Image Compression Through Wavelet Transform Coding

Ronald A. DeVore, Björn Jawerth, and Bradley J. Lucier, *Member, IEEE*

Abstract—**A new theory is introduced for analyzing image compression methods that are based on compression of wavelet decompositions. This theory precisely relates a) the rate of decay in the error between the original image and the compressed image (measured in one of a family of so-called L^p norms) as the size of the compressed image representation increases (i.e., as the amount of compression decreases) to b) the smoothness of the image in certain smoothness classes called Besov spaces. Within this theory, the error incurred by the quantization of wavelet transform coefficients is explained. Several compression algorithms based on piecewise constant approximations are analyzed in some detail. It is shown that if pictures can be characterized by their membership in the smoothness classes considered here, then wavelet-based methods are near optimal within a larger class of stable (in a particular mathematical sense) transform-based, nonlinear methods of image compression. Based on previous experimental research on the spatial-frequency–intensity response of the human visual system, it is argued that in most instances the error incurred in image compression should be measured in the integral (L^1) sense instead of the mean-square (L^2) sense.**

Index Terms—**Image compression, wavelets, smoothness of images, quantization.**

I. Introduction

IMAGE compression methods that employ pyramid encoding, quadrature mirror filters, or so-called wavelet transforms (see [11] for a somewhat mathematical overview) have been successful in providing high rates of compression while maintaining good image quality. In this paper, we present a new mathematical theory for analyzing these wavelet-based compression methods. Our theory precisely relates a) the rate of decay in the error between the original image and the compressed image (measured in one of a family of so-called L^p norms) as the size of the compressed image representation increases (i.e., as the amount of compression decreases) to b) the smoothness of the image in certain smoothness classes called Besov spaces. In particular, our theory bounds the error incurred by quantizing wavelet transform coefficients. We introduce and analyze several algorithms based on piecewise-constant wavelet approximations; additionally, for these algorithms, our theory bounds the errors introduced by quantizing pixels and using fixed-point arithmetic. More generally, we suggest a unified mathematical framework that is useful in analyzing any transform coding method for image compression. We show that if images can be characterized by their membership in the smoothness classes we consider, then wavelet-based methods are near optimal within a general class of stable, transform-based, nonlinear methods of image compression. We argue that psychological data of the spatial-frequency–amplitude response of the human visual system, summarized by the Contrast Sensitivity Threshold curve, can help to choose an image quality metric from the class of L^p metrics; in particular, we argue that the L^1 (mean-absolute) error metric is more appropriate for measuring the error of image compression than the L^2 (mean-square) error metric. Finally, our analysis, which is based on models from nonlinear approximation theory and harmonic analysis rather than from probability theory, provides a direct and practical way to estimate the smoothness of images. In this introduction, we shall put our work into mathematical and practical perspective and give an overview of the remainder of the paper.

By an image, we shall mean a digitized grey scale picture that consists of 2^m by 2^m pixels (typically, $7 \le m \le 11$), each of which takes a value between 0 and $2^n - 1$ (typically, $n = 8$). We shall denote the value of the pixel in row j_1 and the column j_2 of the image by p_j, $j := (j_1, j_2)$.

Whereas transform coding is most often described solely in terms of the discrete pixel values p_j, our analysis is based on interpreting the image as a function f defined on the unit square $I := [0, 1]^2$. We view the image compression problem as one of approximating f by a second (compressed) function $\tilde{f}$. The object of such a compression algorithm will be to represent certain classes of pictures with less information than was used to represent the original pictures. For a lossless algorithm, the original and compressed images will be the same, and the error between them will be zero. We shall generally consider algorithms that introduce differences between the original and compressed images in order to achieve higher compression levels.

While one could associate to each image a function f that is independent of the transform being applied, it seems more natural (and amenable to our analysis) to allow the representation to depend on the transform. For example, when we apply the Haar transform, or any transform whose terms can

Manuscript received February 15, 1991; revised September 1, 1991. This work was supported in part by the National Science Foundation (Grants DMS-8803585, DMS-8922154, and DMS-9006219), the Air Force Office of Scientific Research (Contract 89-0455-DEF), the Office of Naval Research (Contracts N00014-90-1343, N00014-91-J-1152, and N00014-91-J-1076), the Defense Advanced Research Projects Agency (AFOSR Contract 90-0323), and the Army High Performance Computing Research Center.

R. A. DeVore and B. Jawerth are with the Department of Mathematics, University of South Carolina, Columbia, SC 29208.

B. J. Lucier is with the Department of Mathematics, Purdue University, West Lafayette, IN 47907.

IEEE Log Number 9104594.

be interpreted as being constant on square subdomains of pixels, we shall associate to the image the function $f(x)$ defined for $x := (x_1, x_2)$ in I by

$$f(x) := p_j, \qquad \text{for} \quad \frac{j_1}{2^m} \le x_1 < \frac{j_1 + 1}{2^m}$$
$$\text{and} \quad \frac{j_2}{2^m} \le x_2 < \frac{j_2 + 1}{2^m}.$$

Thus, the discrete Haar transform of the pixel data yields the same coefficients as the continuous Haar transform of the continuous function f. We shall follow the same principle when analyzing other transform methods that are based on mathematical expansions of functions (such as the discrete cosine transform and wavelet transforms), and associate to each set of pixel values the function

$$f = \sum_{k \ge 0,\, j = (j_1, j_2)} c_{j,k} \phi_{j,k}, \tag{1.1}$$

whose coefficients $c_{j,k}$ are the result of the discrete transform applied to the pixel data. (For the discrete cosine transform, for example, the functions $\phi_{j,k}$ are the products of cosines in x and y, and f is the trigonometric polynomial that satisfies $f(j/2^m) = p_j$.) Therefore, a transform depends both on the choice of representation functions $\phi_{j,k}$ and the method of determining the coefficients $c_{j,k}$. If the functions $\phi_{j,k}$ are redundant (or, equivalently, if they are linearly dependent), then there may be more than one way to calculate the coefficients $c_{j,k}$, and so to represent the function f by an expansion of the form (1.1). However, a given compression algorithm begins by fixing such a representation, i.e., by calculating the coefficients in a fixed specific way.

To repeat, the transform associates to the given pixel values a new sequence of numbers $c_{j,k}$ that are interpreted, by (1.1), as the coefficients of the expansion of a function f, which we take to be the representation of the image. Given the transform, the algorithm then calculates quantized coefficients $\tilde{c}_{j,k}$, and the compressed function takes the form

$$\tilde{f} = \sum \tilde{c}_{j,k} \phi_{j,k}. \tag{1.2}$$

The method of quantizing coefficients involves applying a strategy, which we consider fixed, that depends on one or more parameters (number of coefficients, global picture quality, local picture quality, etc.), which are allowed to vary. We store or transmit a coded representation of the coefficients $\tilde{c}_{j,k}$, typically through some type of entropy coding. Once we decide on an algorithm, we can apply it not only to representations of images but to any function f for which the continuous decomposition (1.1) can be calculated.

The description (1.2) is general enough to include discrete cosine transform coding, pyramid encoding, multiresolution schemes based on wavelets or box splines, etc. We shall concentrate on the latter methods in which the representation functions $\phi_{j,k}$ typically have a characteristic frequency of $O(2^k)$ and are supported in a square of side length $O(2^{-k})$; $j = (j_1, j_2)$ will be a multiindex indicating the location of the support of $\phi_{j,k}$. For many methods there is a single function ϕ such that for all j and k, $\phi_{j,k}(x) := \phi(2^k x - j) = \phi(2^k(x - j/2^k))$.

In providing a mathematical framework for image compression algorithms, one confronts the following fundamental questions.

1) In what metric should the error be measured?
2) How should one measure the efficiency of algorithms?
3) For which pictures does an algorithm give good results?
4) Is there an optimal level of compression that cannot be exceeded within a given class of compression algorithms and pictures?
5) What are near-optimal algorithms?

One must decide how to measure the error between f and $\tilde{f}$. Some researchers have used the mean-square error

$$\| f - \tilde{f} \|_{L^2(I)} := \left(\int_I | f(x) - \tilde{f}(x) |^2 \, dx \right)^{1/2},$$

usually without *a priori* justification. In practice, one desires a metric that parallels the human visual system, with the hope that image differences judged to be large by the human eye are mathematically large and image differences which, for whatever reason, are insignificant to the eye will have small size in the error metric. There are many possible choices of such a metric; we shall investigate (somewhat arbitrarily) the use of the $L^p(I)$ norms with $0 < p < \infty$ as error metrics. These norms, defined by

$$\| f - \tilde{f} \|_{L^p(I)} := \left(\int_I | f(x) - \tilde{f}(x) |^p \, dx \right)^{1/p},$$

include as special cases the mean-square error and the mean-absolute error

$$\| f - \tilde{f} \|_{L^1(I)} := \int_I | f(x) - \tilde{f}(x) | \, dx.$$

The parameter p gives added flexibility, in that the relative sizes of the component functions $c_{j,k}\phi_{j,k}$ with contrast $c_{j,k}$ and frequency 2^k, given by $\| c_{j,k}\phi_{j,k} \|_{L^p(I)}$, can be changed by varying the parameter p. In other words, varying p allows us to change the relative importance of contrast and frequency in measuring the size of basic functions. We argue in Section IV-A that 1) within the scale of L^p spaces, 2) with a compression scheme that keeps the low and middle frequency information, and 3) to be consistent with data from the contrast sensitivity threshold (CST) curve, the choice of p that best matches the properties of the human visual system is $p = 1$. We give examples in Section V-B which show that attempting to minimize the error in $L^1(I)$ leads to more pleasing pictures than in $L^2(I)$.

While we believe that the scaling given by the $L^1(I)$ metric (namely, $\|\phi_{j,k}\|_{L^1(I)} = 4\|\phi_{i,k+1}\|_{L^1(I)}$ when $\phi_{j,k}(x) = \phi(2^k x - j)$ for some ϕ) is the correct one for high-frequency representation functions, the same scaling holds for spaces other than $L^1(I)$, such as the Sobolev space $W^{-1/2,2}(I)$ and the Hardy space $H^1(I)$. Even though it may be true that for

a given family of representation functions $\phi_{j,k}$ we have

$$\|\phi_{j,k}\|_{L^1(I)} = \|\phi_{j,k}\|_{W^{-1/2,2}(I)} = \|\phi_{j,k}\|_{H^1(I)},$$

it is *not* true for arbitrary f that

$$\|f\|_{L^1(I)} = \|f\|_{W^{-1/2,2}(I)} = \|f\|_{H^1(I)},$$

because the different norms of f are calculated by combining in different ways the norms of the representation functions making up f. The CST curve by itself does not address how the eye sees *combinations* of representation functions, so other experiments will be needed to provide this new information.

After deciding on a space X whose metric $\|\cdot\|_X$ will be used to measure the error between f and $\tilde{f}$, we address the question of how to measure the efficiency of a given algorithm. Recall that an algorithm depends on three things: the choice of representation functions $\phi_{j,k}$, the method of calculating the coefficients $c_{j,k}$ (which together we call the transform), and the quantization strategy. A given algorithm generates different compressed functions $\tilde{f}$ depending on the parameters of the quantization strategy.

We shall evaluate algorithms by comparing the error $\|f - \tilde{f}\|_X$ to the number of nonzero quantized coefficients $\tilde{c}_{j,k}$. (In practice, one is more interested in the number of bits necessary to represent the quantized coefficients $\tilde{c}_{j,k}$. The two measures correlate well in practice; see Section V-B.) Suppose that a compression algorithm produces a family $\{\tilde{f}\}$ of compressed images corresponding to different parameters in the quantization strategy. We introduce for this algorithm the error function

$$a_N(f)_X := \inf_{\tilde{f} \text{ has} \le N \text{ coefficients}} \|f - \tilde{f}\|_X. \qquad (1.3)$$

In other words, a_N measures the compression error that is achieved if the number of coefficients in the compressed function does not exceed N.

Given two algorithms with their respective errors $\tilde{a}_N$ and $\bar{a}_N$, one would obviously say that the first is better than the second if

$$\tilde{a}_N(f)_X \le \bar{a}_N(f)_X,$$

for each function f. It is very unlikely that such a relationship would hold, since each reasonable algorithm is good for some pictures and not so good for others. A more meaningful comparison is to consider the class of functions f for which $a_N(f)_X$ decays at a prescribed rate as N get large. For example, we shall call the α class of an algorithm, $\alpha > 0$, the set of functions f that satisfy

$$a_N(f)_X = O(N^{-\alpha}), \qquad \text{as } N \text{ tends to infinity}.$$

We say that one algorithm is better than another if its α class contains the other's for some range of α. One can describe, within this framework, optimal compression algorithms.

Before doing so, we ask the following question: If the $\phi_{j,k} = \phi(2^k \cdot - j)$ are fixed, how smooth are functions that can be approximated to $O(N^{-\alpha})$ with $\le N$ coefficients by algorithms that use the functions $\phi_{j,k}$. For the spaces $X = L^p$ and for many classes of representation functions $\phi_{j,k}$, DeVore, Jawerth, and Popov [13] have shown that, roughly speaking,

$$\inf_{\text{all algorithms using } \phi_{j,k}} a_N(f)_{L^p(I)} = O(N^{-\alpha/2}) \Leftrightarrow f \in B_q^\alpha(L^q(I)), \qquad (1.4)$$

where $q = 1/(\alpha/2 + 1/p)$ and the Besov space $B_q^\alpha(L^q(I))$ consists of functions that have α bounded "derivatives" in $L^q(I)$. (The previous statement is inaccurate; a complete definition of Besov spaces is given in Section III-B, and a precise statement of (1.4) is given in Section II.) In particular, this is true for box splines [3] (with piecewise constant approximations as a special case) when $0 < p \le \infty$ (see [13] for $p < \infty$ and [14] for $p = \infty$) and orthogonal wavelets [20] (of which the Haar transform is a special case) when $1 < p < \infty$. So, if we consider N, the number of coefficients in the representation of $\tilde{f}$, to be a measure of the amount of information one must use to represent the compressed image, one can hope to achieve a particular rate of error decay in $L^p(I)$, if and only if f is in a specific Besov smoothness space $B_q^\alpha(L^q(I))$. In proving this theorem, DeVore *et al.* provide specific algorithms for each set of functions $\phi_{j,k}$ that give the optimal rate of convergence. One should note that for a wide class of representation functions $\phi_{j,k}$, the optimal selection of coefficients results in the same α classes. Nonetheless, there are various constants hidden in the big-O notation that may determine whether one set of representation functions is better in practice than others.

The equivalence (1.4) suggests that membership in Besov spaces $B_q^\alpha(L^q(I))$ is an appropriate way of classifying images, in that we can check the effectiveness of a given compression algorithm by seeing how it performs on functions in $B_q^\alpha(L^q(I))$. However, it is of practical interest to measure smoothness in the spaces $B_q^\alpha(L^q(I))$ only if common types of images are in these spaces. (For example, one can easily show that linear wavelet approximations to f converge in $L^2(I)$ at a rate $O(N^{-\alpha/2})$ if f is in a Sobolev space $W^{\alpha,2}(I)$, but no image with a jump discontinuity in intensity across a one-dimensional curve is in this space if $\alpha \ge 1/2$.) For approximations in $L^1(I)$, we show empirically in Section V-B that common head-and-shoulders and outdoor images are in $B_q^\alpha(L^q(I))$ for $\alpha \approx 0.5$. This can be interpreted in two ways: Our methods are of practical interest, and the previous successes of pyramid schemes can be explained by our analysis.

To say that a function $f \in B_q^\alpha(L^q(I))$ has enough smoothness to be approximated to $O(N^{-\alpha/2})$ in $L^p(I)$ by functions of the form $\tilde{f}$, does not, in and of itself, explain how to find an algorithm to achieve this. The latter rests on two main issues: finding a suitable representation (1.1) (i.e., correctly calculating the coefficients for the given value of p) and the method of quantizing the coefficients (which will also depend on p). We shall discuss these issues in a general setting of wavelet based transform methods in Section II. In Section III, we consider in more detail transform methods based on approximation by piecewise constant functions. This corresponds to decompositions (1.1) where the $\phi_{j,k}$ are Haar functions or are characteristic functions of cubes.

In Section II, we briefly describe how wavelet decompositions obtained by multiresolution can be incorporated into compression algorithms. The issues we study in Section II include: different basis functions $\phi_{j,k}$, different methods of choosing the coefficients $c_{j,k}$, how to choose the coefficients $\tilde{c}_{j,k}$ in the compressed image $\tilde{f}$ of (1.2) if the compression error is to be measured in the $L^p(I)$ norm, and the relationship between the compression error and the smoothness of pictures. We present a general transform algorithm based on wavelet decompositions that has (in a certain mathematical sense) optimal compression. We also discuss a general way to measure the optimality of compression algorithms by using the mathematical concept of n-widths. In particular, we note that our methods are optimal in a wider class of methods than those in which $\tilde{f}$ is represented by (1.2). DeVore and Yu [16] have shown that if 1) the smoothness of f is measured in a Besov space $B_q^\alpha(L^q(I))$ with $1/q = \alpha/2 + 1/p$, and 2) $\tilde{f}$ is derived from *any* approximation process that has a certain "continuous selection" property (roughly speaking, changing f a little changes the representation of $\tilde{f}$ only a little) then

$$\sup_{\|f\|_{B_q^\alpha(L^q(I))} \le 1} \inf_{\tilde{f} \text{ has } N \text{ parameters}} \|f - \tilde{f}\|_{L^p(I)} \ge CN^{-\alpha/2}.$$

In other words, *any* approximation process satisfying the continuous selection property can achieve at most an approximation rate of $O(N^{-\alpha/2})$. So if one believes that the Besov space norm of an image completely characterizes the smoothness of that image, and if one is willing to limit algorithmic considerations to methods that satisfy the continuous selection property, the family of methods we describe has optimal order accuracy.

As an application of the theory in Section II, we consider approximation by piecewise constants in Section III. The construction of compression methods, in this case, is related to approximation by constants on cubes, which is discussed in Section III-A. Among other things, we show that median operator is a good method of constant approximation for all L^p, $0 < p \le \infty$ while averaging is good only for $p > 1$. This helps us explain when Haar functions should be used in (1.1) (namely, for $p > 1$) and the theoretical advantages of clipping and median transforms in image compression. This will be important for our image applications when approximating in $L^1(I)$.

In Section III-B, we introduce the function spaces containing the images that can be approximated well by wavelet transform coding, and we relate previous results about equivalent norms of functions in $B_q^\alpha(L^q(I))$ based on the size of the coefficients $c_{j,k}$ in the representation (1.1). In Section III-C we analyze the effects of pixel quantization (spatial averaging and rounding to discrete pixel values) in these function spaces. In Section III-D, these results are used to derive a family of algorithms that give optimal compression in a certain mathematical sense. In particular, we derive error bounds for these algorithms that give a compressed function $\tilde{f}$ with no more than $O(\mathcal{N})$ coefficients that satisfies

$$\|f - \tilde{f}\|_{L^p(I)} \le O(\mathcal{N}^{-\alpha/2}),$$

whenever $f \in B_q^\alpha(L^q(I))$. Because of (1.4), the order of this approximation is optimal within a large class of transform algorithms. Further in Section III-D, we apply our theory to the example of *progressive transmission*, and show that one can achieve faster convergence for many images by using our techniques rather than using previously suggested ones [7]. In Section III-E we give several specific examples of piecewise constant transforms that satisfy the theory of the previous section. In Section III-F, we briefly mention high-order wavelet approximations that satisfy the theory of Sections II and III.

In Section IV-A, we discuss our interpretation of the contrast sensitivity threshold curve that leads us to believe that within the scale of $L^p(I)$ spaces one should measure the error in $L^1(I)$ for image compression. In Section IV-B, we show that because most images have spatial discontinuities in intensity, one can expect $\alpha < 2/p$. Therefore, for $p = 1$ we can, *a priori*, expect at most second order smoothness for images. This section also discusses how to estimate empirically the smoothness of images; for the images we have tested in Section V-B one has $0.3 \le \alpha \le 0.6$. Later, in Section IV-C, we interpret coefficient quantization levels in terms of approximation in $L^p(I)$. For example, our analysis shows that if the interval between quantized coefficients doubles when the frequency doubles, then the approximation is effectively in $L^2(I)$; if the quantization interval quadruples when the frequency doubles then the approximation is in $L^1(I)$.

Section V contains computational results. In Section V-A we describe the implementation of several compression algorithms that satisfy the assumptions of the theory in Section III. In Section V-B, we contrast the results of algorithms that attempt to minimize the error in $L^2(I)$ and $L^1(I)$. Further, we use (1.4) to estimate *a posteriori* whether various images are in $B_q^\alpha(L^q(I))$ and, if so, to estimate their $B_q^\alpha(L^q(I))$ norm. Finally, we report on the empirical relationship observed between the number of nonzero coefficients $\tilde{c}_{j,k}$ in $\tilde{f}$ (our theoretical measure of compression) and the number of bytes required to encode the representation of those coefficients.

The Appendix contains the proofs of theorems in Section III.

II. Wavelet Decompositions and Compression

A. Wavelet Decompositions

We shall describe in this section a generic method for obtaining decompositions (1.1). Specifically, we give our own perspective of multiresolution analysis as introduced by Meyer [20], Mallat [19], and Daubechies [11] in their construction of orthogonal wavelets. While the mathematical framework for image compression developed in the following sections is not limited to this case, it does form the primary examples of our theory. It will be convenient to begin by discussing infinite expansions (1.1) that hold for all functions

in the space $L^p(\mathbb{R}^d)$, $0 < p < \infty$ (all uniformly continuous functions f when $p = \infty$) where $d = 1, 2, \cdots$. The expansions we discuss will include as special cases the orthogonal wavelets as developed by Meyer [20], Mallat [19], and Daubechies [11], B-spline and box spline expansions (cf. Chui [8]), and the more recent nonorthogonal wavelets (sometimes called prewavelets) of Battle [1] and Chui-Wang [9]. Techniques other than multiresolution can lead to function decompositions of the type considered here; for example, Frazier and Jawerth [17] have derived function decompositions that are based on the Calderon reproducing formula.

The starting point for our construction is the refinement equation

$$\phi(x) = \sum_{j\in\mathbb{Z}^d} a_j \phi(2x - j) \tag{2.1}$$

where $(a_j)_{j\in\mathbb{Z}^d}$ is a finite sequence of real numbers. (Here $\mathbb{Z}$ denotes the integers, and $\mathbb{Z}^d$ denotes the set of multiindices $(j_1, \cdots, j_d)$ with $j_i \in \mathbb{Z}$.) There are various sufficient conditions [11] and [10] on the sequence (a_j) that guarantee that there is a (unique) solution to (2.1). We shall assume that such a function ϕ exists, has compact support, and that the integer translates $\phi(\cdot - j)$, $j \in \mathbb{Z}^d$, of ϕ are linearly independent.

Let $V := \operatorname{span}\{\phi(\cdot - j) \mid j \in \mathbb{Z}^d\}$. By dilation, we obtain a scale of spaces $V_k := \operatorname{span}\{\phi(2^k \cdot - j) \mid j \in \mathbb{Z}^d\}$, $k \in \mathbb{Z}$. From (2.1), it follows that $V_{k-1} \subset V_k$, $k \in \mathbb{Z}$.

We fix p and look for expansions (1.1) that hold for all $f \in L^p(\mathbb{R}^d)$. Let P be a bounded projector from $L^p(\mathbb{R}^d)$ onto V. By dilation, we obtain uniformly bounded projectors P_k of $L^p(\mathbb{R}^d)$ onto V_k for each $k \in \mathbb{Z}$. That is, $P_k := D_k P D_{-k}$ with $D_k f(x) := f(2^k x)$ a dilation operator. For example, in the case $p = 2$, we could let P be the orthogonal projector onto V, which is the best $L^2(\mathbb{R}^d)$ approximation to f from V. We shall assume that $\bigcup V_k$ is dense in $L^p(\mathbb{R}^d)$. It then follows that $P_k f \to f$ in $L^p(\mathbb{R}^d)$ for each $f \in L^p(\mathbb{R}^d)$. Indeed, for any $g_k \in V_k$,

$$\begin{aligned} \|f - P_k f\|_{L^p(\mathbb{R}^d)} &= \|(I - P_k) f\|_{L^p(\mathbb{R}^d)} \\ &= \|(I - P_k)(f - g_k)\|_{L^p(\mathbb{R}^d)} \\ &\le C \|f - g_k\|_{L^p(\mathbb{R}^d)}. \end{aligned}$$

The right side of the last inequality tends to 0 with $k \to \infty$ because of the denseness of $\bigcup V_k$.

We shall also assume that $\|P_k f\|_{L^p(\mathbb{R}^d)} \to 0$ as $k \to -\infty$ for each $f \in L^p(\mathbb{R}^d)$. For example, in the case $p = 2$, it is easy to see that this follows if P_k is the L^2 projection and $\bigcap V_k = \{0\}$. Now $P_k f - P_{k-1} f$ is an element of V_k and therefore can be expressed as a linear combination of the $\phi(2^k x - j)$, $j \in \mathbb{Z}^d$. Therefore, we have

$$f = \sum_{k\in\mathbb{Z}} (P_k f - P_{k-1} f) = \sum_{j\in\mathbb{Z}^d,\, k\in\mathbb{Z}} c_{j,k}(f) \phi_{j,k}, \tag{2.2}$$

which is the analogue of (1.1).

Examples of functions ϕ that satisfy a refinement equation (2.1) are $\phi = \chi_\Omega$ with $\Omega := [0, 1]^d$ and multivariate box splines and B-splines, which are piecewise polynomial functions with compact support (see for example the monograph of Chui [8]).

We should emphasize that there is redundancy in (2.2) since the set of $\phi(2^k x - j)$, $j \in \mathbb{Z}^d$, $k \in \mathbb{Z}$, are not linearly independent. Orthogonal wavelets eliminate redundancy, among their many other attractive properties. The univariate orthogonal wavelets of Daubechies are obtained from a univariate function ϕ (called a "mother wavelet") that satisfies (2.1) for $d = 1$ and whose translates $\phi(\cdot - j)$, $j \in \mathbb{Z}$, form an orthonormal system. The orthogonality condition is equivalent to $\sum_{j\in\mathbb{Z}} a_j a_{j+2k} = 2\delta_0(k)$ where δ_0 is the Kronecker δ, which is one for $k = 0$ and 0 for all other integers k. The existence of such functions ϕ with compact support and arbitrary differentiability is the main result of Daubechies [11]. For P we take the orthogonal projection of $L^2(\mathbb{R})$ onto V. Then, for $k \in \mathbb{Z}$, $P_k - P_{k-1}$ is the orthogonal projection onto the space $W_k := V_k \ominus V_{k-1}$, which is the orthogonal complement of V_{k-1} in V_k. The spaces W_k are obtained from $W := V_1 \ominus V_0$ by dilation: $W_k := \{f(2^k \cdot) \mid f \in W\}$, $k \in \mathbb{Z}$. We then have that $L^2(\mathbb{R}) = \bigoplus_{k\in\mathbb{Z}} W_k$ with W_k the dilated spaces.

Mallat [19] has shown that there is a function ψ whose translates $\psi(\cdot - j)$, $j \in \mathbb{Z}$, form an orthonormal basis for W. If ψ is to have compact support, it is uniquely given (up to translation) by

$$\psi(x) := \sum_j (-1)^j a_{1-j} \phi(2x - j). \tag{2.3}$$

It follows that $P_k f - P_{k-1} f$, which is in W_k, can be expressed in terms of the L^2 normalized translates $\psi_{j,k} := 2^{k/2} \psi(2^k \cdot - j)$, $j \in \mathbb{Z}$:

$$P_k f - P_{k-1} f = \sum_{j\in\mathbb{Z}} c_{j,k} \psi_{j,k},$$

where $c_{j,k} := \int_{\mathbb{R}} f \psi_{j,k}$. In summary, we obtain the decomposition, valid for all $f \in L^2(\mathbb{R})$:

$$f = \sum_{j,k\in\mathbb{Z}} \langle f, \psi_{j,k} \rangle \psi_{j,k}. \tag{2.4}$$

The orthogonal wavelet decomposition (2.4) is also valid for functions in $L^p(\mathbb{R})$ with convergence in $\|\cdot\|_{L^p(\mathbb{R})}$ provided $1 < p < \infty$. Indeed, from the compactness of the functions $\psi_{j,k}$, it follows that P_k is an L^p bounded projector from $L^p(\mathbb{R})$ onto V_k. The assumption on denseness of $\bigcup V_k$ then gives that $\|f - P_k f\|_{L^p(\mathbb{R})} \to 0$ for all $f \in L^p(\mathbb{R})$, $1 \le p < \infty$. Moreover, $\|P_k f\|_{L^p(\mathbb{R})} \to 0$, $k \to -\infty$, if $p > 1$. These results also hold for $p = \infty$ if L^∞ is replaced by the space of uniformly continuous functions on $\mathbb{R}$. However, it is important to note that the decomposition (2.4) *does not* hold for $p = 1$ since orthogonality gives $\int P_k f = \int f$ and therefore $\|P_k f\|_{L^1(\mathbb{R})}$ does not tend to 0 as $k \to -\infty$. The importance of these remarks is that orthogonal wavelet decompositions are not suitable when compression is desired in the L^1 metric. We have more to say on this later. However, orthogonal wavelet decompositions are valid for the Hardy space $H^1(\mathbb{R})$, which is sometimes used as a substitute for $L^1(\mathbb{R})$, and for the Hardy spaces $H^p(\mathbb{R})$, $0 < p < 1$. Among

other properties, functions in the Hardy space have mean value zero, so $\| P_k f \|_{H^p(R)}$ tends to zero as $k \to -\infty$.

Nonorthogonal wavelet decompositions can be obtained by beginning with other bases for the space W, such as with B-splines.

Orthogonal wavelet decompositions in higher dimensions $d > 1$ are usually obtained by taking tensor products. For example, for $d = 2$ (the case of interest to us), if ϕ is a univariate mother wavelet and ψ is its corresponding orthogonal wavelet given by (2.3), then the three functions $\psi^{(1)}(x, y) := \psi(x)\psi(y)$, $\psi^{(2)}(x, y) := \psi(x)\phi(y)$, and $\psi^{(3)}(x, y) := \phi(x)\psi(y)$ form by translation and dilation an orthogonal basis for $L^2(\mathbb{R}^2)$. That is, each $f \in L^2(\mathbb{R}^2)$ can be represented as in (2.4) where ψ is now any of the three functions $\psi^{(j)}$, $j = 1, 2, 3$.

We next discuss how one obtains decompositions (1.1) to be used in conjunction with compression algorithms; the same idea applies to orthogonal wavelets. One chooses the dyadic level m corresponding to the picture size and finds a "representation" of the picture

$$f = \sum_{j \in \mathbb{Z}} a_{j,m} \phi(2^m \cdot - j). \tag{2.5}$$

One possible choice for the coefficients $a_{j,m}$, but certainly not the only one, is to take $a_{j,m} = p_j$. However, some additional values of $a_{j,m}$ are needed near the boundary of the picture (i.e., near the boundary of I) corresponding to the functions $\phi_{j,m}$ with $j2^{-m}$ not in I that nonetheless contribute to the picture. We do not discuss this issue further here but refer the interested reader to [12], where the analogous question for surface compression is discussed.

Once (2.5) is found, the representation (1.1) is simply

$$\begin{aligned} f = P_m f &= P_0 f + \sum_{k=1}^{m} (P_k f - P_{k-1} f) \\ &= \sum_{j \in \mathbb{Z},\, k \geq 0} c_{j,k} \phi(2^k \cdot - j). \end{aligned} \tag{2.6}$$

The coefficients in (2.6) can be computed recursively by using the refinement equation. For orthogonal wavelets, for example, one uses a fast wavelet transform that is analogous to the fast Haar transform. One never computes ψ explicitly (although ψ is easy to recover numerically) because all computations can be done in terms of the refinement coefficients (a_j). Namely, one creates filters for the various operations needed for the decomposition (2.2). One filter, L, computes the projection $P_0(f)$ of an element $f \in V_1$ onto V_0. Its adjoint, L^*, rewrites a function $f \in V_0 \subset V_1$ in terms of the basis $\phi_{j,k}$ of V_1; corresponding filters H and H^* project an element $f \in V_1$ onto W and rewrite elements in W in terms of the $\psi_{j,1}$. In this way, the description of wavelet transforms includes those for quadrature mirror filters when there are such underlying functions ϕ and ψ for these filters.

B. Compression

To compress the representation of a digitized image, we first choose a decomposition (1.1) with respect to some wavelet basis of a function f representing the image. We then choose new coefficients $\tilde{c}_{j,k}$ for the compressed approximation (1.2). We fix a value of p with $0 < p < \infty$ and measure compression error in the $L^p(I)$ metric. In [13], a method for choosing the coefficients $\tilde{c}_{j,k}$ was given that is optimal in a certain mathematical sense. We give a slightly more general version of their algorithm that is more useful for image compression.

Algorithm 1 (Generalized Transform Coding Algorithm): Given a parameter value ϵ (which controls the error of compression), and a representation (1.1) of the function f, we choose quantized coefficients $\tilde{c}_{j,k}$ that satisfy

$$\| (c_{j,k} - \tilde{c}_{j,k}) \phi_{j,k} \|_{L^p(I)} \leq \epsilon.$$

We assume that $\| c_{j,k} \phi_{j,k} \|_{L^p(I)} < \epsilon$ implies $\tilde{c}_{j,k} = 0$. Our transformed picture is

$$\tilde{f} := \sum_{k \geq 0,\, j \in \mathbb{Z}^2} \tilde{c}_{j,k} \phi_{j,k}.$$

To describe the sense in which the above algorithm is optimal, we introduce the Besov smoothness spaces, which are described in more detail in Section III-B. The Besov space $B_q^\alpha(L^q(I))$ is a collection of functions with a common smoothness in $L^q(I)$. For the time being, it is enough to think of this space as functions with α derivatives in $L^q(I)$; we emphasize, however, that $\alpha > 0$ is not necessarily an integer and that q may be less than one, so a function f in $B_q^\alpha(L^q(I))$ may not have any true derivatives, even in the distributional sense.

For a fixed value of α, there is one particular value of q that is important for compression in $L^p(I)$; it is given by $q^{-1} = \alpha/2 + p^{-1}$, where the "2" arises because we are dealing with compression in two dimensions. We quote the following results from [13], which are valid under some restrictions (described at the end of this section) on the wavelet ϕ, the value of p, and the decomposition (1.1). (Even though the algorithm presented here is slightly more general than that in [13], the proof presented there applies without any alteration whatsoever to the following theorem.)

Theorem 1 (Error Bounds): For $0 < \alpha$, there exist constants C_1 and C_2 such that for all $f \in B_q^\alpha(L^q(I))$ with $1/q = \alpha/2 + 1/p$, for all $N = 1, 2, \cdots$, and for $\epsilon := N^{-1/q}$, Algorithm 1 gives a function $\tilde{f}$ with the following properties.

1) The number, $\mathcal{N}$, of nonzero coefficients $\tilde{c}_{j,k}$ satisfies

$$\mathcal{N} \leq C_1 N \| f \|^q_{B_q^\alpha(L^q(I))}. \tag{2.7}$$

2) The error $f - \tilde{f}$ satisfies

$$\| f - \tilde{f} \|_{L^p(I)} \leq C_2 N^{-\alpha/2} \| f \|^{q/p}_{B_q^\alpha(L^q(I))} \tag{2.8}$$

and

$$\| f - \tilde{f} \|_{L^p(I)} \leq C_1^{\alpha/2} C_2 \mathcal{N}^{-\alpha/2} \| f \|_{B_q^\alpha(L^q(I))}. \tag{2.9}$$

Thus, this theorem says that functions in $B_q^\alpha(L^q(I))$ can be approximated by Algorithm 1 with an $L^p(I)$ error not

exceeding $CN^{-\alpha/2}$ with the approximating function $\tilde{f}$ having at most N coefficients. The error estimate of this theorem can be improved slightly in the sense that functions in the space $B_q^\alpha(L^q(I))$ can be characterized by their approximation error in Algorithm 1 (see [13]). If $a_N(f)_{L^p(I)}$ is defined by (1.3) for Algorithm 1, then we have

$$\sum_{N=1}^{\infty} \left[N^{\alpha/2} a_N(f)_{L^p(I)}\right]^q \frac{1}{N} < \infty \Leftrightarrow f \in B_q^\alpha(L^q(I)). \tag{2.10}$$

Thus, functions in $B_q^\alpha(L^q(I))$ have a little better approximation by Algorithm 1 than the heretofore mentioned $O(N^{-\alpha/2})$ because of the convergence of the series in (2.10). This is the precise statement of (1.4). In particular, from (2.10) one can derive an equivalent norm for $B_q^\alpha(L^q(I))$:

$$\|f\|^q_{B_q^\alpha(L^q(I))} \approx \|f\|^q_{L^q(I)} + \sum_{N=1}^{\infty} \left[N^{\alpha/2} a_N(f)_{L^p(I)}\right]^q \frac{1}{N}. \tag{2.11}$$

Algorithm 1 is optimal in the following sense: If a second compression algorithm using the same wavelet spaces V_k gives a compression error $\bar{a}_N(f)_{L^p(I)}$, then, for each $f \in B_q^\alpha(L^q(I))$, we have

$$\|f\|^q_{L^q(I)} + \sum_{N=1}^{\infty} \left[N^{\alpha/2} a_N(f)_{L^p(I)}\right]^q \frac{1}{N} \le C_0 \left(\|f\|^q_{L^q(I)} + \sum_{N=1}^{\infty} \left[N^{\alpha/2} \bar{a}_N(f)_{L^p(I)}\right]^q \frac{1}{N} \right),$$

with C_0 independent of f.

We have mentioned that the above results hold under certain conditions on ϕ. For the error estimates (2.8), (2.9), one needs in addition to the usual properties of the function ϕ and the projectors P_k, that the space V contains all polynomials of total degree $< \alpha$. This latter condition can be restated in terms of the Fourier transform of ϕ: $\hat{\phi}(\omega)$ should have a zero of multiplicity $> \alpha$ at $2k\pi$, $k \neq 0$ (see [5]). The projectors P_k are chosen to be bounded on $L^p(I)$ and $L^q(I)$. These projectors then lead to the decomposition (1.1). For the characterization results (2.11) one needs additional properties of ϕ, the most important of which is that ϕ should have slightly more smoothness than membership in $B_q^\alpha(L^q(I))$ (see [13] for details). With this, (2.11) holds for all $0 < p < \infty$.

In the case of orthogonal wavelets, (2.9) is valid for all $1 < p < \infty$ (it fails for $p \le 1$ for the reasons mentioned earlier) provided again that V contains all polynomials of degree $< \alpha$.

There is another way to measure the optimality of approximation processes, based on the mathematical concept of n-widths, that we feel may have useful application in further work on compression. Let $\mathcal{M}_n$ be an n-dimensional (nonlinear) manifold of functions from $L^p(I)$. This means that each function $M \in \mathcal{M}_n$ is determined by n real parameters, which we denote by $a := (a_1, \cdots, a_n)$. We can, therefore, denote the elements of $\mathcal{M}_n$ by $M(a)$, $a \in \mathbb{R}^n$. If K is a set of functions, we say that a mapping $\bar{a}$ of K into $\mathcal{M}_n$ is a continuous selection for K if $\bar{a}$ is continuous with respect to some topology on K. This means that whenever f and g are close, the parameters $\bar{a}(f)$ and $\bar{a}(g)$ are close. If $\bar{a}$ is such a selection, then one can think of $M(\bar{a}(f))$ as an approximation to f. The error in approximating the class of functions K by this procedure is

$$E(K, \bar{a}, \mathcal{M}_n) := \sup_{f \in K} \|f - M(\bar{a}(f))\|_{L^p(I)}.$$

The functions in K that are approximated most poorly by our selection procedure into M determine this error.

The n-width of K is defined by

$$d_n(K)_{L^p(I)} := \inf_{\bar{a}, \mathcal{M}_n} E(K, \bar{a}, \mathcal{M}_n).$$

In other words, the n-width measures the maximum error of the best manifolds for the approximation of the elements of K.

We take for K the unit ball of $B_q^\alpha(L^q(I))$, i.e., the collection of all functions in $B_q^\alpha(L^q(I))$ with $\|f\|_{B_q^\alpha(L^q(I))} \le 1$. The n-width of K (in the univariate case $d = 1$) was determined in [16]. An argument similar to that give in [16] would show that

$$C_0 n^{-\alpha/2} \le d_n(K) \le C_1 n^{-\alpha/2} \tag{2.12}$$

with C_0, C_1 absolute constants.

We can view a compression algorithm based on a wavelet decomposition (1.1) as a method of approximation from the nonlinear manifold consisting of all functions $S = \sum_{j,k \in \Lambda} a_{j,k} \phi(2^k \cdot - j)$, where Λ is a set of at most n indicies j, k. This is a manifold of dimension $3n$ with the parameters j, k and the coefficients $a_{j,k}$. According to Theorem 1, we can approximate the elements of K by using Algorithm 1 and achieve the optimal error of (2.12). Although the selection of indicies and coefficients given in this algorithm does not have the continuous selection property (it is probable that Algorithm 1 can be recast as a continuous selection by considering continuous parameterizations), (2.12) shows that no compression algorithm based on a continuous selection from a nonlinear manifold can give a better approximation order for all the functions in $B_q^\alpha(L^q(I))$ than that provided by Algorithm 1.

III. Mathematics of Wavelet Approximations

In this section, we discuss the simplest application of the mathematical theory of the previous section to image compression, namely wavelet transform coding methods based on piecewise constants. In particular, we shall consider the wavelet decompositions of the previous section for the case when V is the space of all piecewise constants on dyadic cubes of sidelength one with vertices at the integers. In this case, the mother wavelet $\phi = \chi_I$, $I := [0, 1]^2$, and the Daubechies orthogonal wavelet is $\psi := \chi_{[0,1/2)} - \chi_{[1/2,1)}$, which gives the two-dimensional Haar functions.

If we fix a value of p and measure the error of compression in the L^p metric, then we should begin with a decomposition (2.2) that is valid for all $f \in L^p$. Therefore, the

projector P used in this decomposition should at a minimum be bounded on L^p. Moreover, the construction [13] of optimal algorithms for a given value of α requires that P produce a piecewise constant approximation to f in the metric L^q, $q^{-1} = \alpha/2 + p^{-1}$, with special approximation properties. This leads us to discuss in Section III-A when constant functions are "best" or "near best" approximations in $L^q(I)$.

We have already pointed out the importance of Besov spaces in our mathematical framework for compression. In Section III-B, we give the definition of these spaces and consider some of their properties important for compression. In Section III-D, we recount Algorithm 1 for the special case of piecewise constant approximation. As an example, we study a version of the progressive transmission of coefficients that satisfies the hypotheses of our algorithm. In Section III-E, we discuss in detail the effect of choosing different projectors and different basis functions for the wavelet decompositions.

Finally, in Section III-F, we discuss high-order generalizations of the piecewise constant approximations analyzed in the previous sections.

The proofs of the several theorems in this section are given in the Appendix.

A. Near-Best Approximations

For the unit square interval $I := [0, 1]^2 \subset \mathbb{R}^2$, any exponent $0 < p \le \infty$, and any function $f \in L^p(I)$, we let $\mathbb{Q}_p f$ denote a best $L^p(I)$ constant approximation to f, that is

$$\| f - \mathbb{Q}_p f \|_{L^p(I)} = \inf_{c \in \mathbb{R}} \| f - c \|_{L^p(I)}.$$

For example, $\mathbb{Q}_2 f$ is the average of f over I,

$$\mathbb{Q}_\infty f = \frac{\sup_{x \in I} f(x) + \inf_{x \in I} f(x)}{2},$$

and $\mathbb{Q}_1 f$ is a *median* of f on I, where a median is defined to be any number m for which

$$|\{x \in I \mid f(x) \ge m\}| \ge 1/2 \quad \text{and}$$
$$|\{x \in I \mid f(x) \le m\}| \ge 1/2.$$

For other values of p a *best* $L^p(I)$ approximation is not always easy to find, so we consider instead *near-best* constant approximations $\mathbb{P}_p f$ that satisfy, for some constant C,

$$\| f - \mathbb{P}_p f \|_{L^p(I)} \le C \| f - \mathbb{Q}_p f \|_{L^p(I)}. \tag{3.1}$$

Of course, if f is not constant on I, then the right side of (3.1) is nonzero, and *any* approximation $\mathbb{P}_p f$ is near-best for some constant C. However, we shall consider families of near-best approximations for which the constant C is fixed in advance.

We note that if $\mathbb{P}_p f$ is near-best in $L^p(I)$, then it is near-best in any $L^q(I)$, with $q > p$; cf. [15]. Furthermore, the following theorem shows that $\mathbb{Q}_1$ is near-best, not only for $q \ge 1$, but for all q.

Theorem 2: For each $q \in (0, \infty]$ there exists a constant C_q such that for all $f \in L^q(I)$,

$$\| f - \mathbb{Q}_1 f \|_{L^q(I)} \le C_q \| f - \mathbb{Q}_q f \|_{L^q(I)}.$$

An examination of the proof shows that the theorem is true for order parameters other than the median, such as the first and third quartiles of f. (A first quartile of f is any number ξ that satisfies

$$|\{x \in I \mid f(x) \ge \xi\}| \ge 3/4 \quad \text{and}$$
$$|\{x \in I \mid f(x) \le \xi\}| \ge 1/4,$$

while a third quartile of f is any number ζ that satisfies

$$|\{x \in I \mid f(x) \ge \zeta\}| \ge 1/4 \quad \text{and}$$
$$|\{x \in I \mid f(x) \le \zeta\}| \ge 3/4.$$

For bilevel images (halftones, for example), where black is represented by 0 and white is represented by 1, the median of f is particularly easy to evaluate — it is just the most common pixel value in I. Note that the average $\mathbb{Q}_2 f$ is *not* a near-best approximation if $q < 1$: If we let $f := \chi_{[0, 2^{-j}]^2}$, $j > 1$, then $\mathbb{Q}_2 f = 2^{-2j}$ on I and

$$\| f - \mathbb{Q}_2 f \|_{L^q(I)} \ge 2^{-2j},$$

whereas

$$\| f - \mathbb{Q}_q f \|_{L^q(I)} \le \| f \|_{L^q(I)} = \left(\int_{[0, 2^{-j}]^2} 1^q \, dx \right)^{1/q} = 2^{-2j/q},$$

so that

$$\frac{\| f - \mathbb{Q}_2 f \|_{L^q(I)}}{\| f - \mathbb{Q}_q f \|_{L^q(I)}} \ge 2^{2j(1-q)/q} \to \infty$$

as $j \to \infty$. However, if we round $\mathbb{Q}_2 f$ to the nearest of the values 0 or 1 then for bilevel images we end up with the median, which is a good approximation for any $L^q(I)$, $0 < q < \infty$. This fact can be generalized in Theorem 3.

Theorem 3: Assume that $N > 0$ and mutually disjoint (measurable) sets $I_j \subset I$ are given for $j = 0, \cdots, N$ such that $I = \bigcup_{j=0}^{N} I_j$ and for all $x \in I$,

$$f(x) = \sum_{j=0}^{N} j \chi_{I_j}(x);$$

i.e., f takes only finitely many integer values on I. If we define $\tilde{\mathbb{Q}}_2 f$ to be $\mathbb{Q}_2 f$ rounded to the nearest integer, then for each $0 < q < 1$, there exists a constant $C_{N,q}$ such that

$$\| f - \tilde{\mathbb{Q}}_2 f \|_{L^q(I)} \le C_{N,q} \| f - \mathbb{Q}_q f \|_{L^q(I)}.$$

The previous theorem has the interesting consequence that although the Haar transform using exact arithmetic does not result in optimal order approximation in $L^1(I)$ of functions in $B_q^\alpha(L^q(I))$, $1/q = 1 + \alpha/2$, that take on finitely many integer values, the Haar transform using rounded integer arithmetic *does* result in optimal order approximations in $L^1(I)$; see Section III-E.

We show by the following argument that the set of near-best approximations is convex if we are willing to change the parameter C. Assume that f_i, $i = 1, 2$, are near-best $L^q(I)$ approximations with constant C_i, respectively. Then for any $\alpha \in (0, 1)$ and $q < 1$,

$$\begin{aligned}\|f - (\alpha f_1 + (1-\alpha)f_2)\|_{L^q(I)}^q &= \|\alpha(f-f_1) + (1-\alpha)(f-f_2)\|_{L^q(I)}^q \\ &\le \|\alpha(f-f_1)\|_{L^q(I)}^q + \|(1-\alpha)(f-f_2)\|_{L^q(I)}^q \\ &= \alpha^q\|(f-f_1)\|_{L^q(I)}^q + (1-\alpha)^q\|(f-f_2)\|_{L^q(I)}^q \\ &\le (\alpha^q C_1^q + (1-\alpha)^q C_2^q)\|f - \mathbb{Q}_q f\|_{L^q(I)}^q.\end{aligned}$$

Therefore, we can take the new constant $C := (\alpha^q C_1^q + (1-\alpha)^q C_2^q)^{1/q}$.

For example, if we let f_1 and f_3 be the first and third quartiles, respectively, of f on I, then

$$\mathbb{P}_q f := \max(f_1, \min(f_3, \mathbb{Q}_2 f)),$$

which is a convex combination of f_1 and f_3, is a near-best $L^q(I)$ approximation to f for any q with a constant C that depends only on q. We shall use this approximation in some of our later examples.

These results are extended in an obvious way for square intervals I in $\mathbb{R}^2$ that are not the unit interval.

B. Equivalent Norms for Besov Spaces

In this section, we recount results of [15] that give equivalent norms for Besov spaces. These results will be used in Section III-D to bound the error in certain transform methods of coding.

We are interested in functions f in Besov spaces $B_q^\alpha(L^p(I))$, with $\alpha > 0$, $0 < p < \infty$, $0 < q \le \infty$. Roughly speaking, a function in $B_q^\alpha(L^p(I))$ has α "derivatives" in $L^p(I)$, while the parameter q measures more subtle gradations in smoothness. The usual Besov space norm is defined as follows. Fix $\alpha > 0$ and an integer $r > 0$. Define the rth forward difference with parameter $h \in \mathbb{R}^2$ by

$$\Delta_h^0 f(x) := f(x), \text{ and for } k = 1, \cdots, r,$$

$$\Delta_h^k f(x) := \Delta_h^{k-1} f(x+h) - \Delta_h^{k-1} f(x). \tag{3.2}$$

Next define the rth modulus of smoothness in L^p,

$$\omega_r(f, t)_p := \sup_{|h| \le t} \left(\int_{I_{rh}} |\Delta_h^r f(x)|^p \, dx \right)^{1/p},$$

where $I_{rh} := \{x \in I \mid x + rh \in I\}$. (In other words, $\Delta_h^r f(x)$ is defined, if and only if $x \in I_{rh}$.) The space $B_q^{\alpha, r}(L^p(I))$ consists of all functions f for which

$$\|f\|_{B_q^{\alpha,r}(L^p(I))} := \|f\|_{L^p(I)} + \left(\int_0^\infty [t^{-\alpha}\omega_r(f, t)_p]^q \frac{dt}{t} \right)^{1/q} < \infty \tag{3.3}$$

when $q < \infty$ and

$$\|f\|_{B_\infty^{\alpha,r}(L^p(I))} := \|f\|_{L^p(I)} + \sup_{t>0} [t^{-\alpha}\omega_r(f, t)_p] < \infty.$$

Note that when $q = \infty$, we require $\omega_r(f, t)_p$ to decay at least as fast as $O(t^\alpha)$ as $t \to 0$, whereas when $q < \infty$, we require $\omega_r(f, t)_p$ to decay at a slightly faster rate. When q or p are less than one, $\|\cdot\|_{B_q^{\alpha,r}(L^p(I))}$ does not satisfy the triangle inequality, so it is not, strictly speaking, a "norm," but only a quasi-norm, for which there exists a constant C such that for all $f, g \in B_q^{\alpha, r}(L^p(I))$,

$$\|f + g\|_{B_q^{\alpha,r}(L^p(I))} \le C(\|f\|_{B_q^{\alpha,r}(L^p(I))} + \|g\|_{B_q^{\alpha,r}(L^p(I))}).$$

The spaces $B_q^{\alpha, r}(L^p(I))$ and $B_q^{\alpha, r'}(L^p(I))$ are related in the following way. First note that because of (3.2),

$$\omega_{r+1}(f, t)_p \le \begin{cases} 2\omega_r(f, t)_p, & 1 \le p, \\ 2^{1/p}\omega_r(f, t)_p, & 0 < p < 1. \end{cases} \tag{3.4}$$

We can conclude from (3.3) and (3.4) that

$$f \in B_q^{\alpha, r}(L^p(I)) \Rightarrow f \in B_q^{\alpha, r+1}(L^p(I)); \tag{3.5}$$

by induction f will be in $B_q^{\alpha, r'}(L^p(I))$ for any $r' > r$. On the other hand, it can be shown that if both r and r' are strictly greater than α, then $B_q^{\alpha, r}(L^p(I)) = B_q^{\alpha, r'}(L^p(I))$; furthermore $\|\cdot\|_{B_q^{\alpha,r}(L^p(I))}$ and $\|\cdot\|_{B_q^{\alpha,r'}(L^p(I))}$ are equivalent norms. Thus, without confusion, we define the Besov space $B_q^\alpha(L^p(I))$ to be $B_q^{\alpha, r}(L^p(I))$ for any $r > \alpha$.

As one particular application, we are interested in the smoothness of bilevel (black-and-white) images. A bilevel image can be represented as the characteristic function f of a set S, with $f(x) = 1$ where $x \in S$ and $f(x) = 0$ where $x \notin S$. As a partial converse to (3.5), the following theorem shows that if a characteristic function f is in $B_q^\alpha(L^p(I)) \equiv B_q^{\alpha, 2}(L^p(I))$ for $1 \le \alpha < 2$, then f is also in the space $B_q^{\alpha, 1}(L^p(I))$.

Theorem 4: If f takes only the values 0 and 1 in I, $1 \le \alpha < 2$, and $f \in B_q^\alpha(L^p(I)) \equiv B_q^{\alpha, 2}(L^p(I))$, then $f \in B_q^{\alpha, 1}(L^p(I))$ and $\|f\|_{B_q^{\alpha,1}(L^p(I))} \le 2^\alpha \|f\|_{B_q^{\alpha,2}(L^p(I))}$.

It is not true for *all* f that $\|f\|_{B_q^{\alpha,1}(L^p(I))} \le 2^\alpha \|f\|_{B_q^{\alpha,2}(L^p(I))}$. A simple example is $f(x, y) := x$, for which $\|f\|_{B_q^{\alpha,2}(L^p(I))} = \|f\|_{L^p(I)}$. We have with $h = (t, 0)$,

$$\begin{aligned}\omega_1(f, t)_p &= \left(\int_{I_h} |\Delta_h^1 f(x)|^p \, dx \right)^{1/p} \\ &\approx \left(\int_I |t|^p \, dx \right)^{1/p} = t;\end{aligned}$$

substituting this into (3.3) shows that this f is not in $B_q^{\alpha, 1}(L^p(I))$ for any $\alpha > 1$. Roughly speaking, only functions that are linear combinations of characteristic functions have any hope of being in $B_q^{\alpha, 1}(L^p(I))$ for $\alpha > 1$; see, however, Section IV-B.

In the following sections, we shall consider approximations in $L^p(I)$, $0 < p < \infty$, of functions in $B_q^{\alpha, r}(L^q(I))$, with

$$\frac{1}{q} = \frac{\alpha}{2} + \frac{1}{p}.$$

DeVore and Popov [15] showed that $B_q^{\alpha,r}(L^q(I))$ is continuously embedded in $L^p(I)$, written $B_q^{\alpha,r}(L^q(I)) \hookrightarrow L^p(I)$. This means that there exists a constant C such that for all $f \in B_q^{\alpha,r}(L^q(I))$,

$$\|f\|_{L^p(I)} \le C\|f\|_{B_q^{\alpha,r}(L^q(I))}.$$

We remark that this embedding is not compact, in contrast to the embedding of the Sobolev space $W^{\alpha,2}(I)$, $\alpha > 0$, into $L^2(I)$, for example.

DeVore and Popov have given equivalent norms for $B_q^{\alpha,r}(L^q(I))$ in terms of certain sequence spaces. For the rest of this section we shall recount this theory for $r = 1$; there is a similar theory, which we discuss in Section III-F, for any $r > 0$.

Let $I_{j,k}$ be the square with sidelength 2^{-k} with lower-left corner at the point $j/2^k$, with the multiindex $j = (j_1, j_2) \in \mathbb{Z}_k^2 := \{j \in \mathbb{Z}^2 \mid 0 \le j_1 < 2^k, 0 \le j_2 < 2^k\}$, and let $\phi_{j,k}$ be its characteristic function ($\phi_{j,k}(x) = 1$ if $x \in I_{j,k}$, and $\phi_{j,k}(x) = 0$, otherwise). Fix a constant C such that for each $k \ge 0$ and $j \in \mathbb{Z}_k^2$, $d_{j,k}\phi_{j,k}$ is a near-best $L^q(I_{j,k})$ approximation to f with constant C from the set of all functions $d\phi_{j,k}$. For each $k \ge 0$, let $S^k(f) := \sum_{j \in \mathbb{Z}_k^2} d_{j,k}\phi_{j,k}$ be a "level k" approximation to f ($S^{-1}(f) := 0$), and define the coefficients $c_{j,k}$ by

$$S^k(f) - S^{k-1}(f) = \sum_{j \in \mathbb{Z}_k^2} c_{j,k}\phi_{j,k}. \tag{3.6}$$

Here we have used the fact that we can rewrite

$$\phi_{(j_1,j_2),k-1} = \phi_{(2j_1,2j_2),k} + \phi_{(2j_1+1,2j_2),k} + \phi_{(2j_1,2j_2+1),k} + \phi_{(2j_1+1,2j_2+1),k}.$$

We can write

$$f = \sum_{k \ge 0,\, j \in \mathbb{Z}_k^2} c_{j,k}\phi_{j,k}, \tag{3.7}$$

where the sum converges in $L^q(I)$, because the functions $d_{j,k}\phi_{j,k}$ are near-best approximations in $L^q(I_{j,k})$ to f. A particular result from [15] is that whenever $q^{-1} = \alpha/2 + p^{-1}$, as we assume, the quasi-norm $\|f\|_{B_q^{\alpha,1}(L^q(I))}$ is equivalent to the sequence quasi-norm

$$\|f\|_{\mathcal{S}} := \left(\sum_{k \ge 0,\, j \in \mathbb{Z}_k^2} \|c_{j,k}\phi_{j,k}\|_{L^p(I)}^q \right)^{1/q}, \tag{3.8}$$

when $\alpha < 2/p$; that is, there exist positive constants C_1 and C_2 such that for all $f \in B_q^{\alpha,1}(L^q(I))$

$$C_1\|f\|_{\mathcal{S}} \le \|f\|_{B_q^{\alpha,1}(L^q(I))} \le C_2\|f\|_{\mathcal{S}}.$$

Furthermore, in every case where f is written as a sum (3.7) (even if the coefficients $c_{j,k}$ are *not* calculated as the difference of near-best approximations (3.6)) the quasi-norm $\|f\|_{B_q^{\alpha,1}(L^q(I))}$ is bounded by a constant multiple of (3.8).

C. Pixel Quantization

In this section, we show that an image f formed by pixel quantization of an intensity field F inherits the smoothness of F in the Besov spaces $B_q^\alpha(L^q(I))$, $1/q = \alpha/2 + 1/p$, $p > 0$ and $0 < \alpha < 2/p$. We shall use this result in Section IV-B to help justify our claim that images have little smoothness.

One can model the formation of the digitized image as follows. One begins with a spatially varying intensity field F defined on the square $I = [0, 1]^2$, normalized to range between 0 and 1. On each dyadic square $I_{j,m}$ of size $2^{-m} \times 2^{-m}$, an average or other projection of the intensity F is taken to give a value $\mathbb{P}_{j,m}F$. Then, we round $\mathbb{P}_{j,m}F$ to the nearest value of $i/2^n$ to give a pixel value $p_{j,m}$. (In this section, we shall consider pixels to take values in the set $\{i2^{-n} \mid 0 \le i < 2^n\}$.) We work with the digitized image

$$f = \sum_{j \in \mathbb{Z}_m^2} p_{j,m}\phi_{j,m}. \tag{3.9}$$

If one uses smooth wavelets then some other method of constructing f from the pixel values $p_{j,m}$ must be devised.

The following lemma bounds the norm in $B_q^{\alpha,1}(L^q(I))$ of functions of the form (3.9) in terms of their $L^\infty(I)$ norm.

Lemma 1: Assume that $p > 0$, $0 < \alpha < 2/p$, and $1/q = \alpha/2 + 1/p$. Then there exists a constant C such that for each f of the form (3.9),

$$\|f\|_{B_q^{\alpha,1}(L^q(I))} \le C\|f\|_{L^\infty(I)}2^{\alpha m}. \tag{3.10}$$

It follows immediately from Lemma 1 and (3.5) that

$$\|f\|_{B_q^\alpha(L^q(I))} \le C\|f\|_{L^\infty(I)}2^{\alpha m}.$$

It is shown in Section IV-B that if f is not constant then it is not in $B_q^\alpha(L^q(I))$ for $\alpha \ge 2/p$, so the lemma is in some sense sharp. Also, this shows that *any* image is in $B_q^{\alpha,1}(L^q(I))$, with a bound for its norm that increases exponentially in m and α.

We can use the previous lemma to prove the following theorem, which compares the $B_q^{\alpha,1}(L^q(I))$ norm of f to the same norm of F.

Theorem 5: Assume that $p > 0$, $0 < \alpha < 2/p$, and $1/q = \alpha/2 + 1/p$. Then a) if $1 < p < \infty$ and $\mathbb{P}$ is the $L^2(I)$ projection (i.e., the average of the intensity), or b) if $0 < p < \infty$ and $\mathbb{P}$ is a near-best projection operator in $L^q(I)$, there exists a C such that for all F,

$$\|f\|_{B_q^{\alpha,1}(L^q(I))} \le C\|F\|_{B_q^{\alpha,1}(L^q(I))} + C2^{\alpha m}2^{-n}.$$

Thus, the norm of the image f is bounded by a constant times the norm of the original intensity distribution F plus a small amount caused by rounding the pixels to the values $i2^{-n}$. Thus, our estimates in Section V of the Besov space norm of our test images do *not* grossly underestimate the smoothness of these images. See Section IV-B for further discussion of how we estimate the smoothness of images.

Theorem 5 also indicates how the number of grey-scale levels should increase as one increases the spatial resolution of a digital imaging device in order to keep $\|f\|_{B_q^{\alpha,1}(L^q(I))}$ bounded when F is in $B_q^{\alpha,1}(L^q(I))$. Roughly speaking, we would require $n \ge \alpha m + C$. In Section V, we estimate the global smoothness of our images to range from 0.3 to 0.6, while $n = 8$ and $m = 9$.

D. The Generalized Wavelet Transform Coding Algorithm with Piecewise Constant Approximations

In this section, we present the piecewise constant version of the generalized wavelet transform coding Algorithm 1 for image compression. We then present an error analysis that is based on the theory of Sections III-A and III-B. Finally, this analysis is applied to the example of progressive transmission of coefficients.

Given our previous mathematical framework, the algorithm and error bounds are easy to state. We shall assume that f is in the space $B_q^{\alpha,1}(L^q(I))$, $\alpha > 0$ (which is equivalent to the Besov space $B_q^{\alpha}(L^q(I))$ when $0 < \alpha < 1$), $0 < q < \infty$, and $q^{-1} = p^{-1} + \alpha/2$ with $0 < p < \infty$. This corresponds to approximating in $L^p(I)$ pictures with smoothness in $B_q^{\alpha,1}(L^q(I))$.

Algorithm 2 (Generalized Transform Coding Algorithm): Choose a positive integer N and numbers $0 < p < \infty$ and $0 < \alpha$. Let q satisfy $q^{-1} = p^{-1} + \alpha/2$. Write f as (3.7), where the coefficients $c_{j,k}$ are calculated from differences of near-best approximations S^k. Choose quantized coefficients $\tilde{c}_{j,k}$ that satisfy

$$\|(c_{j,k} - \tilde{c}_{j,k})\phi_{j,k}\|_{L^p(I)}^q \le \frac{1}{N}. \tag{3.11}$$

We assume that $\|c_{j,k}\phi_{j,k}\|_{L^p(I)}^q < 1/N$ implies $\tilde{c}_{j,k} = 0$. Our compressed picture is

$$\tilde{f} := \sum_{k \ge 0,\, j \in \mathbb{Z}_k^2} \tilde{c}_{j,k}\phi_{j,k}.$$

Theorem 6 (Error Bounds): For each $0 < \alpha$ and $0 < p < \infty$ there exist constants C_1 and C_2 such that for all $f \in B_q^{\alpha,1}(L^q(I))$ with $1/q = \alpha/2 + 1/p$.

1) The number, $\mathcal{N}$, of nonzero coefficients $\tilde{c}_{j,k}$ satisfies

$$\mathcal{N} \le C_1 N \|f\|_{B_q^{\alpha,1}(L^q(I))}^q. \tag{3.12}$$

2) The error $f - \tilde{f}$ satisfies

$$\|f - \tilde{f}\|_{L^p(I)} \le C_2 N^{-\alpha/2} \|f\|_{B_q^{\alpha,1}(L^q(I))}^{q/p} \tag{3.13}$$

and

$$\|f - \tilde{f}\|_{L^p(I)} \le C_1^{\alpha/2} C_2 \mathcal{N}^{-\alpha/2} \|f\|_{B_q^{\alpha,1}(L^q(I))}. \tag{3.14}$$

Example (Progressive Transmission): The important point about the above theorem is that the method of choosing approximate coefficients $\tilde{c}_{j,k}$ affects most strongly the amount of smoothness required in an image to achieve a rate of convergence of $\mathcal{N}^{-\alpha/2}$. The following example, which uses a strategy of *progressive transmission*, should illustrate this point.

We begin with a wavelet decomposition (3.7) of a function f representing our image. (Either orthogonal wavelets or the specific method using piecewise constant approximations given above will suffice.) Progressive transmission, as put forth, e.g., in [7], is a strategy of successively sending coefficients $c_{j,k}$ of f to a receiver, who progressively reconstructs the picture and who may decide when enough detail has been achieved and the transmission can stop.

We consider two orders in which to transmit the coefficients. First, as has been suggested several times, one can simply send the coefficients at the coarsest level first, in some fixed lexicographical order, and when the store of coefficients at one level is exhausted, one moves on to the next finer level (or greater k). If we denote by $\bar{f}$ the picture reconstructed using $\mathcal{N}$ coefficients sent by this strategy, then it is not difficult to show that for any $1 < p < \infty$,

$$\|f - \bar{f}\|_{L^p(I)} \le C\mathcal{N}^{-\alpha/2} \|f\|_{B_p^{\alpha}(L^p(I))}, \tag{3.15}$$

and this bound is essentially sharp (i.e., no numbers $\beta < \alpha$ or $q < p$ can be substituted in the norm of f on the right side of (3.15)). The inequality (3.15) can be interpreted as saying that in order for the error in the reconstructed image, as measured in $L^p(I)$, to decay at a rate $O(\mathcal{N}^{-\alpha/2})$, one requires, roughly speaking, α derivatives in $L^p(I)$.

In contrast, we propose a different ordering that satisfies (3.11). In our ordering, we transmit coefficients in decreasing order of the values of $\{\|c_{j,k}\phi_{j,k}\|_{L^p(I)}\}$; i.e., given $\mathcal{N}$ we have that

$$\tilde{c}_{j,k} = \begin{cases} c_{j,k}, & \text{for the } \mathcal{N} \text{ biggest values of } \|c_{j,k}\phi_{j,k}\|_{L^p(I)}, \\ 0, & \text{otherwise.} \end{cases}$$

It is not difficult to show that this choice of $\tilde{c}_{j,k}$ satisfies (3.11) and implicitly defines N. When, as usual, we denote by $\tilde{f}$ the reconstructed image using *this* order in which to send coefficients, we have

$$\|f - \tilde{f}\|_{L^p(I)} \le C\mathcal{N}^{-\alpha/2} \|f\|_{B_q^{\alpha}(L^q(I))},$$

where now $q = 1/(\alpha/2 + 1/p) = p/(1 + \alpha p/2)$. To achieve an approximation order of $\mathcal{N}^{-\alpha/2}$ in $L^p(I)$ one now needs only α derivatives in $L^q(I)$, and q is now less than p. This can be interpreted in two ways. First, more functions f can be approximated to order $\mathcal{N}^{-\alpha/2}$ by the second method, which requires f to have α derivatives in $L^q(I)$, than the first, which requires f to have α derivatives in $L^p(I)$. Second, if an image has at most β derivatives in $L^q(I)$ and α derivatives in $L^p(I)$, then, because $q < p$, we have $\beta \ge \alpha$, and any difference between β and α is strictly reflected in the rate of decay in the error in the reconstructed images. Asymptotically, our ordering is better.

We conducted an experiment to test our theory at moderate compression levels. Our experimental setup is as follows. We write the image f in terms of a pure Haar transform; see Section III-E. Since our images have 8 bits per pixel, the highest-frequency coefficients can be represented using at most 10 bits, the next highest frequency coefficients in 12 bits, etc. For the lexicographical ordering, we start transmitting coefficients at the coarsest level, counting the number of bits that are needed to send each coefficient without any entropy coding. Therefore, it takes 26 bits to send each of the first four coefficients, 24 bits to send each of the next 12, etc.

In order to construct the approximate pictures according to the new ordering, we first sorted the set $\{\|c_{j,k}\phi_{j,k}\|_{L^2(I)}\}$ in decreasing order, and transmitted the coefficients in this order. For each coefficient, we transmitted an extra 18-bit

number that indicated the location (and, incidentally, the number of bits in the coefficient), and then the value of the coefficient itself. Given a picture transmitted using the old ordering in B bits, we transmitted just enough coefficients in the new ordering to transmit at least B bits.

The results for the L^2 norm are indicated in Table I. For the same number of bits we achieve a smaller $L^2(I)$ error (measured in grey scales), and, more importantly, our pictures look better, as the examples given in Fig. 1 shows.

TABLE I
ERROR IN PROGRESSIVE TRANSMISSION

Old Ordering			New Ordering		
Bits	Coefficients	$\|f-\tilde{f}\|_{L^2(I)}$	Bits	Coefficients	$\|f-\tilde{f}\|_{L^2(I)}$
19112	1364	25.58	19138	533	22.38
68264	5460	18.87	68272	2019	15.30
240296	21844	13.15	240314	7577	9.53
830120	87380	7.87	830144	27550	5.13

E. Examples of Methods for Grey Scale and Bilevel Images

We shall next analyze some methods for image compression in the framework of the mathematical analysis of the previous sections.

Let ϕ denote the characteristic function of the unit interval $I=[0,1]^2$: $\phi(x)=1$ for x in I, and $\phi(x)=0$ for $x\notin I$. By dilation and translation, we obtain the functions $\phi_{j,k}(x):=\phi(2^k x-j)$ for $j=(j_1,j_2)\in\mathbb{Z}^2$ and $k\geq 0$. Thus, $\phi_{j,k}$ is the characteristic function of the square $I_{j,k}$ with sidelength 2^{-k} and lower-left corner at the point $j/2^k$, with the multiindex $j\in\mathbb{Z}_k^2:=\{j\in\mathbb{Z}^2\mid 0\leq j_1<2^k,\ 0\leq j_2<2^k\}$. We recall that we represent the image by the function

$$f:=\sum_{j\in\mathbb{Z}_m^2} p_j\phi_{j,m}.$$

For each algorithm, we compute, for each $k\geq 0$ and $j\in\mathbb{Z}_k^2$, a local projection

$$d_{j,k}\phi_{j,k}:=\mathbb{Q}f\mid_{I_{j,k}}\phi_{j,k},$$

which approximates f in the interval $I_{j,k}$. For each $k\geq 0$, we let $S^k:=\sum_{j\in\mathbb{Z}_k^2} d_{j,k}\phi_{j,k}$, and define the coefficients $d'_{j,k}$ by

$$f=S^0(f)+\sum_{k=1}^{m}\left(S^k(f)-S^{k-1}(f)\right)=\sum_{k=0}^{m}\sum_{j\in\mathbb{Z}_k^2} d'_{j,k}\phi_{j,k}.$$

Here we have used the refinement equation (2.1) to rewrite

$$\phi_{(j_1,j_2),k-1}=\phi_{(2j_1,2j_2),k}+\phi_{(2j_1+1,2j_2),k}+\phi_{(2j_1,2j_2+1),k}+\phi_{(2j_1+1,2j_2+1),k}.$$

For computational reasons we shall examine also the Haar transform of the representation for f. We let

$$\Psi(x):=\begin{cases}0, & -\infty<x<0,\\ -1, & 0<x<\frac{1}{2},\\ 1, & \frac{1}{2}<x<1,\\ 0, & 1<x<\infty,\end{cases}$$

and

$$\Phi(x):=\begin{cases}0, & -\infty<x<0,\\ 1, & 0<x<1,\\ 0, & 1<x<\infty.\end{cases}$$

The four basis functions for the local Haar representation of functions are

$$\psi^{(1)}(x,y):=\Phi(x)\Psi(y),\qquad \psi^{(2)}(x,y):=\Psi(x)\Phi(y),$$
$$\psi^{(3)}(x,y):=\Psi(x)\Psi(y),\qquad \psi^{(4)}(x,y):=\Phi(x)\Phi(y).$$

Again, by dilation and translation, we obtain $\psi_{j,k}^{(i)}:=\psi^{(i)}(2^k\cdot-j)$, $i=1,\cdots,4$, $j\in\mathbb{Z}^2$, and $k\geq 0$. Then, we can rewrite

$$\begin{aligned}&d'_{(2j_1,2j_2),k}\phi_{(2j_1,2j_2),k}+d'_{(2j_1+1,2j_2),k}\phi_{(2j_1+1,2j_2),k}\\&\quad+d'_{(2j_1,2j_2+1),k}\phi_{(2j_1,2j_2+1),k}\\&\quad+d'_{(2j_1+1,2j_2+1),k}\phi_{(2j_1+1,2j_2+1),k}\\&=\frac{c_{j,k-1}^{(1)}}{4}\psi_{j,k-1}^{(1)}+\frac{c_{j,k-1}^{(2)}}{4}\psi_{j,k-1}^{(2)}+\frac{c_{j,k-1}^{(3)}}{4}\psi_{j,k-1}^{(3)}\\&\quad+\frac{c_{j,k-1}^{(4)}}{4}\psi_{j,k-1}^{(4)}.\end{aligned}$$

The coefficients $c_{j,k}^{(i)}$ are determined by the identity

$$\begin{pmatrix}c_{j,k-1}^{(1)}\\ c_{j,k-1}^{(2)}\\ c_{j,k-1}^{(3)}\\ c_{j,k-1}^{(4)}\end{pmatrix}=\begin{pmatrix}-1&-1&1&1\\ -1&1&-1&1\\ 1&-1&-1&1\\ 1&1&1&1\end{pmatrix}\begin{pmatrix}d'_{(2j_1,2j_2),k}\\ d'_{(2j_1+1,2j_2),k}\\ d'_{(2j_1,2j_2+1),k}\\ d'_{(2j_1+1,2j_2+1),k}\end{pmatrix},\tag{3.16}$$

so that if all the $d'_{j,k}$ are integers, then so are the $c_{j,k}^{(i)}$. With this transformation, we have

$$f=\frac{1}{4}\sum_{k\geq 0}\sum_{j\in\mathbb{Z}_k^2}\sum_{i=1}^{4} c_{j,k}^{(i)}\psi_{j,k}^{(i)}+d_{0,0}\phi_{0,0}.$$

Since each of the four functions $\psi^{(i)}$ play a similar role, we shall often omit the superscript where no confusion will result.

Example 1: Haar transform for grey scale images. The Haar transform of an image f can be described as follows. The local approximations $\mathbb{Q}f$ are taken to be $\mathbb{Q}_2 f$, the averages of f on each interval $I_{j,k}$. The coefficients $d'_{(2j_1,2j_2),k}$, $d'_{(2j_1+1,2j_2),k}$, $d'_{(2j_1,2j_2+1),k}$, and $d'_{(2j_1+1,2j_2+1),k}$ in each 2×2 block satisfy

$$d'_{(2j_1,2j_2),k}+d'_{(2j_1+1,2j_2),k}+d'_{(2j_1,2j_2+1),k}+d'_{(2j_1+1,2j_2+1),k}=0,\tag{3.17}$$

so we have that all the coefficients $c_{j,k}^{(4)}$ except $c_{0,0}^{(4)}$ (which is four times the average intensity of the image over the entire square) are zero. The basis elements $\{\psi_{j,k}^{(1)},\psi_{j,k}^{(2)},\psi_{j,k}^{(3)}\}$ form an orthogonal basis for $L^2(I)$, and if we ignore the super-

(a)

(b)

(c)

Fig. 1. Example of progressive transmission with the Haar transform: (a) Lenna compressed using the coarsest to finest lexicographical ordering and 240 296 bits, (b) Lenna compressed using the best $L^2(I)$ ordering and 240 314 bits, and (c) Lenna compressed by ordering the terms of the wavelet decomposition in $L^1(I)$, with 240 320 bits.

scripts we can write

$$f = \frac{1}{4} \sum_{k \geq 0,\, j \in \mathbb{Z}_k^2} c_{j,k} \psi_{j,k} + d_{0,0} \phi_{0,0}, \qquad (3.18)$$

for any $f \in L^2(I)$. Because the representation functions form a basis, this representation is unique; furthermore, the representation is not redundant, in that there are precisely as many coefficients $c_{j,k}$ as pixels p_j.

If $q := 1/(\alpha/2 + 1/p) \geq 1$ then $\mathbb{Q}_2$ is a bounded linear projection on $L^q(I)$ (and is, therefore, near-best in $L^q(I)$) and the theory of the previous section applies to show that if $\tilde{f} = \frac{1}{4} \sum_{k \geq 0,\, j \in \mathbb{Z}_k^2} \tilde{c}_{j,k} \psi_{j,k}$ with coefficients $\tilde{c}_{j,k}$ quantized according to (3.11) then

$$\| f - \tilde{f} \|_{L^p(I)} \leq C_p \mathcal{N}^{-\alpha/2} \| f \|_{B_q^{\alpha,1}(L^q(I))}, \qquad (3.19)$$

where $\mathcal{N}$ is the number of nonzero coefficients $\tilde{c}_{j,k}$. (The fact that we used the representation (3.18) does not change the theorem.) In fact, DeVore, Jawerth, and Popov [13] show that for any $1 < p < \infty$, (3.19) is true for the Haar system, or any system of orthogonal wavelets. In their estimates, C_p tends to infinity as p tends to 1, and in fact (3.19) does not hold for $p = 1$, as the following example shows.

We let f be the characteristic function of the interval $[0, 2^{-J}]^2$. It is easy to show from the definition that $f \in B_q^{\alpha,1}(L^q(I))$, $q^{-1} = 1 + \alpha/2$, for any $\alpha < 2$, and from the equivalence of $\| f \|_{B_q^{\alpha,1}(L^q(I))}$ and $\| f \|_{\mathcal{F}}$, we can take $\| f \|_{B_q^{\alpha,1}(L^q(I))} = \| f \|_{L^1(I)} = 2^{-2J}$. If we do the calculation we find that

$$\| c_{0,k}^{(i)} \psi_{0,k}^{(i)} \|_{L^1(I)} = \| f \|_{L^1(I)}, \qquad \text{for } i = 1, 2, 3 \text{ and } k < J.$$

In addition, $\psi_{0,k}^{(i)}$ is nonzero on $[0, 2^{-k}]^2$. Therefore, if the sum for $\tilde{f}$ omits any of the $3J$ terms $c_{0,k}^{(i)} \psi_{0,k}^{(i)}$ for $i = 1, 2, 3$ and $k < J$, we have

$$\| f - \tilde{f} \|_{L^1(I)} \geq C \| f \|_{L^1(I)} \geq C \| f \|_{B_q^{\alpha,1}(L^q(I))}.$$

Since J is arbitrary, (3.19) cannot hold for arbitrary $\mathcal{N}$ when $p = 1$. This argument signifies that if we decide to measure the error of image compression in the $L^1(I)$ norm, then the Haar transform does not lead to optimal algorithms for the class $B_q^{\alpha}(L^q(I))$, $q^{-1} = 1 + \alpha/2$.

For images, f takes values only from the set $0, \cdots, 2^n - 1$, and one would not usually calculate the values of $d_{j,k}$ to infinite precision. We consider a new method where

$$d_{j,k} := \tilde{\mathbb{Q}}_2 f_{j,k},$$

i.e., the coefficients $d_{j,k}$ are now given the value of the average rounded to the nearest integer value. A nontrivial argument shows that the coefficients $d'_{j,k}$ now satisfy

$$-3 \leq d'_{(2j_1, 2j_2), k} + d'_{(2j_1+1, 2j_2), k} + d'_{(2j_1, 2j_2+1), k} + d'_{(2j_1+1, 2j_2+1), k} \leq 3,$$

so we see from (3.16) that the coefficients $c_{j,k}^{(4)}$ are no longer zero but take values between -3 and 3. Again, if we ignore the superscripts we can write

$$f = \frac{1}{4} \sum_{k \geq 0,\, j \in \mathbb{Z}_k^2} c_{j,k} \psi_{j,k}$$

for any f associated to an image. Now the representation functions $\{\psi^{(1)}_{j,k}, \psi^{(2)}_{j,k}, \psi^{(3)}_{j,k}, \psi^{(4)}_{j,k}\}$ are no longer linearly independent, and there are $4/3$ times as many coefficients $c_{j,k}$ as there are pixels p_j. However, the extra coefficients all have $|c_{j,k}| \le 3$, so they add little to the number of bytes needed to represent a picture, and indeed, if the quantization interval is greater than 8, then the extra quantized coefficients are all zero. At any rate, the exact Haar transform generates $10\frac{2}{3}$ bits per pixel, whereas our redundant, modified Haar transform using the projector $\tilde{\mathbb{Q}}_2$ generates 11 bits per pixel.

By Theorem 3, our new functions S^k are now locally near-best approximations in $L^q(I)$ for any $q < 1$ with a near-best bound that depends on q and 2^n. Thus, by our theory, the new decomposition does satisfy for any $0 < p < \infty$,

$$\|f - \tilde{f}\|_{L^p(I)} \le C_{n,p} \mathcal{N}^{-\alpha/2} \|f\|_{B^{\alpha,1}_q(L^q(I))},$$

$$0 < \alpha, \qquad \frac{1}{q} = \frac{\alpha}{2} + \frac{1}{p}.$$

It is interesting to note that the Haar transformation defined using real arithmetic does *not* generate near-optimal approximations in $L^1(I)$ to functions in $B^{\alpha,1}_q(L^q(I))$, whereas the Haar transform as it is more likely to be implemented, with rounded integer arithmetic, *does* generate near-optimal approximations.

We comment here about one final stability consideration for the projections $\tilde{\mathbb{Q}}_2 f$: How many bits accuracy must one maintain so that $\tilde{\mathbb{Q}}_2 f$ is a near-best approximation in $L^q(I)$ for $0 < q < 1$? Theorem 3 assumes that the intermediate averages are calculated perfectly, which for an image will require that two bits of accuracy must be added for each level k, $k = 1, \cdots, m$. In the following theorem, we give an upper bound on the number of bits that are needed to the right of the binary point to maintain stability.

We shall work under the following assumptions. Let K denote the number of bits to the right of the binary point. We assume that the pixel values p_j, $j = (j_1, j_2) \in [0, 2^m - 1]^2$, are integers between 0 and $2^n - 1$, inclusive. We set $a_{j,m} = p_j$ for $j \in \mathbb{Z}^2_m$, and for $k = m, \cdots, 1$, we calculate

$$a_{j,k-1} = \text{round}_K\Big(\tfrac{1}{4}\big(a_{2j_1,2j_2),k} + a_{(2j_1+1,2j_2),k} + a_{(2j_1,2j_2+1),k} + a_{(2j_1+1,2j_2+1),k}\big)\Big).$$

Here the summation and multiplication are assumed to be computed exactly, and round_K is the operator that takes any real number x and rounds it to have K bits to the right of the binary point, i.e.,

$$\text{round}_K(x) = \text{round}(2^K x)/2^K.$$

(It does not matter whether round(1/2) = 0 or round(1/2) = 1.) We then set $d_{j,k} = \bar{\mathbb{Q}}_2 f_{j,k} := \text{round}(a_{j,k})$ for all j, k.

Theorem 7: If $m \le 2^{K-1} + \lfloor K/2 \rfloor$ then for all $0 < q \le 1$ and for all $n > 0$, there exists a constant $\bar{C}_{n,q}$ such that for all $f = \sum_j p_j \phi_{j,m}$,

$$\|\bar{\mathbb{Q}}_2 f_{j,k} - f\|_{L^q(I_{j,k})} \le \bar{C}_{n,q} \|\mathbb{Q}_q f_{j,k} - f\|_{L^q(I_{j,k})},$$

for all $0 \le k \le m$ and $j \in \mathbb{Z}^2_k$. (Here, $\lfloor x \rfloor$ is the largest integer less than or equal to x.)

For example, if the $a_{j,k}$ are rounded at each step to 5 bits to the right of the binary point, then $K = 5$ and $\bar{\mathbb{Q}}_2 f$ is a near-best approximation if $m \le 18$, which is sufficient for any purpose. If the pixel data takes values between 0 and 255 then the numbers $a_{j,k}$ can be calculated to this accuracy using 16 bit, signed, integer arithmetic. If one assumes 4 bits to the right of the binary point, then one requires $m \le 10$, and 10 bit pixel data can be processed using 16 bit, unsigned, integer arithmetic.

Example 2: Clipped average transform for grey scale images. The integer Haar transform of the previous section achieves near-optimal approximation, but the constant $C_{n,p}$ depends on the number of grey scale levels 2^n. If we want a transform that achieves near-optimal approximation to functions in $B^{\alpha,1}_q(L^q(I))$ where the constant does not depend on the number of grey scales, we could approximate f on each interval $I_{j,k}$ by taking

$$d_{j,k} := \max\big(f^1_{j,k}, \min\big(f^3_{j,k}, \tilde{\mathbb{Q}}_2 f_{j,k}\big)\big),$$

where $\tilde{\mathbb{Q}}_2 f_{j,k}$ is the rounded average of f on $I_{j,k}$ and $f^1_{j,k}$ and $f^3_{j,k}$ are the first and third quartiles, respectively, of f on $I_{j,k}$. By Section III-A, $S^k := \sum_{j \in \mathbb{Z}^2_k} d_{j,k} \phi_{j,k}$ is a locally near-best approximation to f in $L^q(I)$ for any $0 < q < \infty$, and the constant depends only on q and not on the number of possible grey scales. (Actually, we can make an arbitrarily large error in calculating $\mathbb{Q}_2 f_{j,k}$ because our final approximation is always between the two near-best approximations $f^1_{j,k}$ and $f^3_{j,k}$.) Note that when $k = m$ or when $k = m - 1$, $d_{j,k} = \tilde{\mathbb{Q}}_2 f_{j,k}$, because the average of four numbers is always between the largest and smallest of these numbers. In addition, if f is nearly affine (linear) on $I_{j,k}$ with a nonzero gradient, then the average of f will most likely be between the first and third quartiles. Whenever $d_{j,k-1} = \tilde{\mathbb{Q}}_2 f_{j,k-1}$, the coefficient $c^{(4)}_{j,k}$ will still be between -3 and 3. If, however, the average is outside the interval defined by the quartiles, $c^{(4)}_{j,k}$ could be quite large, increasing the number of bits necessary to represent $\tilde{f}$. In this instance we have again introduced more redundancy, but have achieved a transform that is more stable in $L^p(I)$ for $p \le 1$. For the images we use in Section V, there are very few times when $\tilde{\mathbb{Q}}_2 f_{j,k} \ne d_{j,k}$, and these occurrences have very little effect either on the error or the compressed image.

Note that in Example 1 we could calculate the transform in $O(2^{2m})$ time, where there are 2^{2m} pixels. However, to calculate the first and third quartiles $f^1_{j,k}$ and $f^3_{j,k}$ on all dyadic intervals $I_{j,k}$, one might simply sort the entire array of pixels using mergesort, which takes $O(m2^{2m})$ time. Alternately, we could exploit the fact that there are only 2^n different grey scale values to calculate the quartiles in $O(2^n 2^{2m})$ time. (In fact, when an interval contains more than

2^n pixels, it is faster simply to compute a table which contains, for each possible grey scale value, the number of pixels that take on that value.) To make a fair comparison of the different techniques one would need to know explicitly the constants hidden in the various big-O estimates.

Example 3: Median transform for grey scale images. For this transform one would choose

$$d_{j,k} := \mathbb{Q}_1 f_{j,k},$$

where $\mathbb{Q}_1 f_{j,k}$ is the median value of f on $I_{j,k}$. This transform is very much like the clipped average transform, in that it results in near-optimal transforms in $L^p(I)$ for $0 < p < \infty$, with constants that depend on p but not on the number of grey scale values 2^n; furthermore, it takes somewhat more time to calculate than the rounded Haar transform. For this transform it is not clear if the rewriting rule (3.16) will result in lower entropy of the coded coefficients, because most often the median of f on an interval $I_{j,k}$ will not be the average of f on the same interval. Where f is smooth, however, and is approximately an affine function on $I_{j,k}$, then $\mathbb{Q}_1 f_{j,k}$ and $\mathbb{Q}_2 f_{j,k}$ should be close, so (3.16) may be of some benefit.

Example 4: Median transform for bilevel images. When one compresses a bilevel image where each pixel takes one of only two values, the median operator $\mathbb{Q}_1 f_{j,k}$ on each interval $I_{j,k}$ is simple to calculate—it is the most common pixel value in $I_{j,k}$. Assume for simplicity that the two pixel values are 0 and 1. Then the coefficients $d_{j,k}$ will take on values 0 or 1, and $d'_{j,k}$ will take values from the set $\{-1, 0, 1\}$. If we are interested in compressing facsimile images then we shall not quantize these coefficients in any way, and we send the coefficients $d'_{j,k}$ in the order $d'_{0,0}$, $\{d'_{j,1}\}$, $\{d'_{j,2}\}$, etc., and successively reconstruct S^0, S^1, S^2, etc. The coefficient $d'_{j,k}$ will be zero if the most common pixel in $I_{j,k}$ is the same as the most common pixel in the interval $I_{i,k-1}$ that contains $I_{j,k}$, and it will be nonzero otherwise. So if we reconstruct the picture as we receive the coefficients $d'_{j,k}$, we need send only the absolute value of $d'_{j,k}$, which, if nonzero, will indicate that we must change the intensity of f on $I_{j,k}$ to the opposite of f on the background interval $I_{i,k-1}$.

Bilevel images that are in $B_q^\alpha(L^q(I)) \equiv B_q^{\alpha,2}(L^q(I))$ for $1 \le \alpha < 2$ can be approximated to order $(\mathcal{N}^{-\alpha/2})$ with $O(\mathcal{N})$ piecewise linear pieces. (Approximation by piecewise linear functions is discussed in the next section.) What is interesting, however, is that they can be approximated to the same order by piecewise constants, because any bilevel image in $B_q^{\alpha,2}(L^q(I))$ is also in $B_q^{\alpha,1}(L^q(I))$ by Theorem 4. This means that

$$\|f - \tilde{f}\|_{L^p(I)} \le C_p \mathcal{N}^{-\alpha/2} \|f\|_{B_q^\alpha(L^q(I))},$$

for $0 < \alpha < 2$ if f is a characteristic function.

This transform, which substitutes for 2^{2m} bits representing pixels p_j four-thirds as many bits representing coefficients $d'_{j,k}$, will not succeed in compressing the image unless the entropy of the coefficients is less than the entropy of the original. Such a condition is equivalent, heuristically, to there being spatial correlations among the pixels in f that are removed by the transformation.

F. Methods with $\alpha > 1$

Wavelet approximation methods following the framework of Section II that use piecewise polynomials of degree $< r$ can be defined for all $r > \alpha > 0$. The definitions and algorithms are slightly more complicated than the piecewise constant approximations. A summary of the methods and results for $r = 2$ (piecewise linear approximations) are collected in this section. This is not much of a practical restriction, because it is easily shown that images that have a discontinuity in intensity across a line must have $\alpha < 2$ for approximation in $L^p(I)$, $p \ge 1$; see Section IV-B.

We start with a linear *hat function*, which is a special case of so-called *box splines* introduced by de Boor and DeVore [3]. Let ϕ be the continuous, piecewise linear function with support in $[-1, 1]^2$ whose derivative is discontinuous along the lines $x_1 = -1, 0, 1$, $x_2 = -1, 0, 1$, $x_1 - x_2 = -1, 0, 1$, and which takes the values one at $x = 0$ and zero for $x = j$, $j \in \mathbb{Z}^2$, $j \ne 0$. For $k \ge 0$ define $\phi_k(x) := \phi(2^k x)$ and for $j \in \tilde{\mathbb{Z}}_k^2 := \{j \in \mathbb{Z}^2 \mid 0 \le j_1 \le 2^k, 0 \le j_2 \le 2^k\}$ define $\phi_{j,k}(x) := \phi_k(x - j/2^k)$.

One must now choose locally near-best, level k approximations

$$P_k f := \sum_{j \in \tilde{\mathbb{Z}}_k^2} d_{j,k} \phi_{j,k}$$

to f. This can be done by finding locally near-best, discontinuous, piecewise linear approximations in $L^q(I_{j,k})$ (again $\mathbb{Q}_1$ will do the job for all q; see [6]), and then using *quasi-interpolants* (see [4]) to project the discontinuous approximations onto the space spanned by $\phi_{j,k}$, $j \in \tilde{\mathbb{Z}}_k^2$. One again calculates $c_{j,k}$ from

$$P_k f - P_{k-1} f = \sum_{j \in \tilde{\mathbb{Z}}_k^2} c_{j,k} \phi_{j,k}$$

and chooses $\tilde{c}_{j,k}$ as in Algorithm 1. By the results in Section II, Theorem 1 still holds, now for $0 < \alpha < 2$ and $0 < p < \infty$.

Similar wavelet decompositions can be based on the Daubechies wavelets of high order, D_4, for example.

In a similar way, Algorithm 2 and Theorem 6 can be extended to any finite α by considering approximation by piecewise polynomials of higher and higher degree. As a practical matter, we can prove that $\alpha < 2$ for any image with intensity discontinuous across a curve, and we don't believe that approximations of order higher than linear are warranted globally.

IV. Applications to Image Compression

A. Images should be Approximated in L^1

One of the first issues to be decided when applying the previous mathematical framework to image compression is, "In which space $L^p(I)$ should one approximate an image f?" In other fields, such as computer-aided design, the answer is often simple: If one wishes to compress the representation of a function f that defines a segment of the surface of a mechanical part, then one must maintain accuracy in

$L^\infty(I)$—for two parts to fit together, they must be machined to a given maximum tolerance.

For image compression the answer is not so obvious, and depends on the sensitivity of the human observer to details at different frequencies and contrasts. One aspect of this relationship is summarized by the contrast sensitivity threshold (CST) curve, which can be describe as follows. Using the notation of the previous section, it is clear that the difference $S^k - S^{k-1} = \sum_{j \in \mathbb{Z}_k^2} c_{j,k}\phi_{j,k}$ represents the detail present at a feature size of 2^{-k}, or equivalently, at a frequency of 2^k. (For simplicity we assume that the representation functions satisfy $\phi_{j,k}(x) = \phi(2^k x - j)$ for some ϕ.) Consider an image grey $+ c \sin(2^k x)$ of an oscillating high-frequency pattern on a grey background in $x - y$ space with the contrast c adjusted so that the pattern is just visible to a human observer. Let us now increase the frequency and consider the pattern grey $+ c \sin(2^{k+1} x)$. Because the first pattern was just barely visible, the second pattern will not be distinguishable from grey, and we must increase the contrast to a value $\hat{c} := \hat{c}(k, \text{grey})$ before the second pattern becomes visible. If our approximation-theory model of an efficient compression algorithm is applied to the eye, then we can say that for all i and j,

$$\| c\phi_{i,k} \|_{\text{eye}} = \| \hat{c}\phi_{j,k+1} \|_{\text{eye}}.$$

In other words, to the eye, the inclusion of $c\phi_{i,k}$ in an approximate image is equally noticeable as the inclusion of $\hat{c}\phi_{j,k+1}$ (when both are at the threshold of what can be distinguished by people). As indicated, the new contrast level $\hat{c}$ depends on the background level of grey, and also on the frequency. Now, if we wish to choose a value of p such that measuring the error in $L^p(I)$ would be the same as the error the eye "sees," the following relationship should hold:

$$\| c\phi_{i,k} \|_{L^p(I)} = \| \hat{c}\phi_{j,k+1} \|_{L^p(I)}. \tag{4.1}$$

The function $\hat{c}(k, \text{grey})$ is called the CST curve; see, for example, [18]. Fig. 2 presents one representation of the data in the CST curve. In this pattern, the frequency of oscillation increases exponentially from left to right and the contrast decreases exponentially (at the same rate) from bottom to top. What concerns us here is the slope of the (purely vision-system-generated) curve between the oscillatory pattern in the lower middle and right of the figure and the grey region in the upper right corner, which has low contrast and high frequency. We are interested in the question of how much we must increase $\hat{c}$ from c when we double the frequency, and for which value of p does (4.1) hold? Well beyond the middle frequency ranges where contrast sensitivity is highest (this property *defines* "middle frequency," of course!), the curve has slope about -2; definitely the slope is steeper than -1. A slope of -2 is consistent with most renderings of this curve. This means that if the frequency is doubled, one must quadruple the contrast to still see the oscillations. The only value of p for which (4.1) holds under these assumptions on c and $\hat{c}$ is $p = 1$, because $\| \phi_k \|_{L^1(I)} = 4 \| \phi_{k+1} \|_{L^1(I)}$. Thus we conclude that the high-frequency sensitivity of the human visual system is consistent with our mathematical model when we choose approximations in $L^1(I)$. If the slope of the curve had been approximately -1, then the contrast would have doubled when the frequency doubled, which indicates from (4.1) that the image error should be measured in $L^2(I)$. But this is not the case.

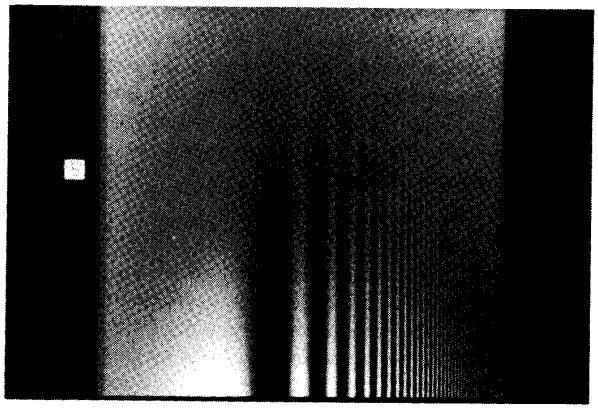


Fig. 2. Oscillating pattern whose frequency increases exponentially from left to right and whose contrast decreases exponentially (at the same rate) from bottom to top. Contrast sensitivity threshold curve (for an individual) is the imaginary curve between the grey regions in the upper left and right regions and the oscillations in the middle and lower regions.

The choice of the space $L^p(I)$ has implications in other areas also. To get $O(\mathcal{N}^{-\alpha/2})$ approximation in $L^p(I)$ with $O(\mathcal{N})$ nonzero coefficients, one needs α "derivatives" in $L^q(I)$ with $q = 1/(\alpha/2 + 1/p)$. Increasing p (from one to two, say) increases q, which implies that one needs more smoothness in the image for good approximation. In addition, if one uses orthogonal wavelets as basis elements $\phi_{j,k}$, then one may be tempted to choose the best $L^2(I)$ approximation to f. However, if the L^2 and L^1 approximations with identical levels of data compression are put side by side, it is quickly clear that the features that are saved by the L^2 approximation but are left out by the L^1 approximation are far less noticed by the human eye than the features left out by L^2 and included by L^1. To put it briefly, the $L^1(I)$ approximation looks better than the $L^2(I)$ approximation. Such comparisons are made in the section on the computational results.

B. Images have Little Smoothness in $B_q^\alpha(L^q(I))$

To apply the mathematical framework previously introduced, one characterizes the smoothness of images by their inclusion in Besov spaces $B_q^\alpha(L^q(I))$. We now give *a priori* bounds on the smoothness of any image with spatial discontinuities in intensity (that is, almost *any* image), and in the section on computation we calculate some *a posteriori* estimates of the smoothness of certain images. The experimental estimates, which are lower than the theoretical upper bounds, indicate that globally, piecewise constant approximations give as good a *rate* of approximation as piecewise polynomial approximations of higher degree. In this section, we assume a certain familiarity with the definitions and results of Section III-B.

Let us assume for example that an image consists precisely of darkness on the left ($f(x_1, x_2) = 0$ for $x_1 < 1/2$) and light on the right ($f(x_1, x_2) = 1$ for $x_1 \geq 1/2$). Because

$$\Delta_h^r f(x) = \sum_{k=0}^{r} (-1)^{r-k} \binom{r}{k} f(x + kh),$$

we know that $|\Delta_h^r f(x)| \leq 2^r \sup |f(x)| \leq 2^r$. On the other hand, if $|x_1 - 1/2| > r|h|$ then $|\Delta_h^r f(x)| = 0$, and if $h := (t, 0)$ then $|\Delta_h^r f(x)| \geq 1$ for $|x_1 - 1/2| < rt$. Thus, for small t,

$$\begin{aligned}\omega_r(f, t)_q &= \sup_{|h|<t} \left(\int_{I_{rh}} |\Delta_h^r f(x)|^q \, dx \right)^{1/q} \\ &\approx C \left(\int_{|x_1 - 1/2| \leq rt} 1 \, dx \right)^{1/q} \\ &\approx C t^{1/q},\end{aligned}$$

while for $r|h| > \sqrt{2}$, I_{rh} is empty, so $\omega_r(f, t)_q$ is constant for $t > \sqrt{2}/r$. So we have

$$\begin{aligned}|f|_{B_q^\alpha(L^q(I))} &= \left(\int_0^\infty \left[t^{-\alpha} \omega_r(f, t)_q \right]^q \frac{dt}{t} \right)^{1/q} \\ &= \left(\int_0^{\sqrt{2}/r} \left[t^{-\alpha} \omega_r(f, t)_q \right]^q \frac{dt}{t} \right. \\ &\quad \left. + \int_{\sqrt{2}/r}^\infty \left[t^{-\alpha} \omega_r(f, \sqrt{2}/r)_q \right]^q \frac{dt}{t} \right)^{1/q} \\ &\approx \left(\int_0^{\sqrt{2}/r} \left[t^{-\alpha} C t^{1/q} \right]^q \frac{dt}{t} \right. \\ &\quad \left. + \int_{\sqrt{2}/r}^\infty \left[t^{-\alpha} \omega_r(f, \sqrt{2}/r)_q \right]^q \frac{dt}{t} \right)^{1/q}.\end{aligned}$$

The second integral is always finite, whereas the first integral is finite only for $\alpha < 1/q$.

To relate this now to approximation in $L^p(I)$, we recall that

$$\frac{1}{q} = \frac{\alpha}{2} + \frac{1}{p}, \tag{4.2}$$

so that f is in $B_q^\alpha(L^q(I))$ subject to (4.2), if and only if

$$\alpha < \frac{1}{q} = \frac{\alpha}{2} + \frac{1}{p}$$

or

$$\alpha < \frac{2}{p}.$$

That is, if we wish to approximate f in $L^1(I)$, then $f \in B_q^\alpha(L^q(I))$, $1/q = \alpha/2 + 1$, at most for $\alpha = 2$, and we can achieve a rate of approximation by wavelets *of any smoothness* of at most $O(\mathcal{N}^{-1})$, where $\mathcal{N}$ is the number of nonzero coefficients in the wavelet approximation. Similarly, if we wish to approximate f in $L^2(I)$, then $\alpha < 1$, and we can achieve a rate of approximation by wavelets of at most $O(\mathcal{N}^{-1/2})$.

We can use these estimates and the results of Section III-C on pixel quantization to estimate empirically the smoothness of the intensity field F underlying the image f constructed from the quantized pixels $p_{j,m}$. Let us assume, for example, that F is in $B_\tau^\beta(L^\tau(I))$ for some $\beta > 1$, with $1/\tau = \beta/2 + 1/p$. Then $F \in B_q^\alpha(L^q(I))$, $1/q = \alpha/2 + 1/p$, for every $\alpha < 1$, since $B_\tau^\beta(L^\tau(I))$ is embedded in $B_q^\alpha(L^q(I))$ when $\beta > \alpha$. Thus, Theorem 5 implies that the image f contructed from F by pixel quantization will be in $B_q^\alpha(L^q(I)) = B_q^{\alpha,1}(L^q(I))$, and $\|f\|_{B_q^\alpha(L^q(I))} \leq C\|F\|_{B_q^\alpha(L^q(I))} + C 2^{\alpha m} 2^{-n}$. Thus, the image f is at least as smooth as the underlying intensity function F in $B_q^\alpha(L^q(I))$.

Now, it was shown in Section III that f is in $B_q^\alpha(L^q(I)) = B_q^{\alpha,1}(L^q(I))$, if and only if (roughly speaking) $\|f - \tilde{f}\|_{L^p(I)} \leq C\mathcal{N}^{-\alpha/2}$ when approximated by a piecewise-constant wavelet function $\tilde{f}$. Thus, we can estimate the smoothness of a image f by estimating the rate of decay of $\|f - \tilde{f}\|_{L^p(I)}$. If, for example, on a log-log graph of error versus number of nonzero coefficients $\mathcal{N}$, one observes that the data lie on a straight line with slope -0.3, one could reasonably assume that

$$\|f - \tilde{f}\|_{L^p(I)} \approx C\mathcal{N}^{-0.3}, \tag{4.3}$$

and that f and F are in $B_q^\alpha(L^q(I))$ for $\alpha \approx 0.6$.

Perhaps more importantly, F and f are *not* in $B_q^\alpha(L^q(I))$ for larger α; if f were in $B_q^{0.8}(L^q(I))$, for example, then we would have observed

$$\|f - \tilde{f}\|_{L^p(I)} \approx C\mathcal{N}^{-0.4},$$

while in this example we are assuming that we observe only (4.3), a slower rate of convergence. In other words, if the observed rate of convergence of our piecewise-constant wavelet approximation is approximately $\mathcal{N}^{-\alpha/2}$ with $\alpha < 1$, then f and F can have smoothness at most α.

In Section V, we have carried out this computation for four test images. In each case, we have observed convergence rates ranging from $O(\mathcal{N}^{-0.3})$ to $O(\mathcal{N}^{-0.15})$, which strongly suggests that the smoothness of the images ranges from 0.3 to 0.6. We find that these estimates of α correlate well with our subjective estimate of how smooth each image is. We have carried out these tests on other images, including a library of fingerprints and some satellite images; in all cases the smoothness of the images was between 0.3 and 0.6.

For large values of $\mathcal{N}$, the fact that f is, in fact, piecewise constant will cause the error to decay very rapidly, so one is interested in the decay rate for relatively small values of $\mathcal{N}$, or equivalently, at relatively high compression rates. Alternately, one could estimate the Besov space norm $\|f\|_{B_q^{\alpha,1}(L^q(I))}$ from the sequence norm

$$\|f\|_{\mathcal{S}} := \left(\sum_{k \geq 0,\, j \in \mathbb{Z}_k^2} \|c_{j,k} \phi_{j,k}\|_{L^p(I)}^q \right)^{1/q},$$

as in Section III-B; we do not do this.

C. How to Choose Quantization Levels

To transmit images along a communications channel, one sends integer codes that represent the quantized coefficients of the transformed image. Here we discuss how to choose quantization levels based on the criterion introduced by the generalized transform coding algorithm of Section III-D.

Specifically, for some $N > 0$, and a given representation

$$f = \sum_{k \geq 0,\, j \in \mathbb{Z}_k^2} c_{j,k} \phi_{j,k},$$

one chooses quantized coefficients $\tilde{c}_{j,k}$ such that

$$\| (c_{j,k} - \tilde{c}_{j,k}) \phi_{j,k} \|_{L^p(I)}^q \leq \frac{1}{N},$$

or

$$\| (c_{j,k} - \tilde{c}_{j,k}) \phi_{j,k} \|_{L^p(I)} \leq \frac{1}{N^{1/q}}. \tag{4.4}$$

Because $\| \phi_{j,k} \|_{L^p(I)} = 2^{-2k/p}$, (4.4) says that

$$| c_{j,k} - \tilde{c}_{j,k} | \leq \frac{2^{2k/p}}{N^{1/q}}, \tag{4.5}$$

or

$$\frac{N^{1/q}}{2^{1+2k/p}} | c_{j,k} - \tilde{c}_{j,k} | \leq \frac{1}{2}. \tag{4.6}$$

Inequality (4.5) implies that when approximating f in $L^p(I)$, one should choose a quantization separation for $c_{j,k+1}$ that is $2^{2/p}$ times the quantization separation for $c_{j,k}$. Thus, to approximate f in $L^2(I)$, one should reduce by one the number of bits one sends of $\tilde{c}_{j,k}$ for each level k. To approximate in $L^1(I)$, coefficients $\tilde{c}_{j,k+1}$ should have two fewer significant bits than $\tilde{c}_{j,k}$. (Conversely, if one follows these quantization procedures, no matter how they were arrived at, one is in fact approximating f in the appropriate $L^p(I)$ class.) In practice, we suggest setting $\tilde{c}_{j,k} = c_{j,k}$ for k less than some fixed level K, and then reducing the number of bits in $\tilde{c}_{j,k}$ for higher k according to the formula (4.5). Inequality (4.6) suggests that, equivalently, one could use the integer code

$$\text{code}_{j,k} := \text{round}\left(\frac{N^{1/q}}{2^{1+2k/p}} c_{j,k} \right)$$

to represent $\tilde{c}_{j,k}$, which is recovered by

$$\tilde{c}_{j,k} = \frac{2^{1+2k/p}}{N^{1/q}} \text{code}_{j,k}.$$

The set $\{\text{code}_{j,k}\}$ is then compressed using some type of entropy coding.

V. Computational Results

In this section, we discuss both the implementation (Section V-A) and results (Section V-B) of our generalized wavelet compression Algorithm 2 using piecewise constant approximations. We shall examine six algorithms discussed in Section III-E; each algorithm will use as a projection $\mathbb{P}$ either the rounded average of the pixels in an interval, the rounded average clipped to the quartile values of the pixels in an interval, or the median of the pixels in an interval. Three of the algorithms will use the Haar rewrite rule (3.16), while three will not. We shall report on their performance in $L^1(I)$ and $L^2(I)$, and we shall estimate the smoothness of our test images based on the rate of approximation that we achieve.

A. Implementing Algorithm 2

Here we briefly outline implementations of Algorithm 2 described in Section III-E. The theoretical aspects of the algorithms are discussed at some length in Section III-E, so we limit our discussion here to implementation questions. The images to which we applied the algorithms generally had 512×512 pixels with 256 grey scales, so, to be specific, we describe our implementations for this case.

We consider pixels to take integer values between 0 and 255. We first describe how we calculated (approximate) averages needed for the first two projections (rounded averages and rounded averages clipped to quartiles). For each square interval $I_{j,k}$ with sidelength 2^{9-k} pixels, $k = 9, \cdots, 0$, and lower-left corner pixel indexed by $2^{9-k} j := 2^{9-k}(j_1, j_2)$, we calculate the (approximate) average value of the function f on $I_{j,k}$: for all j, $a_{j,9} = p_j$, and for $k = 9, \cdots, 1$ and j,

$$a_{(j_1,j_2),k-1} = \text{round}_5 \left(\frac{1}{4} \left(a_{(2j_1,2j_2),k} + a_{(2j_1+1,2j_2),k} + a_{(2j_1,2j_2+1),k} + a_{(2j_1+1,2j_2+1),k} \right) \right), \tag{5.1}$$

where round_5 rounds real numbers to the closest number of the form $K 2^{-5}$. In effect, the averages are computed in fixed-point arithmetic with 5 bits to the right of the binary point. Because the data are represented using 8 bits, all the intermediate calculations can be carried out using 16 bit, signed, integer arithmetic.

For Example 1 of Section III-E, we set the coefficients $d_{j,k} = \text{round}(a_{j,k})$, which, like the pixels themselves, take integer values between 0 and 255. By Theorem 7, we have kept enough binary digits in our computation so that these rounded averages are near-best approximations to f on each interval $I_{j,k}$. For each $j \in \mathbb{Z}_{k-1}^2$, $k = 9, \cdots, 1$, we then set

$$d'_{(2j_1,2j_2),k} = d_{(2j_1,2j_2),k} - d_{j,k-1},$$

$$d'_{(2j_1+1,2j_2),k} = d_{(2j_1+1,2j_2),k} - d_{j,k-1},$$

$$d'_{(2j_1,2j_2+1),k} = d_{(2j_1,2j_2+1),k} - d_{j,k-1},$$

and

$$d'_{(2j_1+1,2j_2+1),k} = d_{(2j_1+1,2j_2+1),k} - d_{j,k-1}.$$

It is not very difficult to show that $-192 < d'_{j,k} < 192$ and that

$$-4 < d'_{(2j_1,2j_2),k} + d'_{(2j_1+1,2j_2),k} + d'_{(2j_1,2j_2+1),k} + d'_{(2j_1+1,2j_2+1),k} < 4. \tag{5.2}$$

For the second and third algorithms, we sort the pixels in each interval $I_{j,k}$ into increasing order $\tilde{p}_i$, $0 \le i < 2^{2(9-k)}$. For the second algorithm, the first and third quartiles are taken to be $\tilde{p}_{i_1}$ and $\tilde{p}_{i_3}$, respectively, where $i_1 = \frac{1}{4}2^{2(9-k)} - 1$ and $i_3 = \frac{3}{4}2^{2(9-k)}$. For each j and k, $d_{j,k}$ is calculates as $\max(\tilde{p}_{i_1}, \min(\tilde{p}_{i_3}, \text{round}(a_{j,k})))$. In this case, we have $-256 < d'_{j,k} < 256$ and

$$-256 < d'_{(2j_1,2j_2),k} + d'_{(2j_1+1,2j_2),k} + d'_{(2j_1,2j_2+1),k} + d'_{(2j_1+1,2j_2+1),k} < 256. \quad (5.3)$$

For the third algorithm, we take $d_{j,k}$ to be the median $\tilde{p}_{i_2}$, where $i_2 = \frac{1}{2}2^{2(9-k)}$, and do not bother with the averages at all. When taking medians we obtain the new coefficient bounds $-256 < d'_{j,k} < 256$ and

$$-1020 \le d'_{(2j_1,2j_2),k} + d'_{(2j_1+1,2j_2),k} + d'_{(2j_1,2j_2+1),k} + d'_{(2j_1+1,2j_2+1),k} \le 1020. \quad (5.4)$$

The first three algorithms set $c_{j,k} = d'_{j,k}$ and write

$$f = \sum_{k\ge 0,\ j\in\mathbb{Z}_k^2} c_{j,k}\phi_{j,k},$$

where $\phi_{j,k}$ is the characteristic function of $I_{j,k}$. The last three algorithms use the rewrite rule (3.16) to write

$$f = \frac{1}{4}\sum_{k\ge 0}\sum_{j\in\mathbb{Z}_k^2}\sum_{i=1}^{4} c_{j,k}^{(i)}\psi_{j,k}^{(i)} + d_{0,0}\phi_{0,0}.$$

In order to apply some type of lossless entropy encoding to the quantized coefficients, we need information about the range of the coefficients $c_{j,k}^{(i)}$. Whenever the range of the projection $\bar{\mathbb{P}}f$ is $0,\cdots,2^n - 1$ (the same as the range of the pixels), then the range of the $c_{j,k}^{(i)}$ for $i = 1, 2, 3$ does not depend on the projection used. This is because (3.16) can be written as

$$\begin{pmatrix} c_{j,k-1}^{(1)} \\ c_{j,k-1}^{(2)} \\ c_{j,k-1}^{(3)} \\ c_{j,k-1}^{(4)} \end{pmatrix} = \begin{pmatrix} -1 & -1 & 1 & 1 \\ -1 & 1 & -1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix} \cdot \left[\begin{pmatrix} d_{(2j_1,2j_2),k} \\ d_{(2j_1+1,2j_2),k} \\ d_{(2j_1,2j_2+1),k} \\ d_{(2j_1+1,2j_2+1),k} \end{pmatrix} - d_{j,k-1}\begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \right].$$

It is easily seen that the range of the first three coefficients is $-2^{n+1} + 2, \cdots, 2^{n+1} - 2$, independent of $d_{j,k-1} = \bar{\mathbb{P}}f \mid_{I_{j,k-1}}$. It is only for the fourth coefficient that the projection plays any part; for the three projections that we use, the bounds (5.2), (5.3), and (5.4) hold for the fourth coefficient.

We have described our wavelets and the various methods for calculating the coefficients $c_{j,k}$, which together form the transforms we use. For each algorithm we now choose a quantization strategy that depends on the $L^p(I)$ error metric that we may wish to apply. The quantization strategy is parametrized by the maximum quantization interval q. The quantization intervals q_k for the coefficients $c_{j,k}$ are chosen as $q_9 = q$ and $q_k = \max(1, \text{round}(q_{k+1}/2^{2/p}))$, $k = 8,\cdots,0$. The quantized coefficients $\tilde{c}_{j,k}$ are then taken to be $q_k \times \text{round}(c_{j,k}/q_k)$, where, to save a few more bits in the final compressed data, we always round numbers of the form $K + \frac{1}{2}\,\text{sgn}(K)$ (which occur quite often in practice) to K. We could very well have used truncation rather than rounding and set $\tilde{c}_{j,k} = q_k \times \text{trunc}(c_{j,k}/q_k)$; since the results are very similar we do not report the conclusions separately here.

B. Computations

In Figs. 3–6, we present various images that have been compressed using the projections $d_{j,k} = \text{round}(a_{j,k})$ (i.e., the projector $\bar{\mathbb{Q}}_2$ of Section III-E), the rewrite rule (3.16), and the quantization strategy associated with approximation in $L^1(I)$ (i.e., $p = 1$). We report the number of nonzero coefficients $\tilde{c}_{j,k}$ and the number of bytes used to represent the coefficients $\tilde{c}_{j,k}$ after encoding by a 10 bit, conditional, adaptive, binary, arithmetic coder derived from work in [21]. (We encode all the information needed to reconstruct $\tilde{f}$, i.e., the values and locations of the coefficients $\tilde{c}_{j,k}$.) Each figure presents for one commonly-used 512 by 512 image the original and compressed images with $q = 128$, 256, and 512. (These images are the green components of RGB color images, so in some cases, Fig. 3, for example, they are quite dark.) The images can be grouped together, in that the first and second images are rather easy to compress, while the last two are more difficult, because of their lack of smoothness.

We have argued on the basis of the CST curve that images should be compressed in $L^1(I)$ rather than $L^2(I)$ if we are to minimize the human perception of the compressed image error. In Figs. 7–10, we compare images compressed in $L^1(I)$, with $q = 1024$, with images compressed in $L^2(I)$, with q determined so that the final size of the encoded $L^2(I)$ coefficients was as close as possible to the size of the encoded $L^1(I)$ coefficients. We claim that at least for Figs. 7 and 8 the $L^1(I)$ picture looks better; this effect is also seen at other compression levels, but does not seem to be as marked for the more complex pictures.

We now report on various systematic experiments to investigate the effects of the different projections and the rewrite rule (3.16) on compression. The results for the four pictures were similar; in Figs. 11 and 12 we compare the $L^1(I)$ and $L^2(I)$ error to the number of nonzero coefficients $\tilde{c}_{j,k}$ for each of the six methods when applied to lenna. Three things should be noticed. First, it seems to make little practical difference which projection or rewrite rule we use. Second, the results when the coefficients are not clipped or clipped to quartiles are essentially the same; this type of clipping was rarely invoked. This implies, very roughly, that for our test images the rounded average projection was almost as stable as the rounded average clipped to quartiles, despite the different bounds of Theorem 2, which depends only on the space $L^q(I)$, and Theorem 7, which depends on both $L^q(I)$ and the number of grey scales. Therefore, the extra computa-

Fig. 3. Lenna, compressed using the $L^1(I)$ quantization strategy: (a) original; (b) compressed with $q = 128$ (20 236 coefficients, 14 604 bytes); (c) compressed with $q = 256$ (12 068 coefficients, 8925 bytes); (d) compressed with $q = 512$ (7001 coefficients, 5404 bytes).

tion to compute the grey scale quartiles on each dyadic subinterval of $[0, 1]^2$ in order to compute the more stable clipped average projection operator does not seem to be worth the trouble. Third, the advantage of having small values of $c_{j,k}^{(4)}$ when the rounded average is used as a projection and the rewrite rule (3.16) is used (i.e., what is closest to the classical Haar transform method) seems to result in slightly fewer coefficients for a given error than the other methods. Because of this, we shall use this combination of projection and rewrite rule when comparing the differences in performance among the various images.

We next compare the error $\|f - \tilde{f}\|_{L^p(I)}$ to $\mathcal{N}$, the number of nonzero coefficients in $\tilde{f}$. (We normalize f so that black pixels are zero and white pixels are one.) Fig. 13 presents the case $p = 1$; i.e., the error was measured in $L^1(I)$ and the L^1 quantization strategy was used. When there are over 100 000 coefficients (high image quality), the fact that the image is in fact piecewise constant leads to a very rapid decrease in the error with any increase in the number of coefficients, so for the moment we concentrate on the part of the graphs with between 100 and 30,000 coefficients. In this range, the graphs for all the images are almost linear, so that to a very good approximation $\|f - \tilde{f}\|_{L^p(I)} \approx C\mathcal{N}^{-\beta}$ for different values of C and β. Because of (1.4) and (3.14) we can use this information to estimate the smoothness of the images: we estimate $f \in B_q^\alpha(L^q(I))$, $\alpha \approx 2\beta$ and $1/q = \alpha/2 + 1$, and $\|f\|_{B_q^\alpha(L^q(I))} \approx C$. The results (using the eight leftmost data points for each image) are reported in Table II. The first two images have a Besov space smoothness of $\alpha \approx 0.6$; $\alpha \approx 0.35$ for the other images. (The correlation coefficient indicates the goodness of linear fit on a log-log scale.) Because in all cases $\alpha < 1$, these figures suggest that piecewise constant wavelet approximations achieve the highest *rate* of approximation for image compression in the $L^1(I)$ metric. That the latter two images have significantly less smoothness than the first two images expresses mathematically what may be concluded on a purely subjective basis simply by looking at them.

The corresponding graphs when approximating in $L^2(I)$ are given in Fig. 14. Interpreting this data is more difficult. On the one hand, it seems that the rate of error decay in $L^2(I)$ is sometimes greater than that in $L^1(I)$. On the other

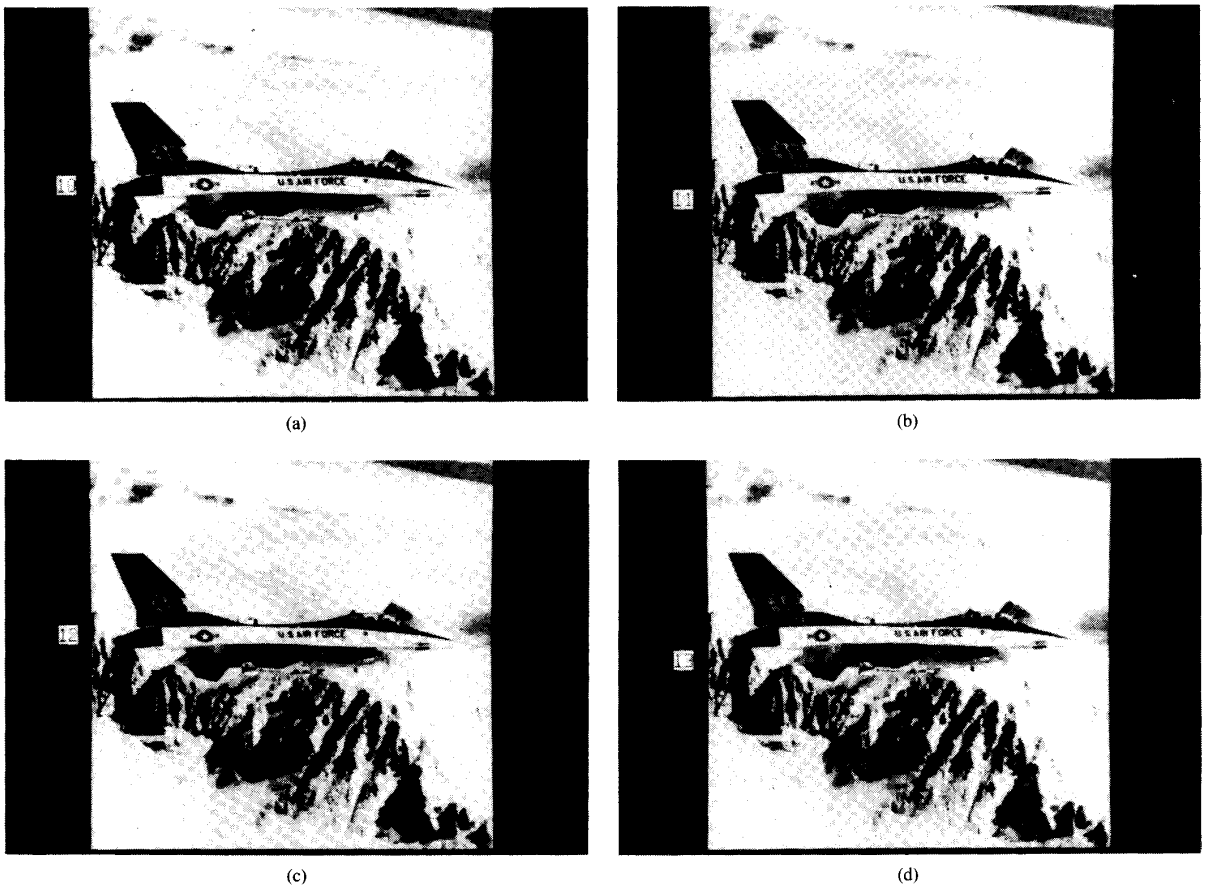


Fig. 4. F-16, compressed using the $L^1(I)$ quantization strategy: (a) original; (b) compressed with $q = 128$ (24 309 coefficients, 18 420 bytes); (c) compressed with $q = 256$ (14 613 coefficients, 11 376 bytes); (d) compressed with $q = 512$ (8613 coefficients, 6736 bytes). (The left-most column of F-16 contains black pixels, with the value zero, thereby reducing the amount of compression that we achieve.)

hand, this contradicts the fact that $f \in B_q^\alpha(L^q(I))$ with $1/q = \alpha/2 + 1/2$ implies that $f \in B_s^\alpha(L^s(I))$ with $1/s = \alpha/2 + 1$, so that any convergence rate achievable in $L^2(I)$ is also achievable in $L^1(I)$. (Of course, it is possible that no higher rate is achievable in $L^1(I)$.) It could happen that the fast convergence rate for large numbers of coefficients observed in the $L^1(I)$ approximation kicks in much earlier with $L^2(I)$ approximation, say around 1,000 coefficients. At any rate, we present a summary of the data in Table III, which was computed with the three leftmost data points in each graph.

Our theory bounds the number of nonzero coefficients $\tilde{c}_{j,k}$, when what is of practical interest is the number of bytes needed to represent these coded coefficients. We compare these two measures of compression in Fig. 15. We calculate $349525 \approx \frac{4}{3} \times 2^{18}$ coefficients $\tilde{c}_{j,k}$; except when we used median clipping together with the rewrite rule (3.16), all the coefficients can be represented using 10 bits. To algebraically encode these coefficients we used a 10 bit, conditional, adaptive, binary, arithmetic coder derived from work in [21]. The coder as implemented achieves a compression rate of at most 2,000 to one, even if all the coefficients are zero. (We did not add information to the coder about the maximum possible number of bits in the coefficients $\tilde{c}_{j,k}$ given the quantization level q_k at level k. We did, however, reorder the coefficients [in an image-independent way] to improve the performance of the coder.) Because of this, the coder overhead is relatively large when there are fewer than 1,000 nonzero $\tilde{c}_{j,k}$, so we did not include these data points in Fig. 15. We also left out the data points with median filtering and the Haar rewrite rule. A linear fitting to the log-log data shows that $N = 1.102\mathcal{N}^{0.958}$, where N is the number of bytes in the encoded coefficients, with a correlation coefficient of 0.998. We postulate that the slight upward curve in the graph when nearly all the $\tilde{c}_{j,k}$ are nonzero arises because there are then no small coefficients, as there usually are when there are only few nonzero coefficients, so it takes more bits to encode the larger $\tilde{c}_{j,k}$.

ACKNOWLEDGMENT

The authors thank B. W. Cleveland for coding the entropy coder, writing our X display software, and rendering the figures.

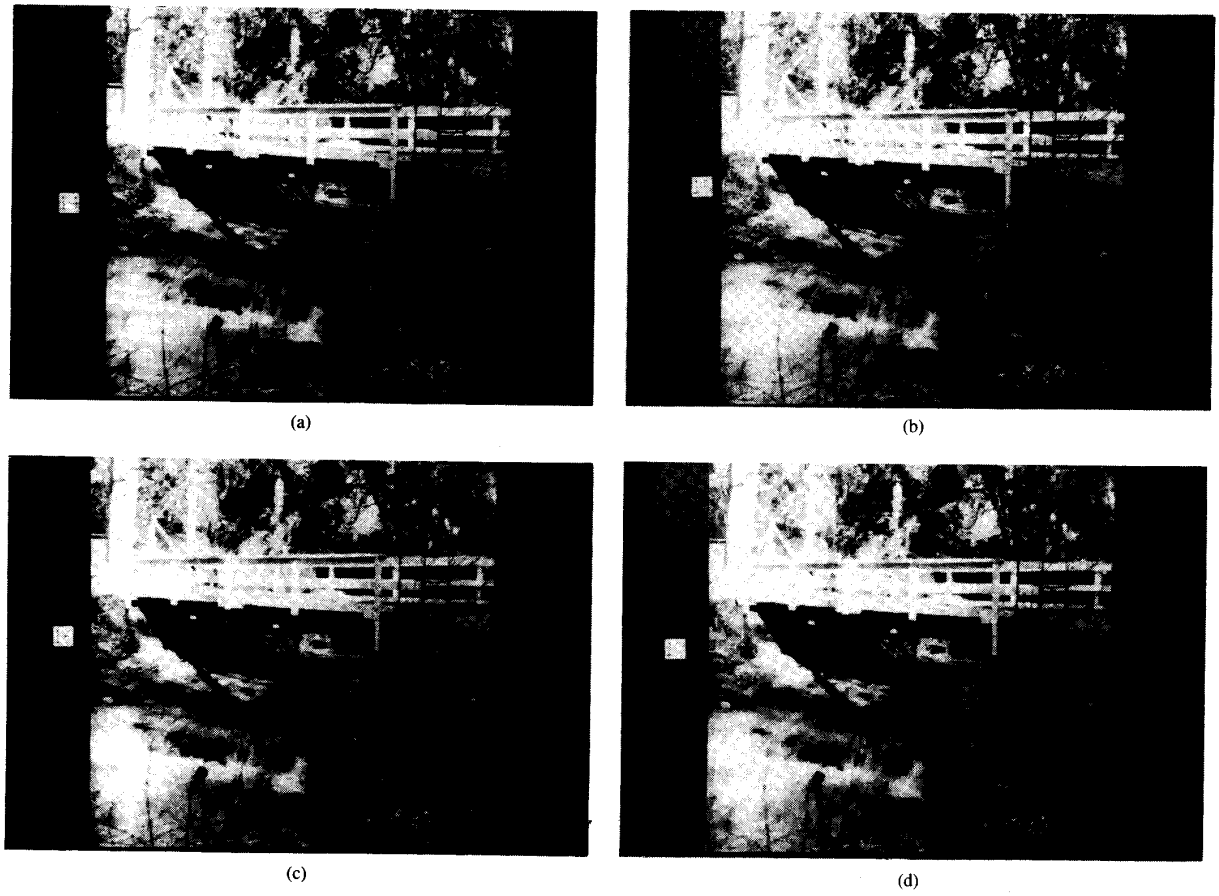


Fig. 5. Bridge, compressed using the $L^1(I)$ quantization strategy: (a) original; (b) compressed with $q = 128$ (44 599 coefficients, 28 917 bytes); (c) compressed with $q = 256$ (23 286 coefficients, 15 292 bytes); (d) compressed with $q = 512$ (11 928 coefficients, 8069 bytes).

Appendix

Proofs of Theorems

Theorem 2: For each $q \in (0, \infty]$ there exists a constant C_q such that for all $f \in L^q(I)$,

$$\| f - \mathbb{Q}_1 f \|_{L^q(I)} \le C_q \| f - \mathbb{Q}_q f \|_{L^q(I)}.$$

Proof: Because, by the previous remark, the theorem is true for $q \ge 1$, we concentrate on $0 < q < 1$. First, it is clear that $\mathbb{Q}_1 f$ is defined for $f \in L^q(I)$, because medians are defined for all measurable functions, not just integrable ones. Next, we show that $\mathbb{Q}_1$ is a bounded (nonlinear) operator on $L^q(I)$. Assume, without loss of generality, that $\mathbb{Q}_1 f = m$ is greater than zero, where m denotes a median of f on I. Then we have

$$\begin{aligned} \| f \|^q_{L^q(I)} = \int_I | f(x) |^q \, dx &\ge \int_{f(x) \ge m} f(x)^q \, dx \\ &\ge \int_{f(x) \ge m} m^q \, dx \\ &\ge \frac{1}{2} m^q = \frac{1}{2} \| \mathbb{Q}_1 f \|^q_{L^q(I)}. \end{aligned}$$

Therefore, for any f,

$$\| \mathbb{Q}_1 f \|_{L^q(I)} \le 2^{1/q} \| f \|_{L^q(I)}.$$

Although $\mathbb{Q}_1$ is a nonlinear mapping, it *is* linear with respect to the subtraction of constant functions, that is, for any v such that $v(x) \equiv c$ for all $x \in I$, $\mathbb{Q}_1(f - v) = \mathbb{Q}_1 f - \mathbb{Q}_1 v = \mathbb{Q}_1 f - v$ for some choice of $\mathbb{Q}_1 f$ and $\mathbb{Q}_1(f - v)$. Therefore,

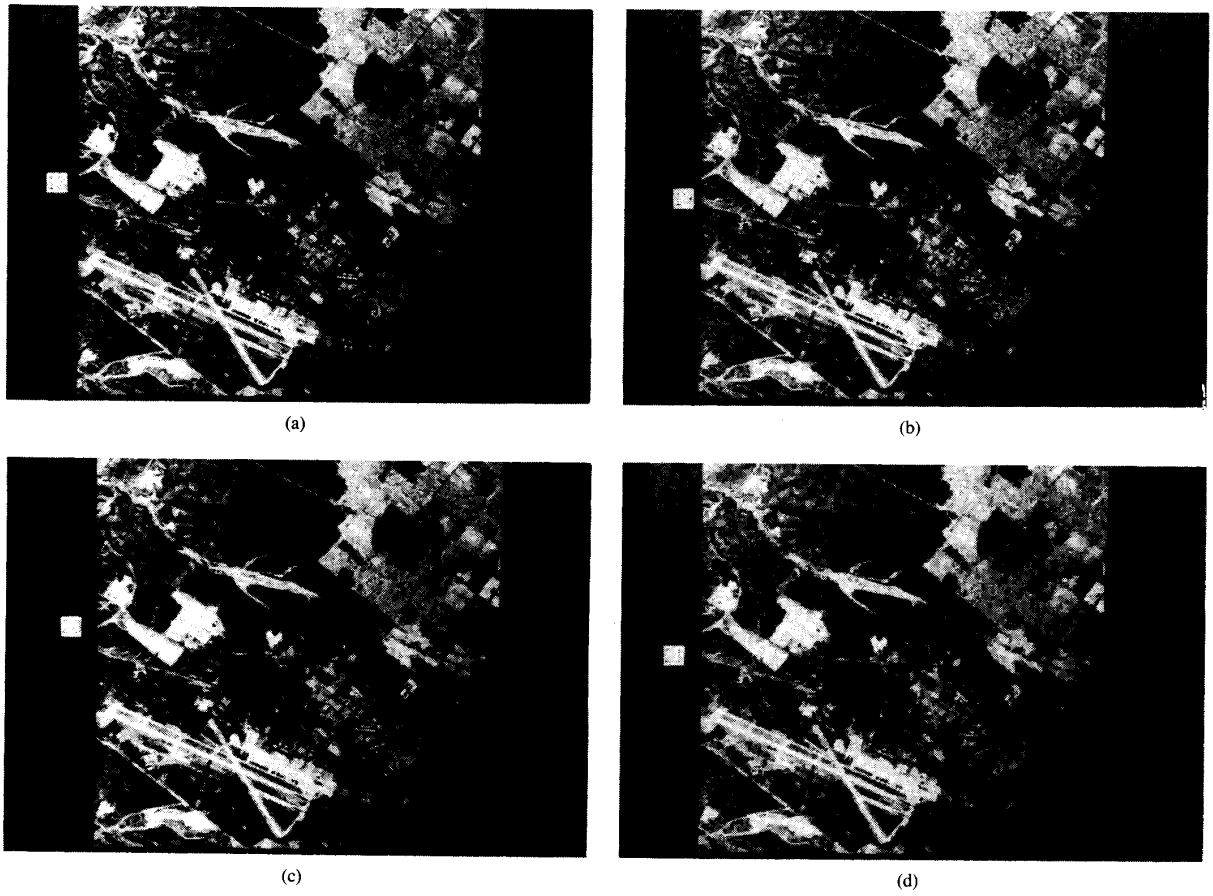
$$\begin{aligned} \| \mathbb{Q}_1 f - f \|^q_{L^q(I)} &\le \| \mathbb{Q}_1 f - \mathbb{Q}_q f \|^q_{L^q(I)} + \| \mathbb{Q}_q f - f \|^q_{L^q(I)} \\ &= \| \mathbb{Q}_1 (f - \mathbb{Q}_q f) \|^q_{L^q(I)} + \| \mathbb{Q}_q f - f \|^q_{L^q(I)} \\ &\le 2 \| \mathbb{Q}_q f - f \|^q_{L^q(I)} + \| \mathbb{Q}_q f - f \|^q_{L^q(I)} \\ &= 3 \| \mathbb{Q}_q f - f \|^q_{L^q(I)}. \end{aligned}$$

Therefore, $\| \mathbb{Q}_1 f - f \|_{L^q(I)} \le 3^{1/q} \| \mathbb{Q}_q f - f \|_{L^q(I)}$. □

We remark that a more delicate argument shows that we can take $C_q = 2^{\frac{1}{q} - 1}$; see [6].

Theorem 3: Assume that $N > 0$ and mutually disjoint (measurable) sets $I_j \subset I$ are given for $j = 0, \cdots, N$ such that $I = \bigcup_{j=0}^{N} I_j$ and for all $x \in I$,

$$f(x) = \sum_{j=0}^{N} j \chi_{I_j}(x);$$



(a) (b) (c) (d)

Fig. 6. Airport, compressed using the $L^1(I)$ quantization strategy: (a) original; (b) compressed with $q = 128$ (47 548 coefficients, 30 630 bytes); (c) compressed with $q = 256$ (23 979 coefficients, 15 529 bytes); (d) compressed with $q = 512$ (12 195 coefficients, 8134 bytes).

i.e., f takes only finitely many evenly spaced values on I. If we define $\tilde{\mathbb{Q}}_2 f$ to be $\mathbb{Q}_2 f$ rounded to the nearest integer, then for each $0 < q > 1$, there exists a constant $C_{N,q}$ such that

$$\|f - \tilde{\mathbb{Q}}_2 f\|_{L^q(I)} \le C_{N,q} \|f - \mathbb{Q}_q f\|_{L^p(I)}.$$

Proof: We know that

$$\|\mathbb{Q}_1 f - f\|_{L^q(I)} \le C_q \|\mathbb{Q}_q f - f\|_{L^q(I)},$$

so if we show

$$\|\mathbb{Q}_1 f - \tilde{\mathbb{Q}}_2 f\|_{L^q(I)} \le C_{N,q} \|\mathbb{Q}_1 f - f\|_{L^q(I)},$$

then

$$\begin{aligned}\|\tilde{\mathbb{Q}}_2 f - f\|^q_{L^q(I)} &\le \|\tilde{\mathbb{Q}}_2 f - \mathbb{Q}_1 f\|^q_{L^q(I)} + \|\mathbb{Q}_1 f - f\|^q_{L^q(I)} \\ &\le C_{N,q} \|\mathbb{Q}_q f - f\|^q_{L^q(I)}.\end{aligned}$$

We can always take $\mathbb{Q}_1 f$ to be integer valued.

If $\tilde{\mathbb{Q}}_2 f = \mathbb{Q}_1 f$ then we are done. Otherwise, assume $\tilde{\mathbb{Q}}_2 f = \mathbb{Q}_1 f + M$, $M > 0$. (The argument is the same when $M < 0$.) Because $\tilde{\mathbb{Q}}_2 f$ is $\mathbb{Q}_2 f$ rounded to the nearest integer, we must have that $\mathbb{Q}_2 f \ge \mathbb{Q}_1 f + M - 1/2$. Therefore, because $f \le N$ and f takes only integer values, the set $\Omega := \{x \in I \mid f(x) \ge \mathbb{Q}_1 f + 1\}$ has measure

$$|\Omega| \ge (M - 1/2)/N.$$

Thus,

$$\|\tilde{\mathbb{Q}}_2 f - \mathbb{Q}_1 f\|_{L^q(I)} = M$$

and

$$\|\mathbb{Q}_1 f - f\|_{L^q(I)} \ge \left(\int_\Omega 1^q \, dx\right)^{1/q} \ge \left[(M - 1/2)/N\right]^{1/q}.$$

Consequently,

$$\begin{aligned}\|\tilde{\mathbb{Q}}_2 f - \mathbb{Q}_1 f\|_{L^q(I)} &\le M/(M - 1/2)^{1/q} N^{1/q} \|\mathbb{Q}_1 f - f\|_{L^q(I)} \\ &\le (2N)^{1/q} \|\mathbb{Q}_1 f - f\|_{L^q(I)}.\end{aligned}$$

□

Theorem 4: If f takes only the values 0 and 1 in I, $1 \le \alpha < 2$, and $f \in B^\alpha_q(L^p(I)) \equiv B^{\alpha,2}_q(L^p(I))$, then $f \in B^{\alpha,1}_q(L^p(I))$ and $\|f\|_{B^{\alpha,1}_q(L^p(I))} \le 2^\alpha \|f\|_{B^{\alpha,2}_q(L^p(I))}$.

Proof: We claim

$$\begin{aligned}|\Delta^1_{2h} f(x)| &= |f(x + 2h) - f(x)| \\ &\le |\Delta^2_h f(x)| = |f(x + 2h) - 2f(x + h) + f(x)|.\end{aligned} \tag{6.1}$$

(a)

(b)

Fig. 7. (a) Lenna compressed in $L^2(I)$ with $q = 303$ (4545 coefficients, 3587 bytes). (b) Lenna compressed in $L^1(I)$ with $q = 1024$ (4514 coefficients, 3587 bytes).

(a)

(b)

Fig. 8. (a) F-16 compressed in $L^2(I)$ with $q = 340$ (5038 coefficients, 4042 bytes). (b) F-16 compressed in $L^1(I)$ with $q = 1024$ (4925 coefficients, 4026 bytes).

If $f(x + 2h) = f(x)$ then this is obvious. If $f(x + 2h) \neq f(x)$ then regardless of the value of $f(x + h)$, we have $|\Delta^1_{2h} f(x)| = 1 = |\Delta^2_h f(x)|$. Since this is true for all x, $\omega_1(f, 2t)_p \leq \omega_2(f, t)_p$. Substituting this into (3.3) shows that $\|f\|_{B^{\alpha,1}_q(L^p(I))} \leq 2^\alpha \|f\|_{B^{\alpha,2}_q(L^p(I))}$. □

Lemma 1: Assume that $p > 0$, $0 < \alpha < 2/p$, and $1/q = \alpha/2 + 1/p$. Then, there exists a constant C such that for each f of the form (3.9),

$$\|f\|_{B^{\alpha,1}_q(L^q(I))} \leq C\|f\|_{L^\infty(I)} 2^{\alpha m}. \tag{3.10}$$

Proof: We have

$$\|f\|_{L^q(I)} \leq \|f\|_{L^\infty(I)}.$$

This bounds the first term on the right of (3.3) in the correct way. We next want to bound the qth power of the second term,

$$\int_0^\infty \left[t^{-\alpha}\omega_1(f, t)_q\right]^q \frac{dt}{t},$$

where

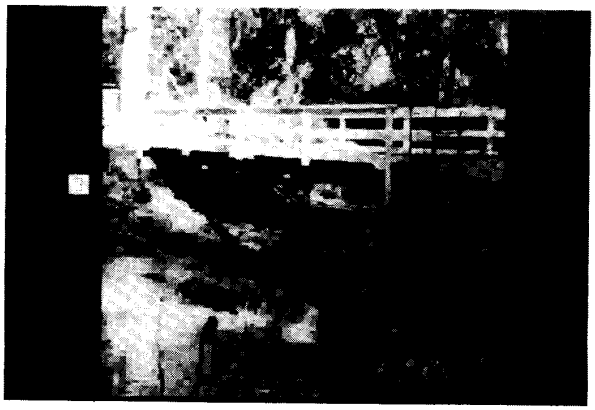
$$\omega_1(f, t)^q_q = \sup_{|h| \leq t} \int_{I_h} |f(x + h) - f(x)|^q \, dx.$$

If x and $x + h$ are in the same square $I_{j,m}$, then $|f(x + h) - f(x)|^q = 0$; otherwise, $|f(x + h) - f(x)|^q \leq 2^q \|f\|^q_{L^\infty(I)}$. The points x and $x + h$ are in the same square unless the line joining x and $x + h$ crosses one of the 2^{m+1} lines $x = i2^{-m}$ or $y = i2^{-m}$, $0 < i < 2^m$, in which case the distance from x to that line must be no greater than $|h|$. Therefore, the set where $|f(x + h) - f(x)|^q$ is nonzero has measure at most $2^{m+1}|h|$; of course, I_h is contained in I, so the measure of I_h is also less than 1. Thus, the measure of the set where $|f(x + h) - f(x)|^q$ is nonzero is no greater than $\min(1, 2^{m+1}|h|)$. Therefore,

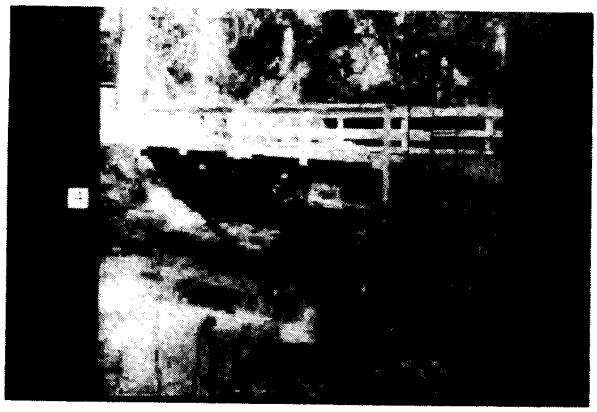
$$\int_{I_h} |f(x + h) - f(x)|^q \, dx \leq 2^q \|f\|^q_{L^\infty(I)} \min\left(1, 2^{m+1}|h|\right).$$

We can conclude that

$$\omega_1(f, t)^q_q \leq 2^q \|f\|^q_{L^\infty(I)} \min\left(1, 2^{m+1}t\right)$$



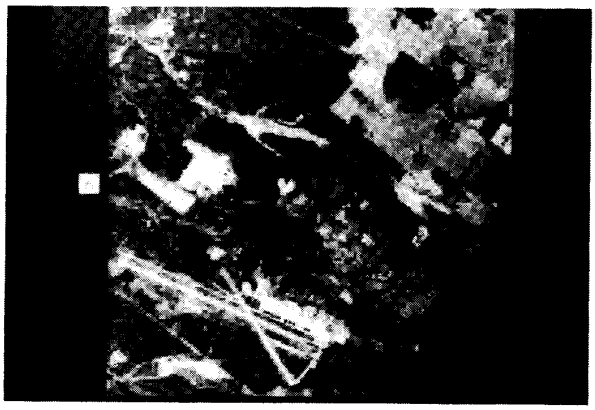
(a)



(b)

Fig. 9. (a) Bridge compressed in $L^2(I)$ with $q = 330$ (5674 coefficients, 4390 bytes). (b) Bridge compressed in $L^1(I)$ with $q = 1024$ (6258 coefficients, 4401 bytes).

(a)



(b)

Fig. 10. (a) Airport compressed in $L^2(I)$ with $q = 318$ (5848 coefficients, 4339 bytes). (b) Airport compressed in $L^1(I)$ with $q = 1024$ (6240 coefficients, 4294 bytes).

and

$$\begin{aligned}
&\int_0^\infty \left[t^{-\alpha}\omega_1(f,t)_q\right]^q \frac{dt}{t} \\
&\quad = \int_0^{2^{-m-1}} \left[t^{-\alpha}\omega_1(f,t)_q\right]^q \frac{dt}{t} \\
&\qquad + \int_{2^{-m-1}}^\infty \left[t^{-\alpha}\omega_1(f,t)_q\right]^q \frac{dt}{t} \\
&\quad \le 2^q \|f\|^q_{L^\infty(I)} \int_0^{2^{-m-1}} t^{-\alpha q} 2^{m+1}\, dt \\
&\qquad + 2^q \|f\|^q_{L^\infty(I)} \int_{2^{-m-1}}^\infty t^{-\alpha q-1}\, dt \\
&\quad = 2^q \|f\|^q_{L^\infty(I)} \left(\frac{2^{(\alpha q-1)(m+1)} 2^{m+1}}{1-\alpha q} + \frac{2^{\alpha q(m+1)}}{\alpha q} \right) \\
&\quad = 2^{(1+\alpha)q} 2^{\alpha m q} \|f\|^q_{L^\infty(I)} \left(\frac{1}{1-\alpha q} + \frac{1}{\alpha q} \right).
\end{aligned}$$

The first integral on the right is finite when $\alpha q < 1$, i.e., $\alpha < 2/p$, and the second is finite since $\alpha q > 0$. The lemma follows. □

Theorem 5: Assume that $p > 0$, $0 < \alpha < 2/p$, and $1/q = \alpha/2 + 1/p$. Then a) if $1 < p < \infty$ and $\mathbb{P}$ is the $L^2(I)$ projection (i.e., the average of the intensity), or b) if $0 < p < \infty$ and $\mathbb{P}$ is a near-best projection operator in $L^q(I)$, there exists a C such that for all F,

$$\|f\|_{B_q^{\alpha\cdot 1}(L^q(I))} \le C\|F\|_{B_q^{\alpha\cdot 1}(L^q(I))} + C2^{\alpha m}2^{-n}.$$

Proof: We can set

$$S^k(F) := \sum_{j\in\mathbb{Z}_k^2} \left(\mathbb{P}_{j,k}F\right)\phi_{j,k},$$

and define $c_{j,k}$ by

$$F = \sum_{k=0}^\infty \left(S^k(F) - S^{k-1}(F)\right) = \sum_{k\ge 0,\, j\in\mathbb{Z}_k^2} c_{j,k}\phi_{j,k}.$$

Then, one can conclude from [17] in case a) and [15] in case b) that the sequence norm

$$\|F\|_{\mathcal{I}} = \left(\sum_{k\ge 0,\, j\in\mathbb{Z}_k^2} \|c_{j,k}\phi_{j,k}\|^q_{L^p(I)} \right)^{1/q}$$

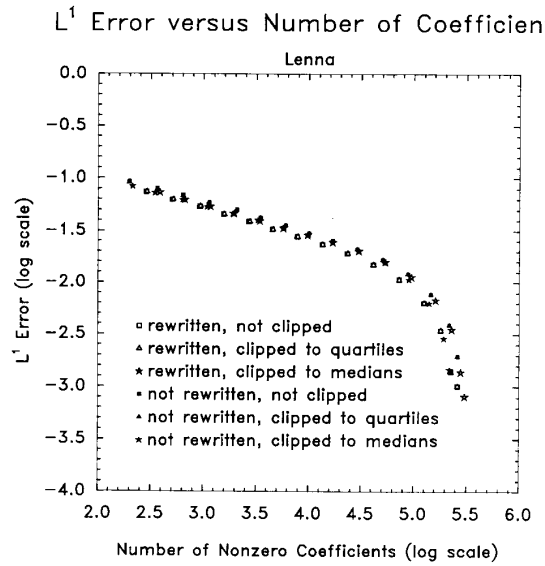


Fig. 11. The $L^1(I)$ error versus the number of nonzero coefficients $\tilde{c}_{j,k}$ for the six methods described in the text applied to Lenna. Here, we quantize in $L^1(I)$ with $q = 2^i$, $i = 1, \cdots, 15$.

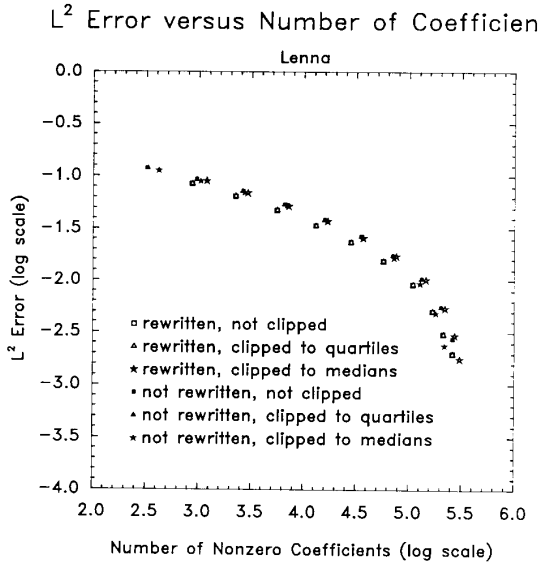


Fig. 12. $L^2(I)$ error versus the number of nonzero coefficients $\tilde{c}_{j,k}$ for the six methods described in the text applied to Lenna. Here, we quantize in $L^2(I)$ with $q = 2^i$, $i = 1, \cdots, 10$.

is equivalent to the Besov space norm $\|F\|_{B_q^{\alpha,1}(L^q(I))}$. From the comment at the end of Section III-B,

$$\|S^m(F)\|_{B_q^{\alpha,1}(L^q(I))} \le C\left(\sum_{k=0}^{m}\sum_{j\in\mathbb{Z}_k^2}\|c_{j,k}\phi_{j,k}\|_{L^p(I)}^q\right)^{1/q};$$

furthermore,

$$\left(\sum_{k=0}^{m}\sum_{j\in\mathbb{Z}_k^2}\|c_{j,k}\phi_{j,k}\|_{L^p(I)}^q\right)^{1/q} \le \|F\|_{\mathcal{S}} \le C\|F\|_{B_q^{\alpha,1}(L^q(I))}.$$

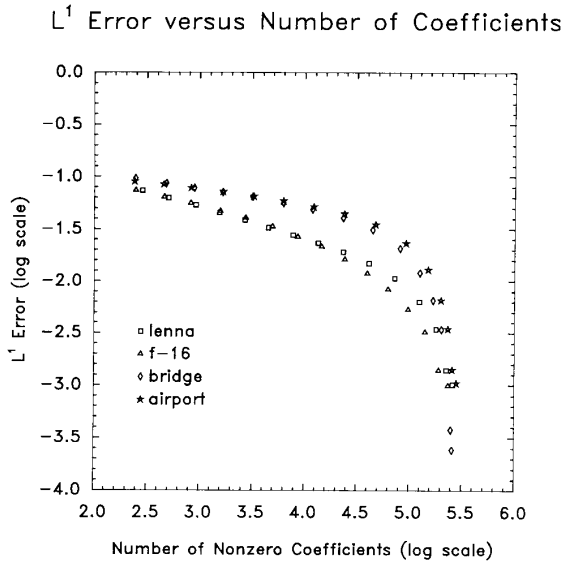


Fig. 13. $L^1(I)$ error versus the number of nonzero coefficients $\tilde{c}_{j,k}$ for the projector $\overline{Q}_2$ and the Haar rewrite rule (3.16) applied to the four images Lenna, F-16, bridge, and airport. Here, we quantize in $L^1(I)$ with $q = 2^i$, $i = 1, \cdots, 15$.

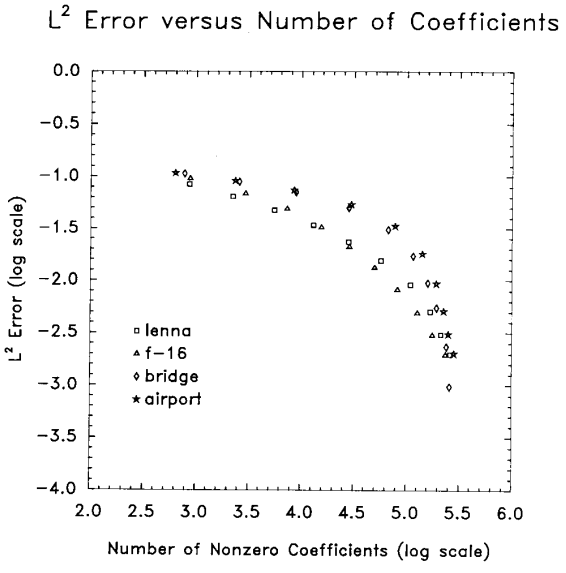


Fig. 14. $L^2(I)$ error versus the number of nonzero coefficients $\tilde{c}_{j,k}$ for the projector $\overline{Q}_2$ and the Haar rewrite rule (3.16) applied to the four images Lenna, F-16, bridge, and airport. Here, we quantize in $L^2(I)$ with $q = 2^i$, $i = 1, \cdots, 10$.

TABLE II
ESTIMATED SMOOTHNESS OF IMAGES: APPROXIMATION IN $L^1(I)$

	Lenna	F-16	Bridge	Airport
Estimated α	0.599	0.597	0.370	0.306
Estimated $\|f\|_{B_q^{\alpha}(L^q(I))}$	0.407	0.405	0.275	0.218
Correlation Coefficient	−0.999	−0.993	−0.994	−0.992

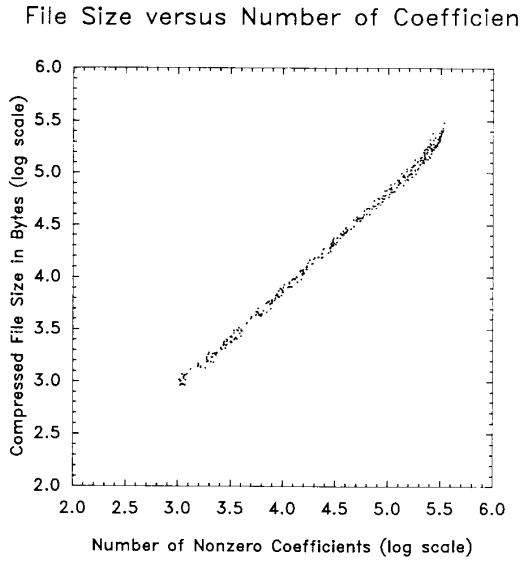


Fig. 15. Compressed file size, in bytes, versus the number of nonzero coefficients. We have compared all the data with more than 1000 coefficients except when median clipping and the Haar rewrite rule (3.16) were applied, a case for which our entropy encoder was not set up.

TABLE III
ESTIMATED SMOOTHNESS OF IMAGES: APPROXIMATION IN $L^2(I)$

	Lenna	F-16	Bridge	Airport
Estimated α	0.618	0.626	0.337	0.287
Estimated $\|f\|_{B_q^\alpha(L^q(I))}$	0.690	0.807	0.330	0.275
Correlation Coefficient	−0.999	−0.997	−0.998	−0.997

Therefore,

$$\begin{aligned}
&\|f\|_{B_q^{\alpha,1}(L^q(I))}\\
&\le C\big(\|S^m(F)\|_{B_q^{\alpha,1}(L^q(I))} + \|f - S^m(F)\|_{B_q^{\alpha,1}(L^q(I))}\big)\\
&\le C\big(\|F\|_{B_q^{\alpha,1}(L^q(I))}\\
&\quad + 2^{\alpha m}\|f - S^m(F)\|_{L^\infty(I)}\big) \qquad \text{by Lemma 1}\\
&\le C\|F\|_{B_q^{\alpha,1}(L^q(I))} + C2^{\alpha m}2^{-n},
\end{aligned}$$

because at each point x, $|S^m(F)(x) - f(x)| \le 2^{-n-1}$. □

Theorem 6 (Error Bounds): For each $0 < \alpha$ and $0 < p < \infty$ there exist constants C_1 and C_2 such that for all $f \in B_q^{\alpha,1}(L^q(I))$ with $1/q = \alpha/2 + 1/p$.

1) The number, $\mathcal{N}$, of nonzero coefficients $\tilde{c}_{j,k}$ satisfies

$$\mathcal{N} \le C_1 N \|f\|^q_{B_q^{\alpha,1}(L^q(I))}. \tag{3.12}$$

2) The error $f - \tilde{f}$ satisfies

$$\|f - \tilde{f}\|_{L^p(I)} \le C_2 N^{-\alpha/2}\|f\|^{q/p}_{B_q^{\alpha,1}(L^q(I))} \tag{3.13}$$

and

$$\|f - \tilde{f}\|_{L^p(I)} \le C_1^{\alpha/2} C_2 \mathcal{N}^{-\alpha/2}\|f\|_{B_q^{\alpha,1}(L^q(I))}. \tag{3.14}$$

Proof for $0 < p \le 1$ and $\alpha < 2/p$:

1) Because of the equivalence of the two norms $\|f\|_{\mathcal{F}}$ and $\|f\|_{B_q^{\alpha,1}(L^q(I))}$, we know

$$\sum_{k\ge 0,\, j\in\mathbb{Z}_k^2} \|c_{j,k}\phi_{j,k}\|^q_{L^p(I)} \le C_1\|f\|^q_{B_q^{\alpha,1}(L^q(I))}. \tag{6.2}$$

Because $\|c_{j,k}\phi_{j,k}\|^q_{L^p(I)}$ must be greater than $1/N$ for $\tilde{c}_{j,k}$ to be nonzero, there are no more than $C_1 N\|f\|^q_{B_q^{\alpha,1}(L^q(I))}$ nonzero coefficients $\tilde{c}_{j,k}$.

2) Let $\gamma_{j,k} := c_{j,k} - \tilde{c}_{j,k}$. By definition of $\tilde{c}_{j,k}$,

$$\|\gamma_{j,k}\phi_{j,k}\|^q_{L^p(I)} \le \frac{1}{N},$$

and either $\tilde{c}_{j,k} = 0$, in which case $\|\gamma_{j,k}\phi_{j,k}\|^q_{L^p(I)} = \|c_{j,k}\phi_{j,k}\|^q_{L^p(I)}$, or $\tilde{c}_{j,k} \ne 0$, in which case $\|c_{j,k}\phi_{j,k}\|^q_{L^p(I)} \ge 1/N \ge \|\gamma_{j,k}\phi_{j,k}\|^q_{L^p(I)}$. Therefore, by (6.2), we have

$$\sum_{k\ge 0,\, j\in\mathbb{Z}_k^2} \|\gamma_{j,k}\phi_{j,k}\|^q_{L^p(I)} \le C_1\|f\|^q_{B_q^{\alpha,1}(L^q(I))}.$$

The coefficients $\gamma_{j,k}$ can be partitioned into sets $\mathcal{I}_1, \cdots, \mathcal{I}_M$, with $M \le C_1\|f\|^q_{B_q^{\alpha,1}(L^q(I))}N$, such that

$$\sum_{\gamma_{j,k}\in\mathcal{I}_n} \|\gamma_{j,k}\phi_{j,k}\|^q_{L^p(I)} \le \frac{2}{N}, \qquad n = 1, \cdots, M. \tag{6.3}$$

This is accomplished simply by sorting $\|\gamma_{j,k}\phi_{j,k}\|^q_{L^p(I)}$ in decreasing order and adding $\gamma_{j,k}$ to $\mathcal{I}_1$ until the sum in (6.3) is greater than $1/N$. The process is repeated with the remaining coefficients added to $\mathcal{I}_n$, $n = 2, \cdots, M$. Because each term is individually less than $1/N$, (6.3) follows. Because each sum is at least $1/N$, the bound on M is immediate.

The functions $f_n := \sum_{\gamma_{j,k}\in\mathcal{I}_n} \gamma_{j,k}\phi_{j,k}$ satisfy

$$\|f_n\|_{B_q^{\alpha,1}(L^q(I))} \le C\Big(\sum_{\gamma_{j,k}\in\mathcal{I}_n} \|\gamma_{j,k}\phi_{j,k}\|^q_{L^p(I)}\Big)^{1/q} \le CN^{-1/q}. \tag{6.4}$$

So,

$$\begin{aligned}
\|f - \tilde{f}\|^p_{L^p(I)} &= \|\sum_{n=1}^M f_n\|^p_{L^p(I)}\\
&\le \sum_{n=1}^M \|f_n\|^p_{L^p(I)} \qquad \text{true for } 0 < p \le 1,\\
&\le C\sum_{n=1}^M \|f_n\|^p_{B_q^{\alpha,1}(L^q(I))}\\
&\qquad \text{because } B_q^{\alpha,1}\big(L^q(I)\big) \hookrightarrow L^p(I)\\
&\le C\sum_{n=1}^M N^{-p/q} \qquad \text{from (6.4)}\\
&\le C\|f\|^q_{B_q^{\alpha,1}(L^q(I))}N^{1-p/q}\\
&\le C\|f\|^q_{B_q^{\alpha,1}(L^q(I))}N^{-\alpha p/2} \qquad \text{by definition of } q.
\end{aligned}$$

Taking pth roots of both sides shows that

$$\|f - \tilde{f}\|_{L^p(I)} \le C\|f\|^{q/p}_{B_q^{\alpha,1}(L^q(I))}N^{-\alpha/2}.$$

Inequality (3.12) implies that $N^{-\alpha/2} \le C_1^{\alpha/2} \mathcal{N}^{-\alpha/2} \| f \|^{\alpha q/2}_{B^{\alpha,1}_q(L^q(I))}$. Substituting this into (3.13) and using the relationship $\alpha q/2 + q/p = 1$ gives (3.14). □

A more sophisticated argument in [13] shows that Theorem 6 holds for all $0 < p < \infty$ and $\alpha > 0$; in this case, however, $C_2 \to \infty$ as $p \to \infty$. This inconvenience can be gotten rid of by using a more complicated algorithm; see [13].

Theorem 7: If $m \le 2^{K-1} + \lfloor K/2 \rfloor$ then for all $0 < q \le 1$ and for all $n > 0$ there exists a constant $\bar{C}_{n,q}$ such that for all $f = \sum_j p_j \phi_{j,m}$,

$$\| \bar{\mathbb{Q}}_2 f_{j,k} - f \|_{L^q(I_{j,k})} \le \bar{C}_{n,q} \| \mathbb{Q}_q f_{j,k} - f \|_{L^q(I_{j,k})},$$

for all $0 \le k \le m$ and $j \in \mathbb{Z}^2_k$. (Here $\lfloor x \rfloor$ is the largest integer less than or equal to x.)

Proof: Recall that

$$a_{j,m} = p_j, \qquad j \in \mathbb{Z}^2_m,$$

and

$$a_{j,k-1} = \text{round}_K \left(\frac{1}{4} \left(a_{(2j_1, 2j_2), k} + a_{(2j_1+1, 2j_2), k} + a_{(2j_1, 2j_2+1), k} + a_{(2j_1+1, 2j_2+1), k} \right) \right),$$

for $j \in \mathbb{Z}^2_k$ and $k = m, \cdots, 1$.

If we can ensure that

$$| a_{j,k} - \mathbb{Q}_2 f_{j,k} | \le \frac{1}{4} \tag{6.5}$$

for all j and k, then we can use the following argument to prove the theorem. If $\bar{\mathbb{Q}}_2 f_{j,k} = \text{round}(a_{j,k})$ is equal to $\mathbb{Q}_1 f_{j,k}$, then we are done because the median is near best in $L^q(I)$ for any $0 < q < \infty$. If not, assume without loss of generality that $\bar{\mathbb{Q}}_2 f_{j,k} > \mathbb{Q}_1 f_{j,k}$, so that $\bar{\mathbb{Q}}_2 f_{j,k} \ge \mathbb{Q}_1 f_{j,k} + M$ for some positive integer M. Thus, we can assume that $a_{j,k} \ge \mathbb{Q}_1 f_{j,k} + M - \frac{1}{2}$, by the definition of round, and by (6.5) we conclude that $\mathbb{Q}_2 f_{j,k} \ge \mathbb{Q}_1 f_{j,k} + M - \frac{3}{4}$. The existence of $\bar{C}_{n,q}$ now follows in the same way as in the proof of Theorem 3.

So we wish to prove (6.5). We first note that

$$a_{j,k} - \mathbb{Q}_2 f_{j,k} = 0,$$

for $k = m - \lfloor K/2 \rfloor, \cdots, m$, because each multiplication by $\frac{1}{4}$ results in two more bits to the right of the binary point, which can be represented exactly using K bits to the right of the binary point. One can show by induction that for $k < m - \lfloor K/2 \rfloor$ we have

$$| a_{j,k} - \mathbb{Q}_2 f_{j,k} | \le 2^{-K-1} \left(m - \lfloor K/2 \rfloor - k \right),$$

because each application of round_K results in an additional error of 2^{-K-1}. The maximum of this quantity, $2^{-K-1}(m - \lfloor K/2 \rfloor)$, occurs when $k = 0$. Therefore, (6.5) holds if

$$2^{-K-1} \left(m - \lfloor K/2 \rfloor \right) \le \frac{1}{4},$$

or

$$m \le 2^{K-1} + \lfloor K/2 \rfloor,$$

as hypothesized. □

References

[1] G. Battle, "A block spin construction of ondelettes," *Commun. Math. Phys.*, vol. 110, pp. 601-615, 1987.

[2] C. de Boor, *A Practical Guide to Splines.* New York: Springer, 1978.

[3] C. de Boor and R. A. DeVore, "Approximation by smooth multivariate splines," *Trans. Amer. Math. Soc.*, vol. 276, pp. 775-788, 1983.

[4] C. de Boor and G. F. Fix, "Spline approximation by quasi-interpolants," *J. Approx. Theory*, vol. 8, pp. 19-45, 1973.

[5] C. de Boor and R.-Q. Jia, "Controlled approximation and a characterization of the local approximation order," *Proc. Amer. Math. Soc.*, vol. 95, pp. 547-553, 1985.

[6] L. Brown and B. J. Lucier, "Best approximations in L^1 are near best in L^p, $p < 1$," *Proc. Amer. Math. Soc.*, to appear.

[7] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Commun.*, vol. 31, pp. 532-540, 1983.

[8] C. K. Chui, *Multivariate Splines,* CBMS-NSF Conference Series. Providence, RI: CBMS-NSF, 1988.

[9] C. K. Chui and J. Z. Wang, "A general framework of compactly supported splines and wavelets," CAT rep. 219, 1990.

[10] W. Dahmen and C. Micchelli, "Subdivision algorithms for the generation of box-spline surfaces," *Comput. Aided Geom. Design,* vol. 1, pp. 191-215, 1984.

[11] I. Daubechies, "Orthonormal basis of compactly supported wavelets," *Commun. Pure Appl. Math.*, vol. 41, pp. 909-996, 1988.

[12] R. A. DeVore, B. Jawerth, and B. J. Lucier, "Surface compression," *Comput. Aided Geom. Design,* to appear.

[13] R. A. DeVore, B. Jawerth, and V. A. Popov, "Compression of wavelet decompositions," *Amer. J. Math,* to appear.

[14] R. A. DeVore, P. Petrushev, and X. M. Yu, "Nonlinear wavelet approximations in the space $C(\mathbb{R}^d)$," in *Proceedings of the US/USSR Conference on Approximation,* Tampa. New York: Springer-Verlag, 1991, to appear.

[15] R. A. DeVore and V. A. Popov, "Interpolation of Besov spaces," *Trans. Amer. Math. Soc.*, vol. 305, pp. 397-414, 1988.

[16] R. A. DeVore and X. M. Yu, "Nonlinear n-widths in Besov spaces," in *Approximation Theory VI: Vol. 1,* C. K. Chui, L. L. Schumaker, and J. D. Ward, Eds. New York: Academic Press, 1989, pp. 203-206.

[17] M. Frazier and B. Jawerth, "A discrete transform and decompositions of distribution spaces," *J. Functional Anal.*, vol. 93, pp. 34-170, 1990.

[18] R. F. Hess and E. R. Howell, "The threshold contrast sensitivity function in strabismic amblyopia: Evidence for a two type classification," *Vision Res.*, vol. 17, pp. 1049-1055, 1977.

[19] S. Mallat, "Multiresolution approximation and wavelet orthonormal bases of $L^2(\mathbb{R})$," *Trans. Amer. Math. Soc.*, vol. 315, pp. 69-87, 1989.

[20] Y. Meyer, "Wavelets and operators," in *Analysis at Urbana I, Proceedings of the Special Year in Modern Analysis at the University of Illinois, 1986-1987,* E. R. Berkson, N. T. Peck, and J. Uhl, eds., London Mathematical Society Lecture Notes Series, vol. 137. New York: Cambridge Univ. Press, pp. 256-365, 1989.

[21] W. B. Pennebaker, J. L. Mitchell, G. G. Langdon, Jr., and R. B. Arps, "An overview of the basic principles of the Q-Coder adaptive binary arithmetic coder," *IBM J. Res. Dev.*, vol. 32, pp. 717-727, 1988.