Context-based Adaptive Image Coding

Christos Chrysafis, Antonio Ortega*
Integrated Media Systems Center
University of Southern California
Los Angeles, CA 90089-2564
chrysafi,ortega@sipi.usc.edu

Abstract

In this paper we present an adaptive image coding algorithm based on novel backward-adaptive quantization/classification techniques. We use a simple scalar quantizer to quantize the image subbands. Our algorithm uses several contexts to characterize the subband data and different arithmetic coder parameters are matched to each context. We show how the context selection can be driven by rate-distortion criteria and how the performance can be improved by replacing the scalar quantization strategy by a entropy-constrained approach. Our results are comparable or better than the recent state of the art with our algorithm also having advantages in terms of simplicity.

1. Introduction

Over the past few years adaptivity has become an essential component of state of the art image coders, in particular those based on wavelets. Several researchers have advocated adapting in various ways the basic components in a wavelet-based image coder, namely, the tree-structured filterbank, the filters themselves, the quantizers and the entropy coders. In this paper we concentrate on the issue of adaptive quantization/entropy coding for a fixed filterbank. The issue of joint adaptation of quantizers and filterbanks [11] is not considered here.

Two main approaches to adaptive quantization have been reported in the recent literature. The first approach relies on a fixed quantization for all coefficients in a given band and a layered transmission of the coefficients using binary or low order (ternary, quaternary) arithmetic coding. Examples include the algorithms of [3, 6, 2]. Context based arithmetic coders were used in [6] while in [3, 2] context information was taken into account by using the zero-tree data structure, which enables the joint transmission of zero-valued coefficient present at the same spatial location across several frequency bands.

The second approach for adaptivity relies on using different quantizers, and thus entropy coders, for different regions of each subband. One example is the work of [1] where a different quantizer is used for each "class" of coefficients, after block-wise classification in each band has been performed. The classification technique used in [1] relied on pre-analyzing the subband data and sending the class assigned to each block as side information. In [10] it was shown that this approach could be extended to a backward adaptation framework, i.e. where the class of each coefficient is determined from previously quantized coefficients in the same band.

In this work we present a novel approach to adaptive quantization of image subbands which can be seen as a combination of both abovementioned methods. We use a fixed uniform threshold quantizer (UTQ) for all the subbands and arithmetic coding of the resulting set of coefficients [1]. Furthermore, as in [10], we use backward adaptive classification to determine which set of probabilities our arithmetic coder will use. Subband coefficients are modelled as Laplacian random variables for which the parameters are explicitly sent to the decoder, so that UTQs in each band can use reproduction levels that are matched to the given Laplacian parameter (see Section2.1). Since several different arithmetic coders (AC) can be used for the quantized coefficients a key issue is that of determining how to assign a coefficient to each AC. To do so we classify current coefficients based on past neighboring quantized coefficients. We generate a predictor based on the neighboring coefficients and select thresholds on the predictor to determine the class. Based on simple assumptions we show that the optimal classification can be approximated by designing an Lloyd-Max quantizer (LMQ) matched to the distribution of the predictor.

We are thus considering a context-based adaptive arith-

*This work was supported in part by the National Science Foundation under grant MIP-9502227 (CAREER).

[1] Note that the size of our alphabet is much larger than in [3, 6, 2]

metic coder similar to that proposed in [9] with the major differences being (i) we operate in the subband domain, rather than the image domain, and (ii) our contexts are determined based on past *quantized* data rather than from the original data as in the lossless compression scheme of [9]. Our approach is simpler than adaptive quantization methods, it may also be better suited to high rates where the layered coding approaches lose some of their benefits.

2 Context-based adaptation

In our algorithm adaptivity is achieved by (i) matching the reproduction levels of the UTQ to the statistics of each band (see Section 2.1), (ii) using different entropy coders depending on the context of the current coefficient (see Section 2.2) and (iii) using adaptive arithmetic coders (see Section 2.3).

2.1 Design of UTQ with dead zone

Assume that in a given subband the wavelet coefficients have Laplacian distribution with known parameter λ:

$$f(\rho) = \frac{1}{2}\lambda e^{-\frac{1}{2}\lambda|\rho|}, \qquad (1)$$

for which we would like to design an optimized UTQ with dead zone, as in Fig. 1. A UTQ has a fixed step size (except for the "dead zone" around zero) and the reproduction levels for each bin should be placed at the centroid of the distribution for that bin, i.e. for the interval $[b_{n-1}, b_n]$ the reproduction level should be[2]:

$$E\{x_n | x_n \in [b_{n-1}, b_n)\} = \frac{\int_{t=b_{n-1}}^{b_n} t f(t) dt}{\int_{t=b_{n-1}}^{b_n} f(t) dt}. \qquad (2)$$

Let $\alpha = b_n - b_{n-1}$ be the length of each interval in UTQ and define θ, θ_0 as:

$$\theta = e^{-\frac{1}{2}\lambda\alpha}, \quad \theta_0 = \frac{1}{2}e^{-\frac{1}{2}\lambda b_0} \qquad (3)$$

where $[-b_0, b_0]$ is the dead zone. The probability that a sample falls into the n^{th} bin $[b_{n-1}, b_n)$ is:

$$p_n = \theta_0 \theta^{|n|}(1-\theta) \qquad (4)$$

so that

$$E\{x_n | x_n \in [b_{n-1}, b_n)\} = b_n + \frac{2}{\lambda} - \frac{\alpha\theta}{1-\theta} \qquad (5)$$

and thus it is easy to find the reproduction levels given the stepsize and the Laplacian distribution parameter. Given the fixed stepsize these will be the optimal reproduction levels. The paramater λ is explicitly sent to the decoder as side information for each subband. In the section to follow we made the choice $b_0 = \alpha$.

[2] In the next equations assume $b_n \geq 0$; the case $b_n < 0$ requires simple modifications

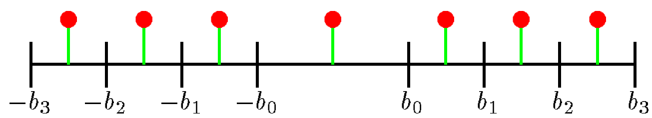

Figure 1. Uniform Thresholded Quantizer with dead zone

2.2 Prediction and Context Selection

Suppose that we have transmitted a number of coefficients $x_k, k = 0 \ldots n$ of the wavelet representation of our image. Based on the past we try to estimate the next coefficient that we need to transmit.

$$\hat{x}_n = \mathcal{P}\{x_n, x_{n-1}, \ldots, x_0\} \qquad (6)$$

Many experiments have shown that traditional linear predicition methods are not very efficient for encoding image subbands. In a linear prediction scheme the difference between the current coefficient and a predictor obtained from previously quantized ones is sent. Since the correlation of the wavelets coefficients tends to be close to zero, and prediction results in doubling of the dynamic range, little gain is in general achieved with this method.

However context information is useful when it comes to adapting the entropy coder as was demonstrated in [9] in a lossless image coding scenario. In this work we use a small neighborhood of previously quantized coefficients to determine, from a finite set of choices, which probability model to use for the entropy coder. The motivation is simple; as shown in [10], when surrounding coefficients are close to zero it is more likely that the next coefficient will also be zero. The coefficient in the same position in the previous band also offers some information about the value of the current coefficient.

In practice we only try to estimate the distribution of the magnitude $|x|$ of the value x we need to transmit. Our predictor has the form:

$$\hat{y} = a_0|y_0| + a_1|y_1| + a_2|y_2| + a_3|y_3|, \qquad (7)$$

where the y_i can be seen in Fig. 3. There is no reason for distinguishing between y_1 and y_2 in the prediction so $a_1 = a_2$. We also found out that for the most of the images $a_0 \approx a_3$ so we set $a_0 = a_3$. In our experiments we used $a_0 = a_3 = 1/10$ and $a_1 = a_2 = 4/10$ which provided good performance for a range of images. Now, based on the value of $\hat{y}$, we can select a different entropy coder for each coeffcient x depending on the $\hat{y}$ obtained from its neighborhood.

Suppose that we have to encode an infinite number of coefficients x and the only information we have about x is an estimate of $|x|$ given by (7). Since the y_i's are quantized they a take a finite number of values and so $\hat{y}$ can only take a finite number of values. It would thus be conceivable to have as many contexts as different values for $\hat{y}$ and to have different entropy coders for each case. However, in a practical scenario, the number of different values that $\hat{y}$ takes might be too large and preclude usage of that many entropy coders. In addition, since the number of inputs x will be finite, there will not be in general enough data to train the entropy coders and we will be faced with a phenomenon called context dilution.

Thus a partition of the interval $[0, \infty)$ is needed for practical reasons. Define the following partition:

$$[0, \infty) = [q_n, q_{n-1}) \cup [q_{n-1}, q_{n-2}) \cup \ldots \cup [q_1, \infty). \quad (8)$$

We assign the pixel x to context Q_i iff $\hat{y} \in [q_i, q_{i-1})$ where the indices decrease from zero to infinity. The question is then how to partition the interval $[0, \infty)$. Note that this is essentially a quantization problem but that the optimal partition need not be the one that minimizes the expected error between the partition and $\hat{y}$. Our goal is to minimize the rate needed to encode the predicted data.

Let $\mathcal{Q}\{\hat{y}\}$ denote the quantized value of $\hat{y}$. This quantization will be efficient if it is such that it minimizes the entropy $H(x|\mathcal{Q}\{\hat{y}\})$. This entropy minimization is equivalent to the maximization of the mutual information $I(x, \mathcal{Q}\{\hat{y}\})$ between x and $\mathcal{Q}\{\hat{y}\}$, or the maximization of the Kullback-Leibler distance $\mathcal{D}(p_{x,\mathcal{Q}\{\hat{y}\}}||p_{\mathcal{Q}\{\hat{y}\}}p_x)$ [5].

So our objective is to make $\mathcal{Q}\{\hat{y}\}$ carry as much information about x as possible. Because of equation (7) we can claim that $\hat{y}$ only carries information about the absolute value of x and no information about the sign. Maximization of the mutual information can be quite involved even if analytic forms of the joint probability density are known so instead we follow a different approach which will not necessarily minimize entropy but at least will minimize the energy of the error. We try to minimize

$$\mathcal{E}(||x| - \mathcal{Q}\{\hat{y}\}|^2) \quad (9)$$

which will lead us to an approximate solution close to the optimal. This approximation is necessary in order to provide an analytical solution, and does not deviate much from the optimal solution, the maximum mutual information (which in ∞ in the case of continuous variables) corresponds to the minimum error (zero) and vice versa.

Minimization of (9) corresponds to the optimal quantizer for $\hat{y}$, given that $\hat{y}$ is an estimate of x. It has to be noted that minimization of (9) is not the same as minimization of $\mathcal{E}((|x| - \mathcal{Q}\{x\})^2)$, but they both have as optimal solution the Lloyd-Max quantizer (LMQ). In our case we only need to make the assumption that the quantization error $\hat{y} - \mathcal{Q}\{\hat{y}\}$ is orthogonal to $|x|$. Under this assumption (9) takes the form:

$$\mathcal{E}((|x| - \mathcal{Q}\{\hat{y}\})^2) = \mathcal{E}((|x| - \hat{y})^2) + \mathcal{E}(|\hat{y} - \mathcal{Q}\{\hat{y}\}|^2) \quad (10)$$

and it can be seen that the optimal choice of $\mathcal{Q}(\hat{y})$ corresponds to the LMQ for $\hat{y}$. We can now model $\hat{y}$ as an exponential random variable, as is consistent with our previously selected model for x. Note that one could explicitly send this parameter but for simplicity in our experiments we choose to assume that $\hat{y}$ has the same variance as x. We observed in our experiments that in general the distributions of $|x|$ and $\hat{y}$ were close. Recall that the Laplacian parameter for each subband is sent to the decoder as side information so that the decoder can calculate the classification thresholds as well.

Sullivan [4] gives a closed form solution for ECSQ for exponential random variables based on the Lambert W function. The above objective function is minimized by the LMQ which can be seen as a particular case of ECSQ. Exponential distributions are memoryless and thus if we have an optimal partition in n intervals we can move to $n+1$ intervals by just adding one more point q_{n+1} and shifting the other points appropriately. So essentially if we have an optimal partition for an exponential distribution of mean one, using N bins we can construct any optimal partition for $n \leq N$ for all exponential distributions. The recursive relation for the optimal partition is:

$$\alpha_{n+1} = v_n + W(-v_n e^{-v_n}) \quad (11)$$

Where:

$$\delta(\alpha) = 1 - \frac{\alpha e^{-\alpha}}{1 - e^{-\alpha}} \quad (12)$$

and $v_n = 1 + \delta(\alpha_n)$, W is the Lambert function which is given in series form[4], and $\alpha_n = \lambda q_n$. The α_n are the lengths of the intervals in our quantizer as in figure (2). More details can be found in [4].

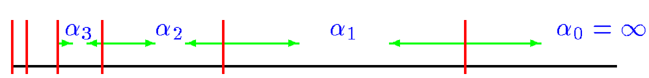

Figure 2. The interval lengths in ECSQ

The optimal number of classes, n, depends on the number of coefficients to be encoded, since excessive number of classes results in context dilution as mentioned earlier. The parameter λ depends only on the statistics of our image and we thus have a simple way of selecting contexts for a given subband based on its size (selection of n) as well as its statistics (selection of λ).

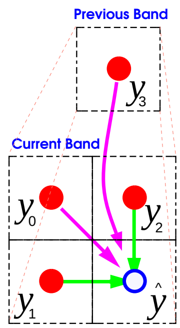

Figure 3. Prediction

We introduce a slight modification of the above classification to take into account the fact that the coefficients y_i are all quantized, i.e. our classification is based on quantized data. For this reason, and in particular at low rates, a very significant fraction of contexts (as high as half the total) will be such that $\hat{y} = 0$. Thus the distribution of $\hat{y}$ is better modeled by a mixture of a continuous and a discrete distribution as in:

$$f_{\hat{y}}(\rho) = \beta\delta(\rho) + (1-\beta)\lambda e^{-\lambda\rho} \tag{13}$$

For pixels x for which $\hat{y} = 0$ a different coder is selected while the classification rules described earlier are applied to those coefficients such that $\hat{y} > 0$. When $\hat{y} = 0$ our classification is based on a larger neighborhood than that of Fig. 3. This modification has proved to be very useful especially at low rates.

2.3 Entropy Coding

Whenever a new band is visited we transmit explicitly the *minimum*, *maximum* and *variance* in this band. Note that this is also equivalent to sending the Laplacian parameter and it allows the decoder to initialize the reproduction levels of the UTQ as well as the classification thresholds. The issue of initialization of the entropy coders is non trivial since different images have different characteristics and explicit initialization may require a significant amount of side information. It is useful to observe that on the top levels of our decomposition the sample distribution is almost uniform, but as we move towards the bottom levels this distribution gets more and more biased. Thus we can use the same look up tables for the entropy coders throughout the whole pyramidal structure. Starting with a uniform distribution on the top level, the distributions "learnt" at a higher level are used to initialize the distribution at lower levels. Thus the statistics are learnt on the fly as we move towards to bottom of our pyramid and no initialization is required for each subband.

3. Description of the Algorithm

The proposed algorithm can be summarized as follows:
Step 1 Compute the wavelet transform for the whole image
Step 2 Apply a UTQ with deadzone with constant step size to all coefficients in all bands. Only the reconstruction levels are different and are set given the Laplacian parameter of each band.
Step 3 Initialize all entropy coders to a uniform distribution.
Step 4 Start scanning all the bands from the low to high resolution in a predetermined order.
Step 5 When a band is first visited send the *maximum*, *minimum* and *variance* of its unquantized coefficients.
Step 6 For each new coefficient define $\hat{y}$ as in equation (7) and decide which entropy coder to use based on $\mathcal{Q}\{\hat{y}\}$.
Step 7 Transmit the codeword closest to x with chosen entropy coder.
Step 8 Continue until all the coefficients have been scanned.

Notice that the whole algorithm is very simple, as no training is required and simple scalar quantizers are used. Classification rules are simple, as is the method to obtain the classification thresholds. The bulk of the complexity comes from computing the wavelet transform rather than from the quantization itself.

In our experiments we used a modified version of the algorithm to take into account both rate and distortion criteria. In the above description all the coefficients were scalar quantized (Step 7), i.e. a given coefficient was assigned to the reproduction level within its bin. Improved performance can be achieved if the quantization is entropy constrained, i.e. we now choose to assign a quantization level to a coefficient if that level minimizes $J = D + \mu R$, where $\mu \geq 0$ is the Lagrange multiplier, R is the rate needed to send that level (based on current statistics of the arithmetic coder corresponding to that coefficient) and D is the distortion for each level. This involves some additional complexity but can be done efficiently since we have an initial guess for the codeword (minimize D) and even a suboptimal solution for the minimization of $J = D + \mu R$ is acceptable. In practice full search versus restricted search gave very small difference in performance if any. The results in the next section include this modification.

4. Experimental Results and Conclusions

Experimental results have shown that linear phase odd-length biorthogonal filters offer advantages in terms of energy compaction and thus compression [8]. We use a 11-13 biorthogonal 2 channel filter bank[8], with a simple modification such that for all the four filters in the filter bank we have $|h_i| = |g_i| = const$. This allow us to use standard bit allocation techniques used in orthogonal filter banks. As far

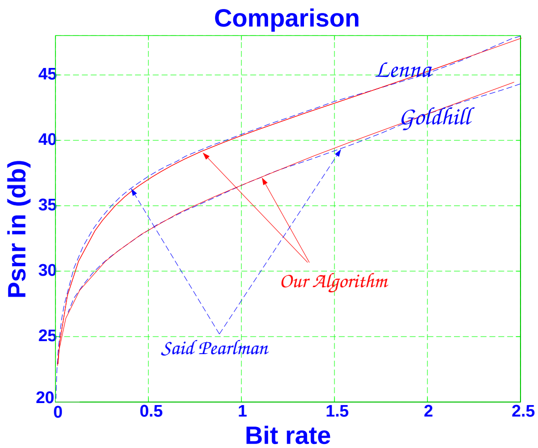

Figure 4. Rate Distortion Curves for the 512×512 Lenna and Goldhill images. Comparison between our algorithm and [2]

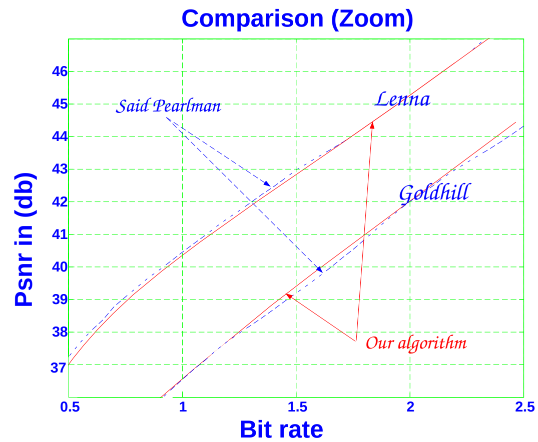

Figure 5. Rate Distortion Curves for the 512×512 Lenna and Goldhill images. Comparison between our algorithm and [2]

as the bit allocation is concerned this filter bank is equivalent to an orthogonal filter bank (see [7] for details.) Bit allocation is required to determine the average number of bits to be used in each band. In figure Fig. 4 we present results on the R-D curves for the lenna and Goldhill images both of size 512x512. We used 11 classes (including the special class for $\hat{y} = 0$).

We compare our performance with one of the best algorithms [2]. In one case our algorithm performs better on the average over different bit rates, while for the other our results are slightly worse. However, for this latter image no algorithm outperforms the other for the whole range of bit rates. Fig. 5 shows a comparison of the two algorithms over a narrower range. The small differences justify our selections and theoretical formulations in the previous sections. We also performed experiments using different classification mechanisms for $\hat{y}$ with our method showing better performance. For example when trying to have equal probability in each bin Q_i as opposed to our previous formulation we got about $0.15db$ less in PSNR for Goldhill at $0.5b/p$ (33.00db versus 33.15db).

Potential benefits of this method compared to the one at [2] are, speed and the fact that there are no tree structures involved so that all the operations can be done sequentially. However our system is not embedded while the one in [2] is. We also verified that our algorithm tended to work better at high rates and indeed could be modified to provide a range of bit rates extending all the way to lossless compression.

References

[1] R. L. Joshi, H. Jafarkhani, T. R. Fisher, N. Farvadin, M. W. Marcellin, and R. H. Bamberger. Comparison of Different Methods of Classification in Subband Image Coding. *Submitted to IEEE Trans. in Image Processing.*, 1995.

[2] A. Said and W. Pearlman. A New Fast and Efficient Image Coder Based on Set Partitioning on Hierarchical Trees. *IEEE Trans. Circuits and Systems for Video Technology*, 6(3):243–250, June 1996.

[3] J. M. Shapiro. Embedded Image Coding Using Zerotrees of Wavelet Coefficients. *IEEE Trans. Signal Processing*, 41(12):3445–3462, December 1993.

[4] G. J. Sullivan. Efficient Scalar Quantization of Exponential and Laplacian Random Variables. *IEEE Trans. Information Theory*, 42(5):1365–1374, September 1996.

[5] T. M. Cover and J. A. Thomas . *Elements of information theory*. Wiley Series in Communications, 1991.

[6] D. Taubman and A. Zakhor. Multirate 3-D Subband Coding of Video. *IEEE Trans. Image Processing*, 3(5):572–588, Sept. 1994.

[7] B. Usevitch. Optimal Bit Allocation for Biorthogonal Wavelet Coding. In *DCC, Data Compression Conference*,, pages 387–395, Snowbird, Utah, March 31 -April 3 1996.

[8] J. Villasenor, B. Belzer, and J. Liao. Wavelet Filter Evaluation for Image Compression. *IEEE Trans. Image Processing*, 2:1053–1060, August 1995.

[9] M. J. Weinberger, J. J. Rissanen, and R. B. Arps. Applications of Universal Context Modeling to Lossless Compression of Gray-Scale Images. *IEEE Trans. Image Processing*, 5(4):575–586, Apr. 1996.

[10] Y. Yoo, A. Ortega, and B. Yu. Adaptive Quantization of Image Subbands with Efficient Overhead Rate Selection. In *Proc. of the Intl. Conf. on Image Proc., ICIP96*, volume 2, pages 361–364, Lausanne, Switzerland, Sept. 1996.

[11] Z. Xiong and K. Ramchandran and M. T. Orchard. Wavelet Packet Image Coding Using Space-frequency Quantization. *IEEE Trans. Image Processing*, Submitted, 1996.