UC San Diego

UC San Diego Previously Published Works

Title

Human Observer Responses to Progressively Compressed Images

Permalink

<https://escholarship.org/uc/item/0rg792mt>

Journal

Signals, Systems & Computers, 1997. Conference Record of the Thirty-First Asilomar Conference on, 1

Authors

Cen, S
Persson, H
Schilling, D
et al.

Publication Date

1997-11-01

Peer reviewed

Human Observer Responses to Progressively Compressed Images*

Song Cen, Håkan Persson, Dirck Schilling, Pamela Cosman, 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

scen@ucsd.edu, dschilli@ucsd.edu, pcosman@ucsd.edu, cberry@tajo.ucsd.edu

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 occasionally 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 apply it to the comparison of several well known progressive algorithms.

1. Introduction

A progressive image compression algorithm represents an image in such a way that the decoder can reconstruct the image with increasing quality as more bits arrive. In order to achieve good performance as measured by PSNR, progressive algorithms have often focused on sending information on the largest DCT or wavelet coefficients first, in order to minimize mean squared error (MSE) distortion at a given bitrate. The algorithm which best achieves this goal is judged by current criteria to be superior to others. For real-world applications of progressive compression such as image transmission over limited-bandwidth communications links (e.g., the Internet), it is important that the user be able to identify the contents of an image as early as possible in its transmission. While browsing a remote database, 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. Measuring the utility of the progressive compression algorithm for a given application should involve issues of human recognition of the reconstructed images.

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 (e.g., [6, 1]). 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 [3]. These methods begin to address the question of image recognition. Our study involves no subjective opinions; it directly assesses image recognition by having observers respond to questions whose answer can only be known by recognizing the image content. We analyze the correctness of the answers as well as the response times, and this can provide a reliable comparison between compression algorithms.

Our comparisons were of three well-known progressive compression algorithms. The popular DCT-based JPEG algorithm is included here as a useful baseline for comparison. A progressive version of JPEG is defined [8], however for the purposes of this study the progression was simulated by displaying successive versions of baseline sequential JPEG at increasing quality levels. The other two progressive coders studied were wavelet zerotree methods: Embedded Zerotree Wavelet (EZW) coding due to Shapiro [5], and Set Partitioning in Hierarchical Trees (SPIHT) due to Said and Pearlman [4].

This paper is organized as follows. Section 2 describes in detail our evaluation framework. The results of the experiments and statistical analysis are presented in Section 3, and we discuss our conclusions and ongoing work in Section 4.

*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.

2. Evaluation Framework

This section describes in detail the proposed experimental and statistical framework for comparing any two progressive compression algorithms.

2.1. Experimental Procedure

A collection of images is selected for which a question with a binary answer can be asked, for example, "Are the persons in this image male or female?" The images are chosen such that the question can be unambiguously and 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 . 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 for each subject. The two sessions are seen one week (or more) apart to minimize intersession learning effects.

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 proceeds in like fashion with the remaining images. 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 confident of the answer.

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 bpp) is displayed for algorithm A is the same as that for algorithm B . For our first experiment, comparing JPEG with SPIHT, frames were spaced 0.02 bpp apart in bit rate and displayed at a rate of 1.33 frames per second. 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 the second experiment, 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 experiment was 2 frames per second.

2.2. 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. An algorithm is superior if it can transmit the necessary information faster, i.e., at a lower bit rate, provided that the error rate is no higher. Thus, statistics are needed to compare bit rates and error rates.

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 $i = 1, \dots, I$ and $j = 1, \dots, J$ be indices 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. Preliminary examination of probability plots of the data suggested 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

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

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 α (1×1) 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 1 in column i for each row involving observer i and zeroes elsewhere, and where Z_2 ($N \times J$) contains a 1 in column j for every row involving image j and zeroes elsewhere. The $(I+J) \times 1$ 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 $\beta_{1i} \sim N(0, \sigma_{obs}^2)$ and $\beta_{2j} \sim N(0, \sigma_{img}^2)$. Finally, ϵ ($N \times 1$) is the residual vector for which it is assumed that $\epsilon_{ij} \sim N(0, \sigma_{res}^2)$.

Using restricted maximum likelihood (REML), estimates are obtained for the coefficient α and for the variance components, σ_{obs}^2 , σ_{img}^2 , and σ_{res}^2 . An estimate of the standard error of $\hat{\alpha}$, $se_{\hat{\alpha}}$, 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. $\hat{\alpha} \pm 1.96 se_{\hat{\alpha}}$, and for its antilog, $\exp(\hat{\alpha} \pm 1.96 se_{\hat{\alpha}})$. Model fitting is performed using the *Splus* function `varcomp` [7].

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. For example, 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.

2.3 Analysis of Observer Mistakes

In addition to the bit rates, the responses of the observers to the questions are recorded and 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. To examine this possibility, 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, those with both members correct, those with algorithm *A* correct and *B* not, those with algorithm *B* correct and *A* not, and those with neither one correct. In the McNemar analysis [2], 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, each would have a binomial distribution with parameter 1/2. As an example of the calculation, 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.

3. Results

3.1. Experiment Parameters

For the results presented in this paper, the following experiment parameters were used:

- Image collections for two different questions were included. (1) "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. (2) "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. forests, fields, underwater.
- Each session consisted of 118 images (60 for question 1, 58 for question 2), for about 45 minutes per session.
- Images were displayed on an SGI O2 workstation with a 20" monitor, in a single window against a black background.
- All images were 256-level greyscale. Their sizes ranged from 160×160 to 640×640 pixels, averaging 420×420 .
- Within a session, sets of 12 images associated with one question were alternated with sets associated with the other question. No observer saw the same image twice on the same day and images were presented in a different random order for each subject.
- For the comparison of JPEG with SPIHT, there were 5 observers. For the comparison of EZW with SPIHT, there were 20 observers. 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.

3.2 Summary of Results

Two experiments were performed: the first comparing JPEG with SPIHT, and the second comparing EZW with SPIHT. Figure 1 shows the log of bit rates for the JPEG-SPIHT comparison. A characteristic of baseline sequential JPEG is that it must transmit a minimum number of bits before anything at all can be displayed (a DC value for each block), which results in the visible skew of the data toward the left of the diagonal at low bit rates. In Figure 2 the data for the EZW versus SPIHT comparison are shown. Visual examination of this plot indicates that these two algorithms are more evenly matched.

When examined quantitatively, SPIHT was found to lead to faster image recognition than either JPEG or EZW. In Figure 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

Figure 1. Scatter plot of JPEG vs. SPIHT data. SPIHT clearly outperforms JPEG.

questions to be answered. In designing new progressive algorithms tuned for fast recognition, information about the image features judged to be important for recognition should be concentrated below these ranges. 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 Figure 3, that the same SPIHT-compressed images required more bits on the average in the experiment where they were compressed against JPEG, than in the experiment where they were compared against EZW. In the JPEG comparison, the frame display rate was faster, so a person's inherent reaction time would be expected to play more of a role. In the comparison against SPIHT, the display rate was slower, and also the frame bit rates were no longer proportional to time, so observers could more carefully respond.

Figure 4 shows the mean bit rate ratios and their confidence intervals. These values can summarize the comparative performance of the algorithms. 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 of mistakes was 16, and the mean value was 4.84 (out of 118 images). Analyzing the paired data with the McNemar statistic, as described above, showed no difference in correctness at the

Figure 2. Scatter plot of EZW vs. SPIHT data. Visual inspection does not provide a clear indication of which algorithm performs better.

5% significance level between algorithms for any of the 25 observers individually, or for the observers in each experiment pooled together. In Figure 5 only the bit rates for erroneous responses in the EZW/SPIHT comparison are plotted. The symmetry of these errors supports our conclusion from the McNemar analysis that errors did not significantly influence our results.

4. Conclusions

We expect that this framework for evaluating progressive transmissions will be useful in a number of ways. For those image databases which are expected to have an important component of progressive fast browsing usage, a choice must be made regarding which progressive algorithm to use. The simplest way to make this choice is through the use of a distortion-rate curve. There are many situations, however, in which distortion-rate curves cannot be meaningfully compared. For example, suppose the database is of human faces, and candidate compression algorithms exist which allocate more bits and higher quality to detected features such as the eyes and mouth. The overall distortion-rate curve for such a region-based compression algorithm could not be sensibly compared against algorithms such as the wavelet zerotree methods. While one could report distortions separately for different regions, actual evaluation of the use-

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

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

fulness of such a progressive region-based scheme would likely require an experiment aimed at measuring and comparing recognition response times. Similarly, perceptually-based compression schemes, which do not attempt to minimize MSE, cannot be expected to prove their worth with a curve of MSE versus bit rate. Likewise, if one had a choice between transmitting a reduced-size image at higher quality, or a larger version at lower quality, the comparison of distortion-rate curves for the images at different resolutions would not be straightforward. In all these types of cases (region-based quality allocation, perceptual quality allocation, multiresolution images), we expect that our proposed framework for measuring and comparing recognition bit rates would be a useful method for comparing algorithms. In this paper, we compared JPEG and SPIHT, and also EZW and SPIHT. The SPIHT algorithm was significantly superior to JPEG for these recognition tasks under this experimental setup. The comparison of SPIHT and EZW also showed that SPIHT allowed image recognition to occur at significantly lower bit rates. In future work we intend to examine perceptually-based algorithms and multiresolution progression under this framework.

We expect that this type of testing will prove useful in other ways as well. It will be of interest to see what bit rates actually tend to correspond to "barely recognizable" quality. By varying the image databases used, and by varying the questions asked about them, one can explore whether perhaps, for fast browsing applications, image quality at

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

ultra-low bit rates should be promoted even at the expense of having worse distortion-rate performance at higher rates.

References

- [1] N. Jayant, J. Johnston, and R. Safranek. Signal compression based on models of human perception. *Proc. IEEE*, 81:1385–1422, Oct. 1993.
- [2] I. McNemar. Note on the sampling errors of the differences between correlated proportions of percentages. *Psychometrika*, 12:153–157, 1947.
- [3] M. G. Ramos and S. S. Hemami. Robust image coding with perceptual-based scalability. In J. Storer and M. Cohn, editors, *Proceedings of the 1997 IEEE Data Compression Conference*, pages 466–487, Snowbird, Utah, March 1997.
- [4] 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.
- [5] J. Shapiro. Embedded image coding using zerotrees of wavelet coefficients. *IEEE Transactions on Signal Processing*, 41(12):3445–3462, December 1993.
- [6] P. Teo and D. Heeger. Perceptual image distortion. In *Proceedings ICIP-94*, volume II, pages 982–986, Austin, Texas, November 1994. IEEE, IEEE Computer Society Press.
- [7] W. Venables and B. Ripley. *Modern Applied Statistics with S-Plus*. Springer Verlag, New York, 1994.
- [8] G. Wallace. The JPEG still picture compression standard. *Communications of the ACM*, 34(4):30–44, April 1991.