$. A quantizer is considered better in the rate-distortion sense for a source random variable X if the expected value of the distortion measure $D = E_X\{d(X–Q[X])\}$ is lower for an equal or lower entropy H of $A[X]$.

2.2. The DZ+UTQ

Define a dead-zone quantizer with uniform threshold values for all steps except the one containing the zero input value, denoted here as a DZ+UTQ (dead-zone plus uniform threshold quantizer), as a quantizer having an index mapping rule $A[x]$ that is based on two parameters: $s > 0$ (the step size for all steps other than the zero-input dead-zone region) and $z \geq 0$ (the ratio of the dead-zone size to the size of the other steps), as follows:

$$A[x] = \text{sign}(x) * \max\left(0, \left\lfloor \frac{|x|}{s} - \frac{z}{2} + 1 \right\rfloor\right) \qquad (1)$$

where the notation $\lfloor \cdot \rfloor$ denotes the smallest integer less than or equal to the argument. The case with $z = 1$ is referred to as a uniform threshold quantizer (UTQ). Quantizers of the UTQ form have good performance for a variety of sources [1], and the DZ+UTQ form is optimal for the Laplacian source [2].

2.3. Optimal reconstruction

Assuming that the source pdf $f(x)$ is symmetric about zero, the optimal reconstruction rule for a symmetric difference-based distortion measure $d(|x–y|)$ is given by [3][4]:

$$\beta[k] = \begin{cases} \min_y{}^{-1} \int_0^{\frac{z\,s}{2}} [d(|x-y|) + d(|y-x|)]\, f(x)\, dx, & k = 0, \\ \text{sign}(k) \min_y{}^{-1} \int_{\frac{z\,s}{2}+(|k|-1)s}^{\frac{z\,s}{2}+|k|\,s} d(|x-y|)\, f(x)\, dx, & k \neq 0. \end{cases} \qquad (2)$$

We assume herein (as would be the case for optimal reconstruction for typical source pdfs and typical distortion measures) that $\beta[0] = 0$. One common distortion measure is the squared-error measure $d(|x–y|) = |x–y|^2$, for which the optimal reconstruction rule uses the conditional mean in each region.

2.4. Single-offset reconstruction

A second form of reconstruction rule will also be discussed here. We call this a *single-offset* reconstruction rule. It is based on an offset parameter Δ (where ordinarily $0 < \Delta \le s / 2$), as follows:

$$\beta[k] = \begin{cases} 0, & \text{for } k = 0, \\ \text{sign}(k)\left[\left(|k| + \frac{z}{2} - 1\right)s + \Delta\right], & \text{for } k \neq 0. \end{cases} \quad (3)$$

A special case of the single-offset reconstruction rule is the *mid-point* reconstruction rule, specified by $\Delta = s / 2$. Mid-point reconstruction is commonly used for convenience, and in the limit as s becomes small, the performance of the mid-point rule becomes optimal under a variety of well-behaved conditions [5].

2.5. Embedded quantization

Define an *embedded* quantizer design (also known as a *progressive* quantizer design) for a DZ+UTQ as an indexed sequence of quantizer mappings $\{ Q_i[\cdot], i = 0, \ldots \}$ such that as i increases, each quantization index mapping function $A_i[x]$ is formed by further segmentation of the regions formed by the classifier $A_{i-1}[x]$. Embedded quantizers can enable *bitstream scalability*, in which a subset of the (entropy-coded) data used to specify a fine representation of the source can be used to obtain a suitable coarser representation. When using an embedded quantizer design, the representation of the source data can be sent in multiple stages, where in each stage the expected quantity of information necessary to be transmitted is the difference in entropy $H_i – H_{i-1}$ of the mapping functions for the two stages.

If $Q_{i-1}[\cdot]$ and $Q_i[\cdot]$ are both DZ+UTQs, this implies that each non-dead-zone step of $Q_{i-1}[\cdot]$ is divided into non-dead-zone steps of $Q_i[\cdot]$, so there is some integer $m_i \ge 0$ such that

$$s_i = s_{i-1} / (m_i + 1), \quad (4)$$

and the dead-zone of $Q_{i-1}[\cdot]$ is divided into a next dead-zone for $Q_i[\cdot]$ plus some other steps to each side of zero, so that there is some integer n_i such that $0 \le n_i \le (m_i+1)\, z_{i-1} / 2$ for which

$$z_{i-1}s_{i-1} - 2n_i s_i = z_i s_i, \quad (5)$$

resulting in

$$z_{i-1} = (z_i + 2n_i) / (m_i + 1). \quad (6)$$

We define the dead-zone ratio to be *stable* if $z_i = z_{i-1}$. Solving Eq. (6) in this case, we obtain the following relation for a stable dead-zone ratio $\hat{z}$ (where the hat denotes stability):

$$\hat{z} = 2n / m. \quad (7)$$

Thus the use of the same $m > 0$ and $n \ge 0$ for every stage can enable maintaining a stable dead-zone ratio of $\hat{z}$ for all $i \ge 0$. Eq. (7) shows that *any rational number* can be achieved as a stable dead-zone ratio by appropriate selection of m and n. This counters a misconception that the author has sometimes encountered in discussions – the belief that $\hat{z} = 2$ (using m and n equal to 1) is the only possible stable dead-zone ratio. In fact an infinite number of alternatives exist.

Starting with any dead-zone ratio z_i, if the value of $m_i > 0$ and $n_i \ge 0$ are held constant for multiple values of i, the result of $j \ge 0$ iterations of Eq. (6) is the relationship

$$z_{i-j} = \hat{z} + (z_i - \hat{z}) / (m+1)^j. \quad (8)$$

This shows that when the same value of m and n are used for multiple stages of embedded quantization, regardless of the dead-zone ratio z_i used at some high bit rate, the dead-zone ratio rapidly approaches $\hat{z}$ as j is increased (i.e., at the earlier, lower bit-rate stages of the embedded quantization design). Thus the dead-zone ratio can only differ substantially from $\hat{z}$ in the final refinement stages of the embedded quantizer operation.

The embedded quantization designs of the JPEG-1992 and JPEG-2000 standards and the still texture object and fine-granularity scalability features of MPEG-4 part 2 are based on designs using $\hat{z} = 2$. The use of $\hat{z} = 2$ or more has also been implicit in a number of non-embedded encoding designs. For example, the spacing of the reconstruction values for predicted regions in the H.261, MPEG-1, MPEG-2, H.263, and MPEG-4 part 2 video coding standards reflects an intent for use of $\hat{z} \ge 2$, as the reconstruction levels are spaced appropriately for mid-point reconstruction with $\hat{z} = 2$ (and altering the decision thresholds to increase optimality for the specified reconstruction values results in the use of an even larger dead-zone ratio).

However, designs based on $\hat{z} = 1$ (or at least $\hat{z} < 2$) are used for the non-embedded form of JPEG-1992; intra DC coefficients in H.261 and H.263; intra DC and AC in MPEG-1, MPEG-2 and H.263 Annex I; all intra DC and some intra AC in MPEG-4 part 2; and all coefficients in H.264/AVC. In these specifications, the reconstruction values are equally spaced around zero and away from zero, so mid-point reconstruction would imply the use of $\hat{z} = 1$ (and widening the dead-zone to $\hat{z} > 2$ for the same reconstruction values would place the non-zero reconstruction values outside of their corresponding classification regions, which would be a very obviously suboptimal quantizer design).

Note that the above standards do not fully specify the quantizer design – each of them allows some variation in the encoder classification rule $A[\cdot]$ and/or the decoder reconstruction rule $\beta[\cdot]$. In fact, in some cases, some parts of these rules are not specified in the standard at all and in others the method provided may not really be intended to represent good practice. (For example, JPEG-1992 provides an embedded reconstruction rule that is effectively a single-offset rule in which Δ does not change with i, which is a rather poor rule approaching or exceeding 6 dB of suboptimality at some rates if i becomes large.)

In practice, we believe that designs using small values of m are the most likely to be of interest for applications, as these provide finer granularity in the bit rates produced by consecutive stages of embedded quantization. We also expect relatively small values of n to be the most typically useful, since large dead-zone ratios appear difficult to justify in rate-distortion analysis for most sources. We therefore focus on the cases $z = 2$ (resulting from $m = 1$ and $n = 1$) and $z = 1$ (an embedded UTQ construction resulting from $m = 2$ and $n = 1$). The case with $z = 0$, sometimes called a *mid-rise* quantizer, is not considered here due to its inability to produce bit rates below 1 bit per sample.

3. THE GENERALIZED GAUSSIAN SOURCE

The generalized Gaussian source pdf is given by

$$f_{GG}(x) = \frac{\nu}{2\Gamma(1/\nu)}\left[\frac{\eta(\nu)}{\sigma}\right]\exp\left\{-\left[\frac{\eta(\nu)}{\sigma}|x|\right]^{\nu}\right\}, \quad (9)$$

where σ is the standard deviation, ν is a shape parameter, and

$$\eta(\nu) = \sqrt{\Gamma(3/\nu) / \Gamma(1/\nu)}. \quad (10)$$

The Laplacian and Gaussian pdfs are special cases of the generalized Gaussian pdf with $\nu = 1$ and $\nu = 2$, respectively.

Members of the generalized Gaussian pdf family with ν in the range of 0.5 to 2 have frequently been used as a model for compression applications [1][2][3][6][7][8][9][10][11][12], particularly including transform-based image and video compression. It is common practice, for example, for the Laplacian source model to be used for the values of non-DC transform coefficients [6][7][8][11] and for the Gaussian source model to be used for the values of DC coefficients [7][8] for image and video coding. Recent theoretical analysis has provided a rationale for this use of the Laplacian model [12].

4. THE LAPLACIAN SOURCE

For the special case of the Laplacian source (a generalized Gaussian source with ν = 1), the quantizer performance can be computed analytically, following the method described in [2]. The Laplacian source pdf is given by

$$f_L(x) = \frac{1}{\sigma\sqrt{2}} e^{-|x|\sqrt{2}/\sigma} . \quad (11)$$

Defining $\alpha = s\sqrt{2}/\sigma$, the DZ+UTQ entropy is then

$$H_L = B(e^{-z\alpha/2}) + e^{-z\alpha/2}\left[1 + B(e^{-\alpha})/(1-e^{-\alpha})\right] \text{ bits,} \quad (12)$$

where the function $B(p)$ for $0 < p < 1$ is given by

$$B(p) = -p\log_2(p) - (1-p)\log_2(1-p) , \quad (13)$$

and the expected distortion for single-offset reconstruction is

$$D_L = \frac{\sigma^2}{2}\left[\gamma(z\alpha/2, 0) + \gamma(\alpha, \delta)\, e^{-z\alpha/2}/(1-e^{-\alpha})\right], \quad (14)$$

where

$$\gamma(a,b) = \int_0^a d(|x-b|)e^{-x}dx , \quad (15)$$

and $\delta = \Delta\sqrt{2}/\sigma$. For the squared-error distortion measure,

$$\gamma(a,b) = (b^2 - 2b + 2)(1-e^{-a}) - ae^{-a}(a-2b+2) , \quad (16)$$

and the optimal reconstruction rule is a single-offset rule with

$$\delta = 1 - \alpha e^{-\alpha}/(1-e^{-\alpha}) . \quad (17)$$

For mid-point reconstruction, $\delta = \alpha/2$. Note that the value of δ for optimal reconstruction approaches $\alpha/2$ as s (and therefore α) approaches zero, confirming that mid-point reconstruction is asymptotically optimal when s is small.

5. QUANTIZER PERFORMANCE

5.1. The Laplacian source

Fig. 1 shows the gain in performance for the use of a quantizer with a dead-zone ratio $z = 1$ versus $z = 2$ when using the squared-error distortion measure and either the optimal reconstruction or mid-point reconstruction rule. When using optimal reconstruction, a DZ+UTQ with $z = 1$ always provides equal or better performance than $z = 2$, and provides a performance improvement up to 0.8 dB, although $z = 2$ appears to dominate current usage practice. When using mid-point reconstruction, although at higher rates a similar amount of performance improvement can be obtained by using $z = 1$ rather than $z = 2$ with mid-point reconstruction as well, there is a region of bit rates below 2 bits per sample in which the performance of DZ+UTQ with $z = 1$ is inferior to that with $z = 2$. This is due to the significant degree of suboptimality of the mid-point reconstruction rule for the case of $z = 1$, as shown by Fig. 2.

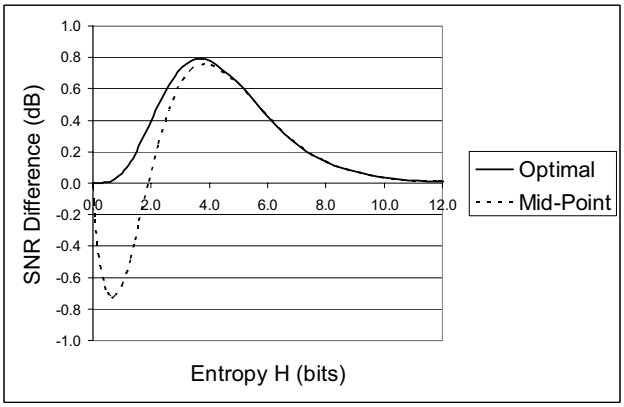

Fig. 1 – Benefit for use of $z = 1$ vs. $z = 2$ with ν = 1.

Fig. 2 shows the loss in performance for the use of mid-point reconstruction rather than optimal reconstruction for dead-zone ratios of $z = 1$ and $z = 2$. Mid-point reconstruction is significantly suboptimal for $z = 1$, causing a performance loss of as much as 0.83 dB at around 0.75 bits per sample, while the penalty diminishes at higher rates. For $z = 2$, mid-point reconstruction is not as harmful, causing a maximum performance penalty of only about 0.08 dB. It can therefore be concluded that the region of performance loss for $z = 1$ relative to $z = 2$ with mid-point reconstruction as shown in Fig. 1 is due to the suboptimality of mid-point reconstruction for $z = 1$.

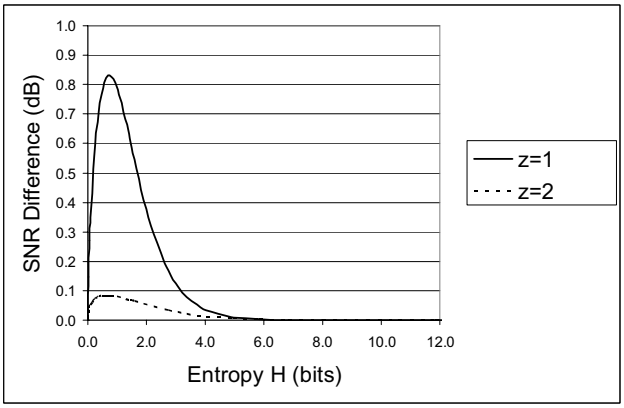

Fig. 2 – Penalty of mid-point vs. optimal rule with ν = 1.

5.2. The Gaussian source

Figs. 3 & 4 correspond to Figs. 1 & 2 for the Gaussian source. Although the amount of performance difference for the Gaussian source varies somewhat from that shown for the Laplacian source, the nature of the performance differences are similar.

5.3. The generalized Gaussian source with ν = 0.5

Figs. 5 & 6 correspond to Figs. 1 & 2 for the generalized Gaussian source with ν = 0.5. Here a somewhat different behavior can be seen. When ν is 0.5 there is a region with a (relatively minor) performance loss (up to about 0.13 dB) for $z = 1$ relative to $z = 2$, even when optimal reconstruction is used. Such a region was not found for the Laplacian or Gaussian cases. The performance characteristics are otherwise roughly similar in nature for ν = 0.5 as for the other two plotted cases.

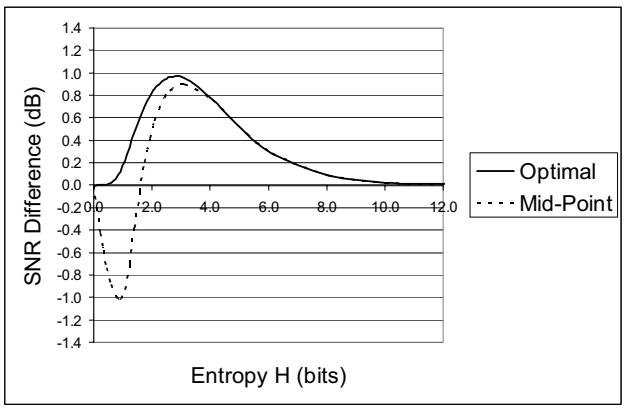

Fig. 3 – Benefit for use of $z = 1$ vs. $z = 2$ with $\nu = 2$.

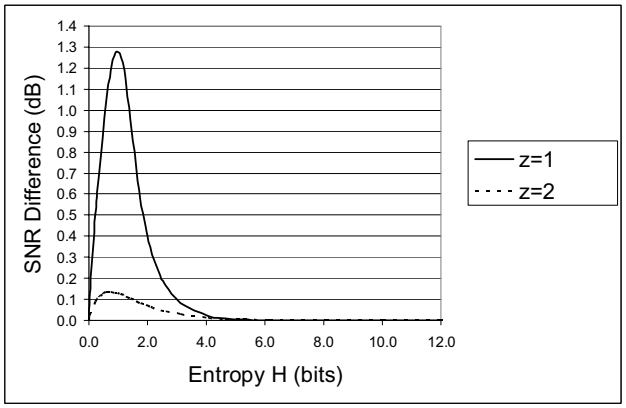

Fig. 4 – Penalty of mid-point vs. optimal rule with $\nu = 2$.

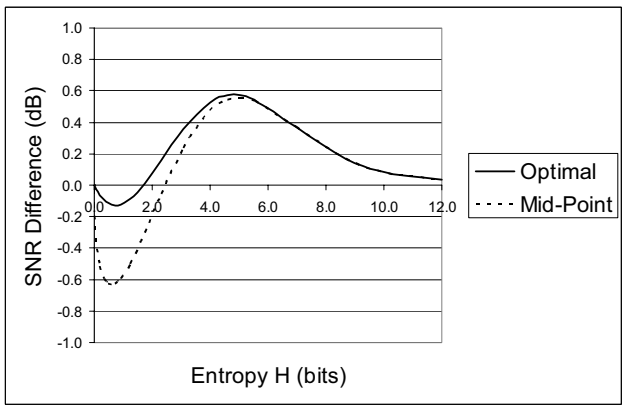

Fig. 5 – Benefit for use of $z = 1$ vs. $z = 2$ with $\nu = 0.5$.

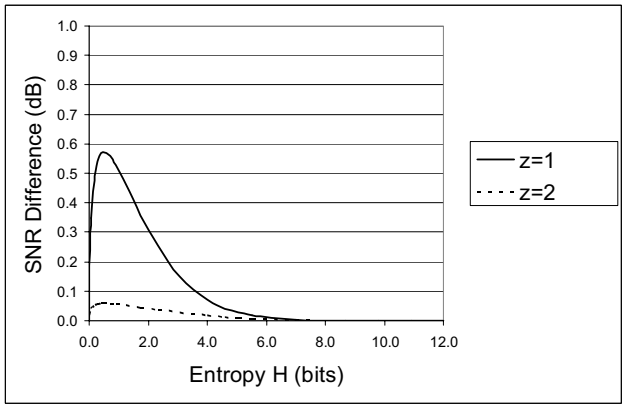

Fig. 6 – Penalty of mid-point vs. optimal rule with $\nu = 0.5$.

6. CONCLUSIONS

We have investigated the rate-distortion performance of DZ+UTQ designs, and have shown that any rational number can be achieved as a stable dead-zone ratio for an embedded quantization design. We have further shown that a significant performance benefit (up to 1 dB) can often be obtained by using a dead-zone ratio of 1 rather than the ratio 2 that appears to be more prevalent in prior embedded quantization use.

7. REFERENCES

[1] N. Farvardin and J. W. Modestino, "Optimum quantizer performance for a class of non-Gaussian memoryless sources", *IEEE Trans. Inform. Theory,* vol. IT-30, No. 3, pp. 485-497, May 1984.

[2] G. J. Sullivan, "Efficient scalar quantization of exponential and Laplacian random variables", *IEEE Trans. Inform. Theory,* vol. IT-42, No. 5., pp. 1365-1374, Sept. 1996.

[3] S. P. Lloyd, "Least squares quantization in PCM", *IEEE Trans. Inform. Theory,* vol. IT-28, No. 2, pp. 7-12, March 1982 (reprint of work originally presented in July 1957).

[4] J. Max, "Quantizing for minimum distortion", *IRE Trans. Inform. Theory,* vol. IT-6, No. 1, pp. 7-12, March 1960.

[5] H. Gish and J. N. Pierce, "Asymptotically efficient quantizing", *IEEE Trans. Inform. Theory,* vol. IT-14, No. 5, Sept. 1968.

[6] A. G. Tescher, "Transform image coding", in *Advances in Electronics and Electron. Physics,* Suppl. 12, Academic Press, New York, pp. 113-115, 1979.

[7] H. Murakami, Y. Hatori, and H. Yamamoto, "Comparison between DPCM and Hadamard transform coding in the composite coding of the NTSC color TV signal", *IEEE Trans. on Commun.,* vol. COM-30, No. 3, pp. 469-479, March 1982.

[8] R. C. Reininger and J. D. Gibson, "Distribution of two-dimensional DCT coefficients for images", *IEEE Trans. on Commun.,* vol. COM-31, No. 6, pp. 835-839, June 1983.

[9] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation", *IEEE Trans. Pattern Anal. and Machine Intell.,* vol. PAMI-11, No. 7, pp. 674-692, July 1989.

[10] F. Müller, "Distribution shape of two-dimensional DCT coefficients of natural images", *IEEE Electronics Letters,* vol. 29, No. 22, Oct. 1993.

[11] R. L. Joshi and T. R. Fischer, "Comparison of generalized Gaussian and Laplacian modeling in DCT image coding", *IEEE Signal Proc. Letters*, vol. SPL-2, No. 5, pp. 81-82, May 1995.

[12] E. Y. Lam and J. W. Goodman, "A mathematical analysis of the DCT coefficient distributions for images", *IEEE Trans. Image Proc.,* vol. IP-9, No. 10, pp. 1661-1666, Oct. 2000.