USC–SIPI REPORT #226

Image Compression with Full Wavelet Transform (FWT)

by

Kwo-Jyr Wong and C.-C. Jay Kuo

December 1992

Signal and Image Processing Institute
UNIVERSITY OF SOUTHERN CALIFORNIA
Department of Electrical Engineering-Systems
3740 McClintock Avenue, Room 400
Los Angeles, CA 90089-2564 U.S.A.

Submitted to Journal of Visual Communication and Image Representation.

Image Compression with Full Wavelet Transform (FWT) *

Kwo-Jyr Wong † and C.-C. Jay Kuo †

December 9, 1992

EDICS 1.6

Abstract

Image compression based on a multiresolution approach has been intensively studied over the last ten years, including the Laplacian pyramid method by Burt and Adelson and the pyramidal wavelet transform (PWT) method by Mallat. In this research, we propose a modified wavelet transform known as the full wavelet transform (FWT) for image compression. By the FWT, we apply recursively the two-scale wavelet decomposition to all subimages so that an image is decomposed into blocks of the same size. It is shown experimentally that energy compaction is achieved in both the spatial and frequency domains via FWT, and can be effectively utilized to achieve high image compression ratio while preserving good image quality. Moreover, entropy coding is used for improve the overall performance. Numerical experiments show that our algorithm has a comparable performance with several existing methods. The relationship between the proposed method and other popular image compression methods such as DCT, PWT and SBC (subband coding) is also discussed.

1 Introduction

Image compression methods based on a multiresolution approach have received a lot of attention over the last ten years. The major advantage of the multiresolution approach is that it provides a graceful degradation between image quality and

*This work was supported by a National Science Foundation Young Investigator Award (ASC-9258396).

†The authors are with the Signal and Image Processing Institute and the Department of Electrical Engineering-Systems, University of Southern California, Los Angeles, California 90089-2564. E-mail: kwong@sipi.usc.edu and cckuo@sipi.usc.edu.

compression ratio and is hence suitable for progressive transmission. The first multiresolution compression method, which is usually known as the *Laplacian pyramid* scheme, was proposed by Burt and Anderson [2]. The basic idea is to decomposition an image into a low resolution image by lowpass filtering and a detailed image which is the difference of the original image and the low resolution image. By recursively performing the decomposition for the lower resolution images, we obtain a sequence of detailed images of different resolutions which can be encoded separately. Improvements of the Laplacian pyramid scheme were considered by Uncer [22]. With recently developed wavelet theory, the application of wavelet transform to image data compression has been studied by many researchers [1], [6], [7], [14], [17], [16], [27]. The wavelet-based methods are very similar to the one proposed by Burt and Anderson except that the lowpass and highpass filters used in the wavelet transform have to satisfy a certain requirements so that the transformed image is in fact obtained via orthogonal transformation. Since the transform is usually performed by using a pyramidal structure proposed by Mallat [16], [17], we call the wavelet-based compression methods the pyramidal wavelet transform (PWT) scheme.

We present a new method for image compression based on a modified wavelet transform called the full wavelet transform (FWT) in this research. With the FWT, we first apply the two-scale wavelet decomposition to the original image and obtain 4 subimages. Then, we apply the two-scale wavelet decomposition to all 4 decomposed subimages and obtain 16 subsubimages. The procedure is performed recursively until a desired level is reached. Thus, an image is decomposed into small blocks of the same size via FWT, where each block corresponds to a particular frequency band (or channel) whereas each transform coefficient in the blocks corresponds to a local spatial region in the original image. We observe experimentally that energy compaction is achieved in both the spatial and frequency domains via FWT. That is, most energy is concentrated in either low frequency blocks or transform coefficients associated with spatial regions with strong variations such as edges or textures. The energy compaction property can be effectively utilized to achieve high image compression

ratio while preserving good image quality.

The study of the human visual system indicates that most important information of an image exists in edges where the gray levels of pixels have a larger variation. Due to the observation, image compression based on edge contour extraction and coding has been examined by many researchers [8], [12], [13]. The resulting methods exploit the spatial localization property of the original image in contrast with transform coding methods which utilize the frequency localization property of the transformed image. Based on the discussion on energy compaction in Section 2.3, one can easily see that the edge information of the original image is compactly summarized in the FWT coefficients.

Following the decomposition procedure, we adopt a very simple scheme for the quantization and coding of FWT coefficients. We call the block consisting the lowest frequency components the d.c. block and all other blocks the a.c. blocks, and the FWT coefficients in the d.c. and a.c. blocks are quantized separately. The d.c. block contains a smoothed and downsampled version of the original image. The gray levels of coefficients in the d.c. block often have a Gaussian density and can be quantized with 6 bits to keep the fidelity. A bit allocation scheme is used to generate a bit assignment map for coefficients in the a.c. blocks, which are then quantized based on the Laplacian or Gaussian densities. Finally, the entropy coding is used to encode the quantized FWT coefficients. While a general theory is lacking at this point, we demonstrate experimentally that the proposed algorithm perform very well.

The paper is organized as follows. We review the conventional pyramidal wavelet transform, introduce the new FWT, and illustrate the energy compaction property of the FWT in Section 2. We examine the quantization and coding of FWT coefficients in Section 3. Numerical experiments are given in Section 4 to demonstrate the performance of the FWT method. In Section 5, we discuss the relationship between our proposed algorithm and three other popular compression schemes, i.e. the DCT (Discrete Cosine Transform), PWT (Pyramidal Wavelet Transform), and SBC (SubBand Coding) schemes. Concluding remarks are given in Section 6.

2 Image Transform via 2-D FWT

2.1 Full Wavelet transform

The wavelet transform provides a multiresolution tool for signal analysis. The two-scale discrete wavelet transform of a sequence $f[n]$ can be viewed as passing the signal through a quadrature mirror filter (QMF) consisting of a low- and high-pass filter pair denoted, respectively, by $h[k]$ and $g[k]$ with $g[k] = (-1)^k h[1-k]$. The forward transform can be written as

$$\begin{aligned} c_k &= \sqrt{2}\textstyle\sum_n h[n-2k]f[n], \\ d_k &= \sqrt{2}\textstyle\sum_n g[n-2k]f[n], \end{aligned}$$

while the inverse transform takes the form

$$f[n] = \sqrt{2}\left(\sum_k h[n-2k]c_k + \sum_k g[n-2k]b_k\right).$$

Additional constraints can be imposed on $h[k]$ and $g[k]$ so that the resulting wavelet representation has some nice properties [20]. Many wavelet transforms with various filter responses $h[k]$ have been proposed. They include the Haar basis, the family of Daubechies bases [5], and the spline wavelet basis [16], [17]. For 2-D signals, one can generalize the above idea by forming the tensor product of 1-D wavelet transforms along horizontal and vertical directions so that the signals can be decomposed into the low-low (LL), low-high (LH), high-low (HL), and high-high (HH) frequency channels. For details on the relationship between the wavelet transform and filter bank theory, we refer to [10] and [23].

There exist many possible ways to generalize the above two-scale wavelet transform to the multiple scale case. The conventional approach is to apply the two-scale decomposition recursively only to the lowest frequency channels. Since it can be efficiently implemented via a pyramidal computational algorithm [16], [17], we call it the pyramidal wavelet transform (PWT). The second approach is to apply the two-scale decomposition to channels which contains a significant amount of energy or information. The structure of the computational algorithm can be well described

by binary and quad trees in the 1-D and 2-D cases, respectively, so that it is called the tree-structured wavelet transform (TWT) [3]. Coifman, Meyer and Wickhauser [4] recently generalized the wavelet basis function to include a library of modulated waveform orthonormal bases called wavelet packets. It turns out that the application of the tree-structured wavelet transform to a signal is equivalent to representing it with a certain wavelet packet basis. Finally, we may consider the recursive application of the two-scale decomposition to all frequency channels, and call it the full wavelet transform (FWT).

2.2 Block Decomposition

Let I_0 denote the original image of size $L \times L$ where $L = 2^l$. By recursively applying the wavelet decomposition r times to I_0, we obtain a transformed image I_r consisting of $R \times R$ ($R = 2^r$) subimages and each subimage contains $M \times M$ ($M = 2^m$ and $l = m + r$) pixels. Each subimage is called a block and is labeled from left to right and top to bottom with 2D indices (k, l) where $0 \leq k, l \leq (R-1)$. The pixel inside a subimage is similarly labeled by 2D position indices (m, n) with $0 \leq m, n \leq (M-1)$. The $c(k, l; m, n)$ is used to denote the value of the transform coefficient located at position (m, n) in block (k, l). Thus, an image can be decomposed into the union of blocks of the same size via the FWT. The FWT decomposition of the Lena image of size 512×512 with the Haar basis and $r = 2$ is given in Fig. 1, where gray levels in each block are normalized to be between 0 and 255 for the ease of visualization. Note that the block $(k, l) = (0, 0)$ carries information of the lowest frequency channel and is quite different from other blocks. For convenience, we call it the *d.c. block* and all other blocks with $(k, l) \neq (0, 0)$ the *a.c. blocks*.

Each block obtained by the FWT corresponds to a frequency band while each transform coefficient corresponds to a local spatial region. By increasing the level number r of the FWT, we can increase frequency resolution at the expense of spatial resolution. There are two relationships of our interest with the above FWT decomposition: the relationship between different blocks and the relationship between different

transform coefficients. Due to the space localization of the FWT, the quantization error is confined within a local area whose size is controlled by M. On the other hand, we want to choose R large enough so that there are sufficient frequency blocks to fully utilize the spatial domain correlation. Thus, we have to consider a balance between the values of R and M. To compress of an image of size 512×512, we find empirically that it is suitable to choose $r = 3$ or 4 so that the block is of size 64×64 or 32×32.

2.3 Energy Compaction

Each block obtained by the FWT corresponds to a frequency band. The energy of block (k, l) can be calculated by

$$E_b(k,l) = \sum_{m=0}^{M-1} \sum_{n=0}^{M-1} |c(k,l;m,n)|^2. \tag{1}$$

Since most images consist primarily of smooth or, equivalently, low frequency components, it is often that the d.c. block contains the highest energy among all blocks. For 512×512 Lena image decomposed into blocks of size 32×32 (or 64×64) with the Daubechies D_4 basis, the d.c. block contains 80.1% (or 90%)of total energy. While the d.c. block provides a smoothed and downsampled version of the original image, the a.c. blocks provide information of edges and textures which are crucial for human visual effect. Thus, even though a.c. blocks contain a relatively small portion of total energy, effective coding of the FWT coefficients in a.c. blocks is important for image quality. To examine the energy compaction effect in the a.c. blocks, we plot the cumulative histogram of these blocks for the Lena image in Fig. 2, where the y-axis is the percentage of total a.c. energy and the x-axis is the percentage of total a.c. blocks. The total numbers of a.c. blocks are 255 (or 63) for blocks of size 32×32 (or 64×64). We can see from the figure that 10% of a.c. blocks have around 87% (or 77%) of total a.c. energy for the case of block size 32×32 (or 64×64).

We next consider the energy compaction property of the FWT coefficients with respect to their spatial positions. The energy of position coefficients at (m, n) in all

a.c. blocks can be expressed as

$$E_p(m,n) = \sum\sum_{k=0,l=0,(k,l)\neq(0,0)}^{R-1,R-1} |c(k,l;m,n)|^2. \quad (2)$$

To examine the energy compaction effect, we plot the cumulative histogram with respect to position indices for the Lena image in Fig. 3, where the y-axis is the percentage of total a.c. block energy and the x-axis is the percentage of total number of spatial positions. The total number of spatial positions is 1024 (or 4096) for block size 32×32 (or 64×64), and we observe from the figure that around 60% (or 80%) of the total a.c. energy comes from only 20% of spatial positions.

Combining results in Fig. 2 and Fig. 3, we also note that a smaller block size achieves better frequency energy compaction but poorer spatial energy compaction, which is in fact consistent with the uncertainty principle. To see the energy distribution of spatial positions, consider the case of block size 64×64. We divide the 4096 spatial positions into three groups, i.e. white, gray and black, and plot them in Fig. 4, where the top 70% a.c. energy is located in the white region, the second 20% is covered by the gray region, and the last 10% is in the black region. As shown in the figure, we see that energy is concentrated in the FWT coefficients corresponding to edges along the hat and face and textures in the hair region. It implies that the FWT coefficients in these regions have larger variances, and have to be quantized with more bits.

3 Bit Allocation and Quantization

Successful image compression methods depend on suitable transforms for energy compaction as well as good bit allocation and quantization schemes for data representation. As illustrated in Fig. 1, the d.c. block provides a low resolution representation of the original image while a.c. blocks primarily provides information about edges and textures. This observation suggests the application of different quantization schemes to d.c. and a.c. blocks.

3.1 D.C. Coefficients

To preserve good quality of the decompressed image, a reasonably accurate representation of the FWT coefficients in the d.c. block is necessary. Since the FWT coefficients in the d.c. block represent a smoothed version of the original image data, its histogram is closely related to that of the original image when the number r of decomposition is small. However, if r becomes larger, the histogram can be approximated by the Gaussian probability density function due to the central limiting theorem, and the coefficients can be effectively quantized with the Lloyd-Max quantizer.

To determine the number of quantization levels, we have studied the relationship between quantization levels and the mean square errors of the decompressed image using only the d.c. block. It is found that the mean square errors in the reconstruction image do not decrease significantly if the number of quantization levels is higher than 64 ($=2^6$) for an original 8-bit image. Thus, 6 bits is used to quantize the d.c. coefficients in our compression experiment. Since the quantization level is high, the quantization error is in fact relatively insensitive to the probability density assumption. Thus, even though the histogram of d.c. coefficients may not have a distribution close to the Gaussian density when $r = 3$ and 4, the Lloyd-Max quantizer with the Gaussian density assumption is still applied.

3.2 A.C. Coefficients

A. Bit Allocation

To assign the number of quantization bits to FWT coefficients $c(k, l; m, n)$ in a.c. blocks, we may group the coefficients according to their block (frequency) indices (k, l) or position (space) indices (m, n), where the distribution of FWT coefficients in each spatial or frequency group is assumed to has a zero mean, since they are obtained by highpass filtering. By assuming the distribution is uniform in one domain, say, the spatial domain, we first concentrate on bit allocation due to the nonuniform distribution in the frequency domain, and then use a weighting function to compensate

the nonuniform spatial distribution.

An optimal bit assignment scheme assigns bits to different groups of data in such a way that the distortion caused by the quantization process is minimized with a fixed total number of quantization bits. The total mean square quantization error in a.c. blocks can be expressed as

$$D_{ac} = \sum_{(k,l)\neq(0,0)} d(k,l),$$

where $d(k,l)$ is the mean square quantization error in block (k,l). According to the rate distortion theory [11], one can express $d(k,l)$ of the form

$$d(k,l) = \epsilon_{k,l}^2 2^{-2B_{k,l}} \sigma_b^2(k,l), \tag{3}$$

where $\epsilon_{k,l}^2$ is the quantization performance factor, $B_{k,l}$ is the bit rate to be used in the block (k,l) and

$$\sigma_b^2(k,l) = \frac{1}{M^2} \sum_{m=0}^{M-1} \sum_{n=0}^{M-1} |c(k,l;m,n)|^2 = \frac{E_b(k,l)}{M^2} \tag{4}$$

is the variance of FWT coefficients at block (k,l). Then, by assuming that the FWT coefficients are uniformly distributed in all spatial positions and using the mean square error as the distortion measure, one can obtain the following optimal bit allocation formula [11]

$$B_{k,l} = \bar{B}_b + \frac{1}{2}\log_2 \sigma_b^2(k,l) - \frac{1}{2(R^2-1)} \sum_{(k,l)\neq(0,0)} \log_2 \sigma_b^2(k,l), \tag{5}$$

where $B_{k,l}$ is the number of bits assigned to coefficients in block (k,l) and $\bar{B}_b$ is the desired average number of bits per block.

However, it is obvious that the FWT coefficients are not uniformly distributed in all spatial positions so that equation (5) has to be modified. Several ways to modify (5) has been discussed in the literature. For different distortion measures, Tribolet and Crochiere [21] proposed to use a frequency weighting function in the above bit allocation formula. For (5), the modified equation assumes the following form:

$$B'_{k,l} = \bar{B}_b + \frac{1}{2}\log_2 \sigma_b^2(k,l)w(k,l) - \frac{1}{2(R^2-1)} \sum_{(k,l)\neq(0,0)} \log_2 \sigma_b^2(k,l)w(k,l), \tag{6}$$

where $w(k,l)$ denotes the weighting function. In the context of nonidentically distributed random variables with distinctive distortion measure, modified bit allocation formula can also be obtained [9].

In the above discussion, the weighting function $w(k,l)$ and the objective function $B'_{k,l}$ have the same parameters k and l. Here, we consider another generalization by considering a weighting function parameterized by another index set (m,n), i.e.

$$\tilde{B}_{k,l,m,n} = \bar{B}_b + \frac{1}{2}\log_2 \sigma_b^2(k,l)w(m,n) - \frac{1}{2(R^2-1)M^2}\sum_{(k,l)\neq(0,0)}\sum_{m,n}\log_2 \sigma_b^2(k,l)w(m,n).$$

In particular, the weighting function $w(m,n)$ is chosen to be the variance of FWT coefficients at position (m,n) in all a.c. blocks

$$w(m,n) = \sigma_p^2(m,n) = \frac{E_p(m,n)}{R^2-1}, \tag{7}$$

where $E_p(m,n)$ is given in (2) and R^2-1 is the total number of a.c. blocks. Thus, we obtain the following bit allocation formula

$$\begin{aligned} B_{k,l,m,n} &= \bar{B}_b + \frac{1}{2}\log_2 \sigma_b^2(k,l) + \frac{1}{2}\log_2 \sigma_p^2(m,n) - \frac{1}{2}\log_2 \rho_b^2 - \frac{1}{2}\log_2 \rho_p^2 \qquad (8)\\ &= C + \frac{1}{2}\left[\log_2 \sigma_b^2(k,l) + \log_2 \sigma_p^2(m,n)\right], \qquad (9)\end{aligned}$$

where C is a constant, ρ_b^2 and ρ_p^2 are, respectively, the geometric means of block and block variances for a.c. blocks, i.e.

$$\rho_b^2 = \left(\Pi_{(k,l)\neq(0,0)}\sigma_b^2(k,l)\right)^{1/(R^2-1)}, \quad \rho_p^2 = \left(\Pi_{m,n}\sigma_p^2(m,n)\right)^{1/M^2}.$$

The bit allocation formula (9) implies that the product of block and position variances is in fact used as a criterion for grouping the coefficients in the FWT domain. Although we do not derive it on a theoretical basis, it serves as a good empirical formula and works pretty well in practice. Some experimental justification of the formula (9) is given below.

It is clear that the FWT coefficients assigned with the same number of quantization bits may belong to different blocks and different positions. Based on the number of

Bit No.	No. of Pixel	Mean	Variance
0	159972	0.00	17.0
1	51979	0.04	75.9
2	27481	-0.08	200.2
3	12711	-0.11	913.2
4	5122	0.73	4538.7
5	709	-8.54	18249.3
6	74	-26.26	82518.2
7	0	0.00	0.0

Table 1: Partitioning of FWT a.c. coefficients with D8 basis, and 9.8:1 compression ratio.

Bit No.	No. of Pixel	Mean	Variance
0	224238	0.01	36.7
1	21808	0.27	429.3
2	8480	-1.39	1733.9
3	2912	-4.18	7510.0
4	610	5.73	31286.5
5	0	0.00	0.0

Table 2: Partitioning of FWT a.c. coefficients with D4 basis and 22:1 compression ratio.

assigned quantization bits, the FWT coefficients in all a.c. blocks can be rearranged into several groups. To verify the efficiency of the bit allocation formula (9), we check whether the variances of these groups meet the form of the rate distortion function as given by (3). Two typical examples are used to illustrate a very good match. Consider the decomposition of the 512×512 Lena image into blocks of size 64×64. We list the variances of each group with two different compression ratios in Tables 1 and 2. One can clearly see that the variance values in each group approximately increase by a factor of 4 as the assigned bit number is increased by 1.

B. Quantization

The distribution of coefficients in the pyramidal wavelet transform domain within

a single frequency band has been studied [17]. Similar results can also be found in the literature of subband coding [24]. These coefficients can be described by the generalized Gaussian density function

$$p(x) = a\epsilon^{-|bx|^c},$$

where

$$a = \frac{bc}{2\Gamma(\frac{1}{c})} \quad \text{and} \quad b = \frac{1}{\sigma_x}(\frac{\Gamma(\frac{3}{c})}{\Gamma(\frac{1}{c})})^{1/2},$$

and where σ_x is the standard deviation of the data. The values $c = 2$ and $c = 1$ correspond to the Gaussian and Laplacian density functions, respectively. In the subband coding context, Westerink et al [24] used both chi-squared (χ^2) and Kolmogorov-Smirnov testing methods to test the distribution of coefficients in the high frequency bands, and suggested the value $c = 0.5$ for the optimum quantizer design. Note that the shape of $p(x)$ with $c < 1$ is sharper than that of the Laplacian density.

To effectively quantize the coefficients, we need to know the distribution of the data in each group. A typical histogram of a.c. coefficients allocated with 1 to 4 bits are shown in Fig. 5. In this work, the Lloyd-Max quantizer is used. The Laplacian density is used for groups with 1 or 2 quantization bits while the Gaussian density is used for groups with more than 2 quantization bits. For 1 and 2 bit quantization group, there are many coefficients with values close to zero. We use 3 and 5 quantization levels to replace 2 and 4 quantization levels, respectively, and then code the quantization levels with entropy coding to reduce the number of quantization bits.

To determine the number of bits assigned to each FWT coefficient at the decoder end, we have to know the constant C and the values of $\log_2 \sigma_b^2(k,l)$ and $\log_2 \sigma_p^2(k,l)$ according to (9). In practice, we quantize $\log_2 \sigma_b^2(k,l)$ and $\log_2 \sigma_p^2(m,n)$ with 8 levels (or 3 bits) and store these values into two matrices of size $M \times M$ and $R \times R$ as side information.

4 Numerical Experiments

We have applied the FWT compression method to the 512 × 512 Lena image, and considered the following design choices:

- wavelet bases: Haar, Daubechies D4, D8 and truncated 17-tape cubic spline bases;
- block sizes: 32 × 32 and 64 × 64.

The quality of the reconstructed image is measured by the peak signal to noise ratio (PSNR) which is defined as

$$\text{PSNR}(dB) = 10 \log_{10} \frac{255^2}{\text{MSE}},$$

where

$$\text{MSE}(\text{ Mean Square Error }) = \frac{1}{L^2} \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} (\hat{x}_{i,j} - x_{i,j})^2,$$

and where $x_{i,j}$ and $\hat{x}_{i,j}$ denote the gray scales of the pixel at position (i, j) of the original and reconstructed images, respectively. To see the performance of the proposed FWT compression method, we plot the PSNR versus compression ratio in Figs 6 and 7, where the compression ratio is computed by also taking the side information needed for image reconstruction into account. The original and several decompressed Lena images are shown in Figs. 8 (a)-(d) for visual judgement.

We compare the performance of the proposed compression scheme with respect to different bases in Fig. 6. We see that the D4 and D8 bases give the best performance. The D4 basis has one extra advantage, i.e. it has an integer multiple of 1/32 on its filter coefficients (h[0]= 11/32, h[1]= 19/32, h[2]= 5/32, h[3]= −3/32) so that the algorithm may be easily implemented with hardware. It is interesting to see that the spline basis does not perform as well as the D4 and D8 bases. This can be explained by that our compression algorithm uses correlation among a.c. blocks and a wavelet basis function with a larger number of taps for convolution may not provide a good space domain energy compaction property.

The performance of the proposed compression scheme with respect to different block sizes is given in Fig. 7. Two block sizes 64×64 and 32×32 have been tested. For this example, the 64×64 block size gives the better result for the compression ratio less than 40 while the 32×32 block size performs better for the compression ratio larger than 40. This is due to a larger amount of side information is needed to store the position and block variances. The two matrices are of sizes 64×64 and 8×8 for the 64×64 block while they are of sizes 32×32 and 16×16 for the 32×32 block.

5 Relationship with Other Compression Methods: Review and Comparison

The FWT compression is closely related to three different types of compression schemes: the Pyramidal Wavelet Transform (PWT), the Discrete Cosine Transform (DCT), and the SubBand Coding (SBC) compression techniques. In this section, we will briefly review each compression scheme, and then discuss the similarities and differences of these schemes with our new method.

5.1 Pyramidal Wavelet Transform (PWT)

The pyramid-structured image compression scheme was first proposed by Burt and Anderson [2]. The basic idea is to decomposition an image into a low resolution subimage by lowpass filtering and a detailed subimage which is the difference of the original image and the low resolution image. For the high frequency detailed subimage, the pixel-to-pixel correlation is removed so that only data with a small dynamic range remain. Besides, since the pixel values usually have a Laplacian-like distribution, the Loyld-Max quantizer can be effectively applied for data compression. The lowpass filtered subimage consists primarily of low frequency components and can be down-sampled as a coarser approximation of the original image. The size-reduced low resolution subimage contains most important information of the original image, but can be encoded less expensively. The above decomposition can be recursively

applied to low resolution subimages, which leads naturally a multiresolution method. To reconstruct the original image, quantized data at different levels can be combined from the coarsest level to the finest level in a sequential way.

A pyramid-structured image compression scheme, which is called the PWT compression here, was also proposed by Mallat [17]. His idea to decompose an image into a low resolution subimage and a detailed subimage is exactly the same as the one of Burt and Anderson. The main difference is on the scheme to achieve image decomposition. The two-scale image decomposition and synthesis is achieved by using the quadrature mirror filter (QMF) banks which guarantee aliasing cancellation and perfect reconstruction. A recursive application of the two-scale decomposition leads to a wavelet transform which possesses a good spatial and frequency localization property and is useful for multiresolution image representation. The detailed information, which exists in the high frequency detailed subimage in Burt and Anderson's work, is now contained by three subimages, each of which has a quarter size of the original image and corresponds to a directional high frequency band. The image data size is preserved the same as the original one via the PWT scheme where as the image data size is increased to 4/3 of the original image size with the Burt and Anderson's scheme. In order to quantize transform coefficients for PWT compression, Mallat modeled the histogram of those coefficients with the following function

$$h(u) = K \times e^{-(|u|/\alpha)^{\beta}},$$

where u denotes the transform coefficient and parameters K, α and β are chosen to give the best data fit. The PWT compression scheme can encode an image with less than 1.5 BPPs (bits per pixel) with few visible distortions [15], [26].

Since the coefficients obtained via the pyramid-structured wavelet transform are of different sizes in different scales, the correlation between them in different scales are hard to evaluate. Besides, quantization rules vary from bands to bands since they have different distributions characterized by different parameters K, α and β. This makes the quantization procedure computationally expensive. In contrast, we

decompose the image with the FWT which results in subimages of the same size. The correlation between transform coefficients is exploited.

5.2 Discrete Cosine Transform (DCT)

Image compression methods based on block transforms have been extensively studied over the last 15 years [25], among which the DCT-based compression scheme is the most popular one and considered to be the state-of-the-art method. Block transform compression is achieved by the energy compaction property via various transforms in the transform domain. For this family of methods, transform coefficients are first obtained by performing an inner products of a set of complete orthonormal basis functions and a series of windowed signals with finite support. They are then effectively quantized by taking advantage of their special statistical distributions such as Laplacian or Gaussian distributions. The major difference between the block transform and PWT (or SBC) compression is in their implementation. For the wavelet transform (or SBC), the image is convolved with an analysis filter bank and followed by a downsampling procedure. The downsampled subimages are either quantized or processed through a further analysis filter bank. The recovering process of the wavelet transform (or SBC) includes a upsampling procedure followed by a synthesis filter bank. Despite the difference, the block transform and the wavelet transform (or SBC) compression are in fact kind of similar at least in the conceptual level. That is, the inner product of the set of basis functions with windowed subimages is equivalent to a filtering process and, therefore, inner products obtained from the same basis function can be viewed as transform coefficients in the same subband. This explains why the block transform compression is sometimes viewed as another kind of SBC technique [19]. The choice of basis functions (or impulse responses of filter banks) in block transform, wavelet transform and subbanding coding are nevertheless very different due to different goals in the original development of these methods.

The DCT-based compression technique has been well developed, and many variants have been proposed. There are two critical steps in their implementations. One

is to locate the regions in the transform domain which contain importment information and assign them with a certain number of bits for quantization which is known as the bit allocation. The other is to quantize the transform coefficients effectively. For bit allocation, it is common to use the variance distribution method [18], where coefficients with larger variance values are selected and bits can be assigned with a certain bit allocation formula. The Loyld-max quantizer is often used for quantization. The DCT-based compression methods perform reasonably well for a compression ratio around 10. However, it is relatively difficult to get a compression ratio much larger than 10. Besides, the compressed image does not degrade gracefully by increasing the compression ratio. That is, if the compression ratio is raised above some threshold value, we may not be able to get any meaningful image via decompression at all.

This limitation of DCT-based methods may be attributed to that the correlation of coefficients in the same position of all blocks is not strong enough. Recall that by using the block DCT transform, we partition the image into small small blocks in the spatial domain and transform them independently. Consider the case that one spatial block consists primarily of a smooth surface while another block contains a textured pattern. Since the frequency components depend on the corresponding spatial blocks, the correlation of the transform coefficients in these two blocks is low. Thus, we conclude that the DCT compression methods take advantage of the frequency localization property of the DCT transform and utilize only the frequency correlation. In contrast, our FWT compression scheme utilizes not only frequency but also spatial correlation.

5.3 SubBand Coding (SBC)

With the subband coding (SBC) technique, we decompose an image into several subimages by using quadrature mirror filter banks so that each subimage has its frequency components concentrated on a particular frequency band called the subband. Then, each subimage is encoded independently via predicative differential coding or vector quantization with the use of the statistical property of data in each subband.

Several advantages have been mentioned in the SBC literatures. First, data in different subbands may be encoded with different methods such as PCM, DPCM and DCT to exploit the maximum redundancy in each subband. Second, the quantization error in each subband is confined to that band only. Third, based on results from human perception research, we can give different weightings to different subbands and allocate the number of bits accordingly with the objective that the decompressed image looks more pleasant with respect to the human visual system.

The major difference between the FWT and SBC is that the FWT utilizes both intra-band and inter-band correlations while the SBC only uses the intra-band correlation. The edge information in an image is usually spread over different subbands, but the correlation between subbands is not utilized at all in the SBC. With the SBC, we want to minimize the correlation between different subbands as much as possible so that it is desirable to design a multiband filter bank which has sharp transition among different subbands. To implement such a filter bank often requires filters of a large number of taps, and the computational cost is higher. In contrast, since the FWT method exploits the correlation between different subbands, there is no need to design filter banks with sharp transition among subbands, and filters with a small number of taps such as Haar,D4, D8 basis functions [5] can be conveniently used. The computational cost can therefore be reduced.

6 Conclusions and Extensions

We presented a new method for image compression based on the full wavelet transform (FWT) combined with a simple quantization scheme and entropy coding in this research. The new method is suitable for progressive transmission due to the nature of multiresolution representation. Besides, it adopts a regular block size so that side information about the data address can be reduced. In particular, we illustrate the energy compaction property of the FWT in both the spatial and frequency domains. This important property can be effectively exploited via novel quantization

and coding. Even though only a simple quantization scheme was used in this research, numerical experiments showed that the proposed algorithm does give a comparable performance with several existing methods.

There exists room to improve the proposed method, say, the use of vector quantization to quantize the FWT coefficients in the a.c. blocks. It is under our current investigation. We showed the energy compaction property of the FWT with numerical data. Theoretical study on this topic is certainly valuable. It also seems attractive to apply the FWT method to compress speech signals.

References

[1] M. Barlaud, P. sole, M. Antonnini, and P. Mathieu, "A pyramidal scheme for lattice vector quantization of wavelet transform coefficients applied to image coding," in *Proc. ICASSP-92*, pp. 401–404, 1992.

[2] P. J. Burt and E. H. Adelson, "The Laplacian Pyramid as a compact image code," *IEEE Trans. on Communications*, Vol. 31, pp. 532–540, Apr. 1983.

[3] T. Chang and C. Kuo, "A wavelet transform approach to texture analysis," in *Proc. ICASSP-92*, pp. 661–664, 1992.

[4] R. R. Ciofman and M. V. Wickerhauser, "Entropy-based algorithms for best basis selection," *IEEE Trans. on Information Theory*, Vol. 38, pp. 713–718, Mar. 1992.

[5] I. Daubechies, "Orthonormal bases of compactly supported wavelets," *Communications on Pure and Applied Mathematics*, Vol. 41, pp. 909–996, Nov. 1988.

[6] P. Desarte, B. Macq, and D. T. M. Slock, "Signal-adapted multiresolution transform for image coding," *IEEE Trans. on Information Theory*, Vol. 38, pp. 897–904, Mar. 1992.

[7] R. DeVore, B.Jawerth, and B. J. Lucier, "Image compression through wavelet transform coding," *IEEE Trans. on Information Theory*, Vol. 38, pp. 719–745, July 1992.

[8] M. Eden and M. Kocher, "On the performance of a contour coding algorithm in the context of image coding. Part II: Coding and contour graphs," *Signal Process*, Vol. 8, May 1985.

[9] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression*, Kluwer Academic Publishers, 1992.

[10] R. A. Gopinath and C. S. Burrus, *Wavelet Transforms and Filter Banks*, ch. Applications to Numerical Analysis and Signal Processing, pp. 603–654, Academic Press, Inc, 1992.

[11] N. S. Jayant and P. Noll, *Digital Coding of Waveforms Principles and Applications to Speech and Video*, Prentice-Hall Inc, 1984.

[12] M. Kocher and M. Kunt, "Image data compression by contour-texture modelling," *SPIE Int. Conf. on the Applications of Digital Image Processing*, Vol. 0, pp. 131–139, Apr. 1983.

[13] M. Kunt, A. Ikonomopoulos, and M. Kocher, "Second-generation image-coding techniques," *Proceedings of the IEEE*, Vol. 73, pp. 549–574, Apr. 1985.

[14] A. S. Lewis and G. Knowles, "Image compression using the 2-D wavelet transform," *IEEE Trans. on Image Processing*, pp. 244–250, Apr. 1992.

[15] S. G. Mallat, "A compact multiresolution representation: the wavelet model," Tech. Rep. MS-CIS-87-69, Department of Computer and Information Science, University of Pennsylvania, Aug. 1987.

[16] S. G. Mallat, "Multifrequency channel decompositions of images and wavelet models," *IEEE Trans. on Acoustic, Speech, and Signal Processing*, Vol. 37, pp. 2091–2110, Dec. 1989.

[17] S. G. Mallat, "A theory for multiresolution signal decomposition: the wavelet representation," *IEEE trans. on Pattern Recognition and Machine Intelligence*, Vol. 11, pp. 674–693, July 1989.

[18] K. R. Rao and P. Yip, *Discrete Cosine Transform, Algorithm, Advantages, Applications*, Academic Press, Inc., San Diego, 1990.

[19] E. P. Simoncelli and E. H. Adelson, *Subband Image Coding*, ch. Subband transforms, pp. 143–192, The Netherlands: Kluwer Academic Publishers, 1991.

[20] G. Strang, "Wavelets and dilation equations: A brief introduction," *SIAM Review*, Vol. 31, pp. 614–627, Dec. 1989.

[21] J. M. Tribolet and R. E. Crochiere, "Frequency domain coding of speech," *IEEE Trans. on Acoustic, Speech, and Signal Processing*, Vol. 27, No. 5, pp. 512–530, 1979.

[22] M. Uncer, "An improved least squares Laplacian pyramid for image compression," *Signal Processing*, Vol. 27, pp. 187–203, 1992.

[23] M. Vetterli and C. Herley, "Wavelets and filter banks: theory and design," *IEEE Trans. on Signal Processing*, Vol. 40, pp. 2207–2232, 1992.

[24] P. H. Westerink, J. Biemond, and D. E. Boekee, *Subband Image Coding*, ch. Subband coding of color images, pp. 192–227, The Netherlands: Kluwer Academic Publishers, 1991.

[25] J. W. Woods, *Subband Image Coding*, The Netherlands: Kluwer Academic Publishers, 1991.

[26] J. W. Woods and S. D. O'Neil, "Subband coding of images," *IEEE Trans. on Acoustic, Speech, and Signal Processing*, Vol. 34, pp. 1278–1288, Oct. 1986.

[27] W. R. Zettler, J. Huffman, and D. C. P. Linden, "Application of Compactly supported wavelets to image compression," in *Proc. SPIE Vol. 1244*, pp. 150–160, 1990.

Figure Captions

Figure 1: 2-level full wavelet transform.

Figure 2: Cumulative histogram of the energy of a.c. blocks.

Figure 3: Cumulative histogram of the energy of a.c. coefficients at the same spatial position.

Figure 4: Energy distribution in different spatial positions in a.c. blocks.

Figure 5: Histograms of a.c. coefficients with (a) 1 (b) 2 (c) 3 and (d) 4 quantization bits.

Figure 6: Performance comparison of different wavelet bases.

Figure 7: Performance comparison of different block sizes.

Figure 8: (a) Original and (b)-(d) decompressed Lena images with (b) compression ratio = 16:1 and PSNR = 32.3, (c) compression ratio = 32:1 and PSNR = 29.56, (d) compression ratio = 69:1 and PSNR = 26.7

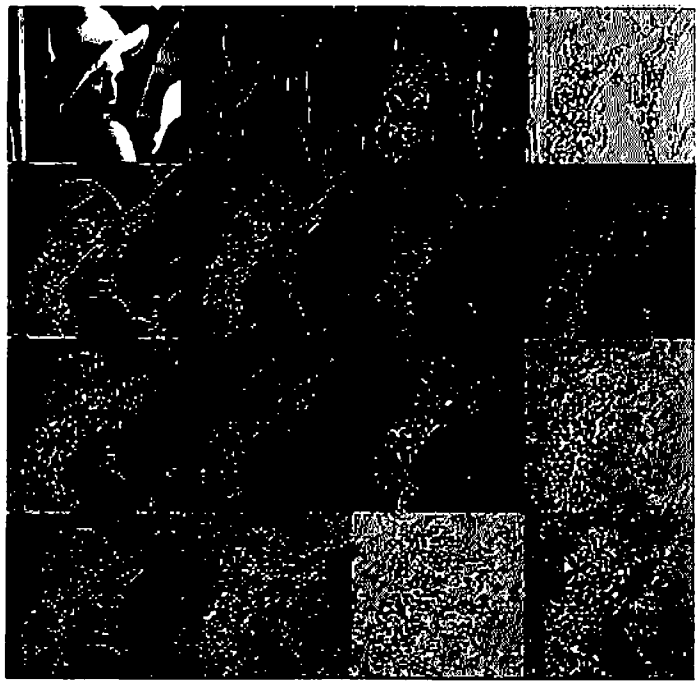

Figure 1: 2-level full wavelet transform.

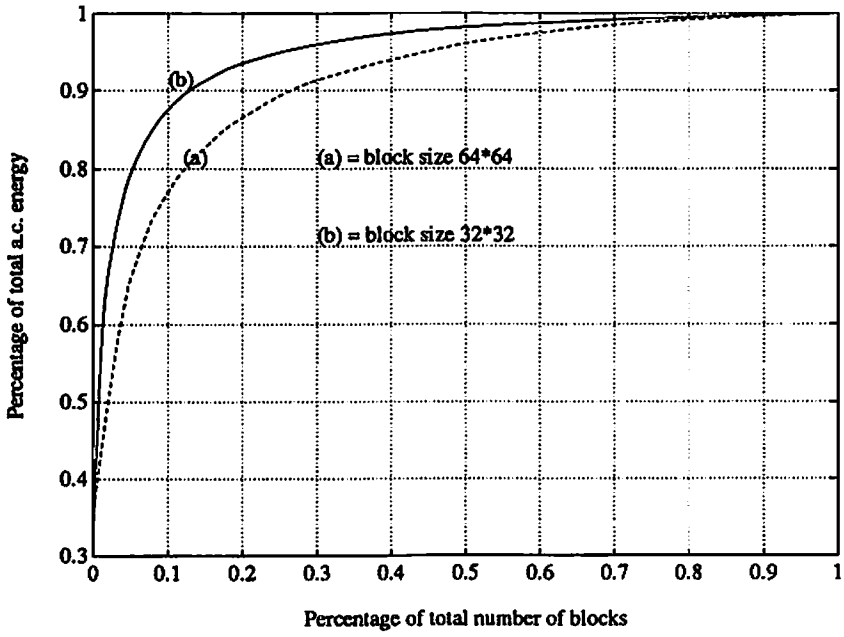

Figure 2: Cumulative histogram of the energy of a.c. blocks.

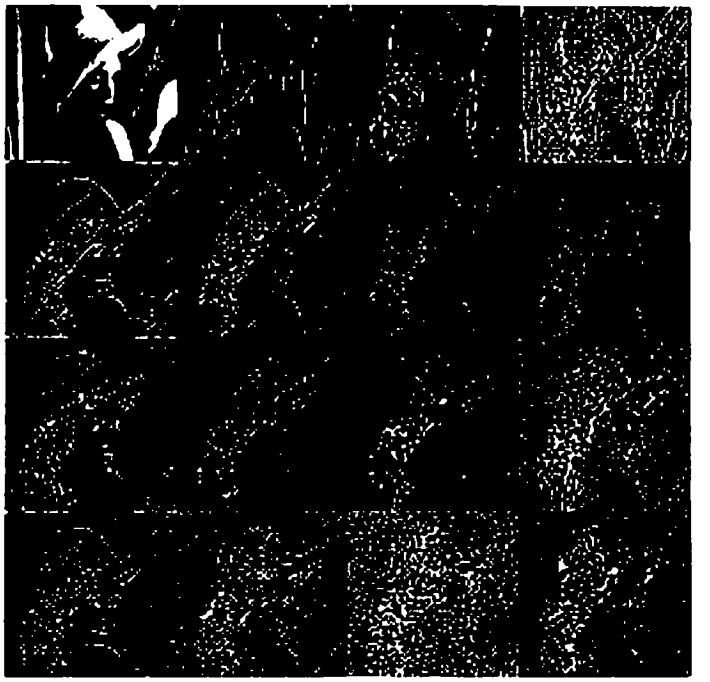

Figure 1: 2-level full wavelet transform.

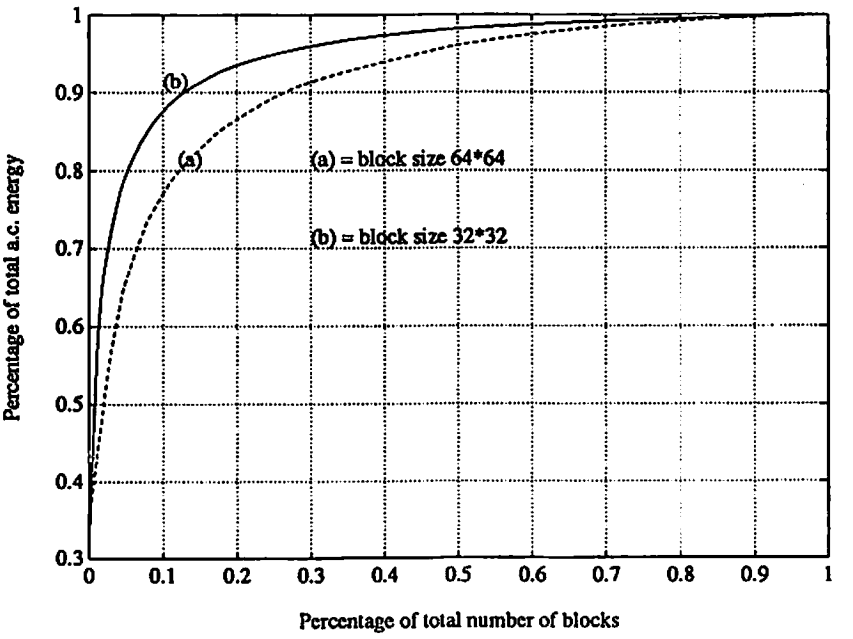

Figure 2: Cumulative histogram of the energy of a.c. blocks.

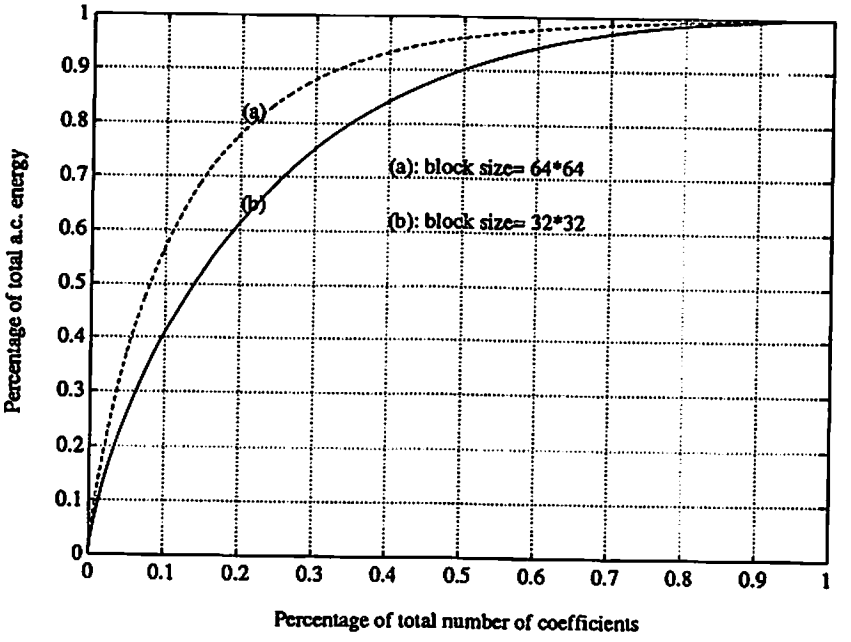

Figure 3: Cumulative histogram of the energy of a.c. coefficients at the same spatial position.

Figure 4: Energy distribution in different spatial positions in a.c. blocks.

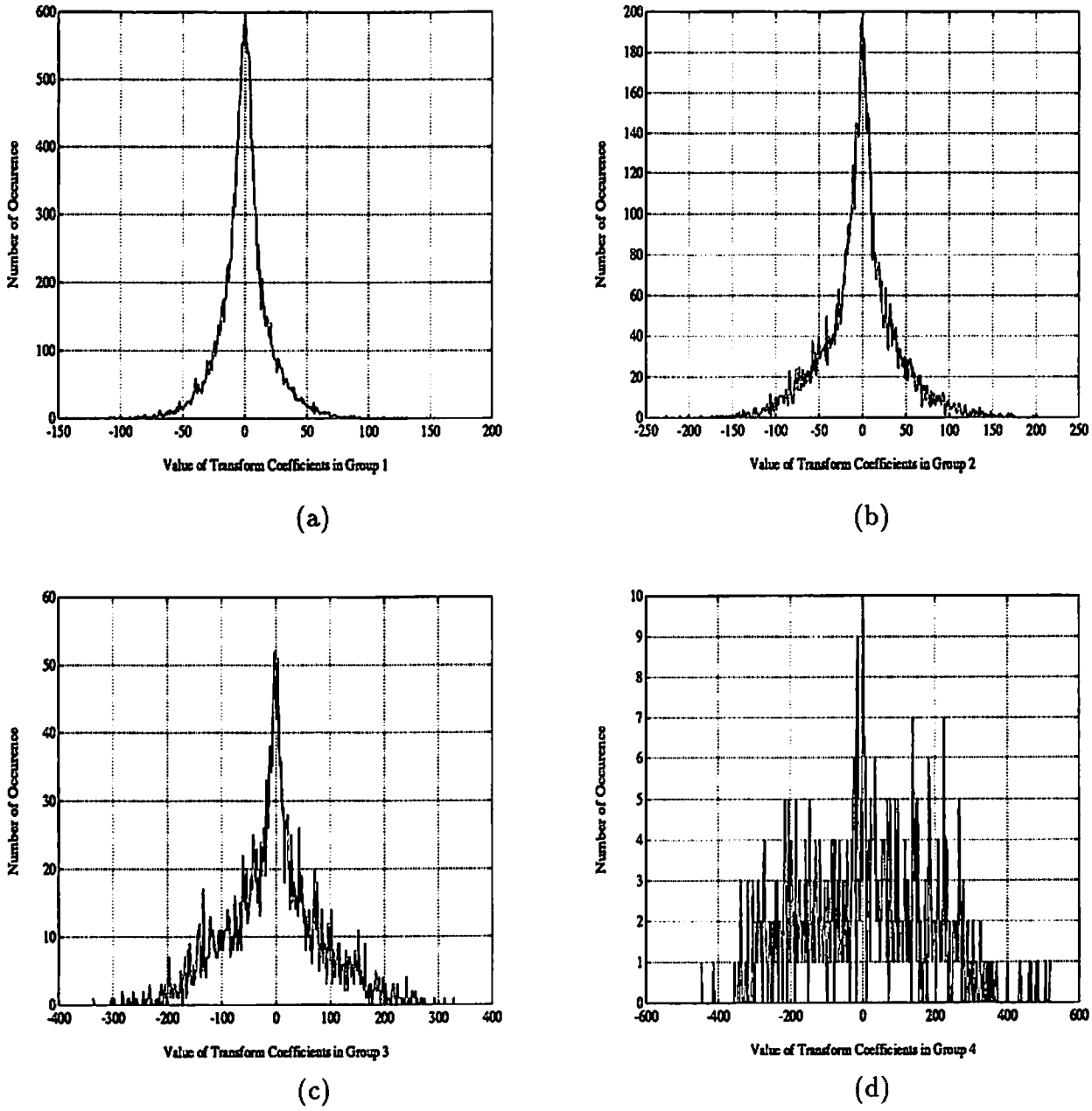

Figure 5: Histograms of a.c. coefficients with (a) 1 (b) 2 (c) 3 and (d) 4 quantization bits.

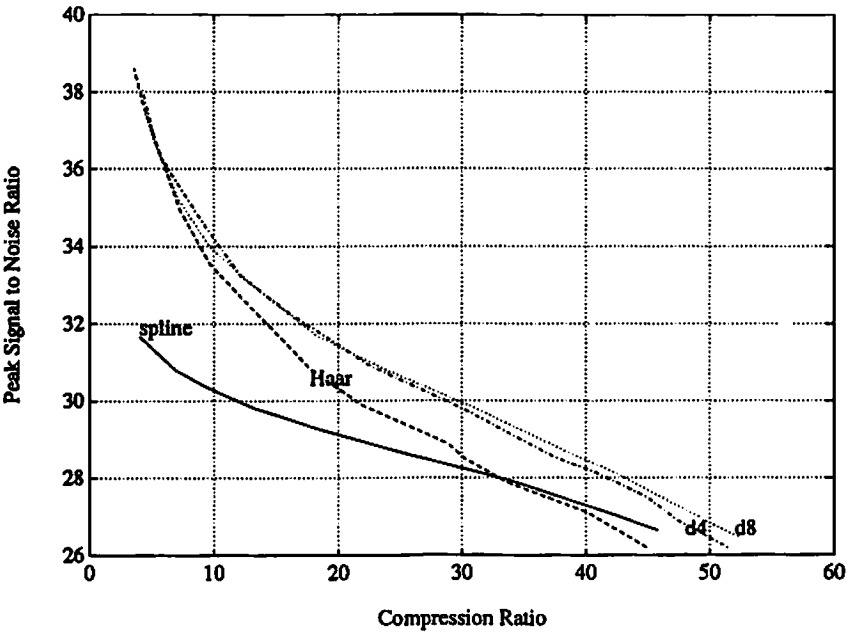

Figure 6: Performance comparison of different wavelet bases.

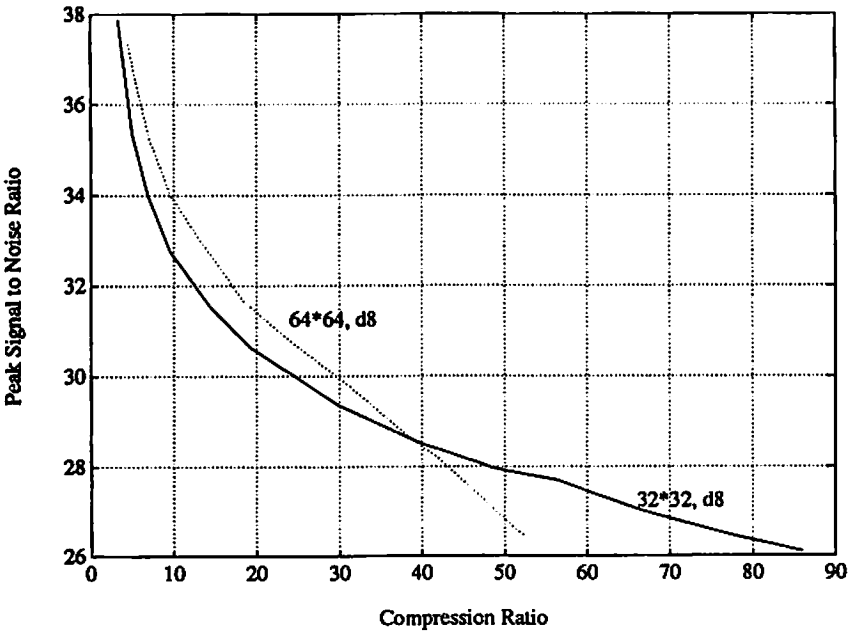

Figure 7: Performance comparison of different block sizes.

(a)

(b)

(c)

(d)

Figure 8: (a) Original and (b)-(d) decompressed Lena images with (b) compression ratio = 16:1 and PSNR = 32.3, (c) compression ratio = 32:1 and PSNR = 29.56, (d) compression ratio = 69:1 and PSNR = 26.7.