Image Recognition in Single-Scale and Multiscale Decoders[*]

Dirck Schilling, Pamela Cosman and Charles Berry[†]
Department of Electrical and Computer Engineering, University of California at San Diego
[†]Department of Family and Preventive Medicine, University of California at San Diego
9500 Gilman Drive, San Diego, California 92093
dschilli@ucsd.edu, pcosman@ucsd.edu, cberry@tajo.ucsd.edu

Abstract

At best, evaluation of image coders using PSNR is of questionable perceptual validity. But for several types of algorithms, including those with spatially scalable decoders, PSNR might not even be computable, and other methods of evaluation must be used. With a simple reordering of the transmitted bit stream, the SPIHT algorithm can be made spatially scalable without any loss in performance or in progressivity. We present experimental results comparing this multiscale SPIHT (MSPIHT) against SPIHT in terms of the bit rates at which viewers recognize objects in the reconstructed images. MSPIHT is also compared with SPIHT images which have been downsampled to the same scale as the MSPIHT images. We show that viewers are able to recognize reduced-scale images, such as those compressed by MSPIHT, substantially earlier than images compressed by SPIHT.

1. Introduction

It is a practical reality that 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 [2]. Displaying an image at reduced scale has the advantage, given that the image can still be recognized, that a smaller image requires fewer bits to transmit and store. Thumbnail images are typically stored separately from their originals, and therefore require additional storage space. If the viewer decides to request the full-sized version of the image, the data for that image is transmitted in addition to the already transmitted thumbnail version.

Another approach to the problem of limited bandwidth is that of progressive transmission. Progressive coders allow the decoder to reconstruct the image with increasing quality as more bits arrive. Popular progressive compression schemes, including SPIHT [3], initially decode a full-sized, blurred version of the original, which gradually comes into focus as the transmission proceeds.

These two approaches can be combined in a *spatially scalable* compression algorithm, such that versions of the image at successively increasing scales can be extracted from the bitstream 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, since the decoder could simply stop doing 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, spatially scalability is usually taken to mean that information about detail scales is not transmitted initially. In zerotree wavelet coders such as SPIHT and EZW [4], information on some coefficients in higher frequency bands is sent before all coefficients in the lowest frequency band have been encoded. So 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 ver-

[*]This work was supported by the National Science Foundation under Grants MIP-9617366 and MIP-9624729 (CAREER), and by the Center for Wireless Communications at UCSD.

sion of the image.

In addition to the basic advantage that spatially 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.

In this paper, we show that, by reordering the transmitted bit stream, the SPIHT algorithm can be made spatially scalable with no loss in performance (PSNR versus bit rate) or in progressivity. We present experimental results comparing this multiscale SPIHT, which we call MSPIHT, against SPIHT in terms of the bit rates at which viewers recognize objects in the reconstructed images. Since there is no generally accepted method for comparing images at different scales using PSNR, we employ the human observer evaluation framework developed in [1]. We show that viewers are able to recognize MSPIHT-compressed images substantially earlier than images compressed by SPIHT.

This paper is organized as follows. In Section 2, we describe the MSPIHT algorithm. Our evaluation of this algorithm using human observers is discussed in Section 3, and we present conclusions in Section 4.

2. Multiscale SPIHT

We now describe the mechanics of our multiscale version of SPIHT, which we call MSPIHT. In MSPIHT, the schedule of scales and bit rates is freely determined; it consists of reordering the SPIHT bit stream with no additional bits required to manage the scales.

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. A single-level decomposition is illustrated in Figure 1. Each layer of blocks in the illustration represents a bit plane, and each column a coefficient whose magnitude is given by the column height. The LL band is indicated by the dark blocks, while the remaining blocks represent coefficients in higher-frequency bands. Both SPIHT and MSPIHT transmit information only about coefficients which exceed the current significance threshold T_i. 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. In MSPIHT, a current scale boundary S_j is defined which excludes any such finer-scale coefficients. Information about these coefficients is deferred until after all coefficients for the coarser scales have been described.

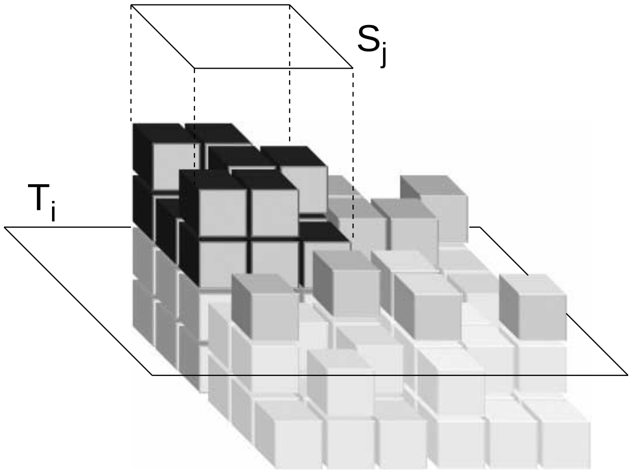

Figure 1. Coefficient bitplanes. SPIHT describes all coefficients exceeding threshold T_i. MSPIHT describes only coefficients above T_i and within scale boundary S_j, deferring remaining coefficients until later.

A scale schedule specifies the bit rates at which each jump to the next larger scale occurs. Since both encoder and decoder know the schedule, no additional bits are required to manage the scale jumps. (If the schedule were unknown to the decoder, it could be transmitted as part of the header 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 Figure 2. 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 6 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 re-examined; 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 non-deferred coefficients. At this point, processing resumes where it left off when the scale jump occurred. This sequence of events is repeated for each scale jump, until the desired final bit rate is reached.

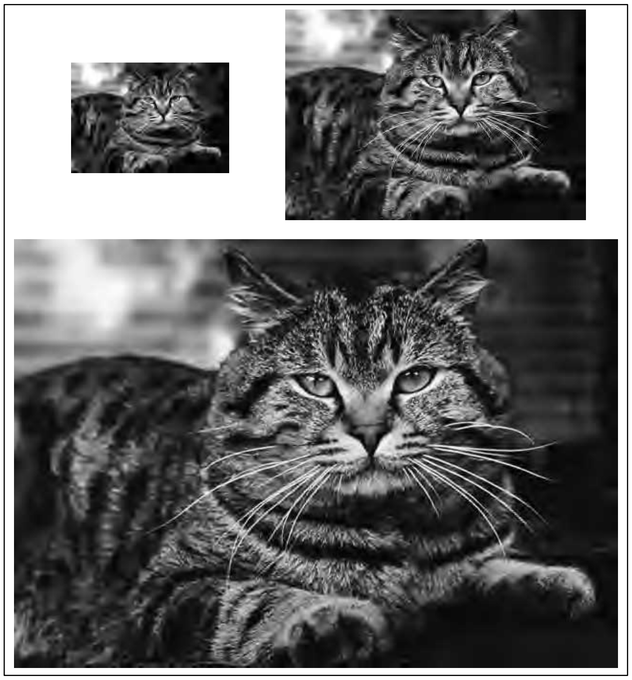

Figure 2. MSPIHT-compressed image at 0.02, 0.09, and 0.30 bpp.

Note that at any given point in the progression, no bits are spent to describe coefficients from scales finer than the current 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. Thus, no bit rate penalty is paid relative to SPIHT for the spatial scalability.

3. Evaluation with Human Observers

Evaluation of image coders using PSNR is not useful for algorithms with spatially scalable decoders. It is not obvious how to compare two images of different sizes using PSNR. However, the human observer evaluation framework described in [1], in which coders are compared on the basis of recognition bit rates, is well suited to this type of evaluation.

Two experiments were performed. The objective of experiment 1 was to compare SPIHT with MSPIHT, and to determine a scale schedule for MSPIHT which performed well. The goal of experiment 2 was to gain some understanding of the causes for MSPIHT's improved recognition performance, noted during experiment 1.

3.1 Comparison of Scale Schedules

For experiment 1, three MSPIHT scale schedules (A, B, C) were prepared (see Figure 3). The experimental procedure for comparing algorithms was similar to [1]. A series of 120 images were displayed progressively to each of 20 observers. Before each image was shown, a question about the image was displayed. While watching the progression, as soon as the observer was reasonably confident that she could correctly answer the question, she hit a key to halt the progression. The bit rate was recorded for each response, as well as whether the correct answer was given.

	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

Figure 3. Scale schedules used for testing MSPIHT.

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. 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. The letters 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.

The images were presented in a different random order for each observer, and each observer saw a given image only once. The algorithm used to compress image j for observer i was selected randomly, subject to the constraint that, for the entire experiment, an equal number of images were compressed by each algorithm. Images were displayed on a 20” monitor, in a single window against a solid background. In order to simulate natural office viewing conditions, the viewing distance was not constrained.

Response bit rates averaged over all observers are presented in Figure 4. Averages were computed for each algorithm over the sets of 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.

	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

Figure 4. Arithmetic mean of recognition bit rates for each algorithm, in bpp. Best performance for each image type is shaded.

3.2 Comparison of MSPIHT, SPIHT and downsampled SPIHT

The results of experiment 1 indicate that MSPIHT allows earlier recognition than SPIHT for several types of images. The amount of improvement depends on the scale schedule employed. We now focus on the potential causes for this improvement. Did observers recognize objects earlier using MSPIHT because MSPIHT defers visually unusable fine-scale information until later, allowing more precise coarse-scale information to be transmitted first? Or was it instead because the objects in the images were seen first at a small size, more suited for rapid recognition by the eye? For example, images contained within the 5.2 degree foveal field might require little eye movement, allowing earlier recognition. Whereas for a large image, the eye has to scan around over the image field, which might slow recognition. If a combination of both effects was responsible, which effect predominated?

A second experiment was performed to investigate these questions. For this experiment, the image sequences processed by SPIHT were downsampled by block averaging to match the image sizes produced by the MSPIHT-C scale schedule (judged to be the best schedule tested for MSPIHT). These sequences of downsampled SPIHT images allowed us to test the psychovisual hypothesis described above. Since the transmitted bitstream for these 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.

Experiment 2 compared SPIHT, MSPIHT-C, and downsampled SPIHT. The same 120 images were displayed progressively to each of 21 new observers, and the same questions were posed as in experiment 1. The bit rate and correctness of each response were recorded. A preliminary analysis of the response bit rates suggested they were drawn from an approximately log-normal distribution. Accordingly, the log of bit rates were used to allow normal theory analysis. The geometric mean of response bit rates was determined for each of the three algorithms in experiment 2. These means were determined by fitting the data to a mixed effects linear model with random effects given by the images and observers, and a fixed effect recording the algorithm used to compress each image. Model fitting was carried out using the $Splus$ function `varcomp` [5]. A difference-of-means analysis was performed for each pairing of the geometric means $r_{MSPIHT-C}$, r_{SPIHT} and $r_{Downsampled\_SPIHT}$. As seen in Figure 5, both MSPIHT-C and downsampled SPIHT outperformed SPIHT with 5% statistical significance in terms of mean response bit rates. The difference between the mean bit rates of MSPIHT-C and downsampled SPIHT, however, was not significant. This was the case for all images taken together, as well as for each of the subtypes tested, i.e., images of animals, vehicles, and images of letters in three different sizes.

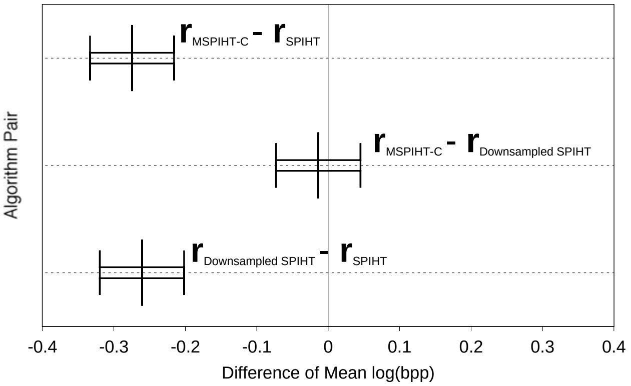

Figure 5. Difference of mean log bit rate for each pair of algorithms.

3.3 Analysis of Observer Mistakes

An analysis was performed to determine the relationship, if any, of each of the compression algorithms with the incidence of observer mistakes, i.e., responses for which the answer given was incorrect. Two questions of interest were, first, whether the incorrect responses could have influenced the overall performance conclusion, and second, whether any of the 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.

Next, the error rates for each algorithm – the percent of total responses that were incorrect – were examined. 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 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.

4. Conclusions

We note that the SPIHT algorithm can be made spatially scalable without any loss in progressivity or in performance, and that this spatially scalable version allows image recognition at lower bit rates for the recognition tasks we tested. Faster recognition was also obtained by downsampling the images produced by SPIHT to the same scales as were chosen for MSPIHT-C's scale schedule. No significant difference in recognition performance or error rates was found between MSPIHT-C and downsampled SPIHT. This appears to indicate that the performance advantage enjoyed by these two methods is primarily due to psychophysical effects related to image size, rather than to the bitstream reordering employed by MSPIHT.

The performance improvement for MSPIHT and downsampled SPIHT is substantial for the images and tasks studied: recognition bit rates averaged more than 26% lower for MSPIHT-C than for SPIHT. This can translate directly into an equivalent savings in storage space, or in transmission bandwidth, given that the goal is content recognition rather than maintaining perfect fidelity.

An embedded, spatially scalable image compression method such as either MSPIHT-C or downsampled SPIHT provides several important benefits:

- Since the algorithm is embedded, the information needed to reconstruct a given scale is contained within the bit stream for all finer scales. No information need be stored or transmitted twice.

- Rather than being limited to a single thumbnail followed by the full-scale version, the image may be displayed at several successive scales as more bits arrive.

- When used with an appropriate scale schedule, a substantial improvement over traditional full-scale progressive methods, in terms of recognition bit rates, can be realized.

Several questions present interesting targets for further study. First, which specific psychophysical effects are responsible for the improved recognition performance of spatially scalable algorithms? Such effects might include the presence of the entire object to be recognized within the foveal field, or the perceived sharpness or clarity of significant edges belonging to the recognized object. Second, is there an optimal size at which objects should be displayed first by an image coder in order to obtain maximum recognition performance? Finally, why do the better-performing scalable methods also appear to cause a higher incidence of observer errors? Answers to this question might suggest modifications to spatially scalable compression algorithms which could mitigate this effect.

Acknowledgements

The authors wish to thank Dr. Kimberly Jameson for many helpful discussions on vision and psychophysics.

References

[1] S. Cen, H. Persson, D. Schilling, P. Cosman, and C. Berry. Human observer responses to progressively compressed images. In *Proceedings of the 31st Asilomar Conference on Signals, Systems and Computers*, volume 1, pages 657–661, Pacific Grove, California, November 1997.

[2] H. Cohen. Retrieval and browsing of images using image thumbnails. *Journal of Visual Communication and Image Representation*, 8(2):226–234, June 1997.

[3] A. Said and W. Pearlman. A new, fast, and efficient image codec based on set partitioning in hierarchical trees. *IEEE Trans. on Circuits and Systems for Video Technology*, 6(3):243–250, June 1996.

[4] J. Shapiro. Embedded image coding using zerotrees of wavelet coefficients. *IEEE Transactions on Signal Processing*, 41(12):3445–3462, December 1993.

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