Efficient Context-Based Entropy Coding for Lossy Wavelet Image Compression

Christos Chrysafis and Antonio Ortega*
Integrated Media Systems Center
University of Southern California
Los Angeles, CA 90089-2564
chrysafi,ortega@sipi.usc.edu
Tel: 213-740-2320, Fax: 213-740-4651

Abstract

In this paper we present an adaptive image coding algorithm based on novel backward-adaptive quantization/classification techniques. We use a simple uniform scalar quantizer to quantize the image subbands. Our algorithm puts each coefficient into one of several classes depending on the values of neighboring previously quantized coefficients. These previously quantized coefficients form contexts which are used to characterize the subband data. To each context type corresponds a different probability model and thus each subband coefficient is compressed with an arithmetic coder having the appropriate model depending on that coefficient's neighborhood. We show how the context selection can be driven by rate-distortion criteria, by choosing the contexts in a way that the total distortion for a given bit rate is minimized. Moreover the probability models for each context are initialized/updated in a very efficient way so that practically no overhead information has to be sent to the decoder. Our results are comparable or in some cases better than the recent state of the art, with our algorithm being simpler than most of the published algorithms of comparable performance.

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 making adaptive 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 [1] is not considered here.

*This work was supported in part by the National Science Foundation under grant MIP-9502227 (CAREER) and in part by the Integrated Media Systems Center, a National Science Foundation Engineering Research Center with additional support from the Annenberg Center for Communication at the University of Southern California, and the California Trade and Commerce Agency.

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 [2, 3, 4]. Context based arithmetic coders were used in [3] while in [2, 4] context information was taken into account by using the zero-tree data structure, which enables the joint transmission of zero-valued coefficients 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 [5] 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 [5] relied on pre-analyzing the subband data and sending the class assigned to each block as side information. In [6] 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 quantizer for all the subbands and arithmetic coding of the resulting set of coefficients[1]. Furthermore, as in [6], we use backward adaptive classification to determine which set of probabilities our arithmetic coder will use. Since several different probability models can be used for the quantized coefficients a key issue is that of determining how to assign each coefficient to a probability model. 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 a Lloyd-Max quantizer (LMQ) matched to the distribution of the predictor.

We are thus considering a context-based adaptive arithmetic coder similar to that proposed in [7] 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 [7]. 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.

While this work was in progress the authors became aware of the work in [8] which follows a similar context based adaptation but uses a different approach for the classification. This alternative approach achieves comparable results.

2 Context-based adaptation

The performance of arithmetic coders (see for example [9]) depends on the *coder* itself, which can efficiently generate a bit-stream out of the input symbols and the estimation of the *probability model* which the coder will use. The coder can achieve an average output code length very close to the entropy corresponding to the probability model it utilizes. Thus if the probability model accurately reflects the statistical properties of the input, arithmetic coding will approach the

[1] Note that the size of our alphabet is much larger than in [2, 3, 4]

entropy of the source. Different probability models will give different compression performance for the same data and thus adaptively learning the probability of the data on the fly will in general be better than a non-adaptive scheme, as it will allow a better approximation to the "true" statistics of the data. In this paper our goal is to further improve the performance of the encoder by modeling the source (the data in each subband) as mixture of pdf's where each distribution occurs after a specific context.

The key issue is then how find an efficient context-based classification i.e. how to determine the probabilistic model to use for a coefficient as a function of its neighbors (see Section 2.1). The arithmetic coder uses the model corresponding to each context and performs on the fly adaptation of each model (see Section 2.2).

2.1 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.

$$\hat{x}_n = \mathcal{P}\{x_n, x_{n-1}, \ldots, x_0\} \tag{1}$$

Many experiments have shown that traditional linear prediction 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 [7] in a lossless image coding scenario. In this work we use a 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 [6], 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:

$$\hat{y} = \sum_{i=0}^{N} a_i |y_i| \tag{2}$$

where the y_i's can be seen in Fig. 1. We chose $N = 13$ in our experiments. Now, based on the value of $\hat{y}$, we can select a different probability model for each coefficient 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 (2). Since the y_i's are quantized they 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 probability models 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 probability

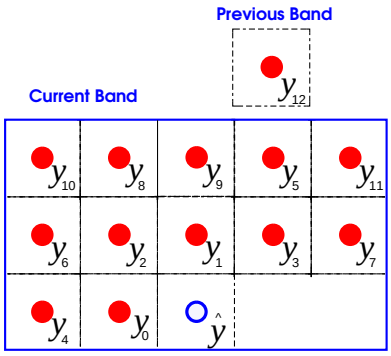

Figure 1: Context used for the evaluation of $\hat{y}$.

models. In addition, since the number of inputs x will be finite, there will not be in general enough data to train the models and we will be faced with a phenomenon called context dilution.

We need to classify $\hat{y}$ in one of a finite and small enough number of classes or equivalently to partition the interval $[0, \infty)$. Define the following partition:

$$[0, \infty) = [q_n, q_{n-1}) \cup [q_{n-1}, q_{n-2}) \cup \ldots \cup [q_1, \infty). \tag{3}$$

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[2] 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 $\mathcal{D}(p_{x,\mathcal{Q}\{\hat{y}\}}||p_{\mathcal{Q}\{\hat{y}\}}p_x)$ [10].

Thus our objective is to make $\mathcal{Q}\{\hat{y}\}$ carry as much information about x as possible. Because of equation (2) 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

$$\mathcal{E}(||x| - \mathcal{Q}\{\hat{y}\}|^2) \tag{4}$$

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

[2] An other way to look at it is as a grouping of several possible values of $\hat{y}$ into a single context.

solution, the maximum mutual information (which is ∞ in the case of continuous variables) corresponds to the minimum error (zero) and vice versa.

Minimization of (4) 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 (4) 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 $|x| - \hat{y}$. Under this assumption (4) takes the form:

$$\mathcal{E}((|x| - \mathcal{Q}\{\hat{y}\})^2) = \mathcal{E}((|x| - \hat{y})^2) + \mathcal{E}(|\hat{y} - \mathcal{Q}\{\hat{y}\}|^2) \tag{5}$$

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 since subband data have been observed to have Laplacian distribution.

Sullivan [11] 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:

$$\alpha_{n+1} = v_n + W(-v_n e^{-v_n}), \tag{6}$$

where:

$$\delta(\alpha) = 1 - \frac{\alpha e^{-\alpha}}{1 - e^{-\alpha}}, \tag{7}$$

and $v_n = 1 + \delta(\alpha_n)$. W is the Lambert function which is given in series form[11], 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 [11].

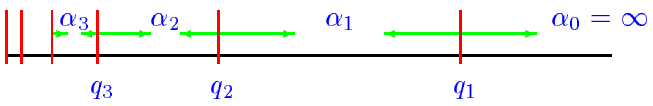

Figure 2: The interval lengths in ECSQ

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).

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 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:

$$f_{\hat{y}}(\rho) = \beta\delta(\rho) + (1-\beta)\lambda e^{-\lambda\rho} \tag{8}$$

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 $\hat{y} > 0$. This modification has proved to be very useful, especially at low rates.

Another issue in this quantization scheme is the selection of the normalization factor $1/\lambda$ (parameter of the exponential). We need to multiply all the quantization step sizes by $1/\lambda$ in order to fit the normalized step sizes to the actual distribution of the data. In our experiments we select this parameter to match the overall distribution of the combined subband data. Our experiments indicate that this is a reasonable choice.

2.2 Entropy Coding

Whenever a new band is visited we transmit explicitly the *minimum* and *maximum* in this band. 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 uniform quantizer (constant step size) to all coefficients in all bands apart from the lower frequency band[3].

Step 3 Initialize all probability models to a uniform distribution.

Step 3 Start scanning all the bands from the low to high resolution in a predetermined order.

Step 4 When a band is first visited send the *maximum* and *minimum* of its quantized coefficients.

Step 5 For each new coefficient define $\hat{y}$ as in equation (2) and decide which entropy coder to use based on $\mathcal{Q}\{\hat{y}\}$. Choose context Q_k iff $\hat{y} \in [q_n, q_{k-1})$ or Q_0 iff $\hat{y} = 0$[4].

Step 6 Using the entropy coder chosen in the previous step, transmit the codeword closest to x in the rate distortion sense.

Step 7 Continue until all the coefficients have been scanned.

[3] For the lower resolution band we use PCM with no entropy coding.

[4] For the first 3 bands we do not have a corresponding coarser resolution band. Thus y_{12} in Fig. 1 would not be available and so we use context structure different from that of Fig. 1.

Notice that the whole algorithm is simple, as no explicit 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 (*Step 6*) we use a rate distortion criterion in order to select the optimal codeword. We 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 estimate of the probability model 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 (the choice that minimizes D) and even a suboptimal solution for the minimization of $J = D + \mu R$ is acceptable. We can thus restrict the search to codewords close to the minimum distortion codeword. In practice there was little difference in performance between this restricted search and the optimal full search. Note that the rate distortion formulation makes our encoder/decoder asymmetric, i.e. all the search complexity is at the encoder while the decoder only needs to use the reproduction level chosen by the encoder.

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 [12]. We use the 23-25 Daubechies biorthogonal 2 channel filter bank, with a simple modification such that for all the four filters in the filter bank we have $\|\mathbf{h}_0\| = \|\mathbf{h}_1\| = \|\mathbf{g}_0\| = \|\mathbf{g}_1\|$. This allow us to use the same quantization step size for all the subbands in the decomposition. The same condition was derived in [13]. As far as the bit allocation is concerned this filter bank is equivalent to an orthogonal filter bank (see [13] for details.) In our algorithm there is no explicit bit allocation involved, but implicitly we do perform bit allocation by selecting the quantization step size. Since this size is the same for the whole image, we need to have the appropriate scaling in all the coefficients in order to get the maximum possible gain in compression. The above normalization of the filters does exactly that.

In figure Fig. 3 we present results on the R-D curves for the Lenna, Goldhill and Barbara images both of size 512×512. We used 12 classes (including the special class for $\hat{y} = 0$). We compare our performance with that of the Said and Pearlman algorithm [4] which has similar complexity to our scheme. On the average our algorithm outperforms [4] for most of the images at various bit rates. In tables 1, 2, 3 we see results for our method and the algorithms in [2, 4, 14, 1], in table 4 we see how our algorithm performs with different filter choices. Note that [4, 14, 1] are using the $7 - 9$ Daubechies biorthogonal filter bank and [2] is using the 9-tap symmetric QMF filter bank.

Our results seem to justify some of the selections and theoretical formulations in the previous sections. We plan to do a more extensive study of these issues. We performed experiments using different classification mechanisms for $\hat{y}$ with our method showing better performance. For example, if we used a constant probability classification method (i.e. one where each bin Q_i has the same probability) instead of the previously introduced approach based on Lloyd-Max Quantization our results where worse. For example a $0.27dB$ loss in PSNR for Goldhill at $0.5b/p$

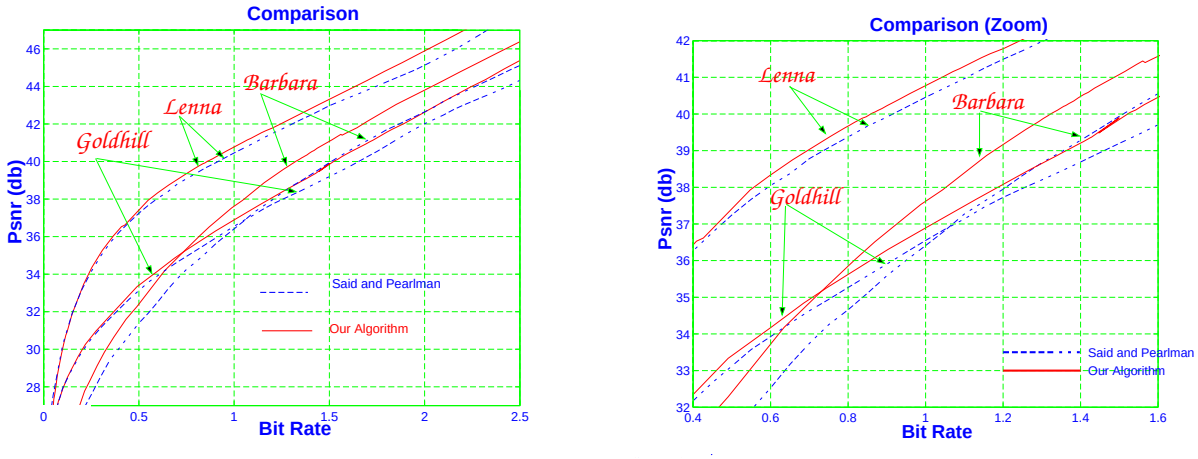

Figure 3: Rate Distortion curves for the 512×512 Lenna, Goldhill and Barbara images. Comparison between our algorithm and the algorithm in [4]

(33.14db versus 33.41db).

Potential benefits of this method compared to the one at [15, 5, 1] are, speed/simplicity 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 [4] 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. At higher rates $\hat{y}$ will be a better estimate of $|x|$ so we will do much better in the classification of the next coefficient. Also by choosing fine enough quantization we expect to be able to use our algorithm for lossless compression.

Rate	EZW [2]	SPIHT[4]	SFQ[14]	WP&SFQ[1]	C/B
0.20 b/p	-	26.64	26.26	27.22	**27.33**
0.25 b/p	26.77	27.57	27.2	28.41	**28.48**
0.50 b/p	30.53	31.39	31.33	**32.63**	32.37
1.00 b/p	35.14	36.41	36.96	**37.69**	37.61

Table 1: Comparison between our method (C/B) and [2, 4, 14, 1] for image Barbara 512×512, (There is no standard *Barbara* image the one used here is the one in [2, 14] the first order entropy of the image was 7.632b/p)

Rate	EZW [2]	SPIHT[4]	SFQ[14]	C/B
0.20 b/p	-	33.16	**33.32**	33.24
0.25 b/p	33.17	34.13	**34.33**	34.31
0.50 b/p	36.28	37.24	37.36	**37.52**
1.00 b/p	39.55	40.45	40.52	**40.80**

Table 2: Comparison between our method (C/B) and [2, 4, 14] for image Lenna 512×512.

Rate	EZW [2]	SPIHT[4]	SFQ[14]	C/B
0.20 b/p	-	29.84	29.86	**29.94**
0.25 b/p	-	30.55	**30.71**	30.67
0.50 b/p	-	33.12	33.37	**33.41**
1.00 b/p	-	36.54	36.70	**36.90**

Table 3: Comparison between our method (C/B) and [4, 14] for image Goldhill 512×512.

References

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

[2] J. M. Shapiro, "Embedded Image Coding Using Zerotrees of Wavelet Coefficients.," *IEEE Trans. Signal Processing*, vol. 41, pp. 3445–3462, December 1993.

[3] D. Taubman and A. Zakhor, "Multirate 3-D Subband Coding of Video," *IEEE Trans. Image Processing*, vol. 3, pp. 572–588, Sept. 1994.

[4] 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*, vol. 6, pp. 243–250, June 1996.

Filter Bank	Bit rate	Psnr
Daubechies 23-25	1.00	**40.80**
Daubechies 21-23	1.00	40.05
Daubechies 19-21	1.00	40.78
Daubechies 11-13	1.00	40.62
Daubechies 7-9	1.00	40.15

Table 4: Comparison of the effect of different filters in the performance of our algorithm. The image used here is Lenna of dimensions 512×512.

[5] R. L. Joshi, V. J. Crump, and T. R. Fisher, "Image Subband Coding Using Arithmetic Coded Trellis Coded Quantization," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 5, pp. 515–523, December 1995.

[6] 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*, vol. 2, (Lausanne, Switzerland), pp. 361–364, Sept. 1996.

[7] 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*, vol. 5, pp. 575–586, Apr. 1996.

[8] S. M. LoPresto, K. Ramchadran, and M. T. Orchard, "Image Coding based on Mixture Modeling of Wavelet Coefficients and a Fast Estimation-Quantization Framework.," in *DCC, Data Compression Conference,*, (Snowbird, Utah), March 25 - March 27 1997.

[9] R. Witten, I. Neal, and J. Cleary, ""Arithmetic Coding for Data Compression"," *Communications of the ACM*, vol. 30, pp. 520–540, June 1987.

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

[11] G. J. Sullivan, "Efficient Scalar Quantization of Exponential and Laplacian Random Variables," *IEEE Trans. Information Theory*, vol. 42, pp. 1365–1374, September 1996.

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

[13] B. Usevitch, "Optimal Bit Allocation for Biorthogonal Wavelet Coding," in *DCC, Data Compression Conference,*, (Snowbird, Utah), pp. 387–395, March 31 -April 3 1996.

[14] Z. Xiong and K. Ramchandran and M. T. Orchard, "Space-frequency Quantization for Wavelet Image Coding.," *IEEE Trans. Image Processing*, vol. Submitted, 1997.

[15] 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.