LOW BIT-RATE TRANSFORM AND WAVELET IMAGE CODING

by

Parthasarathy Sriram

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY
WITH A MAJOR IN ELECTRICAL ENGINEERING

In the Graduate College

THE UNIVERSITY OF ARIZONA

1 9 9 3

INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

U·M·I

University Microfilms International
A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
313/761-4700 800/521-0600

Order Number 9408397

Low bit-rate transform and wavelet image coding

Sriram, Parthasarathy, Ph.D.

The University of Arizona, 1993

U·M·I
300 N. Zeeb Rd.
Ann Arbor, MI 48106

LOW BIT-RATE TRANSFORM AND WAVELET IMAGE CODING

by

Parthasarathy Sriram

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY
WITH A MAJOR IN ELECTRICAL ENGINEERING

In the Graduate College

THE UNIVERSITY OF ARIZONA

1 9 9 3

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Final Examination Committee, we certify that we have read the dissertation prepared by Parthasarathy Sriram

entitled Low bit-rate transform and wavelet image coding

and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy

Dr. Michael W. Marcellin	Date 9/24/93
Dr. Robin N. Strickland	Date 9-24-93
Dr. Jeffrey J. Rodriguez	Date 9-24-93
	Date
	Date

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copy of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

Dissertation Director Dr. Michael W. Marcellin	Date 9/24/93

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotation from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: P. Sriram

ACKNOWLEDGMENTS

Sometimes I think myself as the luckiest guy in the world for personally knowing so many people who have had a positive influence on me. It is almost impossible to name and thank everyone of them individually. In particular, my special thanks to Mr. M. B. Kannan, Mr. K. Divakar, and Dr. R. Ramanujam for showing me what is takes to succeed in life.

I would like to thank my advisor, Dr. Marcellin, for all his help, encouragement, and inspiration in the last four years. My research work has sometimes been very difficult and if not for his contribution, my graduation would still be a dream. I would also like to thank Dr. Robin Strickland and Dr. Jeffrey Rodriguez for being kind enough to review my dissertation and for all their helpful suggestions.

Finally, it is my pleasure to acknowledge the invaluable role played by my colleagues Edward Orcutt and James Kasner and my friend Mariappan Nadar. Their friendship, patience and moral support enabled me to comfortably withstand some difficult times.

TABLE OF CONTENTS

TABLE OF CONTENTS–Continued

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

In this dissertation, adaptive wavelet and transform coding techniques are presented for low bit-rate monochrome and color image coding. The proposed encoders are based on trellis coded quantization. Trellis coded quantization (TCQ) is an effective scheme for quantizing memoryless sources with low to moderate complexity. The TCQ approach to data compression has led to some of the most effective source codes found to date for memoryless sources.

For the transform coder, TCQ is used to encode transform coefficients resulting from applying a 16×16 discrete cosine transform (DCT) to 8-bit gray level and 24-bit color images. For the color images, the red, green, and the blue planes were transformed into NTSC transmission primaries (Y, I, and Q) before the DCT is applied. Both fixed-rate and entropy-constrained systems are considered.

The discrete wavelet transform has recently emerged as a powerful technique for decomposing images into various multi-resolution approximations. We investigate the use of entropy-constrained trellis coded quantization for encoding the wavelet coefficients of both monochrome and color images. The lowest resolution sub-image is encoded using a 4×4 2-D DCT encoder. An integer programming algorithm is employed to allocate the available bit-rate optimally among the subbands.

The objective performance results of our wavelet and transform coders are comparable to or surpass all previous results reported in the literature. The subjective quality of the encoded images is also excellent. In particular, the encoded monochrome images at 0.5 bits/pixel (a compression ratio of 16:1) obtained using our adaptive wavelet coder is almost indistinguishable from the original even when viewed on a high-resolution monitor.

CHAPTER 1

INTRODUCTION

Recent advances in multimedia technology have magnified the demand for efficient image transmission and storage. A digital representation of an image requires a very large amount of data. The goal of image data compression (or image coding) is to minimize the number of information carrying units to represent an image. The mere process of sampling band-limited images where an infinite number of image points per unit sampling area is reduced to a single image sample (or a pixel) without any loss of information can be considered as a simple form of data compression. In digital image processing, each image pixel is also quantized to a fixed but sufficient number of levels such as 256 (8 bits/pixel (bpp)). To store a digital image of size 512×512 pixels with a resolution of 8 bpp requires a memory of 2 Mbits. The basic intention of image data compression is to reduce this memory so that images can be transmitted or stored more efficiently.

The block diagram of a typical digital image communication system is shown in Figure 1.1. An image coder attempts to represent an image with as few bits as possible while preserving the fidelity required for that particular application. Output from an image coder is a bit sequence which represents the input image. The channel coder accepts this bit sequence and modulates it into a signal more suited

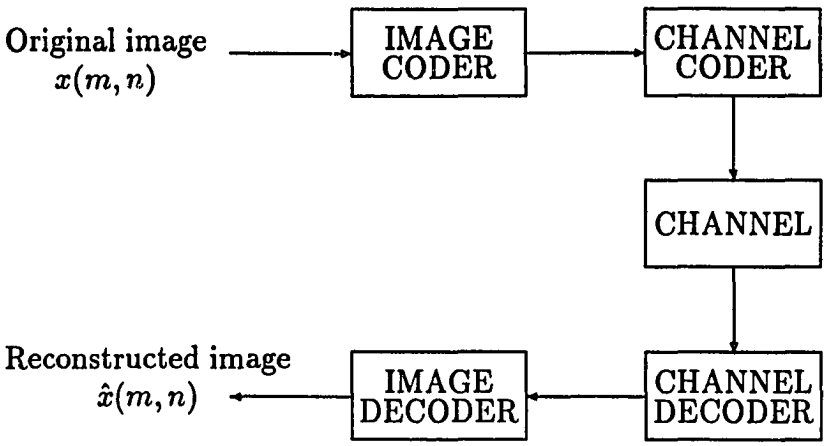

Figure 1.1: Block diagram of a digital image communication system.

for transmission over a analog channel. An analog channel typically introduces some noise and hence, the channel coder adds controlled redundancy for error correction. This modulated signal is converted into a bit stream by the channel decoder. The image decoder inputs this bit sequence and outputs an estimate of the original image. For storage applications, output from the image coder is converted to a format more suitable for storage in a recording medium. In this work, we are primarily concerned about designing image coders and decoders (a noiseless channel is assumed throughout this dissertation).

The fidelity required for the reconstructed imagery varies from application to application. For example, for storing medical images, it may be necessary to preserve all information in the original image. Image coders which do not destroy any information and which allow perfect reconstruction of the original image are called *lossless* image coders. On the other hand, for applications such as digital video, it

is not necessary for an image coder to be lossless. Some amount of distortion can be tolerated as long as it is acceptable for the human viewer. These image coders are commonly referred to as *lossy* image coders. Apparently, the more distortion one can tolerate, the better will be the data compression. The basic intention of a lossy image coder, given a distortion measure, is to minimize distortion for a given encoding rate, or equivalently, for a given distortion, minimize the required encoding rate.

Image coders can also be classified in the following way.

1. Spatial domain coders
2. Frequency domain coders
3. Feature-based coders.

Spatial domain coders:

1. Pulse code modulation (PCM). These coders encode image pixel intensities directly. An acceptable level of image quality can be obtained for most applications at an encoding rate of 4–5 bpp [1]. These coders, in general, do not exploit any redundancy between neighboring samples.
2. Differential pulse code modulation (DPCM). It is well known that neighboring image pixels are highly correlated. DPCM image coders exploit this redundancy by making a prediction of the current pixel intensity using the pixel intensities of previously encoded neighboring samples and

then quantizing the difference between the pixel intensity and its prediction. Consequently, only the new information (residual) is quantized and transmitted. DPCM coders can be made to adapt to the local variations of the image by using adaptive predictors and quantizers. Adaptive DPCM coders are capable of achieving high quality encoding at around 3.4–3.8 bpp [1].

Frequency domain coders:

1. Subband coders. Subband coders decompose an input image into a number of frequency components and encode each of these components separately. This decomposition removes the redundancy in the input and provides a set of uncorrelated components to encode. These coders produce high quality encoding at around 2 bpp.

2. Transform coders. These coders transform an image (or a image block) using an energy preserving transformation to another domain such that maximum information is packed into a smaller number of samples and where the samples are less correlated. Transform coders are capable of producing extremely high quality images at rates about 1 bit/pixel [2],[3].

Feature-based coders decompose an image into a collection of physical entities such as regions or contours. These contours are then encoded according to their

information content. Feature-based coders utilize properties of the Human Visual System (HVS) and are able to achieve excellent quality encoding even at very low rates [4], although at the expense of very high computational burden.

It is a common practice to combine some of these techniques to produce high quality encoding. For example, Woods and O'Neil [5] used 2D-DPCM to encode image subbands, Tanabe and Farvardin [6] investigated the use of both DPCM and the discrete cosine transform (DCT) to encode one of the image subbands in their subband coding system, and so on. Some of these hybrid coders are capable of producing high quality imagery at rates between 0.4 to 0.8 bpp (a compression ratio of 20:1 to 10:1, assuming 8 bpp for the original). Recently, an international image coding standard was proposed by Joint Photographic Experts Group (JPEG) [2] and achieves superior quality encoding for most images at encoding rates greater than or equal to 1 bpp. If the encoding rate of the JPEG coder is reduced below 1 bpp, the quality of the reconstructed imagery drops significantly. Image coding at rates less than 1 bpp is an ongoing research topic and the material presented in this dissertation is directed towards the development of image coders at 1 bpp and below.

In Chapter 2, trellis coded quantization (TCQ) and entropy-constrained TCQ (ECTCQ) are briefly reviewed. TCQ was recently introduced as an efficient scheme for quantizing memoryless sources with low to moderate complexity. Near optimal performance (in a rate-distortion theory sense) for encoding memoryless

sources can be achieved (by using ECTCQ) at all non-negative encoding rates. TCQ is incorporated into a discrete cosine transform coding structure for encoding monochrome and color images in Chapter 3. Chapter 4 investigates the use of entropy-constrained TCQ with the discrete wavelet transform for encoding images. Finally, in Chapter 5, the results are summarized.

CHAPTER 2

TRELLIS CODED QUANTIZATION

2.1 Background

A quantizer is a mapping Q with domain the k-dimensional Euclidean space $\mathbf{R}^k$ and range Y, where $Y = \{y_1, y_2, \cdots, y_L\}$ is a finite subset of $\mathbf{R}^k$ with L elements [7]. Associated with every L-point quantizer in $\mathbf{R}^k$ is a partition $\{P_1, P_2, \cdots, P_L\}$ such that

$$\mathbf{R}^k = \bigcup_{i=1}^{L} P_i \ , \quad P_i \bigcap_{i \neq j} P_j = \phi \tag{2.1}$$

where $P_i = \{x \in \mathbf{R}^k \ ; \ Q(x) = y_i\}$. Hence, a quantizer is uniquely specified by the output set Y and the corresponding partition $\{P_i; i = 1, 2, \cdots, L\}$.

A scalar quantizer is just a one-dimensional quantizer ($k = 1$). Perhaps the simplest scalar quantizer is the uniform scalar quantizer. The input-output characteristic of a uniform mid-rise quantizer is shown in Figure 2.1. Such a quantizer can be described by the following equations [8],

$$\begin{aligned} x_k &= [k - \tfrac{L+2}{2}]\Delta, \quad k = 2, 3, ..., L \\ y_k &= [k - \tfrac{L+1}{2}]\Delta, \quad k = 1, 2, ..., L \end{aligned}$$

$$x_1 = -x_{L+1} = -\infty \tag{2.2}$$

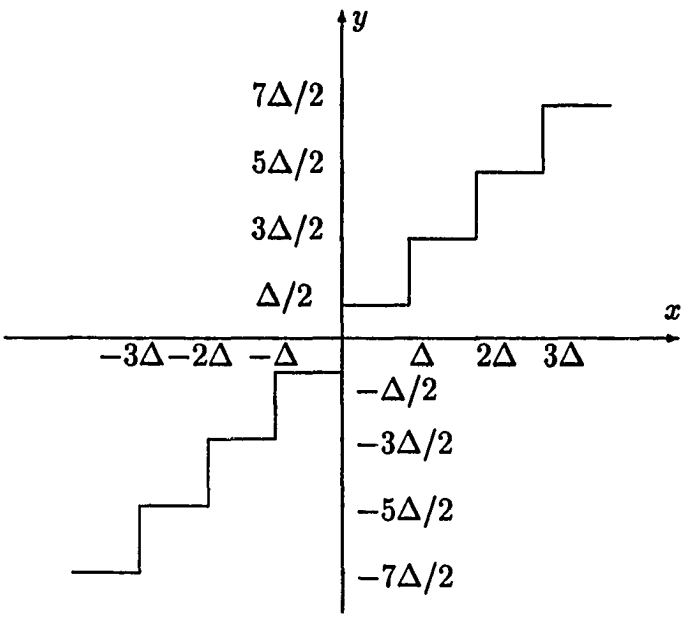

Figure 2.1: Input-Output characteristics of an uniform quantizer.

where the amplitudes x_k are decision levels, y_k are reconstruction levels and Δ is the step size. The partition for this quantizer is given by

$$\{P_1, P_2, \cdots, P_L\} = \{(x_1, x_2), [x_2, x_3), \cdots, [x_L, x_{L+1})\}. \tag{2.3}$$

At each time index, depending on which partition the input sample belongs to, a L-ary number is transmitted to the receiver, typically in binary format. Generally, L is chosen as a power of two. Assuming $L = 2^R$ for some integer R and an equal-length codeword assignment, a bit rate of $R = \log_2 L$ bits/sample is needed to inform the receiver about that index. At the receiver, for an error-free transmission, we have an output

$$\hat{x} = y_k \text{ if } x \in P_k. \tag{2.4}$$

The performance of a quantizer is characterized by its mean-squared error (MSE) performance. Let the input source be modeled as a sequence of realizations of a random variable X with probability density function (pdf) f_X. If $Q = X - Y$ is the quantization error, the MSE (ρ) is given by

$$\rho = \mathrm{E}[Q^2] = \sum_{k=1}^{L} \int_{x_k}^{x_{k+1}} (x - y_k)^2 f_X(x) dx \ . \tag{2.5}$$

The optimum minimum MSE step size Δ_{opt}, for a given input distribution, can be found by minimizing ρ with the constraint of a uniform quantizer. It is customary to report the MSE performance of the quantizer as a signal-to-noise ratio (SNR) which is defined as

$$\mathrm{SNR} = 10 \log_{10}(\sigma_X^2/\rho) \ \mathrm{dB} \tag{2.6}$$

where σ_X^2 is the variance of the input source.

Figure 2.2 shows the distortion-rate function (which gives the minimum possible distortion that can be achieved for a given encoding rate) plotted as SNR in dB along with the performance of the optimal uniform scalar quantizer (OUSQ) for the zero-mean, unit-variance, memoryless Gaussian source [8]. It can be shown that the performance curve for the OUSQ diverges from the distortion-rate function as the encoding rate increases. For input sources with non-uniform pdf's, it is possible to achieve a better performance by having closely spaced partitions in the regions where the pdf is high and by choosing larger quantization intervals in other regions.

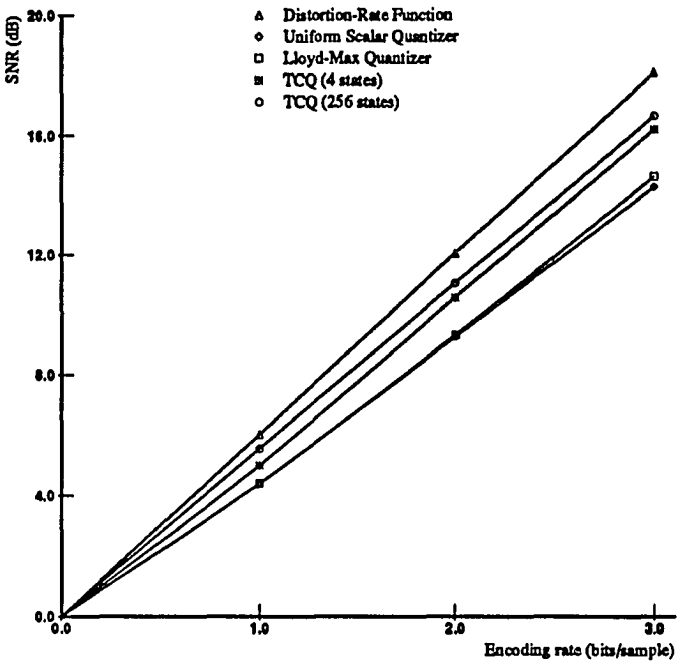

Figure 2.2: Distortion-rate performance for the memoryless Gaussian source.

Optimal non-uniform scalar quantizers can be found by choosing the decision and reconstruction levels such that ρ (as given by Equation 2.5) is minimized for a given rate and input distribution. The necessary conditions for optimality are found by setting the partial derivatives of ρ with respect to the decision and reconstruction levels to zero [8]. The resulting quantizers satisfy the following conditions. (1) The decision levels are at the mid points of their respective reconstruction levels and (2) The reconstruction levels must be the centroid of the pdf in the appropriate interval. Quantizers satisfying these conditions are referred to as Lloyd-Max quantizers (LMQ). If f_X is log-concave, i.e., if $\partial^2 \log f_X(x)/\partial x^2$ is negative, these

conditions are also sufficient [8]. The Laplacian and Gaussian pdf's satisfy the log-concave property and hence, their corresponding Lloyd-Max quantizers are optimal scalar quantizers.

The SNR performance for a memoryless Gaussian source using LMQ is also shown in Figure 2.2. From this figure, one might come to a conclusion that using non-uniform quantization results in only a marginal improvement in performance. However, at higher encoding rates, there is a significant difference in performance. The performance improvement (also known as non-uniform density gain) is as much as 2.6 dB at 7 bits/sample. Even at this high encoding rate, the performance of the LMQ is 4.3 dB away from the distortion-rate function.

The performance of LMQ's can be improved by using variable-length encoding of the quantizer output. Recall from our previous discussion that, at each time index, depending on which partition of the quantizer the input sample belong to, a L-ary number is transmitted to the receiver, typically in binary format. For an equal-length codeword assignment, this requires an encoding rate of $R \geq \log_2 L$ bits/sample. Equal-length codeword assignment, in general, is not the most optimal way (with respect to the average bit rate) of assigning codewords.

A L-point quantizer can be viewed as a discrete source which takes values in $Y = \{y_1, y_2, \cdots, y_L\}$ with probabilities $\{p(y_1), p(y_2), \cdots, p(y_L)\}$ respectively. The

entropy of this source Y is defined as

$$H(Y) = -\sum_{i=1}^{L} p(y_i) \log_2 p(y_i) \text{ bits.}$$

By assigning shorter codewords to more probable outputs and longer codewords to less likely outputs, it is possible to reduce the average encoding rate. It is well known from information theory [9] that $H(Y)$ gives the minimum bound on the average encoding rate for encoding the memoryless source Y. One can come up with variable-length (or entropy) codes whose average codeword length is arbitrarily close to the entropy by using Huffman's technique [9] or Arithmetic coding [10].

Entropy-constrained scalar quantizers (ECSQ) are designed such that the average encoding distortion is minimized, for a given input source, with the constraint that the entropy associated with the quantizer output is below a certain value. That is, the MSE distortion ρ as given by Equation 2.5 is minimized with the constraint that

$$H(Y) \leq H_0. \tag{2.7}$$

Lagrange multiplier techniques can be used to solve the problem of (2.5) and (2.7). Gish and Pierce [11] showed that, for high encoding rates, the optimum ECSQ is an entropy-coded uniform quantizer (ECUQ) regardless of the source pdf. However, at low encoding rates, there could be a large difference in performance between optimal ECSQ and ECUQ [12]. This difference in performance is also a function of the number of quantization levels L and the source distribution. For example, at an

encoding rate of 1 bit/sample, the difference in performance of optimal ECSQ and ECUQ is only 0.24 dB for the memoryless Gaussian source while it is as much as 2.75 dB for the Laplacian source. When the constraint on the number of quantization levels is removed, the resulting ECSQ is known to provide an upper bound to the distortion-rate performance of all memoryless, scalar quantization schemes [12]. For encoding memoryless sources, the performance of ECSQ is within 1.53 dB of the distortion-rate function at all non-negative encoding rates.

The reason why LMQ's fail to achieve a performance close to the distortion-rate function can be easily explained. Let $X^m = [x_1, x_2, \cdots, x_m]$ be a block of m samples from a memoryless source $\mathbf{X}$ with joint density $f_X^m(x) = f_X(x_1, x_2, \cdots, x_m)$. The asymptotic equipartition principle (AEP) [13] suggests that, as m becomes large, f_X^m becomes increasingly localized to a region $\mathcal{P}_m \in \mathbf{R}^m$ inside which it is almost a constant. The shape of this region of high probability is dependent on the source. For example, $\mathcal{P}_m$ is a m-dimensional spherical shell for a Gaussian source and a m-dimensional pyramidal shell for a Laplacian source [14]. Scalar quantizers do not exploit this localization property of the m-dimensional pdf.

It is possible to distribute all the output codevectors to lie inside the region $\mathcal{P}_m$ by using vector quantizers (VQ). In the last few years, several VQ's have sprung up with a MSE performance that is superior to that of a SQ (without entropy coding) for memoryless sources [15]. This performance improvement could be attributed to three different gains [16] [17]. They are

1. Boundary gain. Codebook boundary for the VQ can be chosen such that all its codevectors lie in the high probability region $\mathcal{P}_m$. Uniform scalar quantizers place all its output vectors in a m-dimensional cubic lattice. Boundary gain is dependent on the pdf f_X^m and is independent of the distortion-measure [17]. For a uniform source, since f_X^m is uniform over a m-dimensional cubic lattice, no boundary gain can be realized (at high encoding rates).

2. Granular gain. This gain is achieved by controlling the shape of partitions of the quantizer. Granular gain is dependent on the distortion measure and is independent of the pdf. For the MSE distortion measure, it is well known that it is asymptotically optimal to use m-dimensional spherical partitions instead of m-dimensional cubic partitions used by uniform scalar quantizers.

3. Non-uniform density gain. This gain is achieved by selecting the output vectors to lie closely in a region of higher probability and to lie farther apart in the other regions. AEP suggests that, for a large m, the m-dimensional pdf is roughly a constant in $\mathcal{P}_m$ and the codewords should tend to be uniformly distributed in this region. As a result, this gain disappears for a larger m. As expected, this gain cannot be obtained for a Uniform source.

Note that for non-uniform vector quantizers, it is not easy to distinguish between boundary and non-uniform density gains [16]. The shape and sizes of partitions can be vastly different. It is also fairly difficult to separately evaluate granular and boundary gains that can be achieved by a VQ (for both uniform and non-uniform vector quantizers) at low encoding rates. Eyuboğlu and Forney [17] showed that these two gains can be evaluated separately at high encoding rates and that the ultimate granular gain that can be achieved is upperbounded by 1.53 dB for the MSE distortion. They have also tabulated the boundary and granular gains for different lattice VQ's and trellis quantizers.

TCQ, first introduced by Marcellin and Fischer in [18], can achieve a granular gain of 1.36 dB (of the maximum possible 1.53 dB) using a 256-state trellis with a complexity that is roughly independent of the encoding rate and only a fraction of what is required by most source coders of comparable performance. For example, [18] shows that no VQ of dimension less than 69 can obtain the granular gain achieved by 256-state TCQ. Application of a VQ with such a high dimension is somewhat restricted in practice even at modest encoding rates due to its design and encoding complexity. Both the encoding and design complexity grows exponentially as mr, where m is the vector dimension and r is the encoding rate in bits/sample. Several vector quantization schemes exist with a reduced complexity but at the expense of some performance degradation [14][19].

Although TCQ is able to achieve a significant portion of the ultimate granular gain that can be obtained, it does not realize any boundary gain (at high encoding rates). It can realize some non-uniform density gain by using non-uniform output alphabets. Eyuboğlu and Forney [17] showed that, for Gaussian sources, the ultimate achievable boundary gain is the same as the ultimate gain that can be achieved by variable-length or entropy coding. Fischer and Wang introduced entropy-constrained TCQ (ECTCQ) in [20]. In that work, they showed that the SNR performance of 8-state ECTCQ systems is within 0.5 dB of the distortion-rate function for encoding memoryless sources at rates greater than about 1.5 bits/sample. Subsequently, [21] showed that it is possible to obtain a SNR performance within 0.5 dB of the distortion-rate function at all non-negative encoding rates with ECTCQ by using a slightly different formulation. In this Chapter, we briefly review TCQ and ECTCQ. A more detailed treatment on those two topics can be found in [18], [20], and [21].

2.2 Trellis Coded Quantization

For encoding memoryless sources at R bits/sample, TCQ uses an expanded alphabet size of 2^{R+1} codewords. This expanded codebook is divided into 4 disjoint subsets, with each subset containing 2^{R-1} codewords. These subsets are then assigned to the branches of a suitably chosen trellis. TCQ uses Ungerboeck's amplitude modulation trellises [22]. These trellises have two paths (branches) entering

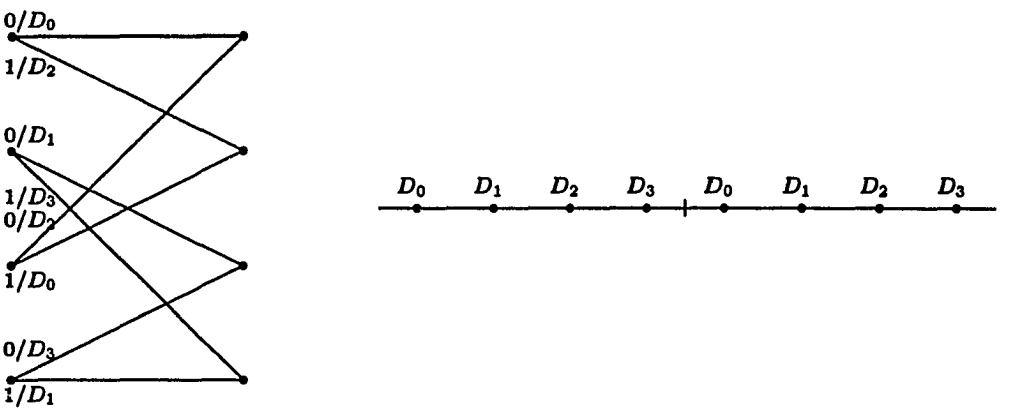

Figure 2.3: A 4-state trellis with subset labeling along with codebook and partition for 2 bit/sample TCQ.

and leaving each node. The justification for choosing this trellis family is provided in [18].

The trellis branch labeling for TCQ is similar to the one used by Ungerboeck's Trellis Coded Modulation [22]. For example, a 4-state trellis with its corresponding subset labeling and codebook partition for a 2 bit/sample TCQ is shown in Figure 2.3. From this figure, it is obvious that, given a "current" state, it would take 1 bit/sample to specify which path was chosen and hence the "next" trellis state. Hence, a path through the trellis can be specified by either a sequence of subsets or 1 bit codewords.

Assume that an input sequence $\mathbf{X} = [x_1, x_2, \cdots, x_n]$ of length n is to be encoded. At any time in the encoding process, for each path emanating from a given node,

scalar quantization is performed to find an element of the subset associated with that particular path that is closest to the current input sample. The squared quantization error introduced by this scalar quantization process is taken as the branch metric. The Viterbi algorithm [23] is then used to find the minimum distortion path through the trellis.

Assume that there is some cost associated with every node of the trellis. The branch cost incurred due to the scalar quantization process is added to the cost associated with the node from which it (the path) emanates from to calculate the overall cost to reach the "next" node if taking that particular path. At each next node, a two-way compare is made to determine the path with the smallest overall distortion (survivor). The survivor associated with each node is retained while the other path is discarded. This recursion is carried out until the end of the data sequence. The survivor associated with the node with the least cost at the end of recursion is just the survivor with the lowest overall distortion. This procedure is described mathematically as follows.

Let $\rho_j^{i-1}, j = 1, 2, \cdots, N$ be the cost associated with node j at time index $i-1$, D_{jk} be the subset associated with the path from node j to node k, d_{jk}^i be the element of subset D_{jk} which is closest to the current input sample x_i, where N is the number of trellis states. The cost incurred by taking the path from node j to

node k is $(x_i - d^i_{jk})^2$ and node costs are found using

$$\rho^i_k = \min_{j\in\{j_1,j_2\}} \{\rho^{i-1}_k + (x_i - d^i_{jk})^2\}$$

where j_1 and j_2 are the nodes from which the two paths to node k emanate. This recursion is carried out for the length of the data sequence. When the end of the data sequence is reached, the node with the smallest cost is found. That is

$$\rho_{\text{opt}} = \min_{j\in\{1,2,\cdots,N\}} \rho^n_j.$$

The survivor associated with ρ_{opt} is the one which minimizes the overall distortion.

There are two popular ways by which the survivor path with the least overall distortion can be mapped into a bit sequence. The first scheme, introduced in [18], makes use of the fact that, given the initial state, it requires 1 bit/sample to specify which path (and hence the subset) was chosen while the remaining $R-1$ bits/sample were used to select an element from that particular subset. These bit sequences are then transmitted through the channel. At the decoder, the bit sequence that is used to specify the path through the trellis is passed through a rate-1/2 (1 input, 2 outputs) convolutional coder. The two output bits from the convolutional coder determine the proper subset while the remaining $R-1$ bits/sample determine the codeword from the chosen subset.

An alternate scheme was proposed in [20] and subsequently used in entropy-constrained TCQ (ECTCQ) systems. This scheme considers the fact that the 4 subsets are disjoint. Hence, for any given state, the next codeword must be chosen

either from the superset $S_0 = D_0 \cup D_2$ or $S_1 = D_1 \cup D_3$. As the subsets are disjoint, specifying an element from a superset uniquely determines which subset it belongs to and hence the "next" trellis state. Since there are 2^R codewords in each superset, a sequence of R bit codewords can be used to specify a path through the trellis. This scheme was shown to be extremely advantageous for use in ECTCQ systems although the performance of these two schemes are identical for fixed-rate systems [20].

Simulation results for encoding a zero-mean, unit variance, memoryless Gaussian source using 4-state and 256-state TCQ is shown in Figure 2.2 [18]. TCQ codebooks were designed using the generalized Lloyd algorithm [24]. It is apparent from this figure that the SNR performance of TCQ is superior to that of a LMQ. A simple 4-state TCQ outperforms the LMQ by as much as 1.57 dB at 3 bits/sample while the performance improvement is 2.02 dB when using the 256-state TCQ.

Figure 2.4 shows the granular gain that can be achieved by TCQ systems as a function of the trellis size. A simple 4-state TCQ can achieve a granular gain of 0.99 dB which is very close to the ultimate granular gain that can be obtained by a 24-dimensional lattice VQ (1.03 dB) [17]. As mentioned in [17] and worth repeating here, a 4-state TCQ is much more easier to implement. The granular gain of 256-state TCQ comes within 0.18 dB of the ultimate granular gain of 1.53 dB that can ever by achieved. By evaluating the asymptotic VQ bound, [18] found

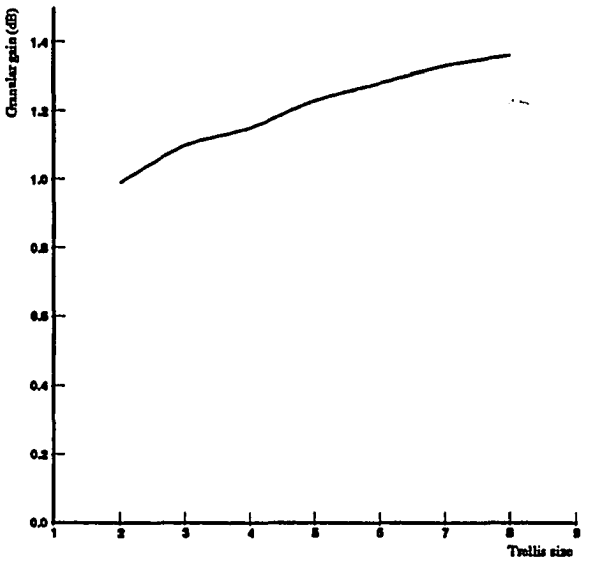

Figure 2.4: Granular gain obtained as a function of trellis size ($\log_2 N$).

out that no VQ of dimension less than 69 can obtain the granular gain achieved by 256-state TCQ.

2.3 Entropy-Constrained Trellis Coded Quantization

Consider a uniform codebook with a step size Δ and a subset partition as shown in Figure 2.5. The trellis branch labeling is the same as before. The number of codewords in each subset is no longer restricted to 2^{R-1} elements for encoding at R bits/sample. For each input sample, the closest codeword in each subset is found and the corresponding squared quantization error is used as the branch metric. Viterbi algorithm is then used to find the minimum MSE path through the trellis.

The ECTCQ encoder proposed in [20] uses the first method discussed in the previous section of assigning bit sequences to codeword sequences. In that scheme,

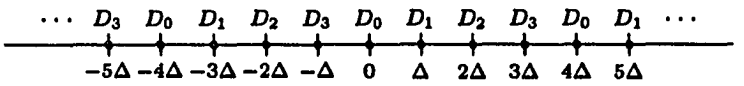

Figure 2.5: A uniform codebook and partition for an ECTCQ system.

1 bit/sample was used to specify a path through the trellis while an estimate of the rate required to specify elements from the chosen subset is given by the conditional entropy of the codebook given the subset. That is, the average encoding rate for this scheme is $1 + H(Y|D)$, where

$$H(Y|D) = -\sum_{i=0}^{3} \sum_{y \in D_i} P(y|D_i) P(D_i) \log_2 P(y|D_i)$$

and $P(y|D_i)$ is the probability of choosing the codeword y given the subset D_i. The minimum possible encoding rate for this scheme is 1 bit/sample (if scalar codebooks are used). Lower encoding rates can be achieved by using multidimensional codebooks. For encoding rates greater than 1.5 bits/sample, [20] reported a MSE performance within 0.5 dB of the distortion-rate function for encoding memoryless sources with smooth densities with 8-state ECTCQ systems.

Marcellin [21] showed that by using the second scheme of allocating bit sequences to codebook sequences, it is possible to encode at rates below 1 bit/sample with ECTCQ using scalar codebooks. In this scheme, since the entire rate is spent on specifying an element from one of of the two supersets S_0 or S_1, an estimate of the average encoding rate is given by the conditional entropy of the codebook given

the superset:

$$H(Y|S) = -\sum_{i=0}^{1} \sum_{y \in S_i} P(y|S_i) P(S_i) \log_2 P(y|S_i). \tag{2.8}$$

The performance of the two above mentioned ECTCQ systems is almost indistinguishable at encoding rates $\geq$ 1.5 bits/sample. However, when the encoding rate is decreased below 1.5 bits/sample, the performance of the system proposed in [21] is superior. Even so, the performance of this ECTCQ system is sub-optimal (in a distortion-rate sense). For example, the MSE performance for encoding the memoryless Gaussian source at 0.5 bits/sample is approximately 2 dB away from the distortion-rate function. [21] claims that this deterioration is due to the use of uniform codebooks and a MSE distortion measure. By using codebooks designed using a modified version of the generalized Lloyd algorithm [25], [21] showed that it is possible for 8-state ECTCQ systems to achieve a MSE performance within 0.5 dB of the distortion-rate function at all non-negative encoding rates. This codebook design algorithm attempts to minimize the MSE of an encoding (but subject to an entropy constraint) by minimizing the cost function

$$J = \mathrm{E}[\rho(x, y)] + \lambda \mathrm{E}[l(y)] \tag{2.9}$$

where x is the data, y is the encoded version of x, $\rho(x, y)$ is the cost (usually MSE) of representing x by y, λ is a Lagrange multiplier, and $l(y)$ is the number of bits used by the variable-length code to represent $\hat{x}$. A useful approximation of $l(y)$ is

$-\log_2 P(y)$. This optimization process considers not only the MSE but also the number of bits used by a variable-length code to represent y.

Figure 2.6 shows the SNR performance of a 4-state ECTCQ system for encoding a zero-mean, unit-variance Gaussian source along with the performance curve for an entropy-constrained scalar quantizer (ECSQ) and the distortion-rate function. The SNR performance of the 4-state ECTCQ system is within 0.55 dB of the distortion-rate function at all non-negative encoding rates. From Figure 2.6, it is also apparent that the performance of ECTCQ is superior to the performance of ECSQ. The performance curve for the 256-state ECTCQ system is within 0.2 dB of the distortion-rate function.

In [21], it was found that for rates greater than 2.5 bits/sample, the optimized codebooks do not provide a significant improvement in MSE over uniform codebooks although there may be some advantage in codebook size. Thus, in our simulations with images, optimized codebooks with 256 elements were used for encoding rates less than or equal to 2.5 bits/sample while uniform codebooks were used for all other rates. Uniform codebooks with at least $2^{(R_i+5)}$ elements were found necessary to achieve good performance for encoding rates greater than 2.5 bits/sample [26]. Smaller codebook sizes may be possible when using optimized (rather than uniform) codebooks.

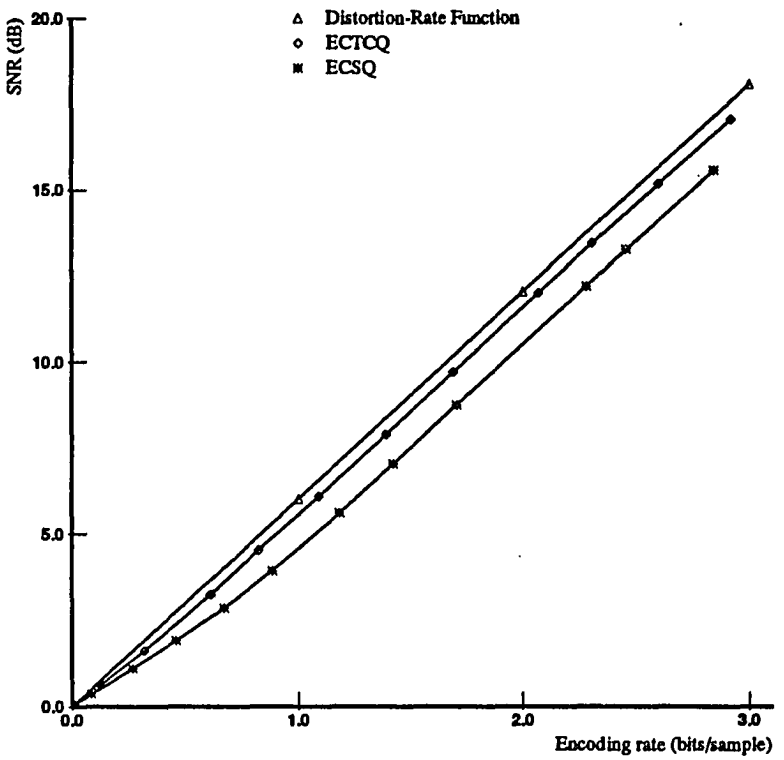

Figure 2.6: Distortion-rate performance of the entropy-constrained systems.

CHAPTER 3

TRANSFORM CODING

3.1 Introduction

Transform coding has proven to be a very effective technique for image coding [27] –[30]. The basic idea of any transform coding scheme is to divide the input image into blocks of pixels. Each block undergoes a two-dimensional transformation to produce an equal-sized block of transform coefficients. There exist different strategies to quantize these transform coefficients (for e.g., see [30]). In most schemes, low-frequency transform coefficients (which, in general, possess most of the energy) are quantized very finely while high-frequency transform coefficients are quantized rather coarsely.

There exist different transforms with good energy compaction properties [8][30]. Among these transforms, the discrete cosine transform (DCT) has emerged as one of the best, partially because of the fact that there exist fast algorithms to implement this transform [31].

The two-dimensional DCT (2-D DCT) of an image block $x(m,n)$ of size $N \times N$ is given by

$$\theta(k,l) = \frac{2}{N}\alpha(k)\alpha(l)\sum_{m=0}^{N-1}\sum_{n=0}^{N-1} x(m,n)\cos\frac{\pi k(2m+1)}{2N}\cos\frac{\pi k(2n+1)}{2N} \qquad (3.1)$$

where the transformed image block $\theta(k,l)$ is also of size $N \times N$ and

$$\alpha(j) = \begin{cases} \frac{1}{\sqrt{2}} & j = 0 \\ 1 & j \neq 0. \end{cases}$$

The inverse 2-D DCT is given by

$$x(m,n) = \frac{2}{N} \sum_{k=0}^{N-1} \sum_{l=0}^{N-1} \alpha(k)\alpha(l)\theta(k,l) \cos \frac{\pi k(2m+1)}{2N} \cos \frac{\pi k(2n+1)}{2N}. \qquad (3.2)$$

Note that the 2-D DCT can be efficiently implemented in two steps, each involving a 1-D DCT.

An efficient transform coder should adapt to nonstationarity among the blocks within each image. A transform coding system can be made adaptive in many different ways [32]. Chen and Smith [27] proposed a transform coder in which each input block is classified into one of four classes depending on their level of image activity. Adaptivity is provided by distributing bits among each class such that more bits are allotted to classes with higher activity.

An adaptive coder was introduced in [28] which reported excellent performance for encoding a color version of the "Lenna" image at 0.4 bpp. This image coder makes use of a threshold coding scheme in which all transform coefficients whose magnitude below the chosen threshold were set to zero. The remaining coefficients were then encoded using the available bit-rate.

In [33], Vaisey and Gersho segmented the input image into sub-blocks that vary in size between 4×4 to 32×32 pixels according to their level of perceptual

activity. They used vector quantization (VQ) to encode the transform coefficients and reported high-quality coding results at rates between 0.35 and 0.7 bpp.

Another adaptive transform coder, proposed in [29], is similar to the system proposed in [27]. Instead of classifying each block into one of four classes, the encoder in [29] estimates the variance of transform coefficients in each block. Using these estimates, they were able to adaptively encode in such a way that each block of the encoded image has the same distortion. The authors claim that keeping a constant block distortion results in encoded images which are perceptually superior.

This work extends the preliminary work described in [34]. TCQ is incorporated into a transform coding structure for encoding monochrome and color images. Both fixed-rate and entropy-constrained TCQ-based systems are investigated. In the monochrome case, the two-dimensional discrete cosine transform (DCT) is used to transform all (non-overlapping) 16×16 blocks of an image. Like coefficients from each block are collected and encoded using an optimum rate allocation method. This rate allocation algorithm requires the transmission of a small amount of side information, consisting of the sample mean of the "DC" coefficient and the sample variance of all transform coefficients. The procedural flow for color images is similar, except for the conversion of the RGB planes into NTSC transmission primaries (Y, I, and Q). The I and Q planes are subsampled (2:1 in each direction) before the DCT is applied.

Iterative codebook optimization algorithms are used to minimize MSE in both the fixed-rate and entropy-constrained TCQ systems. A large set of images (different from those used to test the resulting systems) are used as training data for these algorithms. In the fixed-rate case, the codebooks are designed for minimum MSE, while in the entropy-constrained case, the goal is to minimize MSE, but subject to a constraint on entropy. In each case, a different codebook is generated for each encoding rate used.

3.2 System Description

A block diagram illustrating the procedural flow for the monochrome TCQ transform coder is shown in Figure 3.1. All 16×16 blocks of a monochrome image are transformed using a two-dimensional DCT and like coefficients are collected into sequences of data to be encoded using TCQ. Since there are 256 DCT coefficients per block, there are then 256 sequences of coefficients to be encoded. The length of each of these sequences is 1024 for 512×512 images. Each sequence is normalized by subtracting its mean and dividing by its standard deviation. The normalized version of the sequences are then encoded using TCQ at rates determined by the rate allocation scheme discussed below. At the receiving end, the resulting bit sequence and normalization parameters (side information) are used to reconstruct the quantized coefficients. The inverse DCT is performed to obtain the final reconstructed image.

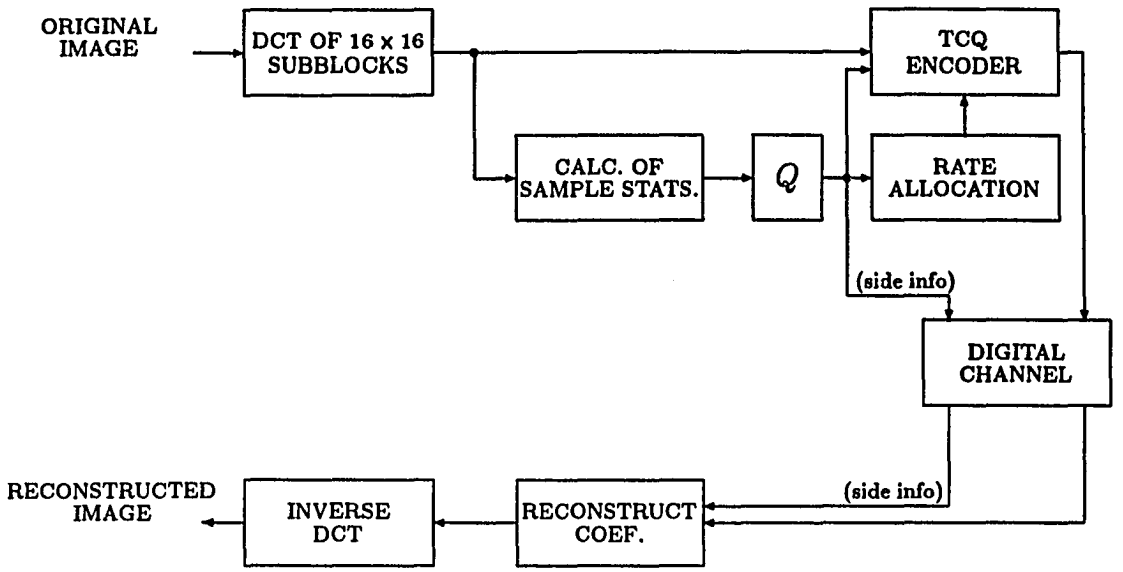

Figure 3.1: TCQ transform coder for monochrome images.

3.3 Rate Allocation and Side Information

The MSE for many source coders can be reasonably modeled by

$$E(R) \approx \alpha\sigma^2 2^{-\beta R} \tag{3.3}$$

where E is the MSE, R is the encoding rate in bits/sample, σ^2 is the variance of the source, and α and β are parameters that depend on the particular encoder structure and the probability density function of the source. It was found that the MSE performance of TCQ can be approximated with a common value of β for each coefficient while the required value for α may vary from coefficient to coefficient [26].

Let $K = 256$ be the number of coefficients obtained by transforming a 16×16 block of image data. The average distortion incurred by encoding the i^{th} coefficient sequence with R_i bits/coefficient can be determined with (3.3). Since the DCT is a unitary transform, the overall average distortion introduced in the image by encoding all K coefficient sequences is given by

$$E = \frac{1}{K} \sum_{i=1}^{K} \alpha_i \sigma_i^2 2^{-\beta R_i}. \tag{3.4}$$

Using Lagrange multipliers to minimize (3.4) subject to the average rate constraint

$$\frac{1}{K} \sum_{i=1}^{K} R_i = R. \tag{3.5}$$

yields the optimum rates [26]

$$R_i = \begin{cases} \frac{1}{\beta} \log_2 \frac{\alpha_i \sigma_i^2}{(\prod_{j=1}^{K'} \alpha_j \sigma_j^2)^{\frac{1}{K'}}} + \frac{RK}{K'} & i = 1, 2, \ldots, K' \\ 0 & i = K'+1, \ldots, K \end{cases}. \tag{3.6}$$

The value of K' is chosen by setting $K' = K$ and computing the rates, R_i, $i = 1, 2, \ldots, K'$ from (3.6). If any rates are negative, K' is decreased by one and the rates are computed again. Each time the variances are sorted so that $\sigma_i^2 \geq \sigma_{i+1}^2$ for all i before using (3.6). This ensures that the coefficients assigned an encoding rate of zero are those with the smallest variances. This process is repeated until all encoding rates are non-negative. The rates obtained from this algorithm are adjusted to integers in the fixed-rate system and to the nearest quarter of a bit in the entropy-constrained system. An alternate approach to rate allocation that can handle a more complicated distortion model than that of (3.3) is given in [35] and used in our subsequent work [36].

As mentioned previously, the sample mean of the DC coefficient (the other means are assumed to be zero) and the sample variance of each coefficient are encoded and transmitted as side information. These parameters were quantized with uniform scalar quantizers using a total of 2056 bits/image, or ≈ 0.008 bpp for 512×512 images. These 2048 bits were distributed among the different side information parameters in an ad hoc manner with more bits being given to the parameters with larger values (as determined by an examination of training data). At high rates (≥ 1 bpp), this side information is a negligible fraction of the overall rate. At lower rates, better performance may be possible through a more sophisticated approach.

In our experiments using TCQ for encoding the transform coefficients, we found that it is important that the initial state of the trellis is not chosen arbitrarily. Fixing the initial trellis state can cause the first few elements of each coefficient sequence to be quantized poorly. For example, if the initial state is fixed as the "top" state in Figure 2.3, the first coefficient is constrained to be quantized to something in the union of subsets D_0 and D_2. Although this makes almost no difference in objective performance, the subjective quality of the upper-left most block can suffer. Hence, the initial state should not be chosen in advance. As a result, the initial state that is ultimately chosen by the Viterbi search needs to be transmitted as side information. For a 4-state trellis, this requires 2 bits/sequence or equivalently, 512 bits/image. Thus, for 512×512 images, the added side information is ≈ 0.002 bpp.

3.4 Codebook design

Training sequence-based algorithms were used to design the TCQ codebooks. The training data were obtained from ten 256 × 256 images. These images were divided into 16×16 blocks and transformed. The DC coefficients were all normalized (using their means and variances) and collected together to be used as a training sequence for fixed-rate and entropy-constrained codebooks. In the fixed-rate case, a codebook was designed for each integer rate up to 8 bits/coefficient, while in the entropy-constrained case, a codebook was designed for each quarter of a bit. The non-DC coefficients were then normalized and grouped together to form a second training sequence for which both fixed-rate and entropy-constrained codebooks were also designed. For the entropy-constrained system, the number of bits required to represent a particular codeword y (i.e., $l(y)$) was computed as

$$\frac{1}{10}\lceil -10 \log_2 P(y|S_i) \rceil \tag{3.7}$$

where $P(y|S_i)$ (the probability of using y given that the superset S_i is used) is estimated from the training data. This yields an upper bound to the number of bits that would be required by a Huffman code operating on a block of codewords with length ten.

The rate-distortion performance of the TCQ designs were used to determine the parameters for the distortion model of Equation (3.3). It was found that for the fixed-rate systems, a value of $\beta = 1.88$ is appropriate for both the DC and

non-DC coefficients, while α should be chosen as 1.54 and 1.67 for the DC and non-DC coefficients, respectively. For the entropy-constrained designs, $\beta = 2.0$ is appropriate for both types of coefficients while $\alpha = 1.13$ and $\alpha = 0.47$ are appropriate for the DC and non-DC terms, respectively.

3.5 Results

3.5.1 Monochrome image coding

Coding simulations were performed for the luminance component of the 512×512 "Lenna" image. The performance of our image coder is reported by tabulating the PSNR which is defined as

$$\text{PSNR} = 10 \log_{10}\left(\frac{255^2}{\text{MSE}}\right) \text{ dB}. \tag{3.8}$$

For color images, the MSE is computed as an average over all three image planes (RGB).

Table 3.1 reports the performance obtained for encoding the 512 × 512 "Lenna" image using both fixed-rate and entropy-constrained TCQ systems with a 4-state trellis. For the purpose of comparison, we also simulated the performance of our system using scalar quantization (SQ) rather than TCQ. At fixed encoding rates, the PSNRs obtained by TCQ ranged from 1.0 to 2.0 dB better than those obtained by SQ. To achieve PSNRs equal to those reported in Table 3.1, the SQ system

Table 3.1: Transform coding results for the monochrome "Lenna" image.
(PSNR values are in dB.)

Fixed-rate		Entropy-constrained	
Rate	PSNR	Rate	PSNR
1.00	35.82	0.98	39.33
0.50	32.54	0.52	35.97
0.25	29.67	0.25	32.49

required 30 to 50 percent higher encoding rates than the TCQ system, with the larger percentages corresponding to the lower encoding rates.

Improvements in PSNR performance of the entropy-constrained system over the fixed-rate systems range from about 1.6 dB to 3.6 dB. As in the fixed-rate case, we also simulated our system using (entropy-constrained) SQ rather than TCQ. We found that to achieve the same PSNR, the SQ system required encoding rates about 10 percent higher than those of the TCQ system. Interestingly, even at equal PSNR, the perceptual quality of the images produced by the TCQ system is significantly better. The images from the SQ system contain considerably more high frequency background noise. To achieve perceptual quality equivalent to that of the TCQ system, the SQ system required an encoding rate at least 30 percent higher than that of the TCQ system. We believe that the reason for this is that the TCQ system does a better job of encoding the high frequency components of the image which are important perceptually, but do not significantly influence the MSE due to their low variance.

The subjective quality of the images produced by our systems are very good with the best results corresponding to the ECTCQ systems. The original 512 × 512 monochrome "Lenna" image along with the encoded image obtained from the ECTCQ system at 0.52 bpp are shown in Figures A.1 and A.2, respectively. As can be seen from these figures, the image quality is quite good with very little high frequency background noise, fuzziness, or blocking. The images obtained for encoding rates of $\approx$ 1.0 bpp are nearly indistinguishable from the originals. At $\approx$ 0.25 bpp, the images have visible artifacts but are still of fairly good quality.

3.5.2 Color image coding

It is well known that the three color planes (red, blue, and green) are highly correlated. To exploit this redundancy, it is a common practice to transform these planes to the NTSC transmission primaries specified as Y, I, and Q [37]. This transform has the added advantage of being compatible with monochrome television (Y component).

It is well known that the human eye is less sensitive to degradation in the chrominance components than to the degradation in the luminance component. As a result, color image coders concentrate on encoding the luminance component more efficiently than the chrominance components [38]. The I and Q planes are subsampled (2:1 in each direction) before applying 2-D DCT. Thus, the dimensions of the Y plane are $N \times N$ while the I and Q planes are each $\frac{N}{2} \times \frac{N}{2}$ where N is 512 for

a 512×512 image. As in the monochrome case, the block size for the DCT is 16×16 for the Y plane, while 8×8 blocks are used for the I and Q planes. The same rate allocation algorithm discussed before may be used, but now $K = 16^2 + 2(8)^2 = 384$.

Side information consisting of the average value for each of the three DC coefficients and the sample variance of all coefficients is transmitted for normalization, rate allocation, and reconstruction of the YIQ coefficients at the receiver as in the monochrome case. The amount of side information used to transmit these quantities totaled 3096 bits/image, or 0.012 bpp for 512×512 images. Before transforming the reconstructed coefficients back to the RGB domain at the receiver, we upsample the I and Q planes (using linear interpolation) by a factor of 1:2 in each direction.

Finally, the results of our color image coding experiments for both fixed-rate and entropy-constrained TCQ are presented in Table 3.2. As in the monochrome case, our system was simulated with SQ in addition to TCQ. Again, to achieve equal PSNR, the fixed-rate SQ required 30 to 50 percent higher encoding rates than fixed-rate TCQ while entropy-constrained SQ required about 10 percent higher encoding rate than entropy-constrained TCQ. Consistent with the monochrome case, the SQ systems required even larger rate increases to achieve perceptual quality equaling that of the TCQ systems. The original "Lenna" image along with the encoded image at 0.49 bpp are shown in Figures A.3 and A.4, respectively.

Table 3.2: Transform coding results for the color "Lenna" image.

(PSNR values are in dB.)

Fixed-rate		Entropy-constrained	
Rate	PSNR	Rate	PSNR
1.00	33.17	1.00	34.82
0.50	31.18	0.49	33.01
0.25	29.03	0.25	30.98

CHAPTER 4

WAVELET CODING

4.1 Introduction

Multi-frequency decomposition schemes are not new in the field of source coding. Subband coding was first introduced by Crochiere, Webber, and Flanagan [39] in 1976 for speech signals. The basic idea of any subband coding scheme is to decompose the input signal into a number of frequency bands (or subbands) using a bank of band-pass filters (analysis stage). Each subband is then decimated and encoded appropriately. At the receiver, the encoded subbands are interpolated and then passed through reconstruction filters (synthesis stage) to obtain the reconstructed signal. This approach, in general, demands the design of sophisticated band-pass filters to minimize the effects of aliasing.

Quadrature mirror filters (QMF) were introduced in [40] and allow alias free reconstruction of the signal in the absence of quantization errors. Vetterli [41] extended the application of QMF's to multi-dimensional signals. Both separable and non-separable extensions were considered, but no coding results were presented. Subsequently, Woods and O'Neil [5] presented the first image coder using subband coding. The input image was split into 16 equal-sized subbands using circular

convolution with 32-tap QMF's designed by Johnston [42]. They used DPCM to encode the image subbands.

Gharavi and Tabatabai [43] proposed another subband coding scheme in which the input image is split into seven un-equal-sized subbands. The lowest frequency subband was encoded using DPCM while other subbands were encoded using memoryless quantizers. Their work was extended for color image coding as well. Since then, a variety of subband coders have emerged, capable of high-quality encoding with bit-rates as low as 0.5 bpp (e.g., [44]–[6]).

There are several advantages to multi-frequency decomposition schemes. Since quantization error variance can be separately controlled in each band, by careful allocation of encoding rate, the overall reconstruction error spectrum can be controlled in such a manner that the reconstructed image is perceptually pleasing. Multi-resolution approximation schemes are also well suited for progressive image transmission [47].

A number of multi-resolution approximation schemes have emerged independently in different fields of engineering and science [48]. Recently, wavelet theory has been recognized as a unifying framework for these multi-resolution techniques [49]–[51]. Wavelets were originally introduced as a family of functions which were derived from translations and dilations of one basic function, referred to as the "mother" wavelet [52].

The basic idea of the discrete wavelet transform (DWT) is that of successive approximation, together with that of "added detail". At each stage, the input signal is decomposed into a coarse approximation signal (which can be considered a low-pass version of the input) and an "added detail" signal (which can be considered a high-pass version). In this regard, the DWT decomposes the input signal into a set of frequency subbands [49].

Wavelet coders for images have been implemented both with scalar quantization [49] and vector quantization [53]. In this chapter, we investigate the use of ECTCQ with the DWT for encoding both monochrome and color images. In order to apply wavelet decompositions to images, we use a separable 2-D DWT in which emphasis is given to the horizontal and vertical directions. For the monochrome case, we evaluate the performance of our wavelet coder for 7-band and 16-band decompositions. In each case, the lowest-frequency sub-image (LFS) is encoded using a 2-D DCT encoder (with a block size of 4×4) while the other sub-images are encoded using ECTCQ for memoryless data. An integer programming algorithm [35] is employed to allocate the available bit-rate optimally among the subbands. A small amount of side information, consisting of the sample mean of the "DC" coefficient and the sample standard deviation of all sub-images to be encoded, are transmitted. The procedural flow for color images is similar, except for the conversion of the RGB planes into NTSC transmission primaries (Y, I, and Q).

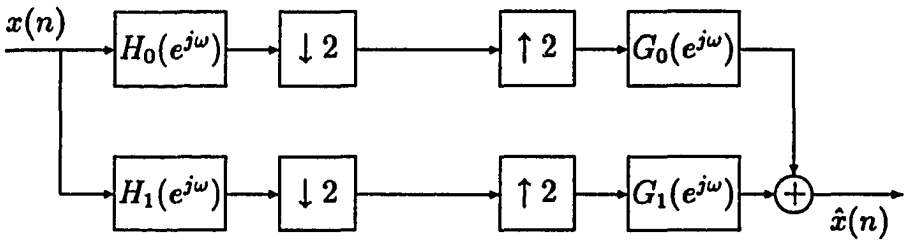

Figure 4.1: A two-band analysis/synthesis system.

4.2 Subband filter banks

Figure 4.1 shows a two-band analysis/synthesis system which is the basic building block for a tree-structured subband coder. In this figure, $\downarrow 2$ represents a downsampler by a factor of 2, $\uparrow 2$ is an upsampler, $H_0(e^{j\omega})$ and $H_1(e^{j\omega})$ are the low-pass and high-pass analysis filters, and $G_0(e^{j\omega})$ and $G_1(e^{j\omega})$ are the low-pass and high-pass synthesis filters, respectively.

The relationship between the input $X(e^{j\omega})$ and the reconstructed output $\hat{X}(e^{j\omega})$ is given by [54]

$$\begin{aligned}\hat{X}(e^{j\omega}) &= \frac{1}{2}\Big[G_0(e^{j\omega})H_0(e^{j\omega}) + G_1(e^{j\omega})H_1(e^{j\omega})\Big]X(e^{j\omega}) \\ &+ \frac{1}{2}\Big[G_0(e^{j\omega})H_0(e^{j(\omega+\pi)}) + G_1(e^{j\omega})H_1(e^{j(\omega+\pi)})\Big]X(e^{j(\omega+\pi)}). \quad (4.1)\end{aligned}$$

Note that the second term in the above equation is responsible for aliasing distortions. For perfect reconstruction (i.e., $\hat{x}(n) = x(n-M)$, M a positive integer to account for processing delay) the analysis and synthesis filters have to satisfy the

following conditions

$$G_0(e^{j\omega})H_0(e^{j\omega}) + G_1(e^{j\omega})H_1(e^{j\omega}) = 2e^{j\omega M}$$

$$G_0(e^{j\omega})H_0(e^{j(\omega+\pi)}) + G_1(e^{j\omega})H_1(e^{j(\omega+\pi)}) = 0. \tag{4.2}$$

Aliasing distortions can be eliminated by choosing the synthesis filters as

$$\begin{aligned} G_0(e^{j\omega}) &= H_1(e^{j(\omega+\pi)}) \\ G_1(e^{j\omega}) &= -H_0(e^{j(\omega+\pi)}). \end{aligned} \tag{4.3}$$

In addition, if the analysis filters are chosen to be frequency shifted versions of one another, i.e., $H_1(e^{j\omega}) = H_0(e^{j(\omega+\pi)})$ and are also constrained to have even lengths, then exact reconstruction requires that

$$H_0^2(e^{j\omega}) - H_0^2(e^{j(\omega+\pi)}) = H_0^2(e^{j\omega}) - H_1^2(e^{j\omega}) = 2e^{j\omega M}. \tag{4.4}$$

Such a filter bank is commonly referred to as a QMF bank [55]. [55] proves that perfect reconstruction is not possible if QMF's have odd lengths. Numerous attempts were made to design linear-phase FIR filters which would approximate the above condition [42][56]. [54] cites two cases in which Equation (4.4) is exactly satisfied. In the first case, filters have to be infinitely long while in the second case, analysis and synthesis filters are simple two-tap filters. Hence, both cases are not useful for practical implementations. Johnston [42] designed linear-phase, higher-order QMF's which approximately satisfied Equation (4.4).

Recently [54], it has been shown that perfect reconstruction can be achieved if the analysis filters are related by

$$H_1(e^{j\omega}) = -H_0(e^{-j(\omega+\pi)})e^{-j\omega M} \tag{4.5}$$

where the filter length ($= M + 1$) is even. The synthesis filters are related to the analysis filters according to Equation (4.3) as before.

The analysis filters chosen according to Equation (4.5) should satisfy

$$|H_0(e^{j\omega})|^2 + |H_1(e^{j\omega})|^2 = 2 \tag{4.6}$$

for perfect reconstruction. Two filters which satisfy Equation (4.6) are said to form a power-complementary (PC) pair. [57] shows that two linear-phase, finite-length FIR filters, $H_0(e^{j\omega})$ and $H_1(e^{j\omega})$, cannot satisfy Equation (4.6) unless they are trivial. PC property is a sufficient condition for perfect reconstruction but not necessary [58]. It is possible to construct non-trivial, finite-length linear-phase filters which achieve perfect reconstruction if the PC property is dropped [59].

4.3 Discrete wavelet transform

DWT can be best explained for the case of an orthogonal representation. A DWT utilizes two functions: the mother wavelet $\tilde{\psi}$ and a scaling function $\tilde{\phi}$. The scaling function $\tilde{\phi}$ can be chosen in such a manner that the translations of (dilated versions of) $\tilde{\phi}$ are orthonormal. That is, for a fixed m,

$$\tilde{\phi}_{mn}(x) = 2^{m/2}\tilde{\phi}(2^m x - n),\ n = \cdots, -1, 0, 1, \cdots$$

forms an orthonormal basis for a vector space, say $\tilde{\mathbf{V}}_m$. Letting m vary results in a sequence of successive approximation spaces. i.e.,

$$\cdots \subset \tilde{\mathbf{V}}_{-2} \subset \tilde{\mathbf{V}}_{-1} \subset \tilde{\mathbf{V}}_0 \subset \tilde{\mathbf{V}}_1 \subset \tilde{\mathbf{V}}_2 \cdots .$$

$\tilde{\mathbf{V}}_m$ is said to have a resolution of m. For each m, $\tilde{\psi}_{mn}$ span a vector space $\tilde{\mathbf{W}}_m$ which is the orthogonal complement of $\tilde{\mathbf{V}}_m$ in $\tilde{\mathbf{V}}_{m+1}$. This is written as

$$\tilde{\mathbf{V}}_{m+1} = \tilde{\mathbf{V}}_m \overset{\perp}{\oplus} \tilde{\mathbf{W}}_m.$$

Consider an arbitrary input function $\mathbf{f}$ at a resolution 0 (i.e., $\mathbf{f} \in \tilde{\mathbf{V}}_0$). The approximation of $\mathbf{f}$ at a lower resolution -1 (say $A_{-1}(\mathbf{f})$) is given by the orthogonal projection of $\mathbf{f}$ onto the vector space $\tilde{\mathbf{V}}_{-1}$. This can be written as

$$A_{-1}(f(x)) = \sum_n \tilde{a}_{-1n}\tilde{\phi}_{-1n}(x)$$

where

$$\tilde{a}_{-1n} = \left\langle f(u), \tilde{\phi}_{-1n}(u) \right\rangle = \int_{-\infty}^{\infty} f(u)\tilde{\phi}_{-1n}(u)\, du. \tag{4.7}$$

Note that since $\mathbf{f} \in \tilde{\mathbf{V}}_0$, $A_0(\mathbf{f}) = \mathbf{f}$. The information lost when going from $A_0(\mathbf{f})$ to the coarser approximation $A_{-1}(\mathbf{f})$ is referred to as the error, or "detail" signal $D_{-1}(\mathbf{f})$ and can be obtained by the projection of $\mathbf{f}$ onto $\tilde{\mathbf{W}}_{-1}$. That is,

$$D_{-1}(f(x)) = \sum_n \tilde{d}_{-1n}\tilde{\psi}_{-1n}(x) \tag{4.8}$$

where $\tilde{d}_{-1n} = \left\langle f(u), \tilde{\psi}_{-1n}(u) \right\rangle$. Since $\tilde{\mathbf{V}}_0 = \tilde{\mathbf{V}}_{-1} \overset{\perp}{\oplus} \tilde{\mathbf{W}}_{-1}$,

$$\begin{aligned} f(x) &= A_0(f(x)) = A_{-1}(f(x)) + D_{-1}(f(x)) \\ &= \sum_n \tilde{a}_{-1n}\tilde{\phi}_{-1n}(x) + \sum_n \tilde{d}_{-1n}\tilde{\psi}_{-1n}(x). \end{aligned} \tag{4.9}$$

Since $(\tilde{\phi}_{0n})_n$ is a basis for $\tilde{\mathbf{V}}_0$, $\mathbf{f}$ can also be written as

$$f(x) = \sum_k \tilde{a}_{0k}\tilde{\phi}_{0k}(x). \tag{4.10}$$

Using Equation (4.10) in Equation (4.7), we get

$$\begin{aligned}
\tilde{a}_{-1n} &= \left\langle f(u), \tilde{\phi}_{-1n}(u)\right\rangle = \int_{-\infty}^{\infty} f(u)\tilde{\phi}_{-1n}(u)\,du \\
&= \int_{-\infty}^{\infty} \sum_k \tilde{a}_{0k}\tilde{\phi}_{0k}(u)\tilde{\phi}_{-1n}(u)\,du \\
&= \sum_k \tilde{a}_{0k}\int_{-\infty}^{\infty} \tilde{\phi}_{0k}(u)\tilde{\phi}_{-1n}(u)\,du \\
&= \sum_k \tilde{a}_{0k}\left\langle \tilde{\phi}_{0k}(u), \tilde{\phi}_{-1n}(u)\right\rangle.
\end{aligned} \tag{4.11}$$

Define $g_0(n) = 2^{-1/2}\left\langle \tilde{\phi}(u/2), \tilde{\phi}(u-n)\right\rangle$. Then

$$\begin{aligned}
\left\langle \tilde{\phi}_{0k}(u), \tilde{\phi}_{-1n}(u)\right\rangle &= \int_{-\infty}^{\infty} \tilde{\phi}(u-k)2^{-1/2}\tilde{\phi}(2^{-1}u-n)\,du \\
&= 2^{-1/2}\int_{-\infty}^{\infty} \tilde{\phi}(u+2n-k)\tilde{\phi}(u/2)\,du \\
&= g_0(k-2n)
\end{aligned} \tag{4.12}$$

Hence

$$\tilde{a}_{-1n} = \sum_k \tilde{a}_{0k}g_0(k-2n) = \sum_k \tilde{a}_{0k}h_0(2n-k) \tag{4.13}$$

where $h_0(n) = g_0(-n)$. Similarly, one can show that

$$\tilde{d}_{-1n} = \sum_k \tilde{a}_{0k}g_1(k-2n) = \sum_k \tilde{a}_{0k}h_1(2n-k) \tag{4.14}$$

with $g_1(n) = 2^{-1/2}\left\langle \tilde{\psi}(u/2), \tilde{\phi}(u-n)\right\rangle = h_1(-n)$. Thus $(\tilde{a}_{-1n})_n$ and $(\tilde{d}_{-1n})_n$ can be found by convolving $(\tilde{a}_{0n})_n$ with h_0 and h_1, respectively and then keeping every

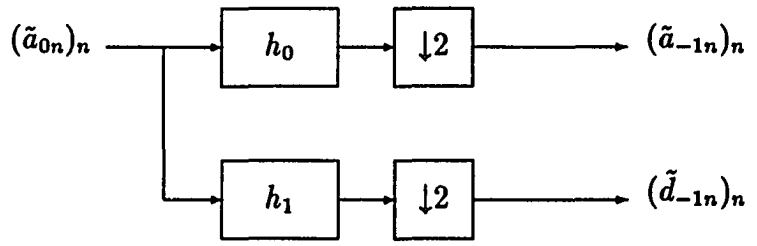

Figure 4.2: Block diagram of a wavelet decomposition.

other sample of the output. h_1 is typically a high-pass filter while h_0 is a low-pass. This algorithm is illustrated by the block diagram shown in Figure 4.2. Approximations at lower-resolutions are obtained by repeated application of this algorithm.

Given $(\tilde{a}_{-1n})_n$ and $(\tilde{d}_{-1n})_n$, $(\tilde{a}_{0n})_n$ can be reconstructed using

$$\begin{aligned}
\tilde{a}_{0n} &= \left\langle f(u), \tilde{\phi}_{0n}(u) \right\rangle \\
&= \left\langle \sum_k \tilde{a}_{-1k}\tilde{\phi}_{-1k}(u) + \tilde{d}_{-1k}\tilde{\psi}_{-1k}(u), \tilde{\phi}_{0n}(u) \right\rangle \\
&= \sum_k \tilde{a}_{-1k}\left\langle \tilde{\phi}_{-1k}(u), \tilde{\phi}_{0n}(u) \right\rangle + \sum_k \tilde{d}_{-1k}\left\langle \tilde{\psi}_{-1k}(u), \tilde{\phi}_{0n}(u) \right\rangle \\
&= \sum_k \tilde{a}_{-1k} g_0(n-2k) \; + \; \sum_k \tilde{d}_{-1k} g_1(n-2k). \qquad (4.15)
\end{aligned}$$

That is, $(\tilde{a}_{0n})_n$ can be reconstructed by interpolating $(\tilde{a}_{-1n})_n$ and $(\tilde{d}_{-1n})_n$ by a factor of two and filtering the resulting signals with g_0 and g_1, respectively. The block diagram shown in Figure 4.3 illustrates this algorithm. Figures 4.2 and 4.3 reveal that discrete wavelet transforms are essentially subband decomposition systems.

Mallat [49] and Daubechies [51] have shown that analysis filters corresponding to orthogonal wavelet bases form PC pairs. Filters obtained from Daubechies's

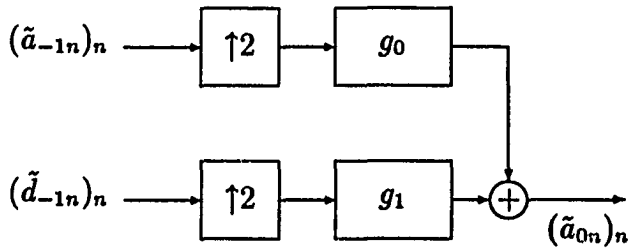

Figure 4.3: Block diagram of a wavelet reconstruction.

wavelets in [51] are non-linear-phase, FIR filters. For image processing applications, one would prefer analysis and synthesis filters to have linear-phase. If the mother wavelet is symmetric, the resulting filters will be linear-phase filters. Wavelets which result in finite-length filters are said to have "compact support". Mallat's orthonormal bases are built using symmetric wavelets with non-compact support.

[51] proved that the only orthonormal basis of compactly supported wavelets for which the associated filters have linear phase is the Haar basis. Filters associated with Haar basis are trivial two-tap filters. It is very important to have filters with good attenuation characteristics to minimize the effect of aliasing due to coding distortions. This contradiction is already well known in the subband coding literature as explained in the previous section.

Our preliminary work with TCQ of image subbands as reported in [60] employed 33-tap filters based on Mallat's wavelet (referred to as Mallat's filters for the reminder of this dissertation). With these filters, the reconstruction of $(\tilde{a}_{0n})_n$ given $(\tilde{a}_{-1n})_n$ and $(\tilde{d}_{-1n})_n$ is not perfect. However, for the 512 × 512 "Lenna" image,

the reconstruction PSNR is 69 dB (in the absence of quantization) for a 4 band decomposition.

Biorthogonal bases for wavelets were recently introduced, independently, by Cohen, Daubechies, and Feauveau [61] and by Vetterli and Herley [62]. In [61], it was shown that it is possible to construct bases that yield finite-length, linear-phase filters with the perfect reconstruction property by relaxing the orthonormality requirement (and hence the PC property).

For the biorthogonal wavelet bases, there exist two hierarchies of bases, one for the analysis stage and one for the synthesis stage. Let ϕ and ψ denote the scaling function and the wavelet necessary for reconstruction. $(\tilde{\phi}_{mn})_n$ is a basis for $\tilde{\mathbf{V}}_m$, but no longer orthonormal. The vector space $\tilde{\mathbf{W}}_m$ is the complement of $\tilde{\mathbf{V}}_m$ in $\tilde{\mathbf{V}}_{m-1}$ (but not an orthogonal complement as before). $A_{-1}(\mathbf{f})$ and $D_{-1}(\mathbf{f})$ can be found the same way as in the orthogonal case. i.e.,

$$\begin{aligned} A_{-1}(f(x)) &= \sum_n \tilde{a}_{-1n}\tilde{\phi}_{-1n}(x) \\ D_{-1}(f(x)) &= \sum_n \tilde{d}_{-1n}\tilde{\psi}_{-1n}(x). \end{aligned}$$

However, the sequence of coefficients $(\tilde{a}_{-1n})_n$ and $(\tilde{d}_{-1n})_n$ are found using

$$\begin{aligned} \tilde{a}_{-1n} &= \left\langle f(u), \phi_{-1n}(u) \right\rangle \\ \tilde{d}_{-1n} &= \left\langle f(u), \psi_{-1n}(u) \right\rangle. \end{aligned} \tag{4.16}$$

Since $(\tilde{\phi}_{0n})_n$ is a basis for $\tilde{\mathbf{V}}_0$, f can also be written as

$$A_0(f(x)) = f(x) = \sum_n \tilde{a}_{0n}\tilde{\phi}_{0n}(x) \tag{4.17}$$

where $\tilde{a}_{0n} = \left\langle f(u), \phi_{0n}(x)\right\rangle$. $\tilde{\phi}_{mn}$ and ϕ_{mn} are related by $\left\langle \tilde{\phi}_{mn}, \phi_{ij}\right\rangle = \delta_{mi}\delta_{nj}$. Using Equation (4.17) in Equation (4.16), we can show that

$$\begin{aligned} \tilde{a}_{-1n} &= \sum_k \tilde{a}_{0k} h_0(2n-k) \\ \tilde{d}_{-1n} &= \sum_k \tilde{a}_{0k} h_1(2n-k) \end{aligned} \tag{4.18}$$

where $h_0(n) = 2^{-1/2}\left\langle \phi(u/2), \tilde{\phi}(u+n)\right\rangle$ and $h_1(n) = 2^{-1/2}\left\langle \psi(u/2), \tilde{\phi}(u+n)\right\rangle$. Since $\tilde{\mathbf{V}}_0 = \tilde{\mathbf{V}}_{-1} \oplus \tilde{\mathbf{W}}_{-1}$, $(\tilde{a}_{0n})_n$ can be reconstructed from $(\tilde{a}_{-1n})_n$ and $(\tilde{d}_{-1n})_n$ using

$$\tilde{a}_{0n} = \sum_k \tilde{a}_{-1k} g_0(n-2k) + \sum_k \tilde{d}_{-1k} g_1(n-2k) \tag{4.19}$$

where $g_0(n) = 2^{-1/2}\left\langle \tilde{\phi}(u/2), \phi(u-n)\right\rangle$ and $g_1(n) = 2^{-1/2}\left\langle \tilde{\psi}(u/2), \phi(u-n)\right\rangle$. It is obvious from Equations (4.18) and (4.19) that the decomposition and reconstruction algorithms remain the same as in the orthogonal case. However, the relationship between analysis and synthesis filters are different. When $\tilde{\phi} = \phi$ and $\tilde{\psi} = \psi$, the two cases become the same.

To achieve perfect reconstruction, the analysis and synthesis filters are chosen according to

$$g_1(n) = (-1)^n h_0(-n+1),\ h_1(n) = (-1)^n g_0(-n+1)$$

$$\sum_n h_0(n) g_0(n+2k) = \delta(k). \tag{4.20}$$

There is a condition which the filters developed using wavelet theory typically satisfy while those developed using classical signal processing techniques do not. Recall that h_0 is derived from the scaling function $\tilde{\phi}$. Given h_0, $\tilde{\phi}$ can be obtained using

$$\tilde{\Phi}(\omega) = k \prod_{p=1}^{\infty} H_0(2^{-p}\omega) \tag{4.21}$$

where $\tilde{\Phi}$ is the fourier transform of $\tilde{\phi}$ and k is a scaling factor. h_0 is chosen in such a way that $\tilde{\Phi}(\omega)$ decays faster than $C(1+|\omega|)^{-\epsilon-0.5}$ as $\omega \rightarrow \infty$ for some $\epsilon > 0$ and constant C. This condition is often referred to as a "regularity" condition. The infinite product in Equation (4.21) need not converge. In such a case, it is very difficult to visualize the subband decomposition using wavelets.

[61] gives a set of conditions for which the infinite product in Equation (4.21) converges pointwise to $\tilde{\Phi}(\omega)$ for all ω such that $\tilde{\phi}$ is a continuous function. Filters which are designed in such a manner that their corresponding scaling function obtained using the above mentioned procedure is a continuous function are said to be "regular". The order of regularity is the number of times $\tilde{\phi}$ is continuously differentiable [48]. This regularity condition is included such that any input function can be decomposed and reconstructed using reasonably smooth basis functions.

There exist filters designed for tree-structured filter banks which are not regular. Rioul and Vetterli [48] showed that the infinite product in Equation (4.21) diverges for Smith and Barnwell's non-linear-phase filters designed in [54] for a perfectly reconstructing filter bank. [48] also showed that the scaling function associated

with Johnston's 8-tap filters, although continuous, are not differentiable (regularity of order 0).

If both h_0 and g_0 are regular, then h_1 and g_1 are also regular because of their relationship to g_0 and h_0, respectively. Using the conditions in Equation 4.20, [61] showed that analysis and synthesis filters with arbitrarily high regularity can be obtained if h_0 and h_1 are related by

$$H_0(\omega)H_1(\omega) = \cos(\omega/2)^{2l}\left[\sum_{p=0}^{l-1}\binom{l-1+p}{p}\sin(\omega/2)^{2p} + \sin(\omega/2)^{2l}R(\omega)\right] \quad (4.22)$$

where, $R(\omega)$ is an odd polynomial in $\cos(\omega)$, $2l = k + \tilde{k}$, and the functions ψ and $\tilde{\psi}$ are $(k-1)$ and $(\tilde{k}-1)$ continuously differentiable, respectively. $H_0(\omega)$ and $H_1(\omega)$ can be chosen in several different ways [61]. We have chosen to use the spline variant family of filters suggested in [53] and [61] with $k = \tilde{k} = 4$. For this selection, h_0 and h_1 are 9-tap and 7-tap filters, respectively. For convenience, this filter family is referred to as 9-tap spline filters for the reminder of this dissertation.

4.4 Image coding application

A block diagram illustrating the procedural flow for a monochrome TCQ wavelet coder is shown in Figure 4.4. The input image is decomposed into a series of sub-images using a 2-D DWT. Since images are spatially limited, the filtering and decimation result in an expansion of data. To circumvent this problem, we have used a generalization of the symmetric extension technique described in [45] and [63]

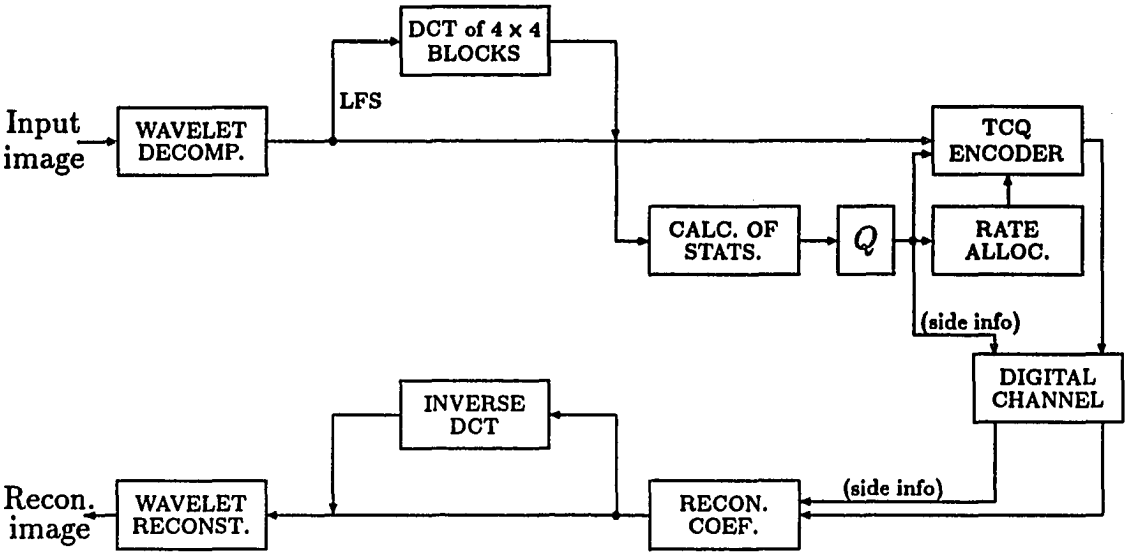

Figure 4.4: Block diagram of a monochrome TCQ wavelet coder.

which allows the amount of data to be reduced to its original size while introducing no distortion.

A similar system using Johnston's QMF's [42] was studied by Tanabe and Farvardin in [6]. They analyzed the statistical properties of the subband data from different images and concluded that the correlation properties of the LFS is similar to that of the original image. All subbands, except the LFS, were found to have negligible interpixel correlation. Based on these facts, they tried using both DPCM and DCT to encode the LFS while other subbands were encoded with no additional processing. The DCT based system was found to be superior.

We have followed their approach and use a 2-D DCT with a block size of 4×4 for encoding the LFS. All "like" DCT coefficients of the LFS are collected into

sequences to be encoded using ECTCQ. Since there are 16 DCT coefficients in each block, there are then 16 sequences of coefficients to be encoded. Each of the High-frequency sub-images (HFS) is also treated as a sequence to be encoded (with no further processing) using ECTCQ.

A small amount of side information, consisting of the sample mean of the "DC" transform coefficient and the sample standard deviation of all sub-images and DCT coefficients, are transmitted. All DCT coefficients and sub-images are normalized by subtracting their mean (all data except the DC transform coefficient are assumed to be zero-mean) and then dividing by their respective standard deviations. The "normalized" transform coefficients and sub-images are then encoded using TCQ at rates determined by the optimum rate-allocation scheme described in a subsequent section. At the receiving end, the resulting bit sequence and normalization parameters (side information) are used to reconstruct the quantized coefficients. The inverse DCT is performed to obtain the reconstructed LFS before the final wavelet reconstruction stage.

4.4.1 Codebook design and rate allocation

A collection of 30 images (different from the "Lenna" image) were used as training data for the optimization algorithm. Three sets of codebooks: one for the DC coefficient, one for the other DCT coefficients and the third for all sub-images

other than the LFS were used. For each set, codebooks were designed for integer multiples of 0.1 bits/sample.

The basic intention of any rate allocation scheme is to appropriately allocate the bits to be used for encoding among the data sequences to be encoded so as to optimize the performance according to some objective cost function. For our wavelet coder, we use a bit allocation algorithm in which the distortion-rate performances of different quantizers are used [35]. This algorithm produces an optimal or very nearly optimal allocation, while allowing the set of admissible bit allocation values to be constrained to a finite set of non-negative numbers.

Specifically, the overall MSE incurred by our coding scheme is given by

$$E = \sum_{i=1}^{K} \alpha_i w_i E_i(r_i) \tag{4.23}$$

where $E_i(r_i)$ is used to denote the distortion-rate performance for encoding the i^{th} data sequence at r_i bits/sample, K is the number of data sequences[1] , and α_i is a weighting coefficient to account for the variability in the size of the sequences. Also, since the biorthogonal synthesis filters, h and g, do not have the same energy, the quantization noise in various subbands will not be equally weighted in the image reconstruction. The scaling factor w_i is introduced to offset this disparity. A detailed treatment of the procedure to find these weighting coefficients for a given set of filters can be found in [64].

[1] K is the number of subbands minus one, plus the number of DCT coefficient sequences. For example, K equals 22 and 31 for 7 and 16-band decompositions, respectively.

In practice, the rate allocation vector $B = (r_1, r_2, ..., r_K)$ is chosen so as to minimize E subject to the constraint that

$$\sum_{i=1}^{K} \alpha_i r_i \leq R \text{ bpp.} \tag{4.24}$$

In [35], it is shown that the solution $B^* = (r_1^*, r_2^*, ..., r_K^*)$ to the unconstrained problem

$$\min_B \left\{ \sum_{i=1}^{K} \alpha_i w_i E_i(r_i) + \lambda \sum_{i=1}^{K} \alpha_i r_i \right\} \tag{4.25}$$

minimizes E subject to $\sum_{i=1}^{K} \alpha_i r_i \leq \sum_{i=1}^{K} \alpha_i r_i^*$. Thus, to find a solution to the constrained problem of Equations (4.23) and (4.24), it suffices to find λ such that the solution to Equation (4.25) yields $\sum_{i=1}^{K} \alpha_i r_i^* \leq R$. A detailed treatment of an algorithm to find the proper λ can be found in [35].

For a given λ, the solution to the unconstrained problem is obtained by minimizing each term of the sum in Equation (4.25) separately. If $V_k = \{p_k, ..., q_k\}$ is the set of allowable rates for the k^{th} quantizer, then r_k^* solves

$$\min_{r_k \in V_k} \left\{ \alpha_k w_k E_k(r_k) + \lambda \alpha_k r_k \right\}. \tag{4.26}$$

4.4.2 Side information

The side information consists of the sample mean of the DC transform coefficient and the sample standard deviation of all data sequences to be encoded. A 16-bit uniform quantizer was used to quantize each of these parameters resulting in $16(K + 1)$ bits/image of side information. In addition, the initial trellis state for

each data sequence needs to be transmitted to the receiver [26]. For a 4-state trellis, this requires $2K$ bits/image. Hence, the overall side information amounts to $(18K+16)$ bits/image which corresponds to 0.002 bpp for a 16-band decomposition of a monochrome, 512×512 image.

4.5 Results

4.5.1 Monochrome image coding

Subband coders which have been proposed in the literature have used both 7-band (pyramidal) and 16-band (tree-structured) decompositions. Westerink, Biemond, and Boekee [65] compared different decomposition schemes in a fixed-rate coding system using QMF's and reported that the best objective performance is obtained when the image is split into 16 equally sized subbands. It is not mentioned in [65], however, if there is any improvement in the subjective quality of the encoded images.

We investigated the performance of our wavelet coder using a 7-band (7B) and a 16-band (16B) decomposition. To obtain the same PSNR value, the 7B system required an encoding rate approximately 15% higher than that of the 16B system. Interestingly, even at equal PSNR, the images obtained from the 16B system are sharper and have less high-frequency background noise than those from the 7B system. One possible explanation for this occurrence is the fact that PSNR is not very sensitive to noise in the HFS because of their low energy content. On the other

hand, these sub-images contain significant edge information and if not quantized efficiently, introduce ringing and high-frequency fuzziness. High-frequency sub-images are encoded more efficiently by the 16B system than by the 7B system. Thus, for a given PSNR, the 16-band decomposition results in an improvement in both the encoding rate and the quality of the reconstructed imagery. All simulations from this point forward, assume the use of a 16-band decomposition.

Recently, Westerink, Biemond and Boekee [66] analyzed the use of QMF's of different lengths on aliasing distortions in a subband image coding application using Johnston's filters and scalar quantization. Comparisons were made at encoding rates of 0.8 and 0.6 bpp. They concluded that from both a MSE calculation and a subjective judgment that aliasing errors can be neglected for filter lengths of 12 taps or more. They also reported that at encoding rates more that than or equal to 0.8 bpp, the effect of aliasing distortions in image subbands is negligible. The encoder in [66] uses Lloyd-Max quantizers to encode the subbands and does not exploit any correlation in the LFS.

Our subband coding system is significantly different than the system in [66]. As a result, their conclusions might not be valid for our system. We simulated our system using Johnston's 8, 16, 24, and 32-tap filters and the results for encoding the 512×512 "Lenna" image for "desired" rates of 0.5 and 0.25 bpp are shown in Table 4.1. The obtained rates are different from the desired rates specified in the

Table 4.1: Performance of Johnston's filters for encoding the "Lenna" image.

8-tap		16-tap		24-tap		32-tap	
Obtained Rate	PSNR	Obtained Rate	PSNR	Obtained Rate	PSNR	Obtained Rate	PSNR
0.49	33.34	0.48	35.26	0.47	36.44	0.47	36.70
0.27	31.68	0.27	33.06	0.27	33.87	0.27	34.01

rate allocation procedure because of the entropy-constrained design of the TCQ systems.

It is obvious from Table 4.1 that the PSNR performance of the 32 and 24-tap systems are superior to the 16 and 8-tap systems. Recall from our discussion in the previous section that a QMF bank using Johnston's filters is not a perfectly reconstructing filter bank even in the absence of quantization distortions. This distortion (QMF distortion) gets bigger as the filter length becomes smaller. QMF distortion is typically not perceptible in the reconstructed image. Hence, the fact that the PSNR of the reconstructed image from the 32-tap system is higher does not necessarily translate to higher quality reconstructed imagery.

Subjective tests revealed that, at both the encoding rates, encoded images obtained from the 8-tap system were the best. At approximately 0.5 bpp, the encoded images from all systems were of extremely high-quality. However, higher-order systems seem to create "ringing" near the edges. At 0.27 bpp, these Gibbs phenomena type errors affected the quality of the encoded images considerably when the long

Table 4.2: Wavelet coding results for encoding the monochrome "Lenna" image.

Designed rate	Obtained rate	PSNR
1.00	0.93	39.85
0.50	0.48	36.61
0.25	0.27	33.77

filters were employed. In smooth regions, all four encoded images seem to have approximately the same amount of perceptible distortion.

We also implemented our subband coder using 9-tap spline filters. Simulation results for encoding the 512 × 512 "Lenna" image are shown in Table 4.2 for "desired" rates of 1.0, 0.5, and 0.25 bpp. Comparing results in Tables 4.2 and 4.1, it is obvious that the performance of spline filters is comparable to the performance of the 24-tap system while it is uniformly better than the performance of 16 and 8-tap systems.

A subjective evaluation of the encoded images revealed that the encoded images obtained from the system employing 9-tap spline filters are marginally better than those obtained from the 8-tap system . Note that this improvement is obtained at the expense of additional computational complexity. For a two-band decomposition of one-dimensional data, generating one output sample from the low-pass and high-pass analysis filters requires 8 multiplies and 8 adds when using 8-tap Johnston's filters (by using the polyphase decomposition technique suggested in [67]) while spline filters require 9 multiplies and 14 adds. This improvement in the subjective

quality could be attributed to the regularity and differentiability of scaling functions associated with the spline filters [53]. More details regarding the importance of regularity and differentiability of scaling functions in an image coding application can be found in [53]. Due to their superior performance, all results from this point forward assume use of the spline filters for the wavelet decomposition.

The subjective quality of the encoded images from our wavelet coder is excellent. The encoded image at an average rate of 0.93 bpp is almost indistinguishable from the original image and the encoded image at 0.48 bpp is extremely good with very little high-frequency background noise or smoothing. There are no visible artifacts even when viewed on a high-resolution monitor. This image is shown in Figure A.5. The encoded image at 0.27 bpp is quite natural looking but has some perceptual distortion.

4.5.2 Comparisons

In comparison with other results from the literature, we find that both our transform and wavelet coder are quite competitive. Simulation results for our transform and wavelet coders are shown in Figure 4.5 along with other results from the literature [68]. For encoding the monochrome version of the 512×512 "Lenna" image at 0.48 bpp, our PSNR value of 36.61 dB is higher than those of the entropy-constrained scalar quantization based subband coder of Tanabe and Farvardin [6] (35.32 dB at 0.45 bpp), the adaptive entropy-constrained TCQ based transform

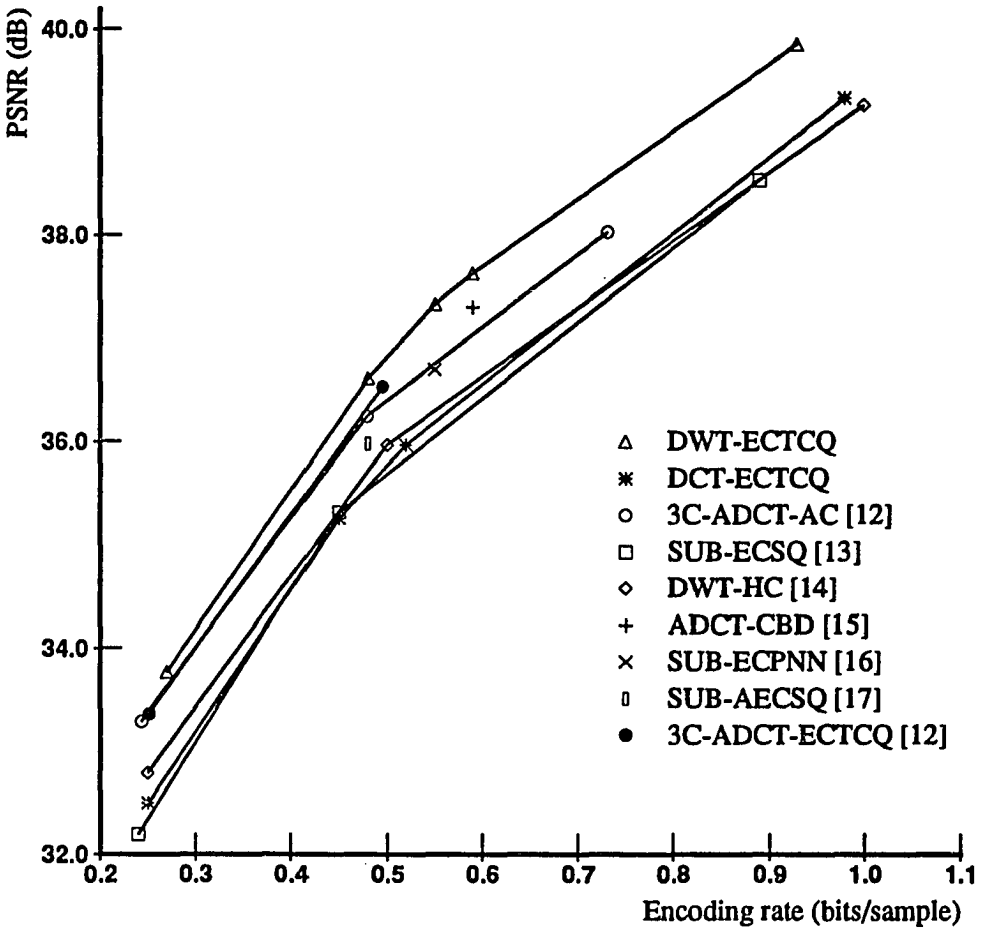

Figure 4.5: Comparisons for encoding the 512 × 512 "Lenna" image.

coder proposed in [69] (36.53 dB at 0.50 bpp), the adaptive entropy coded subband coder of Kim and Modestino [70] (35.98 dB at 0.48 bpp), and the hierarchical image coder of Shapiro [71] (35.97 dB at 0.50 bpp). In [72] and [3], PSNR's of 36.70 dB and 37.3 dB were reported at 0.55 and 0.59 bpp, respectively. When our system is simulated at these rates, PSNR values of 37.33 dB and 37.63 dB are obtained, for improvements of 0.63 and 0.33 dB, respectively.

We also conclude that the performance of our transform coder is quite competitive. Results of [73],[3], and [72] are superior to our transform coder. It should be pointed out however, that all three systems cited above are of high complexity, while the computational burden of our transform coder is fairly low. Unlike [73] and [3], we do not do any kind of intraframe adaptation and the quantization portion of our system requires only a small amount of computation [18].

4.5.3 Color image coding

As in the transform coding case, the RGB planes are transformed into NTSC transmission primaries (Y, I and Q). Each NTSC component is decomposed into 16 equal-sized subbands. The LFS for each component is encoded using a 2-D DCT encoder with a block size of 4×4. The subband coder proposed in [43] discards all HFS associated with the I and Q components while all subbands of the Y component are encoded. At the decoder, the chrominance components are restored to their original size by interpolation.

We investigated the significance of HFS associated with the chrominance components both perceptually and in a MSE sense by implementing our wavelet coder in the following manner:

1. All high-frequency sub-images of I and Q components were encoded (48B),

2. All high-frequency sub-images of I and Q components were discarded (18B).

The performance of the two systems is approximately equal (both objectively and subjectively) at low encoding rates ($\approx$ 0.25 bpp). This is as expected since even for the 48B system, the HFS receive zero encoding rates from the rate allocation algorithm because of their very low variance. At high rates ($\approx$ 1.0 bit/pixel), discarding the HFS associated with the chrominance components causes a significant drop in PSNR and effects the subjective quality of the encoded images in an interesting way. Without side-by-side comparison with the original, the encoded image from the 18B system looks extremely good. However, careful comparison with the original reveals that colors have a lighter, or "washed out" appearance. The 48B system does not suffer from this effect. Hence, in the simulations discussed below, the high-frequency subbands associated with the chrominance components were not discarded.

Simulation results are presented in Table 4.3 for encoding the color version of the 512×512 "Lenna" image at three different encoding rates. For comparison, the performance of our ECTCQ/DCT coder is also shown. It is evident from this table that the PSNR performance of the wavelet coder is superior at all encoding rates. This comparison is not completely fair because the encoder in our ECTCQ/DCT system subsamples the chrominance components by a factor of 2 in each direction before quantization.

Table 4.3: Wavelet coding results for encoding the color "Lenna" image.

System		Design bit rates		
		0.25	0.5	1.0
ECTCQ based wavelet coder	Obtained bit rate	0.24	0.47	1.13
	PSNR	33.44	35.72	39.20
ECTCQ based transform coder	Obtained bit rate	0.25	0.49	1.0
	PSNR	30.98	33.01	34.82

The subjective quality of the encoded images at all three rates is extremely good. In particular, the encoded image at 1.13 bpp is indistinguishable from the original. The encoded image at 0.47 bpp is extremely sharp and devoid of any annoying artifacts. Fuzziness and high-frequency background noise are totally absent even at an average encoding rate of 0.24 bpp. The encoded color "Lenna" image at 0.47 bpp is shown in Figure A.6.

4.6 Perceptual Weighting

One of the most important objectives of an image coding system is to encode images in such a manner that coding distortions are not perceptible. To compress an image such that a human observer cannot perceive coding distortions is not an easy task. One must understand the psychophysics of the human visual system (HVS) very well to achieve this. It is known that the sensitivity of the human eye to perceive distortion is different for different spatial frequencies [74]. We make use

of this information in our subband coder by perceptually weighting each subband according to the sensitivity of the human eye to the energy in that subband. We follow the ideas of Perkins and Lookabaugh [75] for calculating these weighting coefficients and modify the previously discussed rate allocation algorithm to incorporate these coefficients. That is, instead of appropriately allocating among the data sequences to be encoded so as to minimize MSE given by Equation (4.23), the weighted MSE (WMSE), defined as

$$\text{WMSE} = \sum_{i=1}^{K} \alpha_i w_i p_i E_i(r_i), \tag{4.27}$$

is minimized subject to the constraint of Equation (4.24) where p_i is a perceptual weighting coefficient for the i^{th} band.

4.6.1 Calculation of weighting coefficients

To determine the contrast sensitivity of the HVS, experiments using sinusoidal grating patterns were performed in [76]. Sinusoidal gratings are bars of fundamental frequency f_0 displayed against a background intensity B_0 and oriented at an angle of θ to the vertical axis. It can be represented as an intensity pattern by

$$B(x, y) = B_0 + k \cos(2\pi f_0(x \cos\theta - y \sin\theta))$$

where k is the modulation level. For each f_0 with $\theta = 0°$ and a fixed B_0, k was adjusted until the sinusoidal pattern becomes just visible. It was found that for a fixed value of f_0, the visibility threshold depends only on the ratio k/B_0 and not

separately on k and B_0. The contrast of the grating is defined as the ratio k/B_0. Threshold contrast is the contrast below which the grating is not visible. It is traditional to report the visibility threshold as contrast sensitivity (CS), where CS is the reciprocal of the visibility threshold. A typical plot of relative CS (normalized with the maximum value of CS) as a function of f_0 in cycles/degree is shown in Figure 4.6 for a vertical sinusoidal grating ($\theta = 0°$). Many such curves exist in the literature [77]. The important result from these studies is that CS increases linearly from low-spatial frequencies until it reaches a maximum after which it falls of rapidly with increasing frequencies. A closed form expression for relative CS (RCS) obtained by Campbell and Robson in [76] and used in [75] is given by

$$\text{RCS} = 0.6033c^{1.01}10^{-1.18(\log_{10} c)^2} \tag{4.28}$$

where c is the frequency in cycles/degree. CS also depends on the orientation angle of the sinusoidal grating. Campbell, Kulikowski and Levinson [78] investigated the effect of orientation on the visual resolution of gratings and concluded that CS is maximum for horizontal and vertical gratings and decreases with the angle from either axis, to approximately 3 dB at an angle of 45. This conclusion was also found to be valid for the HVS [79]. [75] modeled this angular dependence using

$$S_A(\theta) = -0.0222\theta - 0.001\theta^2 \ , \ 0° \leq \theta \leq 45°$$

where θ is the orientation angle in degrees and S is the angular sensitivity in dB.

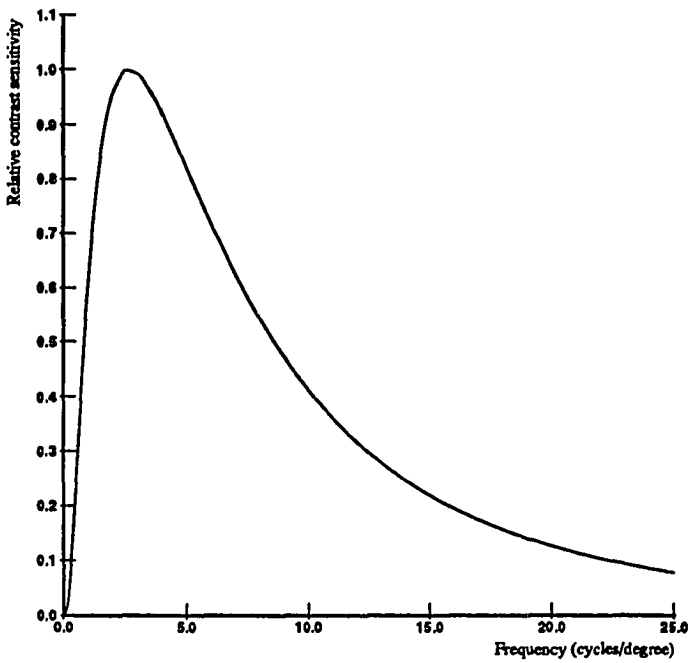

Figure 4.6: Campbell and Robson's relative contrast sensitivity function.

The weighting factor for each subband is calculated using their corresponding center frequency. Consider a discrete, real-valued image of size $M \times N$. Let the display device have a pixel size of $\Delta x \times \Delta y$. The Fourier coefficient $X(k,l)$ $(0 \leq k \leq M/2, 0 \leq l \leq N/2)$ corresponds to a cosine function with wavelength

$$\lambda = \frac{1}{\sqrt{f_x^2 + f_y^2}} = \frac{1}{\sqrt{(\frac{k}{M\Delta x})^2 + (\frac{l}{N\Delta y})^2}}$$

and

$$\theta = \tan^{-1}\frac{f_x}{f_y} = \tan^{-1}\frac{Ml\Delta x}{Nk\Delta y},\ 0 \leq \theta \leq 45$$

where f_x and f_y are the horizontal and vertical orientation frequencies. The algorithm to calculate the weighting coefficients is summarized below.

1. Calculate the wavelength λ and θ corresponding to the center frequency of each subband.

2. $c = \frac{2d}{\lambda}\tan(1/2)$ where d is the distance between the observer and the monitor.
3. $S_\lambda(x) = -4.47x^2 + 8.86x - 4.39$ where $x = \log_e(c)$.
4. $S_A(\theta) = -0.0222\theta - 0.001\theta^2$.
5. $S = S_\lambda(x) + S_A(\theta)$.
6. $p = 10^{S/20}$.

On the monitor used to view images, a 512×512 image is of size 128 mm $\times$ 128 mm. Hence $\Delta x = \Delta y = 0.25$ mm. We set the viewing distance to be 4 times the size of the image (i.e., $d = 512$ mm). We investigated the choice of weighting coefficients associated with the DCT bands by computing them in three different ways. They are

1. All DCT bands were given the same weighting coefficient, the one corresponding to the center frequency of the LFS.
2. Weighting coefficients for the DCT bands were found using the fact that a 4×4 block is of size 1 mm $\times$ 1 mm. For a viewing distance of d, the observers eye subtends an angle α given by (see Figure 4.7)

$$\tan(\frac{\alpha}{2}) = \frac{1/2}{512} = \frac{1}{1024}.$$

Hence, $\alpha = 2\tan^{-1}(1/1024) = 0.1119°$. The function $\cos(\frac{\pi k(2m+1)}{2N})$ will complete k cycles in N samples (and hence in $\alpha°$). As a result, the frequency in cycles/degree is given by $c_k = k/\alpha = 8.9366k$. Similarly, the value of c corresponding to the 2-D DCT coefficient $X_{\text{dct}}(k, l)$ is given by

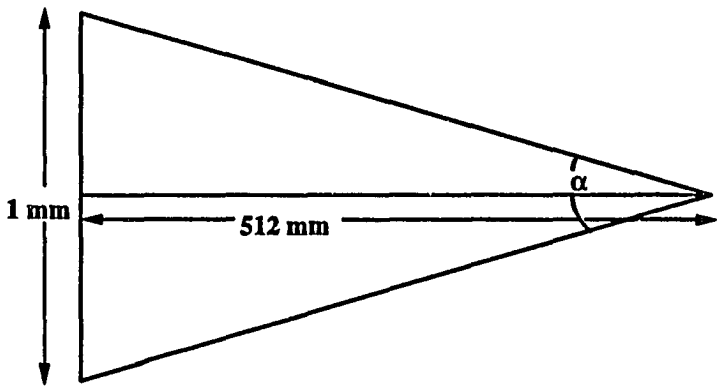

Figure 4.7: Figure used to obtain frequencies in cycles/degree for the DCT bands.

$c = \sqrt{(8.9366k)^2 + (8.9366l)^2} = 8.9366\sqrt{k^2 + l^2}$ while the corresponding angle is given by $\theta = \tan^{-1} l/k$ [73].

3. Frequencies obtained from method (2) were scaled in such a manner that the value of c corresponding to the highest-frequency DCT band is the same as the one corresponding to the highest-frequency DFT coefficient ($X(127, 127)$) in the LFS. The weighting coefficients corresponding to the scaled c's were calculated as before. Note that this scaling will not affect the calculation of θ.

The DC band was always given a weight of 1 because it determines the average intensity of the block. Subjective tests revealed that the encoded images obtained from using weighting coefficients from method (3) appear the best. We obtained PSNR values of 36.37 dB and 33.47 dB at encoding rates of 0.53 and 0.28 bpp, respectively from such a system. Comparing these results with those in Table 4.2, it is evident that perceptual weighting result in a small drop in PSNR values for

approximately the same encoding rate. At approximately 0.5 bpp, it was very difficult to identify any improvement in perceptual quality due to weighting. This is not very surprising as the encoded images are of extremely high quality. However, at an encoding rate of approximately 0.25 bpp, the encoded image from the perceptually weighted system looks better. This improvement in subjective quality gets bigger as the encoding rates become even smaller.

A primary effect of perceptual weighting is to emphasize low frequencies with respect to high frequencies. An image like "Lenna" has little high-frequency content. As a result, the effect of perceptual weighting might be exaggerated. We encoded the "baboon" image and an "Urban" image, both of which have significant high-frequency content, with and without perceptual weighting for "desired" encoding rates of 0.5 and 0.25 bpp. As before, the effect of weighting was very difficult to perceive at 0.5 bpp. However at approximately 0.25 bpp, encoded images from the perceptually weighted system were significantly better. The encoded version of the "baboon" image from the system with no perceptual weighting seem to suffer from "blocking" effects which was totally absent in the weighted version.

Mannos and Sakrison [77] investigated the use of different contrast sensitivity functions in an image coding application and concluded that the function whose closed-form expression given by

$$\mathrm{RCS} = 2.6[0.0192 + 0.114c]e^{-(0.114c)^{1.1}} \tag{4.29}$$

performs the best. This function has a peak at $c = 8$ cycles/degree while Campbell and Robson's [76] RCS function has a peak between 2 to 4 cycles/degree. We utilized the RCS function given in Equation (4.29) to calculate the weighting coefficients and simulated our system using these new coefficients. Subjective performance results revealed that the system utilizing RCS function in Equation (4.28) performed better. This contradiction could be due to the difference in viewing conditions as [77] made all its subjective judgment using photographs.

4.7 Adaptive threshold coding

None of the schemes we have implemented so far do any adaptation other than the bit allocation. It is very well known that statistics of a typical image change from one image block to another. Also, the required fidelity in reproduction for the human viewer changes from pixel to pixel. Visual sensitivity of human observers decreases at and adjacent to large luminance changes i.e., luminance edges. Hence it is highly desirable to design encoding techniques which are capable of adapting to local properties of an image according to the sensitivity of a human observer to perceive distortion.

Visual masking is defined as the action of one visual stimulus on the visibility of another [80]. Studying the effects of visual masking have been an active area of research. Many spatial domain image coders have sprung up in the recent past which make use of visual masking effect of the HVS [80] –[84]. [80] proposed

different masking functions and subsequently used them for encoding images using DPCM. Recently, Safranek and Johnston [85] proposed a visual masking function than can be used with a subband coder. This masking function was obtained from extensive psychovisual experiments for a 16-band decomposition. Using this masking function, they were able to obtain an analytical expression to obtain an estimate of the just-perceptible distortion for each pixel in a given subband. They used DPCM to encode the image subbands.

Subsequently, Baseri and Mathews [86] and Harris and Modestino [87] used the visual masking function proposed in [85] and quantized the image subbands using vector quantization and entropy-constrained scalar quantization, respectively. In this work, we propose an image coder which makes use of the masking function proposed in [85] to adaptively encode image subbands using ECTCQ.

4.7.1 Visual masking function

The visual masking function associated with the $(i,j)^{\text{th}}$ pixel in the k^{th} subband is defined as [85]

$$\text{pt}(i,j,k) = \text{Base}(k)\text{Texenergy}(i,j)^{0.034}\text{Brightcorr}(i,j) \tag{4.30}$$

where Base(k) is the root-mean-squared noise sensitivity threshold for the k^{th} subband which was empirically measured and reported in [85].

It is very well known that humans are more sensitive to noise in the smooth regions than in textured regions. To make use of this property of the HVS, [85]

introduced a texture masking adjustment term which is a function of the "texture energy" at each pixel. Texture energy at the $(i,j)^{\text{th}}$ pixel is evaluated using

$$\begin{aligned}\text{Texenergy}(i,j) &= \sum_{m=1}^{15} \text{RCS}(m)\text{Energy}(i,j,m) \\ &+ \text{RCS}(0)\text{Variance}((i,j),(i+1,j),(i,j+1),(i+1,j+1))\end{aligned}$$

where RCS() is the relative contrast sensitivity function, $\text{Energy}(i,j,m)$ is the squared amplitude of the $(i,j)^{\text{th}}$ pixel in the m^{th} subband and Variance() is the variance of the 4 pixel intensities in the LFS whose locations are given in the above equation.

Noise sensitivity experiments were done in [85] for an uniform background grey level of 127. As explained earlier, the ability of humans to perceive distortion is dependent on the image brightness. For example, noise in very dark areas of an image is less visible than in other parts. A brightness correction factor is included in Equation (4.30) to compensate for this dependence. This brightness correction factor was measured and presented in [85]. We have reproduced their curve by reading the coordinates off the graph and then obtaining a closed form expression using a polynomial fit. The original data along with the polynomial fit are shown in Figure 4.8.

Note that the encoder in [85] employs a separable Generalized QMF bank for the subband decomposition while we use biorthogonal wavelet filter banks. Hence, a QMF correction factor should be included in Equation (4.30). However, to find

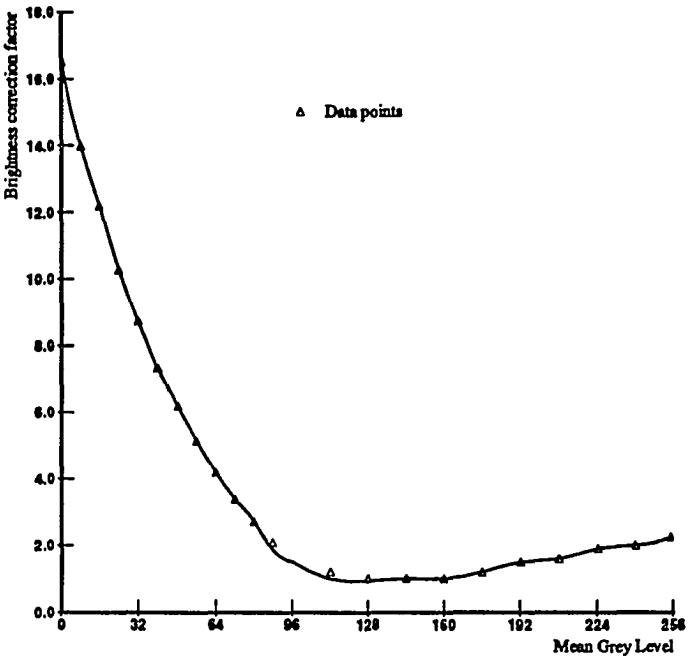

Figure 4.8: Brightness correction curve.

such a factor requires extensive psychovisual studies and hence, is beyond the scope of this work.

4.7.2 Image coding

The input image is decomposed into 16 equal-sized subbands. As before, the LFS is coded using a 2-D DCT encoder. An estimate of the just-perceptible distortion (perceptual threshold) is obtained for each pixel in a subband using Equation (4.30). It is obvious that if the pixel amplitude is less than the perceptible threshold, we need not encode that particular sample. The amplitude of that pixel can be made zero and the resulting distortion would still be imperceptible.

Table 4.4: Sample selection for encoding the 512 × 512 "Lenna" image.

Subband	Number of Samples less than perceptual threshold	Number of blocks	Number of samples encoded
LL-LL	58	1024	16384
LL-LH	8981	976	15616
LL-HL	5883	1012	16192
LL-HH	12723	765	12240
LH-LL	16293	62	992
LH-LH	16195	99	1584
LH-HL	16350	16	256
LH-HH	16358	20	320
HL-LL	15919	216	3456
HL-LH	16346	28	448
HL-HL	16229	87	1392
HL-HH	16370	11	176
HH-LL	16384	0	0
HH-LH	16384	0	0
HH-HL	16384	0	0
HH-HH	16384	0	0

Table 4.4 shows the number of samples in each subband which are below the perceptual threshold for the 512 × 512 "Lenna" image. From this table, it is apparent that there exist subbands which need not be encoded. A side information of 1 bit/subband is required to notify which subbands are to be coded.

Each subband is of size 128 × 128 if the original image is of size 512 × 512. Hence, if we are to directly transmit the location of each pixel in each subband which we wish to encode, it would require 14 bits/sample. This would require a huge amount of side information. To overcome this deficiency, we split each HFS into 4 × 4 sub-blocks. If in a sub-block there exist a sample whose magnitude is

greater than the perceptual threshold associated with that sample, then all the samples in that sub-block are encoded. We need to transmit which sub-blocks are encoded as a side information. Since there are 1024 blocks in each sub-image, it would require 10 bits to transmit the location of each-subblock directly.

Table 4.4 shows the number of blocks that needs be encoded in each subband. Since there are 1024 block in each subband, we would need atmost 1024 bits to transmit this information. However, if the number of blocks is less than or equal to 102, then it is cheaper to transmit the location of sub-blocks that are encoded directly as it requires less than or equal to $102 \times 10 = 1020$ bits/subimage.

Sometimes, it is cheaper to transmit which subblocks are zero-subblocks (sub-blocks in which all elements are set to zero). For example, from Table 4.4, it would take 1024 bits of side information for the LL-LH subband is we are to transmit the locations of the 976 blocks that are to be encoded while it would take only 480 bits to inform which blocks are zero-subblocks. In the later case, we would also need 1 extra bit/subimage to notify if the transmitted index locations are for the zero-blocks. Total side information, including those for the mean, variances and the initial trellis-states, amount to 6423 bits which is approximately 0.03 bpp.

We use an image-dependent distortion measure which is a function of the perceptual threshold at a given pixel in each subband. Let x be the amplitude of the $(i,j)^{\text{th}}$ pixel of the k^{th} subband with $\text{pt}(i,j,k)$ its corresponding perceptual threshold. Perceptual distortion $\text{pd}(x,y)$ between x and y, the encoded version of x is

given by

$$\mathrm{pd}(x,y) = \{|x-y|^2 - \mathrm{pt}^2(i,j,k)\}\mathrm{U}(|x-y| - \mathrm{pt}(i,j,k)) \qquad (4.31)$$

where U() is the unit-step function. Note that the perceptual distortion is zero if the magnitude of the quantization error is less than than the perceptual threshold. When $\mathrm{pt}(i,j,k) = 0$, the perceptual distortion $\mathrm{pd}(x,y)$ is equal to the MSE between x and y. Such a distortion measure has already been used for image coding applications in [81] and [87].

Simulations were performed for encoding the 512×512 "Lenna" image. Since our intention is to minimize the amount of perceptible distortion at a given encoding rate, we have chosen not to report PSNR results for this image coder. The encoded image obtained from the adaptive threshold coder at 0.35 bpp was comparable to the encoded image obtained from our non-adaptive wavelet coder at an encoding rate of 0.5 bpp. This is consistent with the results of Safranek and Johnston [85].

CHAPTER 5

SUMMARY

In Chapter 1, a digital image communication system model was presented and the image coding problem was defined. Chapter 2 was devoted to reviewing the concept of Trellis Coded Quantization (TCQ). It was shown that, by employing entropy-constrained TCQ (ECTCQ), near-optimal performance (in a rate-distortion theory sense) can be achieved for encoding memoryless sources at all encoding rates.

We investigated the application of transform coding to encode monochrome and color images using TCQ in Chapter 3. Specifically, TCQ was used to encode transform coefficients resulting from applying a 16×16 discrete cosine transform (DCT) to 8-bit gray level and 24-bit color images. For the color images, the red, green, and the blue planes were transformed into NTSC transmission primaries (Y, I, and Q) before the DCT was applied. Both fixed-rate and entropy-constrained systems were considered. It was shown that entropy-constrained designs result in a low-complexity system with objective performance rivalling that of many coding schemes from the literature.

Chapter 4 dealt with image coding schemes based on ECTCQ and discrete wavelet transform (DWT). In this chapter, the concept of DWT was briefly reviewed. The relationship between the classical subband filter banks and the wavelet filter banks were also discussed. Performance comparisons were made using a classical quadrature mirror filter bank and 9-tap spline filters that were built using biorthogonal wavelet bases. It was concluded that the encoded images obtained from the system employing 9-tap spline filters appear the best although at the expense of additional computational burden.

The subjective quality of the encoded images from the system employing 9-tap spline filters was excellent. In particular, the encoded images at 0.5 bpp was extremely good with no visible artifacts even when viewed on a high-resolution monitor. The objective performance results of our wavelet coder were shown to be comparable to or surpass all previous results reported in the literature.

It is known that the sensitivity of the human eye to perceive distortion is different for different spatial frequencies. We made use of this information in our wavelet coder by perceptually weighting each subband according to the sensitivity of the human eye to the energy in that subband. It is demonstrated that perceptual weighting improves the quality of the reconstructed imagery.

Finally, we implemented an adaptive wavelet coder that makes use of visual masking functions to estimate the just-perceptible distortion for each pixel in a

subband. Using these estimates, an image-dependent distortion measure was introduced and subsequently used for coding image subbands using ECTCQ. The encoded image obtained from this adaptive coder at 0.35 bpp was found to be comparable to the encoded image obtained from our non-adaptive wavelet coder at an encoding rate of 0.5 bpp. The reconstructed image obtained from the adaptive system at approximately 0.5 bpp was almost indistinguishable from the original.

Appendix A

PHOTOGRAPHS

Figure A.1: Monochrome "Lenna" image (512 × 512).

Figure A.2: Encoded image from transform coder (35.97 dB at 0.52 bpp).

Figure A.3: Color "Lenna" image (512 × 512).

Figure A.4: Encoded color image from transform coder (33.01 dB at 0.49 bpp).

Figure A.5: Encoded image from wavelet coder (36.61 dB at 0.48 bpp).

Figure A.6: Encoded color image from wavelet coder (35.72 dB at 0.47 bpp).

REFERENCES

[1] A. K. Jain, "Image data compression: A Review," *Proc. IEEE* (Invited Paper), vol. 69, pp. 349–389, Mar. 1981.

[2] G. K. Wallace, "The JPEG still-picture compression standard," *Communications of the ACM*, vol. 34, pp. 11–44, Apr. 1991.

[3] W. A. Pearlman, "Adaptive cosine transform image coding with constant block distortion," *IEEE Trans. Commun.*, vol. 38, pp. 698–703, May 1990.

[4] M. Kunt, A. Ikonomopoulos, and M. Kocher, "Second-generation image coding techniques," *Proc. IEEE* (Invited Paper), vol. 73, pp. 549–574, Apr. 1985.

[5] J. W. Woods and S. D. O'Neil, "Subband coding of images," *IEEE Trans. Acoust., Speech, and Signal Proc.*, vol. ASSP-34, pp. 1278–1288, Oct. 1986.

[6] N. Tanabe and N. Farvardin, "Subband image coding using entropy-coded quantization over noisy channels," *IEEE J. Select. Areas in Commun.*, vol. 10, pp. 926–943, June 1992.

[7] A. Gersho, "On the structure of vector quantizers," *IEEE Trans. Inform. Th.*, vol. IT-28, pp. 157–166, Mar. 1982.

[8] N. S. Jayant and P. Noll, *Digital Coding of Waveforms.* Englewood Cliffs, NJ: Prentice Hall, 1984.

[9] M. Mansuripur, *Introduction to Information Theory.* Englewood Cliffs, NJ: Prentice Hall, 1987.

[10] I. H. Witten, R. M. Neal, and J. G. Cleary, "Arithmetic coding for data compression," *Communications of the ACM*, vol. 30, pp. 520–540, June 1987.

[11] H. Gish and J. N. Pierce, "Asymptotically efficient quantizing," *IEEE Trans. Inform. Th.*, vol. IT-14, pp. 676–683, Sep. 1968.

[12] N. Farvardin and J. W. Modestino, "Optimum quantizer performance for a class of non-Gaussian memoryless sources," *IEEE Trans. Inform. Th.*, vol. 30, pp. 485–497, May 1984.

[13] T. Berger, *Rate Distortion Theory: A Mathematical Basis for Data Compression.* Englewood Cliffs, NJ: Prentice Hall, 1971.

[14] T. R. Fischer, "Geometric source coding and vector quantization," *IEEE Trans. Inform. Th.*, vol. 35, pp. 137–145, Jan. 1989.

[15] A. Gersho and R. M. Gray, *Vector quantization and signal compression.* Boston, MA: Kluwer Academic Press, 1992.

[16] R. Laroia, "Design and Analysis of a fixed-rate structured vector quantizer derived from variable-length scalar quantizers," Ph.D. dissertation, University of Maryland, College Park, MD, May 1992.

[17] M. V. Eyuboğlu and G. D. Forney, Jr., "Lattice and trellis quantization with lattice- and trellis-bounded codebooks — high rate theory for memoryless sources," *IEEE Trans. Inform. Th.*, vol. 39, pp. 46–59, Jan. 1993.

[18] M. W. Marcellin and T. R. Fischer, "Trellis coded quantization of memoryless and Gauss-Markov sources," *IEEE Trans. Commun.*, vol. COM-38, pp. 82–93, Jan. 1990.

[19] T. R. Fischer, "A pyramid vector quantizer," *IEEE Trans. Inform. Th.*, vol. 32, pp. 568–583, July 1986.

[20] T. R. Fischer and M. Wang, "Entropy-constrained trellis-coded quantization," *IEEE Trans. Inform. Th.*, vol. IT-38, pp. 415–426, Mar. 1992.

[21] M. W. Marcellin, "On entropy-constrained trellis coded quantization," *IEEE Trans. Commun.*, to appear.

[22] G. Ungerboeck, "Channel coding with multilevel/phase signals," *IEEE Trans. Inform. Th.*, vol. IT-28, pp. 55–67, Jan. 1982.

[23] G. D. Forney, Jr., "The Viterbi algorithm," *Proc. IEEE* (Invited Paper), vol. 61, pp. 268–278, Mar. 1973.

[24] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, pp. 84–95, Jan. 1980.

[25] P. A. Chou, T. Lookabaugh, and R. M. Gray, "Entropy-constrained vector quantization," *IEEE Trans. Acoust., Speech, and Signal Proc.*, vol. ASSP-37, pp. 31–42, Jan. 1989.

[26] M. W. Marcellin, P. Sriram, and K.-L. Tong, "Transform coding of monochrome and color images using trellis coded quantization," *IEEE Trans. Circuits Syst. Video Tech.*, to appear.

[27] W.-H. Chen and C. H. Smith, "Adaptive coding of monochrome and color images," *IEEE Trans. Commun.*, vol. COM-25, pp. 1285–92, Nov. 1977.

[28] W.-H. Chen and W. K. Pratt, "Scene adaptive coder," *IEEE Trans. Commun.*, vol. COM-32, pp. 225–232, Mar. 1984.

[29] W. A. Pearlman, P. Jakatdar, and M. M. Leung, "Adaptive transform tree coding of images," *IEEE J. Select. Areas in Commun.*, vol. 10, pp. 902–912, June 1992.

[30] R. J. Clarke, *Transform Coding of Images.* Orlando, FL: Academic Press, 1985.

[31] W.-H. Chen, C. H. Smith, and S. Fralick, "A fast computational algorithm for the discrete cosine transform," *IEEE Trans. Commun.*, vol. COM-25, pp. 1004–1009, Sep. 1977.

[32] A. Habibi, "Survey of adaptive coding techniques," *IEEE Trans. Commun.*, vol. 25, pp. 1275–1284, Nov. 1977.

[33] J. Vaisey and A. Gersho, "Image compression with variable block size segmentation," *IEEE Trans. Signal Proc.*, vol. 40, pp. 2040–2060, Aug. 1992.

[34] K.-L. Tong, "Transform coding of images using fixed-rate and entropy-constrained trellis coded quantization," M.S. thesis, The University of Arizona, May 1991.

[35] Y. Shoham and A. Gersho, "Efficient bit allocation for an arbitrary set of quantizers," *IEEE Trans. Acoust., Speech, and Signal Proc.*, vol. ASSP-36, pp. 1445–1453, Sep. 1988.

[36] P. Sriram and M. W. Marcellin, "Image coding using wavelet transforms and entropy-constrained trellis coded quantization," submitted to *IEEE Trans. Image Proc.*

[37] W. Frei and B. Baxter, "Rate-distortion coding simulation for color images," *IEEE Trans. Commun.*, vol. 25, pp. 1385–1392, Nov. 1977.

[38] W. K. Pratt, "Spatial transform coding of color images," *IEEE Trans. Commun.*, vol. 19, pp. 980–992, Dec. 1971.

[39] R. E. Crochiere, S. A. Webber, and J. L. Flanagan, "Digital coding of speech in subbands," *Bell Syst. Tech. J.*, vol. 55, pp. 1069–1085, Oct. 1976.

[40] A. Croisier, D. Esteban, and C. Galand, "Perfect channel splitting by use of interpolation/decimation/tree decomposition techniques," in *Conf. Proceedings,* 1976 IEEE Int. Conf. on Inform. Sci. Syst., Patras, Greece, May. 1976.

[41] M. Vetterli, "Multi-dimensional subband coding: Some theory and algorithms," *Signal Processing*, vol. 6, pp. 97–112, Apr. 1984.

[42] J. D. Johnston, "A filter family designed for use in quadrature mirror filter banks," in *Conf. Proceedings,* 1980 Int. Conf. on Acoust., Speech, and Signal Proc., Denver, CO, Apr. 1980.

[43] H. Gharavi and A. Tabatabai, "Sub-band coding of monochrome and color images," *IEEE Trans. Circuits Syst.*, vol. CAS-35, pp. 207–214, Feb. 1988.

[44] P. H. Westerink, D. E. Boekee, J. Biemond, and J. W. Woods, "Subband coding of images using vector quantization," *IEEE Trans. Commun.*, vol. COM-36, pp. 713–719, June 1988.

[45] M. J. T. Smith and S. L. Eddins, "Analysis/Synthesis techniques for subband image coding," *IEEE Trans. Acoust., Speech, and Signal Proc.*, vol. ASSP-38, pp. 1446–1456, Aug. 1990.

[46] S. Nanda and W. A. Pearlman, "Tree coding of image subbands," *IEEE Trans. Image Proc.*, vol. IP-1, pp. 133–147, Apr. 1992.

[47] E. Simoncelli and E. H. Adelson, "Non-separable extensions of quadrature mirror filters to multiple dimensions," *Proc. IEEE*, vol. P-78, pp. 652–664, Apr. 1990.

[48] O. Rioul and M. Vetterli, "Wavelets and signal processing," *IEEE Signal Processing Mag.*, vol. 8, pp. 14–38, Oct. 1991.

[49] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE Trans. Pattern Anal. and Mach. Intel.*, vol. 11, pp. 674–693, Jul. 1989.

[50] S. G. Mallat, "Multifrequency channel decomposition of images and wavelet models," *IEEE Trans. Acoust., Speech, and Signal Proc.*, vol. ASSP-37, pp. 2091–2110, Dec. 1989.

[51] I. Daubechies, "Orthonormal bases of compactly supported wavelets," *Commun. Pure Appl. Math*, vol. 41, pp. 909–996, Nov. 1988.

[52] A. Grossmann and J. Morlet, "Decomposition of Hardy functions into square integrable wavelets of constant shape," *SIAM J. Math.*, vol. 15, pp. 723–736, July 1984.

[53] M. Antonini, M. Barlaud, P. Mathieu, and I. Daubechies, "Image coding using wavelet transform," *IEEE Trans. Image Proc.*, vol. IP-1, pp. 205–220, Apr. 1992.

[54] M. J. T. Smith and T. P. Barnwell, "Exact reconstruction techniques for tree-structured subband coders," *IEEE Trans. Acoust., Speech, and Signal Proc.*, vol. ASSP-34, pp. 434–441, June 1986.

[55] D. Esteban and C. Galand, "Application of quadrature mirror filters to split band voice coding schemes," in *Conf. Proceedings,* 1977 Int. Conf. on Acoust., Speech, and Signal Proc., Hartford, CT, May. 1977.

[56] V. K. Jain and R. E. Crochiere, "Quadrature mirror filter design in the time-domain," *IEEE Trans. Acoust., Speech, and Signal Proc.*, vol. ASSP-32, pp. 353–361, Apr. 1984.

[57] P. P. Vaidyanathan, "On power-complementary FIR filters," *IEEE Trans. Circuits Syst.*, vol. CAS-32, pp. 1308–1310, Dec. 1985.

[58] P. P. Vaidyanathan and P. Q. Hoang, "Lattice structure for optimal design and robust implementation of two-channel perfect-reconstruction QMF banks," *IEEE Trans. Acoust., Speech, and Signal Proc.*, vol. ASSP-36, pp. 81–94, Jan. 1988.

[59] T. Q. Nguyen and P. P. Vaidyanathan, "Two-channel perfect-reconstruction FIR QMF structures which yield linear-phase analysis and synthesis filters," *IEEE Trans. Acoust., Speech, and Signal Proc.*, vol. ASSP-37, pp. 676–690, May 1989.

[60] P. Sriram and M. W. Marcellin, "Wavelet coding of images using trellis coded quantization," in *Conf. Proceedings,* 1992 SPIE Conf. on Visual Commun. and Image Proc., Orlando, FL, Apr. 1992.

[61] A. Cohen, I. Daubechies, and J. C. Feauveau, "Biorthogonal bases of compactly supported wavelets," *Commun. Pure Appl. Math*, vol. 41, pp. 485–560, June 1992.

[62] M. Vetterli and C. Herley, "Wavelets and filter banks: Relationships and new results," in *Conf. Proceedings,* 1990 Int. Conf. on Acoust., Speech, and Signal Proc., Albuquerque, NM, Apr. 1990.

[63] S. A. Martucci, "Signal extension and noncausal filtering for subband coding of images," in *Conf. Proceedings,* 1991 SPIE Conf. on Visual Commun. and Image Proc., Boston, MA, Nov. 1991.

[64] J. W. Woods and T. Naveen, "A filter based bit allocation scheme for subband compression of HDTV," *IEEE Trans. Image Proc.*, vol. IP-1, pp. 436–440, Jul. 1992.

[65] P. H. Westerink, J. Biemond, and D. E. Boekee, "Evaluation of image subband coding schemes," in *Conf. Proceedings,* Europ. Signal Proc. Conf. , Grenoble, France, Sep. 1988.

[66] P. H. Westerink, J. Biemond, and D. E. Boekee, "Scalar quantization error analysis for image subband coding using QMF's," *IEEE Trans. Signal Proc.*, vol. SP-40, pp. 421–427, Feb. 1992.

[67] P. P. Vaidyanathan, "Multirate digital filters, filter banks, polyphase networks, and applications: A tutorial," *Proc. IEEE*, vol. P-78, pp. 56–93, Jan. 1990.

[68] P. Sriram and M. W. Marcellin, "Image coding for telecommunications," in *Conf. Proceedings,* 1993 Int. Phoenix Conf. on Comp. and Commun., Tempe, AZ, Mar. 1993.

[69] N. Farvardin, X. Ran, and C.-C. Lee, "Adaptive DCT coding of images using entropy-constrained trellis coded quantization," in *Conf. Proceedings,* 1993 Int. Conf. on Acoust., Speech, and Signal Proc., Minneapolis, MN, Apr. 1993.

[70] Y. H. Kim and J. W. Modestino, "Adaptive entropy coded subband coding of images," *IEEE Trans. Image Proc.*, vol. IP-1, pp. 31–48, Jan. 1992.

[71] J. M. Shapiro, "An embedded wavelet hierarchical image coder," in *Conf. Proceedings,* 1992 Int. Conf. on Acoust., Speech, and Signal Proc., San Francisco, CA, Mar. 1992.

[72] D. P. de Garrido, W. A. Pearlman, and W. A. Finamore, "Vector quantization of image pyramids with the ECPNN algorithm," in *Conf. Proceedings,* 1991 SPIE Conf. on Visual Commun. and Image Proc., Boston, MA, Nov. 1991.

[73] X. Ran and N. Farvardin, "Low bit-rate image coding using a three-component image model," Systems research center technical report TR 92-75, University of Maryland, College Park, MD, 1992.

[74] A. N. Netravali and B. G. Haskell, *Digital pictures, representation and compression.* NY: Plenum Press, 1988.

[75] M. G. Perkins and T. Lookabaugh, "A psychophysically justified bit allocation algorithm for subband image coding," in *Conf. Proceedings,* 1989 Int. Conf. on Acoust., Speech, and Signal Proc., May 1989.

[76] F. W. Campbell and J. G. Robson, "Application of Fourier analysis to the visibility of gratings," *J. Physiol,* vol. 197, pp. 551–566, 1968.

[77] J. L. Mannos and D. J. Sakrison, "The effects of a visual fidelity criterion on the encoding of images," *IEEE Trans. Inform. Th.*, vol. IT-20, pp. 525–536, July 1974.

[78] F. W. Campbell, J. J. Kulikowski, and J. Levinson, "The effect of orientation on the visual resolution of gratings," *J. Physiol,* vol. 187, pp. 427–436, 1966.

[79] F. W. Campbell and J. J. Kulikowski, "Orientation selectivity of the visual system," *J. Physiol*, vol. 187, pp. 437–445, 1966.

[80] A. N. Netravali and B. Prasada, "Adaptive quantization of picture signals using spatial masking," *Proc. IEEE.*, vol. 65, pp. 536–548, Apr. 1977.

[81] D. K. Sharma and A. N. Netravali, "Design of quantizers for DPCM coding of picture signals," *IEEE Trans. Commun.*, vol. COM-25, pp. 1267–1274, Nov. 1977.

[82] A. N. Netravali, "On quantizers for DPCM coding of picture signals," *IEEE Trans. Inform. Th.*, vol. IT-23, pp. 360–370, May 1977.

[83] C. B. Rubinstein and J. O. Limb, "On the design of quantizers for DPCM coders: influence of the subjective testing methodology," *IEEE Trans. Commun.*, vol. COM-26, pp. 565–572, May 1978.

[84] J. O. Limb and C. B. Rubinstein, "On the design of quantizers for DPCM coders: a functional relationship between visibility, probability and masking," *IEEE Trans. Commun.*, vol. COM-26, pp. 573–578, May 1978.

[85] R. J. Safranek and J. D. Johnston, "A perceptually tuned sub-band image coder with image dependent quantization and post-quantization data compression," in *Conf. Proceedings,* 1989 Int. Conf. on Acoust., Speech, and Signal Proc., Glasgow, Scotland, May 1989.

[86] R. Baseri and V. J. Mathews, "Vector quantization of images using visual masking functions," in *Conf. Proceedings,* 1992 Int. Conf. on Acoust., Speech, and Signal Proc., San Fransisco, CA, Mar. 1992.

[87] C. F. Harris and J. W. Modestino, "Entropy-constrained subband coding of images using a perceptual distortion criteria," in *Conf. Proceedings,* 1993 IEEE Int. Symp. on Info. Theory, San Antonio, TX, Jan. 1993.