Data Compression by Wavelet Transforms

M. Shahshahani
Communications Systems Research Section

A wavelet transform algorithm is applied to image compression. It is observed that the algorithm does not suffer from the blockiness characteristic of the DCT-based algorithms at compression ratios exceeding 25:1, but the edges do not appear as sharp as they do with the latter method. Some suggestions for the improved performance of the wavelet transform method are presented.

I. Introduction

The application of wavelet transforms and multiresolution analysis to data compression has attracted much attention recently. This circle of ideas is closely related to the subband compression and the pyramid encoding techniques. The general idea is to transform and reorganize the data in a hierarchical manner so that the upper levels of this hierarchy (or pyramid) represent the general features of the data or the image and the lower levels supply the details. Generally the higher levels of the pyramid are smaller data sets than the lower levels; however, the coefficients in the latter portion are more correlated than those in the former and are better compressed by the standard lossless compression techniques.

The applications of wavelet representations to practical engineering problems are not limited to source coding. For example, one encounters situations that necessitate selecting a subset of a large data set on the basis of certain characteristics. One may achieve this by browsing through the higher levels of the hierarchy, which comprise a much smaller data set, examining the general features of the data, and making judicious choices. The coefficients in the lower levels of the pyramid may be used for edge detection.

As in other methods of data compression, applications of wavelet transforms to source coding assume a priori knowledge of the tolerable level of information loss and/or the desirable compression ratio. Data compression is achieved by quantizing the transformed data and allocating bits to the different levels of the pyramid of the transformed data in a manner compatible with the constraints and the requirements of the particular application. Naturally, in source coding applications more bits are allocated to an individual coefficient in the higher levels of the pyramid than to one in a lower level. In analogy with the discrete cosine transform (DCT), one may regard the lower levels of the pyramid as the high frequencies and the upper ones as the low frequencies.

The presentation of the general theory of wavelet transforms in Section II is intended for application to data compression. The literature on the subject is often inadequate regarding the implementation of the basic ideas, and the theoretical aspects of the subject seem to be only remotely related to practical engineering problems. It is hoped that the concise and concrete presentation of the wavelet transforms in Section II will make the literature more accessible to interested researchers. In Section III, the practical aspects of image compression by wavelet transforms and the results of the applications are reported. Further research topics for the improvement of the performance of wavelet-transform-based compression algorithms are also suggested. Some of the advantages and disadvantages of the wavelet transforms versus the standard DCT techniques are discussed. However, no definitive judgment can be made at this time regarding their relative merits. While the latter approach has been studied extensively in the past decade, the application of wavelet transforms to image compression has not reached the level of maturity that would warrant definitive assessment of its merits and potential.

II. Wavelet Transforms

The idea of wavelet transforms and their applicability to signal analysis, and especially data compression, is most easily demonstrated by focusing on the one-dimensional case first. A straightforward generalization of the theory to two dimensions for application to image compression is indicated at the end of this section. In this case, a data set is represented by an element of $\mathcal{L} = L^2(\mathbf{R})$. Consider the following sequence of partitions of $\mathbf{R}$:

$$\text{Partition } \mathcal{P}_m : \quad \mathbf{R} = \bigcup_{n=-\infty}^{\infty} I_{n,m}$$

where $I_{n,m} = [2^m n, 2^m(n+1))$, and let $\mathcal{L}_m$ be the subspace of $\mathcal{L}$ consisting of functions that are constant on the intervals $I_{n,m}$. The operator p_m of orthogonal projection on the subspace $\mathcal{L}_m$ is

$$p_m(f)(x) = \frac{1}{2^m} \int_{I_m(x)} f(x)dx$$

where $I_m(x)$ is the unique interval $I_{n,m}$ (m fixed) containing x. The subspaces $\mathcal{L}_m$ have the following properties:

$$\mathcal{L}_{m+1} \subseteq \mathcal{L}_m, \qquad \bigcap \mathcal{L}_m = 0, \qquad \overline{\bigcup \mathcal{L}_m} = \mathcal{L} \qquad (1)$$

Let $\mathcal{E}_m$ denote the orthogonal complement of $\mathcal{L}_{m+1}$ in $\mathcal{L}_m$, then $\mathcal{L}$ admits of the orthogonal direct sum decomposition $\mathcal{L} = \oplus\mathcal{E}_m$. Denote orthogonal projection on $\mathcal{E}_m$ by π_m. Let $A_{a,b}$, where $a \neq 0$ and b are real numbers, denote the affine transformation $A_{a,b}(x) = ax + b$, and define the action of $A_{a,b}$ on a function φ by $A_{a,b}(\varphi)(x) = a^{-1/2}\varphi[(x-b)/a]$. It is convenient to introduce the notation $\varphi_{m,n}(x) = A_{2^m,2^m n}(\varphi)(x)$, for m and n integers, and note that a function $f \in \mathcal{L}$ admits of the expansion

$$p_m(f) = \sum_{n=-\infty}^{\infty} a_n^m \chi_{m,n} \qquad (2)$$

where χ is the indicator function of the interval $[0,1)$, and $a_n^m = \int \chi_{m,n}(x)f(x)dx$. The functions $\chi_{m,n}$ are obtained from the single function χ through the action of a set of affine transformations of the line. For each fixed m,

$$\begin{aligned} \mathcal{L}_m &= \text{ span } \{\chi_{m,n} | n \in \mathbf{Z}\} \\ f \in \mathcal{L}_m &\Longleftrightarrow f(2.) \in \mathcal{L}_{m-1} \end{aligned} \qquad (3)$$

From the expansion (2) one easily obtains the expansion of f following the decomposition $\mathcal{L} = \oplus\mathcal{E}_m$. First observe that

$$\chi_{m+1,n} = \frac{1}{\sqrt{2}}(\chi_{m,2n} + \chi_{m,2n+1})$$

Therefore, $a_n^{m+1} = \frac{1}{\sqrt{2}}(a_{2n}^m + a_{2n+1}^m)$, and after a simple calculation one obtains

$$p_m(f) - p_{m+1}(f) = \frac{1}{2}\sum(a_{2n}^m - a_{2n+1}^m)(\chi_{m,2n} - \chi_{m,2n+1})$$

Now set $\varphi_{m,n} = \frac{1}{\sqrt{2}}(\chi_{m,2n} - \chi_{m,2n+1})$ to obtain the expansion

$$f = \sum_{m,n=-\infty}^{\infty} b_n^m \varphi_{m,n} \qquad (4)$$

where $b_n^m = \frac{1}{\sqrt{2}}(a_{2n}^m - a_{2n+1}^m)$. It is a remarkable fact, and easy to prove, that the functions $\varphi_{m,n}$ are also obtained from the single function $\varphi(x) = \chi(2x) - \chi(2x-1)$ by the action of the set $\mathcal{A} = \{A_{2^m,2^m n} | m, n \in \mathbf{Z}\}$ of affine transformations, and an analogue of condition (3) is valid for the subspaces $\mathcal{E}_m$, namely,

$$\begin{aligned} \mathcal{E}_m &= \text{ span } \{\varphi_{m,n} | n \in \mathbf{Z}\} \\ f \in \mathcal{E}_m &\Longleftrightarrow f(2.) \in \mathcal{E}_{m-1} \end{aligned} \qquad (5)$$

The functions $\{\varphi_{m,n}\}$ form a complete orthonormal set for $\mathcal{L}$. The expansion (4) is an example of orthonormal wavelet expansion, and the coefficients b^m_n are called the wavelet coefficients.

To understand the intuitive meaning of the expansion (4), assume that $f \in \mathcal{L}_m$ for a sufficiently large negative number m. The projection of f on $\mathcal{E}_m$ is $\pi_m(f) = f - p_{m+1}(f)$. Now $p_{m+1}(f)$ is a slightly *smoothed* version of f so that $\pi_m(f)$ represents the details that are missing from the smoothed version $p_{m+1}(f)$. Thus, $\pi_m(f)$ or more precisely, the coefficients b^m_n in the expansion (4) belong to the lowest level of the hierarchy. The process can be repeated with $p_{m+1}(f)$ replacing f, thus leading to a hierarchy of the coefficients of the wavelet expansion of f.

An important feature of the expansion (4) is that the coefficients b^m_n and a^m_n can be computed recursively in a simple manner. For example, to compute a^1_n from a^0_n, substitute expansion (2) for $m = 0$ in the formula for a^1_n to obtain

$$a^1_n = \sum_{k=-\infty}^{\infty} \int \frac{1}{\sqrt{2}}\chi(x-k-2n)\chi(x/2)dx \qquad (6)$$

Therefore, if one defines $\alpha(k)$ as the integral in expansion (6) for $n = 0$, one obtains the formula

$$a^1_n = \sum_{k=-\infty}^{\infty} \alpha(k-2n)a^0_k \qquad (7)$$

Similarly,

$$b^1_n = \sum_{k=-\infty}^{\infty} \beta(k-2n)a^0_k \qquad (8)$$

where $\beta(k) = \frac{1}{\sqrt{2}}\int \chi(x-k)\varphi(x/2)dx$. By a straightforward inductive extension of this calculation, one can express b^{m+1}_n and a^{m+1}_n in terms of a^m_n. The resulting formulae are identical with formulae (7) and (8) with m and $m+1$ replacing 0 and 1, respectively. Therefore, the wavelet coefficients b^m_n and a^m_n can be computed by the filters defined by α and β.

The orthonormal basis $\{\varphi_{m,n}\}$ and the expansion (4) are just one example of an orthonormal wavelet expansion. To obtain other expansions, one has to abstract some of the features of this illustrative example. The essential ingredients of the theory are an orthonormal doubly infinite basis $\{\varphi_{m,n}\}$ for $\mathcal{L}$ such that the functions $\varphi_{m,n}$ are obtained from a single function φ via the action of the set $\mathcal{A}$, and for which condition (5) is valid. For applications, knowledge of the corresponding filters β and α is essential. Since in practical engineering problems the data are normally in digital form, it is important to adapt the theoretical framework of wavelets to the discrete or digital case before discussing other wavelet expansions.

In the digital domain, $l^2(\mathbf{Z})$ replaces $L^2(\mathbf{R})$ as the space of one-dimensional data. One can naturally identify $l^2(\mathbf{Z})$ with $\mathcal{L}_0$, and therefore the theory developed above extends to this case immediately. The only difference is

$$\mathcal{L}_{-j} = \mathcal{L}_0, \quad p_{-j} = id., \quad \text{and} \quad \pi_{-j} = 0 \quad \text{for } j \geq 0 \qquad (9)$$

It follows that formulae (2) through (8) remain valid, provided that the range of the values of m is limited to 0 to ∞. In practice, the domain of n is ($\mathbf{Z}$ mod 2^N) for some integer N. Therefore, $\mathcal{L}_N = \mathbf{R}$, and the linear spaces $\mathcal{L}_m$ are finite dimensional. The bases $\{\chi^m_n\}$ and $\{\chi^{m+1}_n, \varphi^{m+1}_n\}$ for $\mathcal{L}_m = \mathcal{L}_{m+1} \oplus \mathcal{E}_m$ differ by an orthogonal transformation. It follows that the coefficients $\{a^m_n\}$ and $\{a^{m+1}_n, b^m_n\}$ are also related by an orthogonal transformation. This orthogonal transformation, which is the matrix representation of the filters α and β, is given by the $2^N \times 2^N$ matrix with 2×2 diagonal blocks

$$\begin{pmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{pmatrix}$$

This means that given a data set represented by a column vector $(f_0, \ldots, f_{2^N-1})^{tr}$, the application of the above matrix transforms it into a vector $(g_0, \ldots, g_{2^N-1})^{tr}$ with the even-numbered components $(g_0, g_2, \ldots, g_{2^N-2})^{tr}$ representing $p_1(f)$ and the odd-numbered ones $(g_1, g_3, \ldots, g_{2^N-1})^{tr}$ representing $\pi_1(f)$. Here the superscript tr means the transpose of the matrix or vector.

The problem of determining other orthonormal wavelet expansions, and especially the corresponding filters, is discussed in detail in [1]. Of particular interest in practical problems is the case where the functions α and β have small support, i.e., $\alpha(j) = 0 = \beta(j)$ for most j's. It is the knowledge of the functions (or filters) α and β, and not the basis functions themselves, that is essential for applications. In [1], the filters α and β of small support are explicitly determined. The simplest of these filters is the one given above. The next simplest one is the matrix $\mathcal{F}$ given by

$$\mathcal{F} =$$

$$\begin{pmatrix} \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & 0 & 0 & 0 & \cdots & \cdots & 0 \\ \alpha_3 & -\alpha_2 & \alpha_1 & -\alpha_0 & 0 & 0 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & \alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & 0 & \cdots & \cdots & 0 \\ 0 & 0 & \alpha_3 & -\alpha_2 & \alpha_1 & -\alpha_0 & 0 & \cdots & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \ddots & \ddots & & & \vdots \\ \alpha_2 & \alpha_3 & 0 & 0 & \cdots & \cdots & \cdots & 0 & \alpha_0 & \alpha_1 \\ \alpha_1 & -\alpha_0 & 0 & 0 & \cdots & \cdots & \cdots & 0 & \alpha_3 & -\alpha_2 \end{pmatrix} \tag{10}$$

where

$$\alpha_0 = \frac{1+\sqrt{3}}{4\sqrt{2}},\ \alpha_1 = \frac{3+\sqrt{3}}{4\sqrt{2}},\ \alpha_2 = \frac{3-\sqrt{3}}{4\sqrt{2}},\ \alpha_3 = \frac{1-\sqrt{3}}{4\sqrt{2}}$$

One should note that an important feature of the orthonormal wavelet expansion is that the inversion procedure can be implemented by the transpose of the orthogonal matrix representing the filters α and β.

The above theory was limited to one-dimensional data. It can be easily adapted to the two-dimensional case by considering products of the basis functions considered in the one-dimensional case. This is equivalent to carrying out the one-dimensional wavelet transforms in the horizontal and vertical directions. The practical aspects of the two-dimensional wavelet transform are discussed in detail in the next section. Of course, there are orthonormal wavelet expansions that may not be separable, i.e., the basis functions are not products of the basis functions for the one-dimensional case, but they will not be considered in this article.

III. Application to Data Compression

To apply the theory to data compression, one fixes an orthonormal wavelet expansion, or equivalently, the filters α and β. In the work reported here, only the filter defined by the matrix $\mathcal{F}$ was used.[1] An image is represented by a matrix $f = (f_{ij})$, where f_{ij} is the intensity of the pixel (i,j). For each fixed row i, one considers the transform $g_i = F f_i^{tr}$, where f_i is the ith row of the matrix f. The components of g_i with even indices represent $p_1(f_i)$ and those with odd indices represent $\pi_1(f_i)$. It is convenient to reorganize the vector g_i in the form $(g_0, g_2, \ldots, g_{2^N-2}, g_1, g_3, \ldots, g_{2^N-1})$. Now the process is repeated for the columns of the matrix of the transformed rows. After reorganizing, the transformed matrix of pixel intensities takes the form

$$g = \begin{pmatrix} g^{11} & g^{12} \\ g^{21} & g^{22} \end{pmatrix}$$

where each g^{ij} is a $2^{N-1} \times 2^{N-1}$ matrix. Since an image is two-dimensional, the hierarchy of the wavelet coefficients requires some elaboration. The matrix g^{11} represents the smoothed version of the image, while the remaining coefficients are the missing details. The coefficients g^{12} and g^{21} belong to the level of the pyramid immediately below g^{11}, and g^{22} lies at the lowest level of the pyramid. Thus, every application of the wavelet transform generates three levels of hierarchy for a two-dimensional image. The process is then repeated by applying the filters α and β to the $2^{N-1} \times 2^{N-1}$ matrix g^{11} along rows and columns. The resulting coefficients are then reorganized in the form

$$g^{11,11} \rightarrow \{g^{11,12}, g^{11,21}\} \rightarrow g^{11,22} \rightarrow \{g^{12}, g^{21}\} \rightarrow g^{22}$$

with the highest level at the extreme left and the lowest at the extreme right. The process can be repeated. It may be more convenient to organize the coefficients differently in the following form:

$$\begin{tikzcd} & g^{11,11} \arrow[dl] \arrow[dr] & \\ g^{11,12}, g^{11,21} \arrow[d] & & g^{11,22} \arrow[d] \\ g^{12}, g^{21} & & g^{22} \end{tikzcd}$$

One then refers to $g^{11,11}$ as SS (smooth-smooth) level 2, to $g^{11,12}, g^{11,21}$ and g^{12}, g^{21} as the SD (smooth-detail) levels 2 and 1, respectively, and to $g^{11,22}$ and g^{22} as the DD (detail-detail) levels 2 and 1, respectively.

In the application of wavelet transforms to image compression, the coefficients at different levels of the pyramid are not equally significant and, therefore, should be encoded differently. The wavelet coefficients of different levels were examined for several images, and certain patterns were observed. In general, the coefficients at a lower level of the pyramid are better approximated by a Laplacian density function than those at the higher levels. Using the nearest integer truncation, one also notices that the entropies of the coefficients at the lower levels are smaller

[1] It is often unclear from the literature what filter is actually used. The filter used in [3] differs from that defined by $\mathcal{F}$.

than those in the upper ones. Figures 1 through 4 show the distributions of the wavelet coefficients at different levels of the hierarchy for a typical image. An approximating Laplacian density function is given in Figs. 1 through 3. Clearly the coefficients at the highest level (Fig. 4) have a very irregular distribution. Table 1 shows the entropies of the wavelet coefficients for the same image.

The image compression process is done by first computing the coefficients $g^{22}, g^{12}, g^{21}, g^{11,22}, g^{11,12}$, etc. These coefficients are quantized according to a bit allocation scheme similar to the one used for the standard DCT-based algorithms. As noted above, more bits are allocated to the higher levels of the hierarchy than to the lower ones. In the pictures of the peppers compressed by the wavelet transform method (Fig. 1), the coefficients in the lowest level have been set to 0. In practice it was observed that because of the quantization errors inherent in any floating-point computation, it is not desirable to go beyond three or four levels of wavelet transforms. To reconstruct the image, the inverse filter was applied to the coefficients. As noted above, the inverse filter is given by the transpose of the orthogonal matrix defining the filter.

There are several issues involved in the application of wavelet transforms to image compression. The choice of the appropriate wavelet transform may be dictated by the complexity of the image. It has been suggested that different transforms may be more appropriate for different images or even different parts of an image. Some ideas in this direction appear in [2] with apparently very promising results. The problem of bit allocation and quantization of the wavelet coefficients is similar to the analogous problem for DCT-based image compression. It may be possible to take advantage of the regularity of the coefficients at the lower levels of the pyramid and use the Laplacian distribution to allocate bits accordingly. However, the experimental work carried out by the author suggests that the simpler method of truncation to the nearest integer followed by decimation by an appropriate number of bits provides better results. Naturally, fewer bits are allocated to the lower levels of the pyramid than to the upper levels. A different method for quantization is proposed in [3]. These authors suggest that using the L^1 rather than the L^2 norm is more compatible with the human visual perception, and their proposed technique of quantization method is based on minimizing the errors in the former norm.

While a definitive comparison between the DCT-based algorithms and wavelet transform techniques is premature, the tests done by the author suggest some important differences. At higher compression ratios, for example at greater than 25:1, the blockiness in the DCT-based techniques becomes very visible. With the wavelet transform used in the tests, the edges were not as clearly defined as those using the DCT-based techniques, but no blockiness was visible. The rms error of the Joint Photographic Experts Group (JPEG) DCT-based algorithm was smaller than that of the wavelet transform method, but visual preference is not necessarily reflected by the mean square error. Figure 5 shows an original image (peppers) on the upper left corner. The images on the upper right and lower left were compressed using the wavelet transform. The compression ratios were 10:1 and 30:1, respectively. The image in the lower right was obtained by the application of the standard DCT-based JPEG algorithm. Its compression ratio is 30:1. The rms error for the lower left image is about 11.0, and for the one at lower right it is approximately 7.2, even though the blockiness makes it much worse than the one at lower left. The rms error for the image on the upper right is about 9.4. It should be pointed out that other methods, such as fractal algorithms, may produce images that are visually preferable to the DCT-based methods for high compression ratios.

The wavelet transform used in this work is the product of a one-dimensional algorithm with itself; that is, essentially separable into one-dimensional algorithms. It is possible to modify this method to make the horizontal and vertical directions more coupled so that the algorithm becomes truly two-dimensional. The visual effects of such modification are unclear at this time. However, it is reasonable to expect improvements in the clarity of the edges if such techniques are properly employed.

IV. Conclusion

The wavelet transform method provides a new approach to image compression. Although this approach has not performed as well as the DCT-based algorithms in terms of the rms error, it appears to have certain visual advantages especially regarding blockiness.

Acknowledgments

The author wishes to thank Dr. A. Zandi of the University of California at Santa Cruz and IBM (Almaden) for helpful discussions, and Dr. F. Pollara of JPL for providing the DCT-based compressed image of peppers.

References

[1] I. Daubechies,"Orthonormal Bases of Compactly Supported Wavelets," *Communications on Pure and Applied Mathematics*, vol. 41, pp. 909–996, 1988.

[2] R. R. Coifman and M. V. Wickerhausen,"Entropy-Based Algorithms for Best Basis Selection," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 713–718, March 1992.

[3] R. A. DeVore, B. Jawerth, and B. J. Lucier, "Image Compression Through Wavelet Transform Coding," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 719–746, March 1992.

Table 1. Entropies of the wavelet coefficients.

Level	Entropy
DD-1	2.5
SD-1	3.2
DD-2	2.8
SD-2	3.8
DD-3	3.4
SD-3	4.5
DD-4	4.0
SD-4	5.0
SS-4	6.4

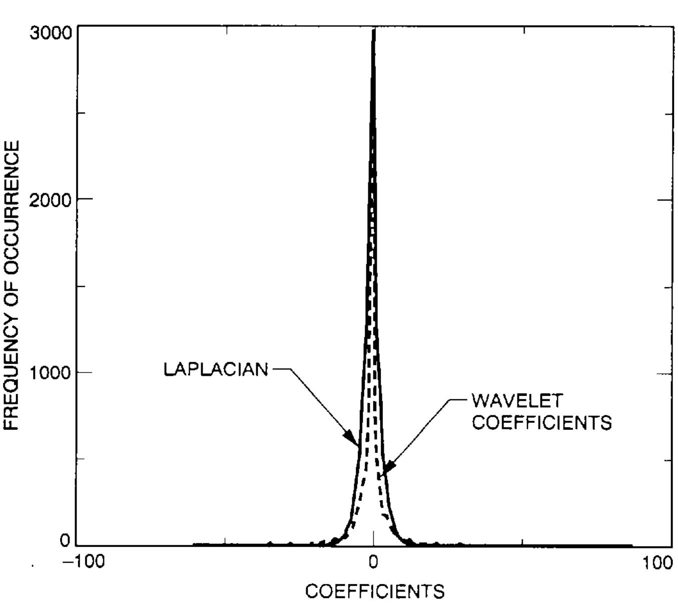

Fig. 1. Level DS-2 wavelet coefficients and Laplacian density.

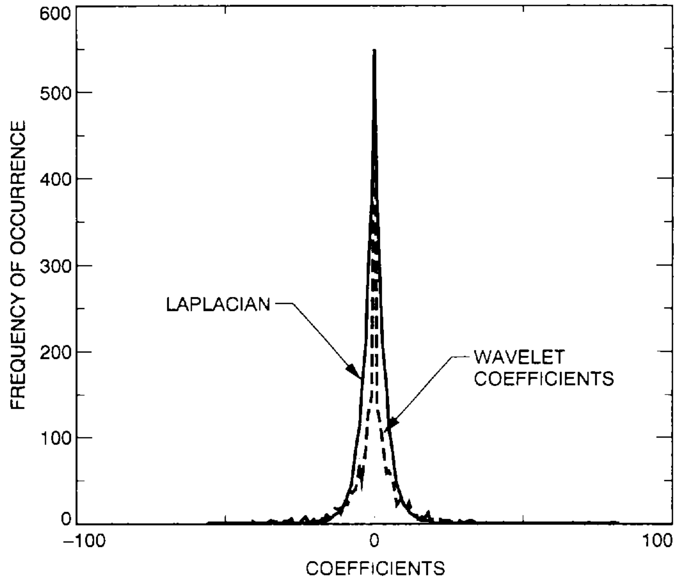

Fig. 2. Level DS-3 wavelet coefficients and Laplacian density.

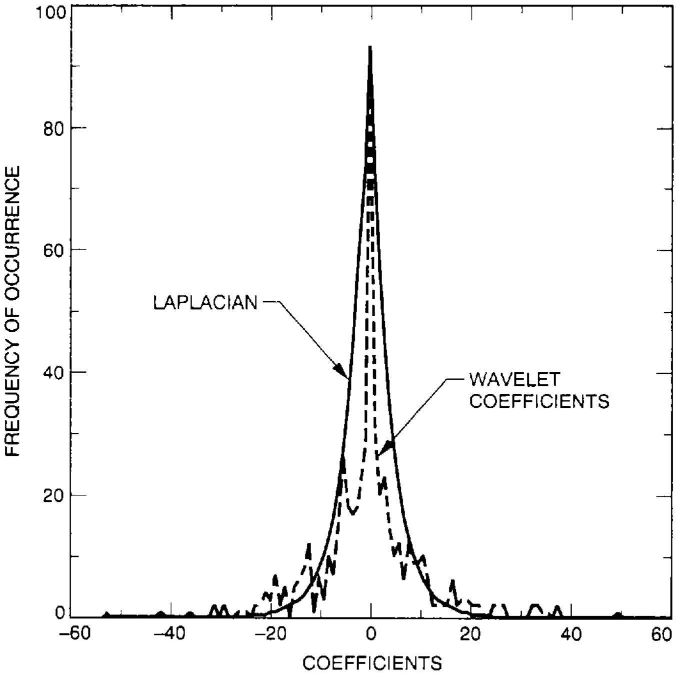

Fig. 3. Level DS-4 wavelet coefficients and Laplacian density.

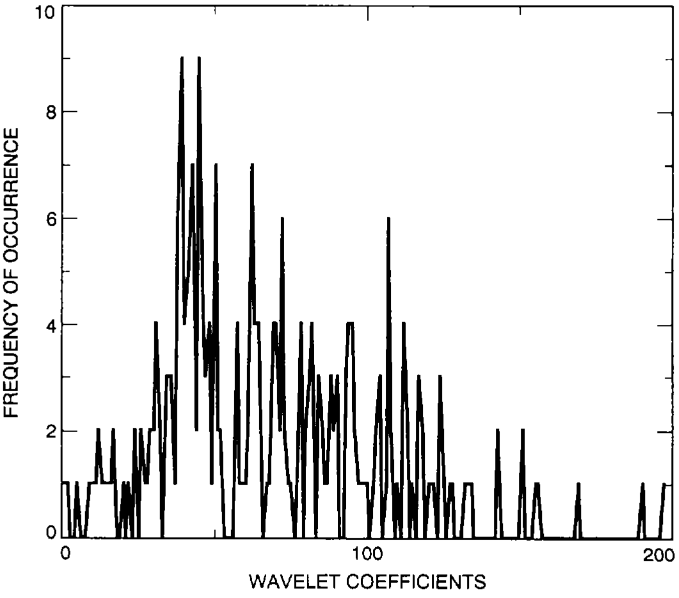

Fig. 4. Level SS-4 wavelet coefficients.

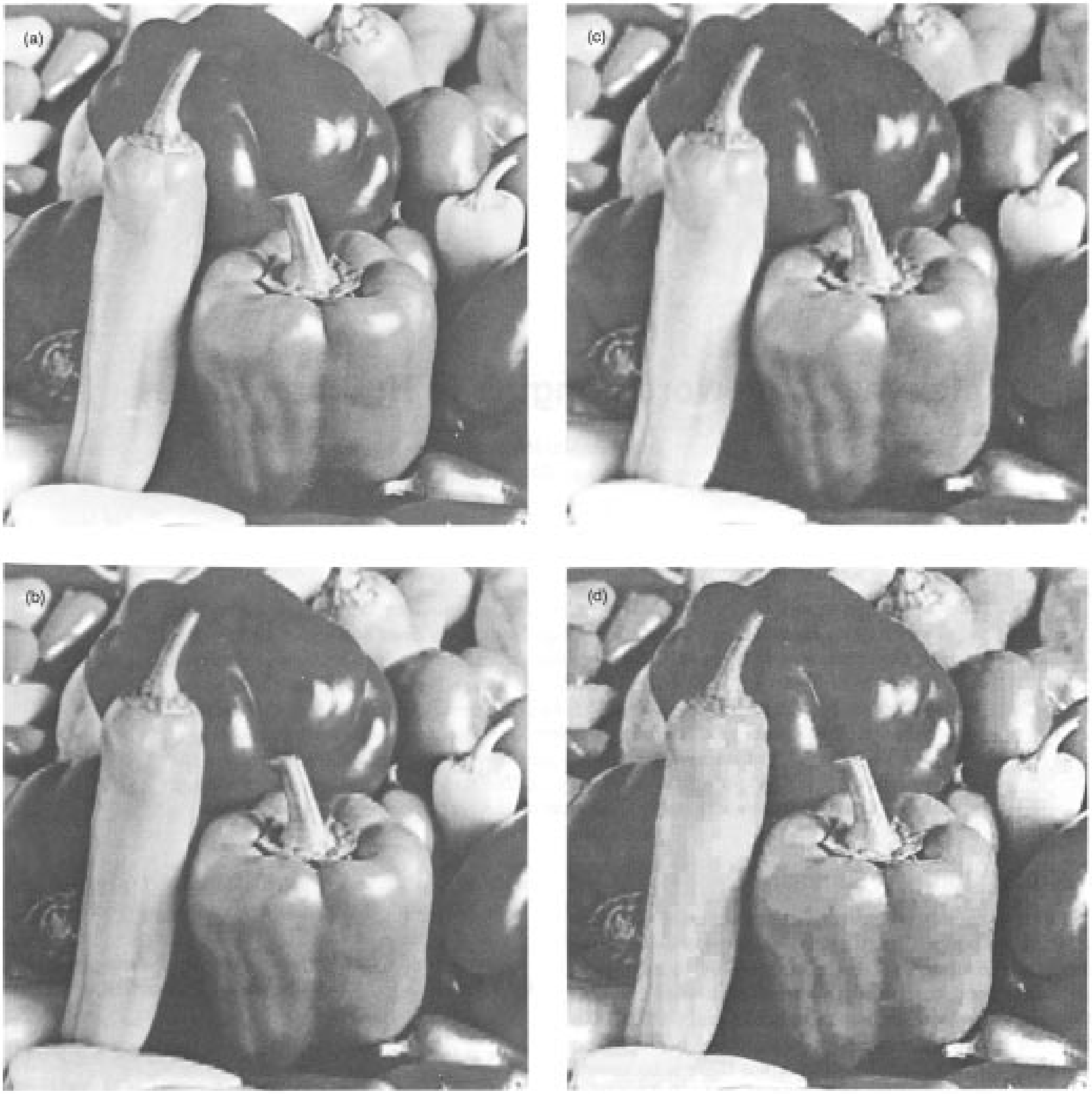

Fig. 5. Image compression: (a) the original uncompressed image; (b) compression ratio of 10:1 by wavelet transform; (c) compression ratio of 30:1 by wavelet transform; and (d) compression ratio of 30:1 by DCT-based JPEG algorithm.