Image Quality Evaluation Based on Recognition Times for Fast Image Browsing Applications

Dirck Schilling and Pamela C. Cosman, *Senior Member, IEEE*

***Abstract*—Mean squared error (mse) and peak signal-to-noise-ratio (PSNR) are the most common methods for measuring the quality of compressed images, despite the fact that their inadequacies have long been recognized. Quality for compressed still images is sometimes evaluated using human observers who provide subjective ratings of the images. Both SNR and subjective quality judgments, however, may be inappropriate for evaluating progressive compression methods which are to be used for fast browsing applications. In this paper, we present a novel experimental and statistical framework for comparing progressive coders. The comparisons use response time studies in which human observers view a series of progressive transmissions, and respond to questions about the images as they become recognizable. We describe the framework and use it to compare several well-known algorithms [JPEG, set partitioning in hierarchical trees (SPIHT), and embedded zerotree wavelet (EZW)], and to show that a multiresolution decoding is recognized faster than a single large-scale decoding. Our experiments also show that, for the particular algorithms used, at the same PSNR, global blurriness slows down recognition more than do localized "splotch" artifacts.**

***Index Terms*—Human image recognition, image quality evaluation, multiresolution coding, progressive image coding, wavelet zerotree coding.**

I. Introduction

THE number of images available on the World Wide Web (WWW) continues to grow, and users are often frustrated by the length of time required to download an image. Fast browsing of image databases is of increasing importance in a number of application areas, including stock photo agencies, geographical information systems, medical databases, law enforcement, and real estate. Often, the image obtained is not the one the user wanted. If the image arriving is recognized as being of no interest, the user can save time by aborting the transmission and jumping to the next item. It is important that the user be able to identify the contents of an image early in its transmission.

With many compression algorithms, the entire compressed bit stream must arrive and be decoded before the decompressed image can be displayed to the viewer. With a progressive image compression algorithm, however, the encoder transmits the bits in an order which allows the decoder to reconstruct the image with increasing quality as more bits arrive. In the past, this progressivity was a property that one paid for dearly, as the total encoded bit stream would require a substantially larger number of bits in order to allow initial portions of the stream to be decodable. Such is the case, for example, with the progressive and hierarchical modes of the JPEG standard (the hierarchical mode can be used to provide progressivity) [1]. Progressive compression algorithms enjoyed a renaissance with the advent of the wavelet zerotree coders (embedded zerotree wavelet (EZW) coding due to Shapiro [2] and set partitioning in hierarchical trees (SPIHT) due to Said and Pearlman [3]) in which the progressivity came with little penalty in the overall distortion-rate performance. Fig. 1 shows an example of the progressively improving image quality provided by SPIHT as the bit stream is decoded.

With many different progressive compression algorithms from which to choose, application designers are in need of appropriate methods for evaluating the comparative performance of various coders. The use of peak signal-to-noise (PSNR) as a performance criterion is problematic. In many cases, it fails to accurately reflect the subtleties of human perception. In addition, for several types of algorithms, including those with spatially scalable decoders, PSNR might not even be computable. Finally, there are applications for which it is not the perceived quality of the decoded image that is of primary importance, but rather the basic recognizability of objects in the image.

A number of methods have been used to evaluate the perceptual distortion caused by lossy compression. One class of methods employs models of human psychovisual response developed by testing specific visual effects [4]–[7]. These models can explain a number of effects such as contrast and orientation masking, but are not yet general enough to predict human understanding of complex real-world images. Other methods rely on subjective opinions, where subjects are asked, for example, which of two images looks better, or whether the primary object in the image has been recognized [8]–[11]. In this paper, we directly assess image recognition by having observers respond to questions whose answer could only be known by recognizing the image content. Although not dealing with progressive compression, a few previous studies have been close in spirit to the work described in this paper; they compare compression algorithms by an objective recognition task in a reasonably realistic simulation of image use. In [12]–[15], still-image compression algorithms are evaluated for diagnostic utility by simulating

Manuscript received June 14, 2000; revised August 22, 2001. This work was supported by NSF Grants MIP-9 617 366 and MIP-9 624 729 (CAREER), by the Center for Wireless Communications at UCSD, and by the CoRe Program of the State of California. The associate editor coordinating the review of this paper and approving it for publication was Dr. M. Reha Civanlar.

D. Schilling was with the Department of Electrical and Computer Engineering, University of California at San Diego, La Jolla, CA 92093-0407 USA. He is now with ViaSat, Carlsbad, CA 92009 USA (e-mail: dirck.schilling@viasat.com).

P. C. Cosman is with the Department of Electrical and Computer Engineering, University of California at San Diego, La Jolla, CA 92093-0407 USA (e-mail: pcosman@code.ucsd.edu; http://www.code.ucsd.edu/cosman/).

Digital Object Identifier 10.1109/TMM.2002.802844.

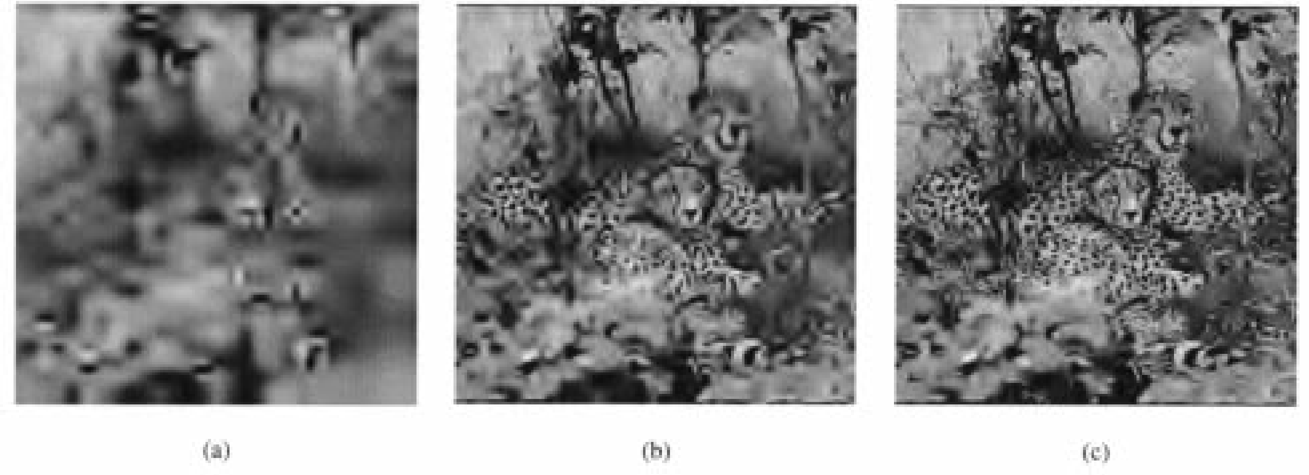


Fig. 1. Example image progression for SPIHT: (a) 0.01 bpp; (b) 0.05 bpp; and (c) 0.10 bpp.

their clinical use by radiologists. In [16], compressed video clips of American Sign Language were compared by deaf subjects for intelligibility.

In our study of recognition times for progressive compression algorithms, we analyze the correctness of the answers as well as the response times. In the first evaluation experiment, comparing JPEG, EZW, and SPIHT, we show that this approach can provide a reliable comparison between progressive algorithms [17]. We then apply our evaluation methodology in a second experiment to demonstrate that images are recognized faster when displayed by a multiresolution decoder than by a decoder that presents images at a single, full-size resolution [18]. In a third experiment, we compare SPIHT with the packetized zerotree wavelet (PZW) coder [19], [20] under lossy channel conditions, and show for these algorithms that, at a given PSNR, global blurriness slows recognition more than do localized blurring artifacts.

Two factors contribute to the performance of a compression algorithm as measured by our experiments: the efficiency with which the algorithm compresses a given item of information, and the psychophysical advantage or disadvantage conferred by displaying that information in a given form. An algorithm focusing on the first factor seeks to present the same visual progression as its competitor, but at a lower bit rate. An algorithm focusing on the second factor might draw upon studies of the human visual system [21]–[23] to prioritize certain spatial frequencies over others, in an effort to provide a more recognizable image at a given bit rate. Our experiments measure the overall comparative performance of algorithms, but do not attempt to identify the contribution of specific psychophysical effects involved in recognition.

This paper is organized as follows. In Section II, we discuss the experimental setup and statistical analysis for these response time studies. In Section III we present the results of comparing JPEG, EZW, and SPIHT (Experiments 1a and 1b). In Section IV, we describe a spatially scalable version of the SPIHT coder, as well as our experimental evaluation of its usefulness for fast recognition (Experiments 2a and 2b). Section V discusses the evaluation of algorithms under lossy channel conditions (Experiment 3), and we present our conclusions in Section VI.

II. Evaluation Framework

This section describes our experimental and statistical framework for simulating fast browsing tasks and comparing any two progressive compression algorithms.

A collection of images is selected for which a multiple choice question can be asked. We have primarily used questions with binary answers, for example, “Do you see males or females in the image?” We also used some artificial images showing a lowercase letter set against a textured background and asked a multiple-choice question: “What letter do you see in the image?” The images are chosen such that the question can be reliably answered when the image is shown at full quality. Several such image collections, each with its own associated question, are combined into an experiment collection. Each image is compressed both by algorithm A and by algorithm B. The method for displaying the images to observers varies slightly depending on the specific experiment. In the method used for our first evaluation, each observer views every image, half in each of two viewing sessions. For each observer, one compressed version of each image is randomly assigned to the first viewing session, and the other version is assigned to the second viewing session. Thus, no observer sees the same image twice on the same day. The images within a given session are presented in a different random order to each observer. The two sessions are seen one week (or more) apart to minimize inter-session learning effects. In the method used for our second and third evaluations, an observer participates in a single viewing session. Each observer sees a given image only once. The images compressed by each algorithm are randomly assigned to observers, under certain restrictions, such that both algorithms are viewed an equal number of times in the aggregate of all observers.

In our experiments, the observers were untrained persons over age 18 drawn from the general university population. They signed informed consent forms, and were paid for their participation. The only requirement was that they have normal or corrected-to-normal vision.

The images selected for our experiments varied in complexity, quality, composition and size, and placement of the object or feature to be recognized. The image sizes varied for some of the experiments, as described in each experiment's discussion. The variation in image size and content is intended to represent to a reasonable degree the variation to be expected in the fast browsing applications addressed by our methodology. All images used in our experiments are 8-bit greyscale. While we expect the evaluation methodology to be applicable to color images as well, several of the compression algorithms tested were available only in greyscale versions, and for this reason color was excluded. Color can provide important cues for recognition, and would be useful to include in future recognition studies.

All of our experiments were carried out at the same workstation under indirect fluorescent lighting typical of an office environment. Observers were allowed to position themselves comfortably with respect to the viewing monitor; the typical viewing distance was about 20 in. Observers in real-life applications employ a variety of conscious and unconscious strategies for image recognition, and it was our intent to create as natural a simulation of a fast-browsing application as possible.

For each image to be viewed, the corresponding question is first displayed on the screen. After reading it, the observer hits a key to begin the progressive display. While watching the progressive display, as soon as the observer is reasonably confident that she can correctly answer the question, she hits a key to halt the progression. The image disappears from the screen, and she enters the answer and goes on to the next image. The time and bit rate required for each response are recorded, as well as whether the correct answer was given. Observers are instructed not to rush, but to answer the question as soon as they are reasonably sure of the answer. We can expect an overall shift of the response times to larger or smaller values depending on how this issue of "reasonably sure" is expressed to the observer. For a medical diagnostic task, the observer could be told that correctness is of the utmost importance, and the observers would tend to wait farther into the progression to answer. In some other application, some incorrect decisions may matter little, and the responses would be faster. However the question is worded, consistency throughout the experiment should ensure a fair comparison of algorithms within a particular bit rate regime.

The progressive transmissions are simulated by displaying image frames at selected bit rates in sequence. For each image, the elapsed time when a given bit rate (in bits per pixel) is displayed for algorithm A is the same as that for algorithm B. For Experiment 1a, comparing JPEG with SPIHT, frames were spaced 0.02 bits per pixel (bpp) apart in bit rate and displayed at a rate of 1.33 frames/s. As we will discuss later in this paper, this relatively slow speed helped to ensure that image recognition time, rather than underlying human reaction time, was being measured [24]. Since both the time and the bit rate spacing of frames were constant, an analogy could be made with the transmission of image data over a fixed-rate channel. Fifty frames were pre-stored for each image, so that the progression could continue out to 1 bpp, ensuring that observers would eventually be able to answer the question with confidence. However, the vast majority of responses were found to occur near the beginning of the progression. Accordingly, for Experiment 1b, in which EZW was compared with SPIHT, the bit rates selected for each frame were spaced evenly on a logarithmic scale. Within the constraints of the total memory usage, this provided greater resolution in bit rate at the very low bit rates, and coarser resolution at the higher rates where fewer response occurred, while still allowing the progression to continue to a sufficiently high final quality if needed. The display rate for this and all later experiments was 2 frames/s.

A. Statistical Analysis

The algorithms are compared both on the basis of the bit rate at which observers answer the posed question for each algorithm and on the frequency of error in the answer. We describe here the statistical methods used in our first evaluation. The same general approach was used in the later evaluations, with some differences which are discussed in the corresponding sections. Denote the bit rate at which observer i answers a question for image j compressed by algorithm A as r_{Aij} and the corresponding bit rate for algorithm B as r_{Bij}, and let $i = 1, \ldots, I$ and $j = 1, \ldots, J$ be indexes for observers and images. For comparing algorithms A and B, we would like to know mean values r_A and r_B, and whether any difference in these values should be deemed statistically significant. Examination of probability plots of the data showed that they approximate the lognormal distribution, which suggests that r_A and r_B be geometric means and that normal theory statistical methods (such as ANOVA) be used to analyze the log-transformed values $\log(r_{Aij})$ and $\log(r_{Bij})$. Three analyses are carried out: one uses the bit rates from algorithm A, one uses the bit rates from algorithm B, and one uses their ratios, i.e., $s_{ij} = r_{Aij}/r_{Bij}$.

These analyses are carried out by fitting the data to the mixed effects linear model [25]

$$Y = X\alpha + Z\beta + \epsilon.$$

In this model, Y is an $N \times 1$ vector containing either $\log(r_{Aij}), \log(r_{Bij})$, or $\log(s_{ij})$, where $N = IJ$ is the total number of observations. X is the design matrix for fixed effects and in this case is just an $N \times 1$ vector of ones with $\alpha(1 \times 1)$ being the mean of Y. Note that $\exp(\alpha)$ is the geometric mean of the bit rate or ratio. Z is the design matrix for random effects and can be partitioned as $Z = (Z_1 \,|\, Z_2)$, where $Z_1(N \times I)$ contains a one in column i for each row involving observer i and zeroes elsewhere, and where $Z_2(N \times J)$ contains a one in column j for every row involving image j and zeroes elsewhere. The $IJ \times 1$ vector of random effects can be partitioned as $\beta = (\beta_1' \,|\, \beta_2')'$ and it is assumed that the random effects are independently distributed as $\beta_{1i} \sim N(0, \sigma^2_{\text{obs}})$ and $\beta_{2j} \sim N(0, \sigma^2_{\text{img}})$. Finally, $\epsilon(N \times 1)$ is the residual vector for which it is assumed that $\epsilon_{ij} \sim N(0, \sigma^2_{\text{res}})$.

Using restricted maximum likelihood (REML), estimates are obtained for the coefficient α and for the variance components $\sigma^2_{\text{obs}}, \sigma^2_{\text{img}}$ and σ^2_{res}. An estimate of the standard error of $\hat{\alpha}, \text{se}_{\hat{\alpha}}$ is obtained from the variance components and this can be used to perform significance tests or to form the 95% confidence interval for α, i.e., $\hat{\alpha} \pm 1.96\ \text{se}_{\hat{\alpha}}$, and for its antilog $\exp(\hat{\alpha} \pm 1.96\ \text{se}_{\hat{\alpha}})$. Model fitting is performed using the *Splus* function `varcomp` [26].

The geometric mean bit rates for responses r_A and r_B give an indication of what compressed bit rates tend to be of interest for recognition responses. The geometric mean of the ratio $s_{ij} = r_{Aij}/r_{Bij}$ can summarize the comparison. If the 95% confidence interval for s_{ij} includes the value one, then neither algorithm can be said to be significantly better than the other by this experiment.

1) Analysis of Observer Mistakes: Observer responses were also examined for correctness. We wish to examine the possibility that more errors occur with one algorithm than with another. It would be possible, in theory, that one algorithm might lead people to make rapid yet incorrect decisions. We use the paired data in which, for a given reader and image, the correctness result for algorithm A is paired with that for algorithm B. For each image in the pair, the observer is either correct or not. There are thus four types of pairs:

1) those with both members correct;
2) those with algorithm A correct and B not;
3) those with algorithm B correct and A not;
4) those with neither one correct.

In the McNemar analysis [27], we concern ourselves with two of the four types: those pairs in which the members differ. If answers are equally likely to be correct whether an image was seen with algorithm A or B, then conditional on the numbers of the other two types, these would have a binomial distribution with parameter $1/2$. For example, one observer saw 118 pairs of images. Of these, both images in the pair were recognized correctly 110 times; for three pairs both versions were recognized incorrectly. Of the remaining five pairs, four times the EZW image was recognized correctly while SPIHT was not, and one time the SPIHT image was recognized correctly while the EZW image was not. The probability that a fair coin flipped five times will produce a heads/tails split at least as great as 4:1 is 0.375, thus this result is not significant.

2) Analysis of Learning Effects: In Experiments 1a and 1b, images were seen twice, once per session. It is thus possible that an observer could remember what was seen in the first session and use this information to answer more quickly, or to answer more correctly, in the second session.

The issue of answering more correctly was addressed by incorporating into the mixed-effects linear model a fixed effect for which algorithm was seen first. The fitted model's coefficient for this session effect provides insight into the session effect's impact. For Experiment 1a (JPEG versus SPIHT), the effect of which algorithm was seen first was not significant. For Experiment 1b (EZW versus SPIHT) the learning effect was small but statistically significant, indicating that observers responded slightly faster in the second session. However, the magnitude of the effect was the same for the two algorithms, and each algorithm was seen first on half of the images, so learning effects did not favor one algorithm over the other.

The issue of answering more quickly was addressed by a McNemar analysis in which the correctness result for Session 1 was paired with the correctness for the same image in Session 2. The comparison was also done broken down by algorithm type. For example, the McNemar analysis was performed for the image pairs seen first by EZW and second by SPIHT, and it was also performed for the image pairs seen first by SPIHT and second by EZW. The reason for examining the data separately by algorithm is that it is possible that seeing a SPIHT image first conveys an advantage in a subsequent viewing of the image compressed by EZW, but that the reverse is not true. The reason for examining the data also in aggregate is that a subtle effect may be found in a larger data set. In no case was a statistically significant difference found. Therefore, we conclude that observers were not making more correct answers on the second session.

In Experiments 2a, 2b, and 3, the observers saw a given image only once, so these learning effects based on image content are not an issue.

III. Experiments 1A and 1B: Comparing JPEG, SPIHT, and EZW

First, we compared JPEG with SPIHT (Experiment 1a); next we compared EZW with SPIHT (Experiment 1b). Although the JPEG standard includes a progressive mode for JPEG [1], we did not use this, but rather created a sequence of frames at progressively higher bit rates using baseline sequential JPEG. JPEG progressive mode uses more bits than baseline sequential does to achieve a given level of precision on the transform coefficients, therefore using a sequence of baseline sequential JPEG frames to simulate a progressive JPEG display will give a somewhat optimistic estimate of the actual recognition times. The demonstrated superiority of the SPIHT algorithm is therefore a conservative conclusion. The following experiment parameters were used:

1) Image collections for two different questions were included:
   - "Do you see males or females in the image?" These images contained one or more clearly visible persons of various ages and races involved in a variety of activities. All persons in the image were of the same sex.
   - "Do you see a single animal or multiple animals in the image?" These images contained a wide range of animals in a variety of natural settings, e.g., forest, field, underwater.
2) Each session consisted of 118 images (59 corresponding to each of the two questions), which required approximately 45 min per session.
3) Images were displayed on a SGI O_2 workstation with a 20″ monitor, in a single window against a black background.
4) All images were 256-level greyscale. Their sizes ranged from 160×160 to 640×640 pixels, averaging 420×420.
5) Images were presented in groups of 12 in a row for a given question, but were randomly mixed within each group from session to session.
6) For the comparison of JPEG with SPIHT, there were five observers, a small but adequate number, given the large difference in recognition bit rates for these algorithms. For the comparison of EZW with SPIHT, there were 20 observers, because we expected the difference in recognition bit rates between these algorithms to be small.

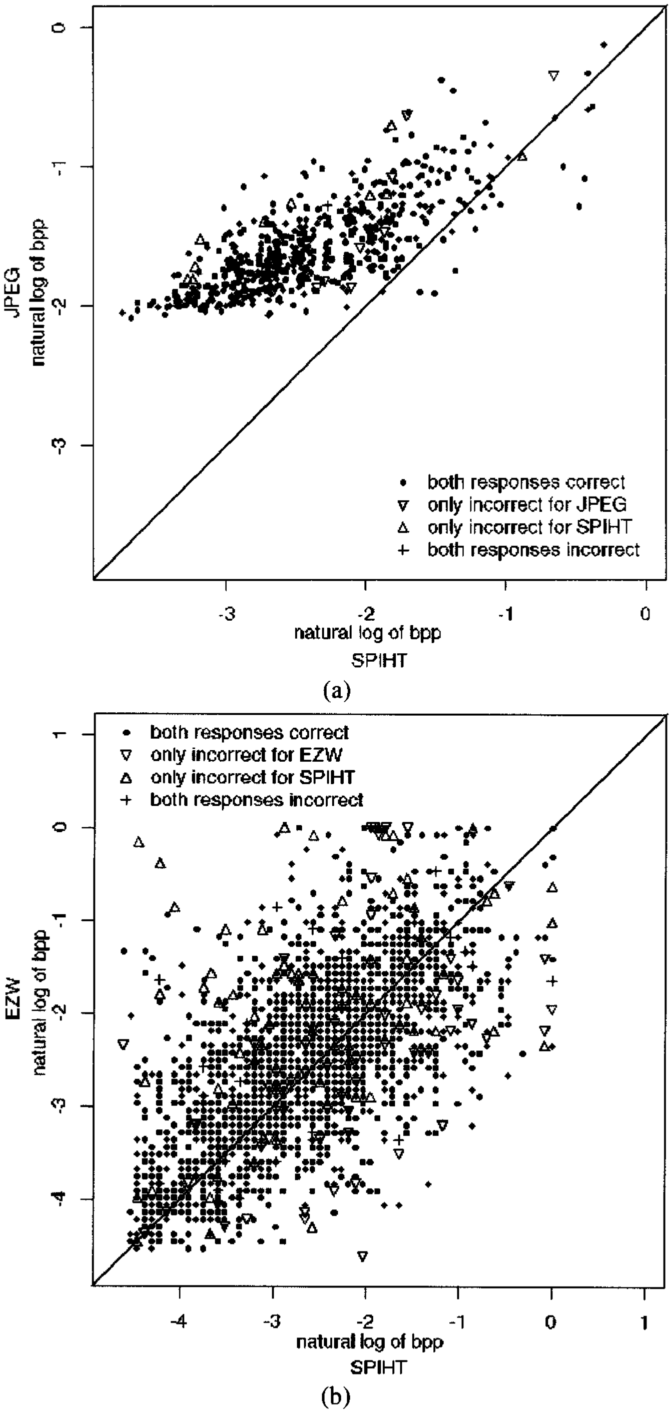


Fig. 2. (a) Scatter plot of JPEG versus SPIHT data. (b) Scatter plot of EZW versus SPIHT data. The plots show the log of the bit rate at which recognition occurred. Visual inspection reveals that SPIHT clearly outperforms JPEG, whereas SPIHT and EZW are more evenly matched.

Fig. 2(a) shows the log of bit rates for the JPEG-SPIHT comparison. Baseline and progressive mode JPEG must transmit a minimum number of bits (a dc value for each block) before anything at all can be displayed, which results in the visible skew of the data toward the left of the diagonal at low bit rates. In Fig. 2(b), the data for the EZW versus SPIHT comparison are shown. Visual examination of this plot indicates that these two algorithms are more evenly matched.

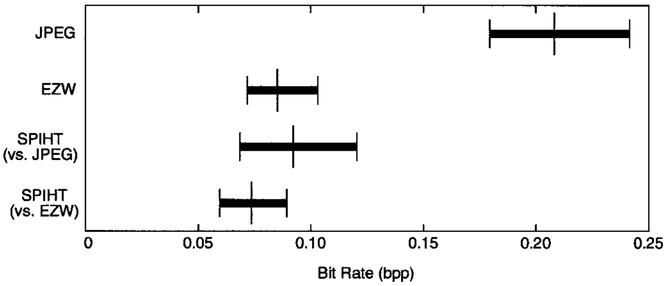


Fig. 3. Mean bit rates for recognition, and their 95% confidence intervals.

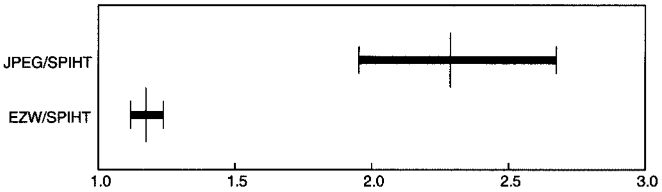


Fig. 4. Mean bit rate ratios and their 95% confidence intervals.

When examined quantitatively, SPIHT was found to lead to faster image recognition than either JPEG or EZW. In Fig. 3, the mean bit rates for recognition and their confidence intervals are shown for each of the algorithms. These values indicate for each algorithm the approximate range at which recognition occurred sufficient to allow the posed questions to be answered. These bit rates can be helpful to designers of new progressive algorithms for fast recognition. When such an algorithm encounters image features judged to be important for recognition, it should attempt to concentrate information about them below these bit rates. The bit rate at which observers answer depends not only on the complexity of the images and of the observation task, but also on the parameters of the experiment. We note in Fig. 3 that the same SPIHT-compressed images required more bits on the average in Experiment 1a, where they were compared against JPEG, than in the Experiment 1b, where they were compared against EZW. Recall that in Experiment 1a, the increase in bit rate from frame-to-frame was proportional to time, whereas in Experiment 1b the bit rate increased logarithmically with time. The difference in SPIHT response rates between the two experiments may be explained by the fact that, in the lower bit rate ranges, the logarithmic spacing allowed observers more time to respond.

Fig. 4 shows the mean recognition bit rate ratios and their confidence intervals. These values can summarize the comparison. In each experiment, SPIHT was found to perform better, in terms of observer recognition, than the algorithm with which it was compared.

Next, the influence of observer errors is examined. The smallest number of incorrect answers given by an observer in a session was zero; the largest number was 16, and the mean value was 4.84 (out of 118 images). Analyzing the paired data with the McNemar statistic showed no difference in correctness at the 5% significance level between algorithms for any of the 25 observers individually, or for the observers in each experiment pooled together. In Fig. 5, only the bit rates for erroneous responses in the EZW-SPIHT comparison are

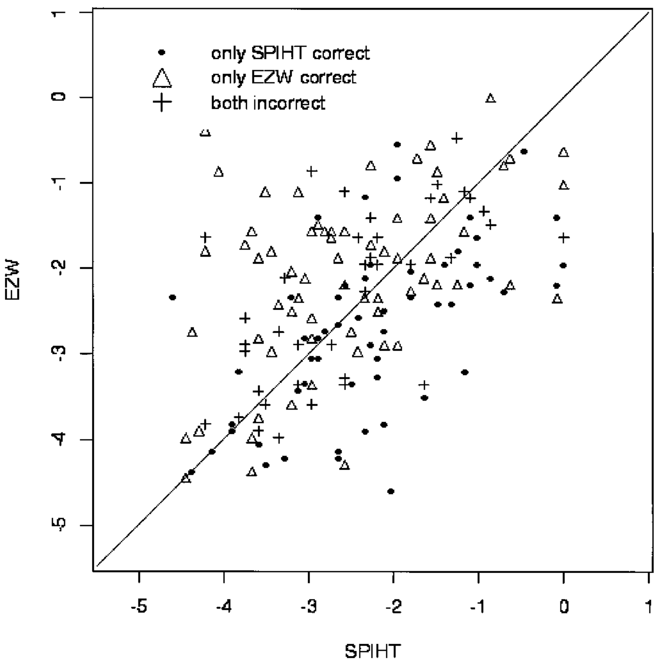


Fig. 5. Bit rates of erroneous responses only, for EZW versus SPIHT. Errors appear evenly distributed between the algorithms.

plotted. The symmetry of these errors supports our conclusion from the McNemar analysis that errors did not significantly influence our results.

IV. Multiresolution Coders

Bandwidth limitations often lead to inconvenient delays while accessing images on the Internet. As a result, thumbnail images have gained wide acceptance as a means of providing viewers with a rapidly available initial preview of a large image [28]. The advantages of thumbnails and of progressive coding can be combined in a spatially scalable progressive algorithm, such that versions of the image at successively increasing scales can be extracted from the bit stream as more bits arrive. That is, when b_1 bits have been received, the decoder can reconstruct an image of a small size, and when a larger number b_2 of bits have arrived, the decoder can reconstruct an image that is either of larger size, or of higher quality at the same size, or perhaps of both higher quality and increased size. In this way, no information need be sent or stored twice. Note that, by this definition, any progressive algorithm can be made spatially scalable simply by downsampling the output image to the desired scale. That is, the b_1 and b_2 bits might both allow reconstruction at a large size, but the b_1 image could simply be downsampled and shown at smaller scale. SPIHT and other zerotree coders based on wavelet decompositions would not even require a separate downsampling step, as the decoder could simply stop the wavelet inverse transform at some level before the final one, and the resulting low-frequency band is essentially a coarse-scale version of the original image.

However, spatial scalability is usually taken to mean that information about detail scales is not transmitted initially. In zerotree wavelet coders such as SPIHT and EZW, information on some coefficients in higher frequency bands is sent before all

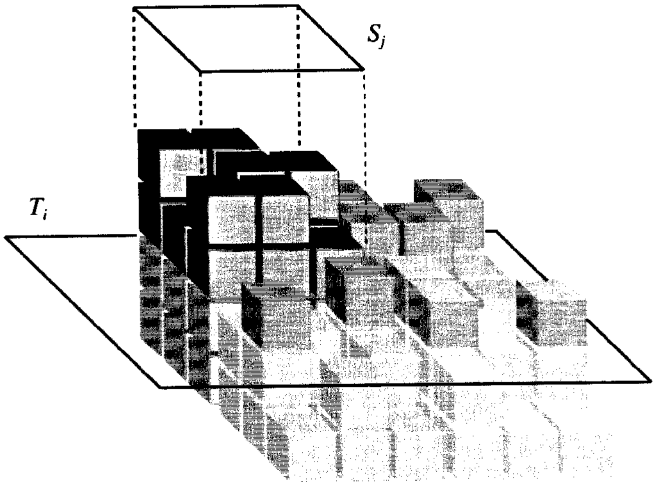


Fig. 6. Coefficient bitplanes. SPIHT describes all coefficents with magnitudes exceeding threshold T_i. MSPIHT describes only coefficients above T_i and within scale boundary S_j, deferring remaining coefficients until later.

coefficients in the lowest frequency band have been encoded. Therefore, according to the more stringent view of spatial scalability, the conventional zerotree coders are not scalable, and even with the less stringent view, these higher frequency coefficients are not used in reconstructing the coarse-scale thumbnail, and therefore represent wasted bits (added cost) when decoding to the coarse-scale version of the image.

In addition to the basic advantage that spatial scalability can lead to bandwidth savings, one might also ask whether an advantage in recognition performance can be gained by displaying images at successive scales. That is, can objects in a small, clear thumbnail image be recognized more readily than in the larger, blurrier full-scale version costing the same number of bits? If so, this would lend an embedded, spatially scalable image coder an additional advantage over traditional full-scale coders for progressive image transmission.

By reordering the transmitted bit stream, the SPIHT algorithm can be made spatially scalable [18], [29]. Compared against SPIHT without arithmetic coding, the spatially scalable SPIHT has no loss in performance (PSNR versus bit rate at the final full size) and retains some progressivity. We refer to this multiscale SPIHT algorithm as MSPIHT. We show that viewers are able to recognize MSPIHT-compressed images substantially earlier than images compressed by SPIHT.

A. Multiscale SPIHT (MSPIHT)

We now describe the mechanics of MSPIHT. Wavelet subbands are each associated with a representation of the image at a given scale. We define a $1/n$-scale image as one where both dimensions are $1/n$ the original dimensions. With a single-level decomposition, the encoder could efficiently describe a $1/2$-scale image to the decoder by transmitting information only about coefficients in the LL band. The remaining bands contain information about frequencies visible in the full-scale image. The SPIHT bit stream has coefficients ordered primarily by magnitude, so some coefficients associated with a fine scale may be transmitted before all coefficients from coarser scales have been described (see Fig. 6). In MSPIHT, information about any

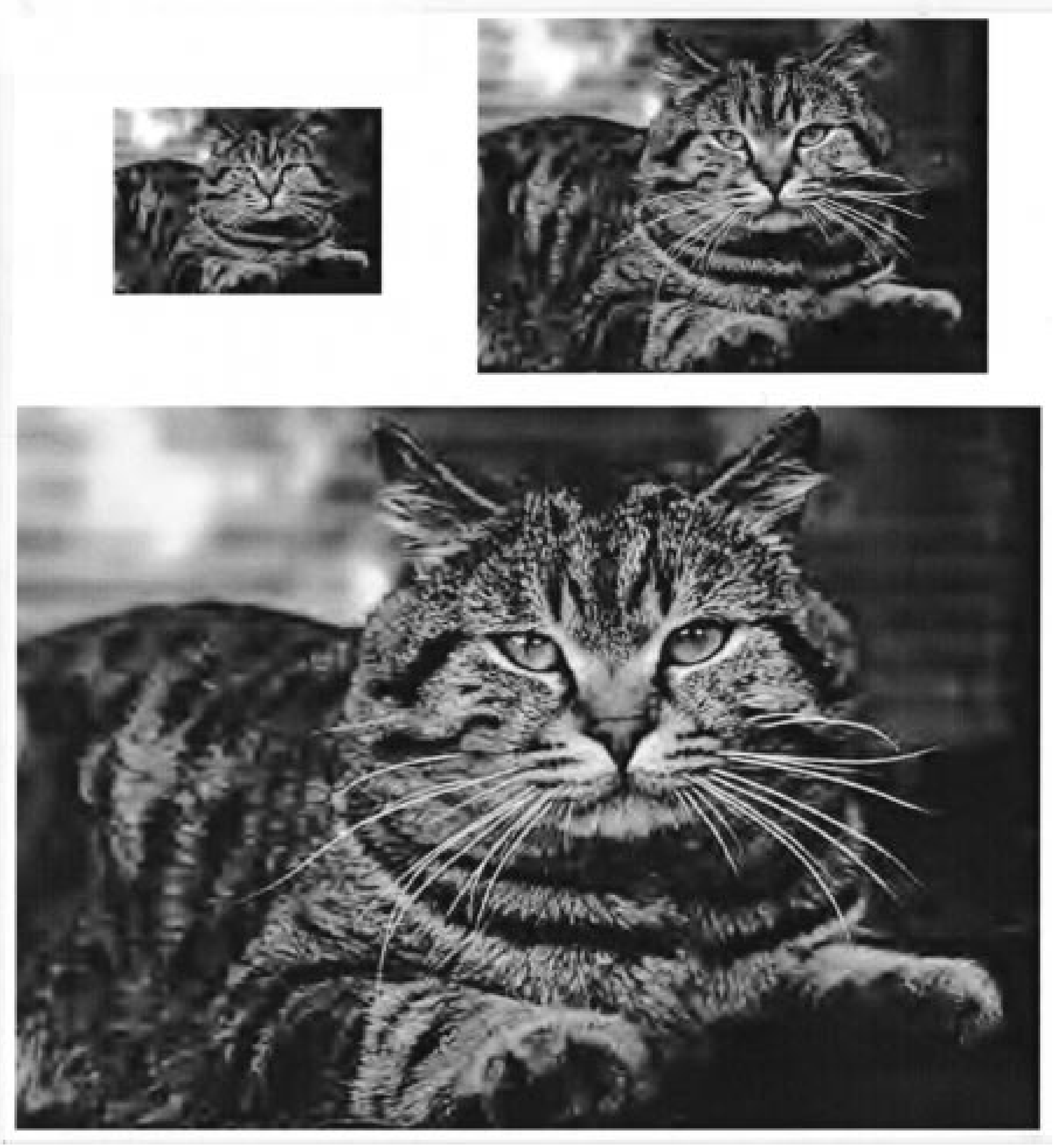


Fig. 7. MSPIHT-compressed image at 0.02, 0.09, and 0.30 bpp.

such finer scale coefficients is deferred until after all coefficients for the coarser scales have been described.

A scale schedule specifies the bit rates at which jumps to the next larger scale occur. Since both encoder and decoder know the schedule, no extra bits are required to manage the scale jumps. (If the schedule were unknown to the decoder, it could be transmitted with a negligible few bytes). For example, the schedule might specify the initial scale as $1/4$. The jump to $1/2$-scale might be scheduled to occur at 0.04 bpp, and the jump to full scale at 0.1 bpp. An example of an MSPIHT progressive display is shown in Fig. 7. Following this scale schedule, MSPIHT begins by performing sorting and refinement passes in the same manner as SPIHT, comparing each coefficient with a significance threshold. However, when a coefficient is examined from a scale larger than $1/4$ (that is, from any of the outer six subbands), it is declared out-of-scale and placed in a deferred list. No bits are transmitted about it at this time, and processing continues as before. When the bit rate reaches 0.04 bpp (jump to $1/2$-scale), the coefficients accumulated in the deferred list are reexamined; those that are now in-scale are removed from the deferred list, and sorted and refined until their significance threshold catches up with the current significance threshold for the nondeferred coefficients. At this point, processing resumes where it left off when the scale jump occurred. These steps repeat for each scale jump, until the desired final bit rate is reached.

Note that at any given point in the progression, no bits are spent to describe coefficients from scales finer than the current

TABLE I
SCALE SCHEDULES USED FOR TESTING MSPIHT

	MSPIHT-A	MSPIHT-B	MSPIHT-C
Starting scale	1/4 scale	1/2 scale	1/4 scale
Jump to 1/2 scale	0.04 bpp	0 bpp	skipped
Jump to full scale	0.10 bpp	0.10 bpp	0.06 bpp

TABLE II
ARITHMETIC MEAN OF RECOGNITION BIT RATES FOR EACH ALGORITHM, IN bpp. BEST PERFORMANCE FOR EACH IMAGE TYPE IS SHADED

	MSPIHT-A	MSPIHT-B	MSPIHT-C	SPIHT
All images	0.0671	0.0751	0.0628	0.0852
Animals/Vehicles	0.0603	0.0607	0.0530	0.0735
Letters	0.0740	0.0896	0.0724	0.0969

one. When the full scale is reached and the coefficients on the deferred list are processed, the distortion and bit rate at that point are precisely the same as for regular SPIHT without arithmetic coding.

B. Experiments 2a and 2b: Comparison of MSPIHT and SPIHT

We first wished to compare SPIHT with MSPIHT, and to determine a scale schedule for MSPIHT which performed well. For Experiment 2a, three MSPIHT scale schedules (A, B, C) were prepared (see Table I). A series of 120 images were displayed progressively to each of 20 observers. Two recognition tasks were included in the experiment. In the first, the observer was asked, "Do you see animals or vehicles in the image?" These images contained a wide range of animals and vehicles in various settings, e.g., forests, underwater, and urban surroundings. The task was intended to represent natural image recognition tasks, particularly those answerable in the lower bit rate ranges. The image widths ranged from 320 to 699 pixels, averaging 510, and the heights ranged from 250 to 576, averaging 391. In the second task, each image contained a single lower case letter in a common font, partially concealed in a variety of noisy and smooth artificial backgrounds. These images were all 512 × 512 pixels. The letters themselves were in three sizes. The observer was asked to identify the letter. This simplified stimulus set was intended to limit the recognition cues available to the observer, and allow comparison of recognition bit rates for stimuli of different sizes.

Response bit rates averaged over all observers are presented in Table II. Averages were computed for each algorithm over the sets o: 1) all images; 2) animal/vehicle images; and 3) letter images. In all cases, SPIHT averaged the slowest recognition (highest bit rates). For the animal/vehicle set, MSPIHT-C yielded an average recognition bit rate 27.9% lower than SPIHT. For the letter set, MSPIHT-C yielded an average recognition bit rate 25.3% lower than SPIHT. For both sets together, MSPIHT-C performed 26.3% better than SPIHT.

This experiment indicates that MSPIHT can allow earlier recognition than SPIHT for several types of images. We now focus on the potential causes for this improvement. Did observers recognize objects earlier using MSPIHT because it defers visually unusable fine-scale information until later, allowing more precise coarse-scale information to be transmitted first? Or is it instead because, with a small image that can be mostly or entirely viewed in the foveal field [30], the observer's eyes do not have to jump around as much in order to scan the image? If a combination of both effects was responsible, which effect predominated? Experiment 2b was performed to investigate these questions. The image sequences processed by SPIHT were downsampled by block averaging to match the image sizes produced by the MSPIHT-C scale schedule. Since the transmitted bit stream for these downsampled SPIHT images was not reordered to defer high frequency information, any advantage the images might yield in recognition bit rate was likely to be due primarily to psychophysical effects related to the size of the objects displayed.

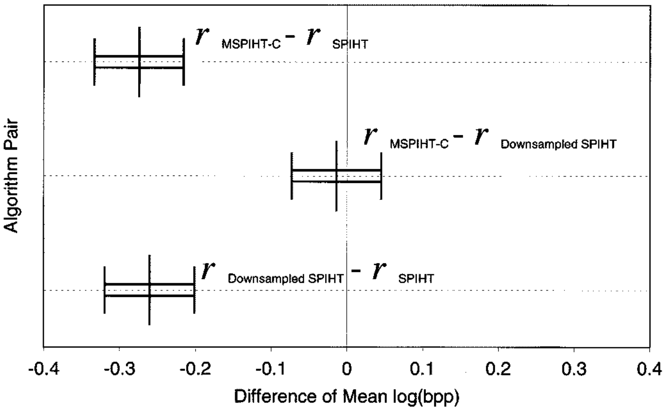


Fig. 8. Difference of mean log bit rate for each pair of algorithms.

To compare SPIHT, MSPIHT-C, and downsampled SPIHT, the same 120 images were displayed progressively to 21 new observers. As shown in Fig. 8, both MSPIHT-C and downsampled SPIHT outperformed SPIHT with 5% statistical significance in terms of mean response bit rates. The difference in mean bit rates of MSPIHT-C and downsampled SPIHT, however, was not significant.

In analyzing observer mistakes, two questions were of interest: whether incorrect responses could have influenced the overall performance conclusion, and whether any algorithms led observers to make more incorrect responses than the others. To answer the first question, the difference of means test was repeated after removing from consideration all images for which any observer had provided an incorrect response (52 of the 120). This shifted the difference-of-means statistics slightly for each algorithm pair, but did not alter the overall conclusions as to relative performance of the algorithms. The error rate for SPIHT was 4.5%; it was 6.8% for MSPIHT-C, and 8.5% for downsampled SPIHT. A two-tailed Wilcoxon signed rank test on paired error counts revealed that both MSPIHT-C and downsampled

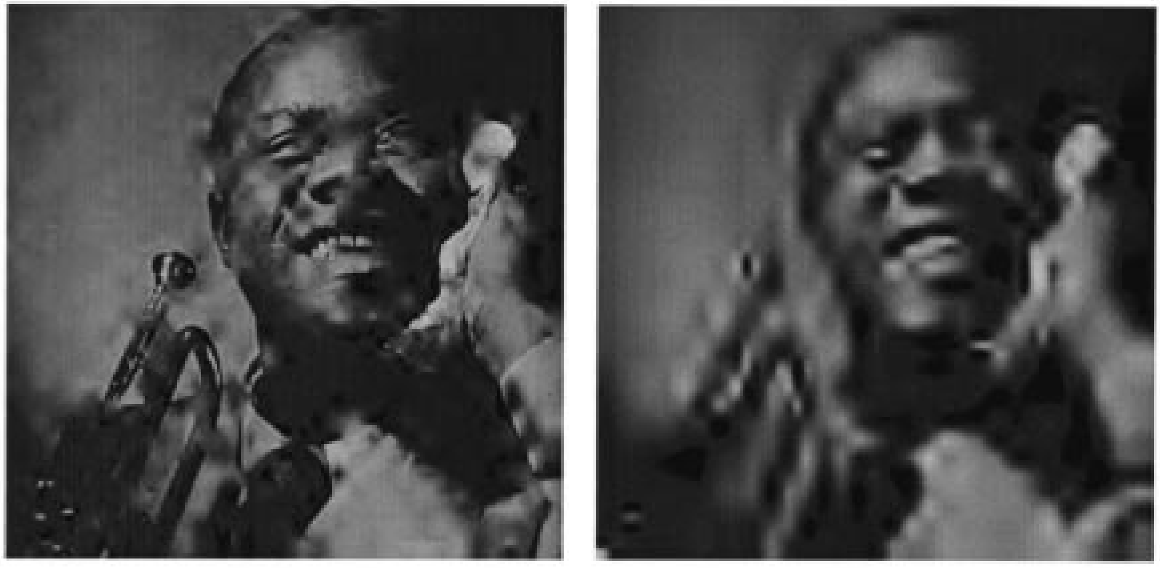


Fig. 9. Example images of PZW (left) and SPIHT (right) at $\mathrm{PSNR} = 23.97$.

SPIHT yielded significantly more errors than SPIHT, but the difference in error counts between MSPIHT-C and downsampled SPIHT was not significant. Finally, by including a fixed effect for response correctness in the difference-of-means analysis described above, it was seen that both MSPIHT-C and downsampled SPIHT remained significantly faster than SPIHT in terms of recognition performance, even when their greater error rates were taken into account.

V. Blurriness Versus Splotchiness: PSNR and Recognition Times

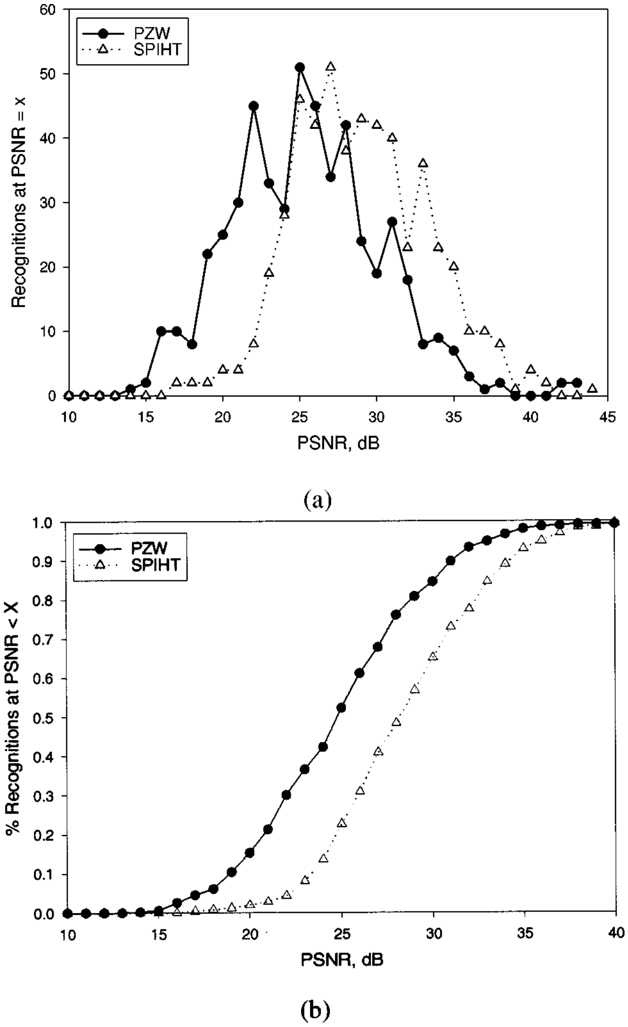
Thus far, we have considered the problem of recognition of progressively transmitted images, where we have assumed an ideal channel. A related problem is that of recognition of images which have been distorted by channel errors. For this problem, rather than the recognition bit rate, we are interested in the minimum quality at which images subject to channel-error distortion can be recognized [31]. Image coders, such as EZW and SPIHT, are vulnerable to channel errors, since a single bit in error can potentially cause the decoder and encoder to lose synchronization for the remainder of the bit stream. This sensitivity to errors has been addressed in a number of different ways, which produce distortions with very different visual appearances. In [32], forward error correction (FEC) is added to the SPIHT bit stream. When an uncorrectable error occurs in this stream, the stream is truncated and a globally blurry image results. In the PZW coder [19], [20], the bit stream consists of independently decodable packets representing spatial patches of the image. This algorithm produces local distortion (splotches) when errors occur. Fig. 9 shows an example of a test image compressed by PZW (and subjected to packet loss) and compressed by SPIHT. The two images have the same PSNR of 23.97.

1) Experiment 3: PZW and FEC-Protected SPIHT: For Experiment 3, we evaluate PZW and FEC-protected SPIHT by showing an observer a sequence of degraded versions of an image at successively increasing PSNRs, rather than at successively increasing bit rates.

A database of 68 greyscale images was collected. Half of the images showed men, and half showed women. All images were of size 512 × 512 pixels. Each of the 68 test images were compressed using the PZW algorithm to a target bit rate of 0.23 b per pixel. The actual bit rate might depart slightly from the target bit rate because of the way PZW fits information into fixed length packets. The target rate of 0.23 bpp led to a high quality decoded image (typically about 40 dB) and required about 180 packets. The channel-degraded versions of these images were produced by dropping some packets and decoding the remainder. Some packets cause more damage than others to the PSNR when dropped. By trying many different random combinations of dropped packets, we created a sequence of (typically) 20 degraded versions of each test image. The sequence of degraded versions had PSNRs ranging between 10 dB and 40 dB, with increments of at least 1 dB between successive images in the sequence. Each image was also compressed by SPIHT at bit rates logarithmically increasing from 0.001 bpp to 0.5 bpp. Twenty versions of the image were saved for each image, and PSNRs for these images also corresponded to a range from 10 dB to 40 dB.

There were 15 observers, each of whom saw each of the 68 images in exactly one sequence, either with the PZW compression or with the SPIHT compression. The selection of PZW or SPIHT was randomized, as was the order in which the images were displayed. Fig. 10(a) shows the number of recognitions (observer responses) that occurred at each PSNR, versus the PSNR, for both SPIHT and PZW. The data look approximately normal. Fig. 10(b) shows the cumulative distribution for these responses as a function of PSNR.

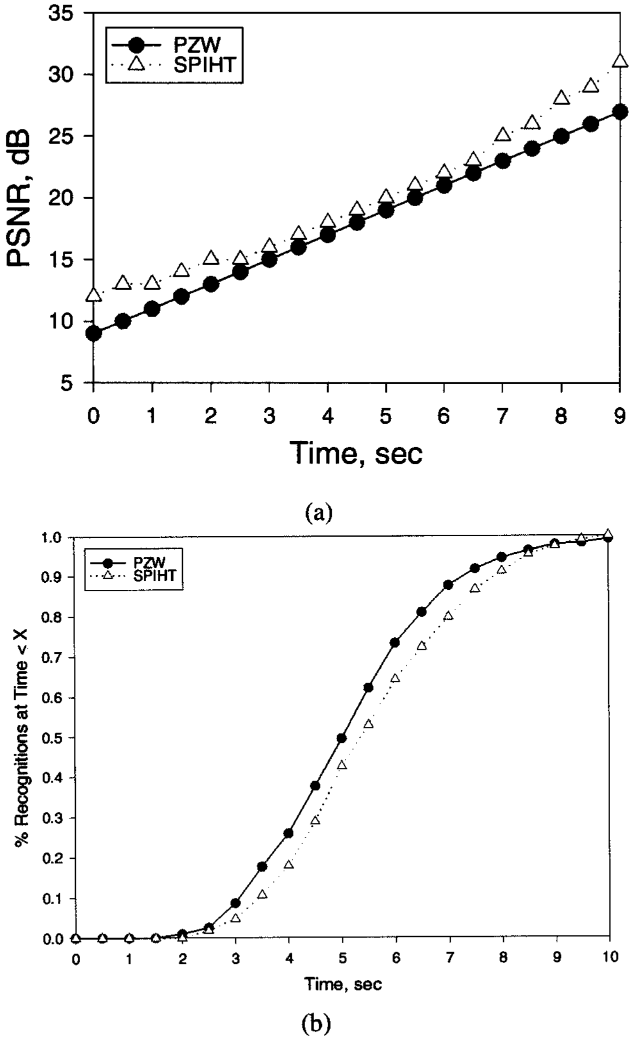
For a given image, the 20 frames compressed by SPIHT were not matched in PSNR, frame-by-frame, to the frames generated by PZW. The SPIHT sequences tended to run at slightly higher PSNRs, as shown in Fig. 11(a) for one particular image in the test set. For all test images, the SPIHT sequence started out



(a)

(b)

Fig. 10. (a) Percentage of observer responses as a function of PSNR for SPIHT and PZW. (b) Cumulative distribution plot of observer responses as a function of PSNR for SPIHT and PZW.



(a)

(b)

Fig. 11. (a) PSNR versus time for SPIHT and PZW for one particular image. (b) Cumulative distribution plot of observer responses as a function of time for SPIHT and PZW.

initially at a higher PSNR. Because of this, one could wonder whether the results displayed in Fig. 10 might merely be reflecting a situation in which observers take a certain more or less fixed amount of time to recognize a given image, or to respond to its display by clicking a mouse button, and that the PZW sequences allow recognition at a lower PSNR simply because those sequences have lower PSNRs initially. That this is not the case is shown by Fig. 11(b), in which the cumulative distribution plot of observer responses is shown as a function of time. It shows that people responded *sooner* in time for the PZW sequences, despite the fact that they were observing *lower* PSNR values during that time.

2) Statistical Analysis: As before, we used a mixed effects linear model (in which the compression algorithm is treated as a fixed effect, and observers and images are treated as random effects) to compare the mean recognition time and PSNR for the two algorithms. The mean PSNR for PZW responses was 25.43 dB, whereas it was 28.97 dB for SPIHT. The 95% confidence interval for the difference of means extended from -3.83 to -3.24. Since the confidence interval does not include zero, we can conclude that the PSNR required for observers to answer the question for the PZW images is significantly less than that required for SPIHT images at the 95% confidence level. When applied to time, the t_{Aij} values were taken to be frame numbers, where frames were shown 500 ms apart. The mean time (frame number) for PZW was 10.72, and was 11.59 for SPIHT. The 95% confidence interval for the difference of mean time extended from -1.17 to -0.58, and again does not include zero. Therefore, we can conclude that observers answered the questions at significantly faster time with PZW, despite the fact that they were answering them at significantly lower PSNR.

The overall error rates for each algorithm were 5.3% for PZW and 5.1% for SPIHT. The Wilcoxon two-sided signed-rank test had a p-value of 0.749 for the comparison of the observer errors, showing that the error rates for the two different algorithms were not significantly different. Observers were also asked for subjective ratings; these results showed similar but not identical trends to the recognition time results [31]. In other work on compressed medical images, large discrepancies between subjective ratings and objective recognition performance have been found [12], [13].

VI. Conclusion

With the proliferation of images on the WWW, and the growing need for fast browsing of remote image databases, increased interest has focussed on progressive compression. Many algorithms explicitly target fast browsing applications [33], [34], [35]; however, performance is still measured using PSNR or subjective ratings, not by simulating fast browsing. In this paper, we have laid out an experimental and statistical framework for such simulations, and we have described the results of a series of such experiments. There are a number of conclusions that we make from this work.

For coders operating with very different principles, such as PZW and SPIHT, there can be a substantial difference in the performance measured by PSNR, subjective ratings and recognition times. Progressive compression algorithms which are intended to be used in a progressive display for fast browsing tasks should be evaluated by simulating a fast browsing task, not by PSNR or subjective ratings.

In a simulation of a fast browsing task, the SPIHT algorithm outperforms JPEG by a substantial margin and EZW by a small margin.

For fast browsing tasks, significantly faster recognition times were achieved by displaying images at a small scale initially, regardless of whether that small scale came from downsampling or by deferring the large-scale information for later in the progressive bit stream.

The average bit rate required for a user to recognize an image and make a decision on it can be quite low, e.g., on the order of 0.05–0.07 bpp for many of the algorithms and image tasks we used. This should be considered when designing algorithms for fast browsing. For example, in [36], detected edges are transmitted first in the image header, and then a progressive wavelet coder is used. Decoding combines the edge information and the progressive data in a subjectively pleasing way. At 0.4 bpp and 0.1 bpp, the decoded example image is subjectively superior to the image produced by the progressive wavelet coder alone. However, for the example image provided, the image header (edge map) by itself takes up 0.052 bpp, and the decoder can display nothing during this time. It is possible that, for a fast browsing task, the subjective superiority of this coder at 0.1–0.4 bpp might be outweighed by its initial handicap in the 0.0–0.05 bpp range where 50% of recognitions may take place.

We examined two particular algorithms that produce global blurriness and localized distortions, and found that, at the same PSNR and for the particular recognition tasks we used, localized distortions allowed faster recognition.

PSNR has found widespread use as an evaluation tool largely because it is easily and cheaply computed. The evaluation methodology described here requires a greater investment in time and expense. It is also most useful when tailored to the specific application for which the compression algorithms are to be evaluated. It is this tailoring, however, which may justify its use for evaluation of algorithms for fast browsing: as our experiments have shown, results obtained from PSNR may be misleading for these applications.

In this paper, we have concerned ourselves with evaluation of compression algorithms directed at fast browsing applications. The emphasis in these applications is on simple recognition tasks, where a decision on whether to continue viewing the image can be made based on a few features available early in the progressive display. In applications where decisions are made based on finer details in the images, other compression algorithms than those described here may be more appropriate. For example, the reduced-scale images employed by MSPIHT allow faster recognition when the image content is suited for display at smaller scales, but this strategy may be unsuitable for written documents. In fact, at the bit rates studied here, textual information is poorly displayed by all of the algorithms discussed in this paper. It is in the nature of progressive transmission that priorities must be set about what information to transmit first, and these priorities will differ depending on the target application. While the algorithms to be evaluated may differ, we expect the experimental methodology presented here to be useful in higher bit rate regimes as well, where recognition of finer details or comprehension of textual content becomes important.

Acknowledgment

The authors would like to thank Prof. C. Berry for advice on statistical analysis, and Prof. K. Jameson for useful discussions on psychovisual phenomena. They are also grateful for the assistance of H. Persson, S. Cen, and N. Serrano.

References

[1] W. B. Pennebaker and J. L. Mitchell, *JPEG Still Image Data Compression Standard*. New York: Van Nostrand, 1993.

[2] J. Shapiro, "Embedded image coding using zerotrees of wavelet coefficients," *IEEE Trans. Signal Processing*, vol. 41, pp. 3445–3462, Dec. 1993.

[3] A. Said and W. A. Pearlman, "A new, fast, and efficient image codec based on set partitioning in hierarchical trees," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, pp. 243–250, June 1996.

[4] P. C. Teo and D. J. Heeger, "Perceptual image distortion," in *Proc. ICIP*, vol. II. Austin, TX, Nov. 1994, pp. 982–986.

[5] N. Jayant, J. Johnston, and R. Safranek, "Signal compression based on models of human perception," *Proc. IEEE*, vol. 81, pp. 1385–1422, Oct. 1993.

[6] V. R. Algazi, Y. Kato, M. Miyahara, and K. Kotani, "Comparison of image coding techniques with a picture quality scale," in *Proc. SPIE Applications of Digital Image Processing XV*, vol. 1771, San Diego, CA, July 1992, pp. 396–405.

[7] H. Barrett, "Evaluation of image quality through linear discriminant models," in *SID'92 Dig. Tech. Papers*, vol. 23, 1992, pp. 871–873.

[8] T. Eude and A. Mayache, "An evaluation of quality metrics for compressed images based on human visual sensitivity," in *Proc. 4th Int. Conf. Signal Processing*, vol. 1, Beijing, China, Oct. 1998, pp. 779–782.

[9] Y. Furusho, K. Kotani, Y. Horita, Y. Kenmochi, and V.-R. Algazi, "Picture quality evaluation model for color coded images: Considering observing points and local feature of image," in *Proc. ICIP*, vol. 4, Kobe, Japan, Oct. 1999, pp. 343–347.

[10] C. Charrier, K. Knoblauch, and H. Cherifi, "Perceptual distortion analysis of color image VQ-based coding," in *Proc. SPIE Very High Resolution and Quality Imaging II*, vol. 3025, San Jose, CA, Feb. 1997, pp. 134–143.

[11] M. G. Ramos and S. S. Hemami, "Robust image coding with perceptual-based scalability," *Proc. IEEE DCC*, pp. 466–487, Mar. 1997.

[12] P. C. Cosman, H. C. Davidson, C. J. Bergin, C. Tseng, L. E. Moses, E. A. Riskin, R. A. Olshen, and R. M. Gray, "Thoracic CT images: Effect of lossy image compression on diagnostic accuracy," *Radiology*, vol. 190, pp. 517–524, 1994.

[13] P. C. Cosman, R. M. Gray, and R. A. Olshen, "Evaluating quality of compressed medical images: SNR, subjective rating, and diagnostic accuracy," *Proc. IEEE*, vol. 82, pp. 919–932, June 1994.

[14] D. P. Chakraborty, “Maximum likelihood analysis of free-response receiver operating characteristic (FROC) data,” *Med. Phys.*, vol. 16, pp. 561–568, 1989.

[15] C. E. Metz, “Basic principles of ROC analysis,” in *Proc. Seminars Nuclear Medicine*, vol. VIII, Oct. 1978, pp. 282–298.

[16] G. Sperling, M. Landy, Y. Cohen, and M. Pavel, “Intelligible encoding of ASL image sequences at extremely low information rates,” *Comput. Vis. Graph. Image Process.*, vol. 31, pp. 335–391, 1985.

[17] S. Cen, H. Persson, D. Schilling, P. Cosman, and C. Berry, “Human observer responses to progressively compressed images,” in *Proc. 31st Asilomar Conf. Signals, Systems, and Computers*, vol. 1. Pacific Grove, CA, Nov. 1997, pp. 657–661.

[18] D. Schilling, P. Cosman, and C. Berry, “Image recognition in single-scale and multiscale decoders,” in *Proc. 32nd Asilomar Conf. Signals, Systems, and Computers*, Pacific Grove, CA, Nov. 1998, pp. 477–481.

[19] J. K. Rogers and P. C. Cosman, “Wavelet zerotree image compression with packetization,” *IEEE Signal Processing Lett.*, vol. 5, pp. 105–107, May 1998.

[20] ——, “Robust wavelet zerotree image compression with fixed-length packetization,” *Proc. IEEE Data Compression Conf.*, pp. 418–427, Mar. 1998.

[21] I. S. Bruner and M. C. Potter, “Interference in visual recognition,” *Science*, vol. 144, no. 3617, pp. 424–425, 1964.

[22] P. Schyns and A. Oliva, “From blobs to boundary edges: Evidence for time- and spatial-scale-dependent scene recognition,” *Psychol. Sci.*, vol. 5, pp. 195–200, July 1994.

[23] E. S. Olds and S. A. Engel, “Linearity across spatial frequency in object recognition,” *Vis. Res.*, vol. 38, pp. 2109–2118, 1998.

[24] R. D. Luce, *Response Times: Their Role in Inferring Elementary Mental Organization*, ser. Oxford Psychology Series. London, U.K.: Oxford Univ. Press, 1986.

[25] G. W. Snedecor and W. G. Cochran, *Statistical Methods*. Ames, IA: Iowa State Univ. Press, 1989.

[26] W. Venables and B. Ripley, *Modern Applied Statistics with S-Plus*. New York: Springer-Verlag, 1994.

[27] I. McNemar, “Note on the sampling errors of the differences between correlated proportions of percentages,” *Psychometrika*, vol. 12, pp. 153–157, 1947.

[28] H. Cohen, “Retrieval and browsing of images using image thumbnails,” *J. Visual Commun. Image Represent.*, vol. 8, pp. 226–234, June 1997.

[29] Z. Xiong, B.-J. Kim, and W. A. Pearlman, “Multiresolutional encoding and decoding in embedded image and video coders,” in *Proc. ICASSP*, vol. 6. Seattle, WA, May 1998, pp. 3709–3712.

[30] B. A. Wandell, *Foundations of Vision*. Sunderland, MA: Sinauer Associates, 1995.

[31] N. Serrano, D. Schilling, and P. C. Cosman, “Quality evaluation for robust wavelet zerotree image coders,” in *Proc. 2nd Ann. UCSD Conf. Wireless Communication*, San Diego, CA, Mar. 1999, pp. 128–134.

[32] P. G. Sherwood and K. Zeger, “Progressive image coding on noisy channels,” *Proc. IEEE DCC*, pp. 72–81, Mar. 1997.

[33] N.-F. Law and W.-C. Siu, “Progressive image coding based on visually important features,” presented at the ICIP, Kobe, Japan, 1999.

[34] Y. Itoh, “An edge-oriented progressive image coding,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 6, pp. 135–142, Apr. 1996.

[35] T.-H. Lan, A. H. Tewfik, and C.-H. Kuo, “Sigma filtered perceptual image coding at low bit rates,” presented at the ICIP, Kobe, Japan, 1999.

[36] A. Mertins, “Image compression via edge-based wavelet transform,” *Opt. Eng.*, vol. 38, pp. 991–1000, June 1999.

Dirck Schilling received the B.Sc. degree in mechanical engineering from Brown University, Providence, RI, in 1983, and the M.S. and Ph.D. degrees in electrical engineering from the University of California at San Diego, La Jolla, in 1997 and 2001, respectively.

He is now with ViaSat, Carlsbad, CA. His research interests include data compression and image and video processing.

Pamela C. Cosman (S’90–M’93–SM’00) received the B.S. degree (with honors) in electrical engineering from the California Institute of Technology, Pasadena, in 1987, and the M.S. and Ph.D. degrees in electrical engineering from Stanford University, Stanford, CA, in 1989 and 1993, respectively.

She was an NSF Postdoctoral Fellow at Stanford University and a Visiting Professor at the University of Minnesota during 1993–1995. Since July 1995, she has been on the faculty of the Department of Electrical and Computer Engineering at the University of California at San Diego, La Jolla, where she is currently an Associate Professor. Her research interests are in the areas of data compression and image processing.

Dr. Cosman was the recipient of the ECE Departmental Graduate Teaching Award (1996), a Career Award from the National Science Foundation (1996–1999), and a Powell Faculty Fellowship (1997–1998). She was an Associate Editor of the IEEE COMMUNICATIONS LETTERS, and was a Guest Editor of the June 2000 Special Issue of the IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS on “Error-Resilient Image and Video Coding.” She was the Technical Program Chair of the 1998 Information Theory Workshop in San Diego. She is a member of Tau Beta Pi and Sigma Xi.