Report Number: WUCS-93-41

1993

Experimental Evaluation of Psychophysical Distortion Metrics for JPEG-Encoded Images

Daniel R. Fuhrmann, John A. Baro, and Jerome R. Cox Jr.

Two experiments for evaluating psychophysical distortion metrics for JPEG-encoded images are described. The first is a threshold experiment, in which subjects determined the bit rate or level of distortion at which distortion was just noticeable. The second is a suprathreshold experiment in which subjects ranked image blocks according to perceived distortion. The results of these experiments were used to determine the predictive value of a number of computer image distortion metrics. It was found that mean-square-error is not a good predictor of distortion thresholds or suprathreshold perceived distortion. Some simple pointwise measures were in good agreement with psychophysical data;... **Read complete abstract on page 2.**

Experimental Evaluation of Psychophysical Distortion Metrics for JPEG-Encoded Images

Daniel R. Fuhrmann, John A. Baro, and Jerome R. Cox Jr.

Complete Abstract:

Two experiments for evaluating psychophysical distortion metrics for JPEG-encoded images are described. The first is a threshold experiment, in which subjects determined the bit rate or level of distortion at which distortion was just noticeable. The second is a suprathreshold experiment in which subjects ranked image blocks according to perceived distortion. The results of these experiments were used to determine the predictive value of a number of computer image distortion metrics. It was found that mean-square-error is not a good predictor of distortion thresholds or suprathreshold perceived distortion. Some simple pointwise measures were in good agreement with psychophysical data; other more computationally intensive metrics involving spatial properties of the human visual system gave mixed results. It was determined that mean intensity, which is not accounted for in the JPEG algorithm, plays a significant role in perceived distortion.

Experimental Evaluation of Psychophysical Distortion Metrics for JPEG-Encoded Images

Daniel R. Fuhrmann, John A. Baro and Jerome R. Cox, Jr.

WUCS-93-41

September 1992

Department of Computer Science
Washington University
Campus Box 1045
One Brookings Drive
St. Louis, Missouri 63130-4899

Experimental Evaluation of Psychophysical Distortion Metrics for JPEG-Encoded Images

Daniel R. Fuhrmann[1], *John A. Baro*[2], *and Jerome R. Cox*[3]

(1) Department of Electrical Engineering
Washington University
St. Louis MO 63130

(2) NASA Classroom of the Future
Wheeling Jesuit College
Wheeling WV 26003

(3) Department of Computer Science
Washington University
St. Louis MO 63130

ABSTRACT

Two experiments for evaluating psychophysical distortion metrics for JPEG-encoded images are described. The first is a threshold experiment, in which subjects determined the bit rate or level of distortion at which distortion was just noticeable. The second is a suprathreshold experiment in which subjects ranked image blocks according to perceived distortion. The results of these experiments were used to determine the predictive value of a number of computed image distortion metrics. It was found that mean-square-error is not a good predictor of distortion thresholds or suprathreshold perceived distortion. Some simple pointwise measures were in good agreement with psychophysical data; other more computationally intensive metrics involving spatial properties of the human visual system gave mixed results. It was determined that mean intensity, which is not accounted for in the JPEG algorithm, plays a significant role in perceived distortion.

This work was supported in part by Southwestern Bell Technology Resources Inc.

1. Introduction

This paper describes a set of experiments designed to improve our understanding of the ways in which the human observer perceives distortion in lossy image coding schemes. The specific goal is the determination of a computable distortion metric for assessing the visual quality of images coded using the Joint Photographic Experts Group (JPEG) algorithm based on the Discrete Cosine Transform (DCT) [1], and for determining quality parameters and/or bit rates in image communication systems.

The need for such a distortion metric has long been recognized in the image processing community. While the mean-square-error (MSE) criterion is normally used for mathematical simplicity in information-theoretic treatments of image compression, such an error measure does not really predict human response. More realistic measures of perceived distortion, even those determined empirically, could play a useful role in rate-distortion studies. Our own motivation in studying such metrics comes from their possible use in variable bit-rate coding schemes for image communication. A reliable metric could be used in a feedback loop to monitor transmitted image quality, and to set image quality parameters (such as the "Q-factor" in the JPEG algorithm) adaptively to achieve a consistent level of image quality. Another motivation might be the incorporation of such criteria into the coding scheme itself; however this is beyond the scope of the present paper.

Let I denote a digital image defined on some discrete set of pixels, and let $\hat{I}$ denote a distorted version of that same image. An image distortion metric is a positive real-valued function $d(I, \hat{I})$ that is a measure of how distorted the image $\hat{I}$ appears relative to I, as perceived by a human observer. This type of metric will be referred to as a *Type I metric* (in order to distinguish it from Type II metrics, as discussed below). There are two properties such a metric should have in order for it to be a useful predictor of human response. First, it should predict the threshold of

distortion detectability, that is, there should be some threshold d_0 such that when $d(I, \hat{I}) < d_0$, the average viewer cannot distinguish between I and $\hat{I}$. Second, for distortion levels above threshold, $d(I, \hat{I})$ should increase monotonically with the perceived amount of distortion. That is, if the observer is presented with image pairs $(I_1, \hat{I}_1)$ and $(I_2, \hat{I}_2)$, and image $\hat{I}_1$ is perceived to be more distorted than $\hat{I}_2$, then we should have $d(I_1, \hat{I}_1) > d(I_2, \hat{I}_2)$.

If the nature of the distortion is well-defined (e.g., it is known that that the image has been compressed using the JPEG algorithm) it is entirely possible that the response of the observer can be predicted from I alone. That is, for certain lossy processing, one might be able to predict which images I would be subject to more or less perceived distortion based on their intrinsic qualities, which could be quantified by a real-valued function of the form $d(I)$. Although this is not a true metric in the mathematical sense, this type of function will be referred to as a *Type II metric*. Such a predictor, if reliable, would have the distinct advantage of allowing one to adjust image quality parameters without having to compute $d(I, \hat{I})$ in a feedback loop. The experiments reported here allowed for the testing of this type of measure as well.

There is a substantial literature on the psychophysics of human vision, and one could expect to draw on that to aid in the determination of good distortion metrics. This has in fact been done, although different researchers have come to different conclusions. One of the aims of the present study was to compare some of the published distortion metrics in a single psychophysical testbed. The present experiments were designed so that any number of metrics can be considered after the data collection phase; for this reason several novel metrics are proposed here.

2. Image Distortion Metrics

2.1. Human Visual System Models

Before proceeding with a description of the actual metrics considered in these experiments, we review some of the salient features of models of the human visual system typically considered in the engineering literature. Good overviews may found in Pratt [2] and Clarke [3], and more detailed treatments are given by Stockham [4], Hall and Hall [5], Limb [6], Sakrison [7], and Tzou [8]. In all of this work there seem to be four phenomena which are predictable and thus useful in an engineering sense.

First, there are non-linearities in the way the visual system responds to incident light. For a wide range of light intensities I, the just-noticeable-difference in intensity ΔI satisfies $\frac{\Delta I}{I} = constant$; this is known as Weber's Law. This suggests a logarithmic relationship between the physical and the "perceived" intensity of light, where the just-noticeable-difference increases with increasing intensity. Other nonlinear relationships have been proposed as well.

Second, there are spatial filtering mechanisms at work - one due to the optics of the eyeball, which is essentially lowpass, another due to the lateral inhibition in the retina which has the effect of applying a spatial highpass filter, and others at later stages of the visual system. The overall bandpass characteristic can be demonstrated via an illusion called the Mach band effect, which is quite striking. Mach bands demonstrate edge enhancement due to lateral inhibition.

The third aspect of vision affecting quality assessment is called spatial masking; this is the suppression of errors or distortion as a result of high image activity or contrast. Although the exact physiological mechanism is not well understood, spatial masking played a prominent role in the psychophysical experiments of Limb [6].

Finally, there is the issue of local versus global attention in the interpretation of images. Since humans attend to visual fields which subtend only 1° - 3° of visual angle, it is not possible

to "take in" all of a large image simultaneously. Complex images are viewed with a series of brief fixations, between which brief, rapid eye movements occur. Some models of how observers integrate the various parts of an image are thus important.

2.2. Type I Metrics

This section describes 5 different metrics of the form $d(I, \hat{I})$ which were considered as candidates to be tested in the psychophysical experiments. A sixth metric of a slightly different form is also described. The following notation will be used throughout: x denotes one pixel in the domain of the image, X is the set of all pixels, and N the number of pixels in the image.

1. Mean-Square-Error (MSE). The MSE metric is given by the average squared difference in gray levels between the original and distorted images:

$$d = \frac{1}{N} \sum_{x \in X} |I(x) - \hat{I}(x)|^2 \quad . \tag{2.1}$$

2. Mean-Square-Error after Non-Linearity (MSENL). Here the two images are passed through a pointwise non-linearity designed to compress the dynamic range of the gray levels. The non-linearity employed here is the cube root, although any other similar concave function such as the μ-law compression function would probably yield comparable results. The mean-square-error of the transformed images is then determined:

$$d = \frac{1}{N} \sum_{x \in X} |I^{1/3}(x) - \hat{I}^{1/3}(x)|^2 \quad . \tag{2.2}$$

In (2.2) it is assumed that the gray levels are defined to lie in the range $0 \leq I(x) \leq 1$.

3. Mannos-Sakrison metric (MANNOS). This metric was proposed by Mannos and Sakrison [9] and studied in a psychophysical experiment not unlike ours. Here one applies a pointwise cube-root non-linearity followed by a spatial bandpass filter to both images, and computes the mean-square-error between the resulting transformed images. The parameters of the bandpass

filter were determined empirically in [9]; it has a radially symmetric 2-dimensional frequency response of the form

$$H(r) = 2.6[0.0192 + 0.114r]\exp[-(0.114r)^{1.1}] \tag{2.3}$$

where r is the magnitude of the frequency (in the 2-dimensional frequency plane) measured in cycles per degree of visual angle subtended. It is important to note that applying this metric requires precise knowledge of the viewing distance and pixel size.

4. The Logarithmic Image Processing metric (LIP). This metric is based on the Logarithmic Image Processing model of Jourlin and Pinouli [10]. The actual metric was proposed by Brailean *et al.* [11] in the context of image restoration. Computation of this metric involves a conversion of both the original and distorted images to the "contrast domain" - each pixel value is replaced by a number representing its local contrast. The mean-square-error between the two contrast-domain images is then taken as the metric. Although computation of this metric is not particularly difficult, it is time-consuming. The details are omitted here for brevity; see [11] for further information.

5. Distortion Contrast (DCON). In this metric one computes the average Michelson contrast between corresponding pixels in the original and distorted images. The expression is

$$d = \frac{1}{N}\sum_{x\in X}\frac{|L(x)-\hat{L}(x)|}{L(x)+\hat{L}(x)} \tag{2.4}$$

where $L(x)$ represents the actual luminance value at pixel location x. This expression can be given in terms of gray level intensity if one knows the relationship between luminance and gray level. For our experimental system (see Sec. 3.2 below) the expression for contrast becomes

$$d = \frac{1}{N}\sum_{x\in X}\frac{|I(x)-\hat{I}(x)|}{I(x)+\hat{I}(x)+23} \quad . \tag{2.5}$$

6. Bit Rate (BITS). This is a metric which is defined only for sub-blocks of an image, and depends on the use of the JPEG DCT-based coding algorithm. If one assumes an overall bit rate

R for an image coded in this way, sub-blocks of the image will exhibit a variable bit rate such that the average over all the sub-blocks is R. It was hypothesized that the bit rate for an image sub-block, relative to the average over the entire image, might be related to the perceived quality of that sub-block. These bit rates are a simple matter to determine if the JPEG algorithm is used to introduce distortion.

2.3. Type II Metrics

The second class of metrics involves those which are not an actual measure of distance between an original and a distorted image. Rather, these functions take as their input the original image, and return a value d which can be used as a measure of a characteristic of an image that may be related to the perceived distortion introduced by a lossy processing scheme, JPEG in this case.

1. Mean Intensity (MI). The first metric of this type is quite simple - it is the average gray level over the image:

$$d = \frac{1}{N} \sum_{x \in X} I(x) \tag{2.6}$$

2. Spectrum Slope (SS). This measure is based on the work of Cargill *et al.* [12] who used a similar scheme in the classification of biomedical images. It was hypothesized that significant high-frequency content in the image will have a masking effect on distortion. In computing this measure, one first computes the 2-dimensional Fourier transform. Then, by averaging the squared Fourier magnitudes over concentric rings in the 2-dimensional frequency plane, a one-dimensional radial power spectrum is obtained. The slope of a line which is fit to this power spectrum, plotted in log-log coordinates, is a measure of the relative high-frequency content of the image (this slope is usually negative, and the less negative it is the greater the high-frequency content).

3. Spectrum Slope over Mean Intensity (SS/MI). This metric is the ratio of metrics 1 and 2 above.

4. Local Contrast (LCON). In this measure one computes the average Michelson contrast between a pixel and its four neighboring pixels. This is a measure of the local "activity" at that pixel in the image. The average of this local contrast is then taken over all the interior pixels of a block.

2.4. The Role of the Q-Factor

Since the focus of the present study is perceived distortion in JPEG-encoded images, it is important to look at the role of a key parameter of that algorithm, known as the "Q-Factor". This is a parameter which is under the control of the system designer and which determines the bit rate and the loss of visual fidelity. More specifically, it determines the bin width used in the quantization of the transform coefficients. The higher the value of Q, the larger the quantization bins, and hence the higher the compression rate and distortion introduced. Typical values of Q are in the range 10-100.

In a sense, Q could be considered a Type I metric - it increases with increasing distortion, and one might hypothesize that there is a fixed value of Q for which the distortion is just noticeable, over a wide range of images. Although this is not the case, as discussed below, it is worthwhile to examine relationships between the metrics described above and Q, particularly in threshold experiments. It will be shown that the threshold Q is highly correlated with certain Type II metrics, a result with significant engineering implications.

3. Experimental Methods

There were two psychophysical experiments, one threshold and one suprathreshold. The objective of the threshold experiment was to determine the threshold of distortion at which the observer can just distinguish a distorted image from the original. In the suprathreshold experiment, observers were asked to assess the relative quality of a set of image sub-blocks which were coded to a constant bit rate (0.5 bits per pixel). As pointed out in the Introduction, a good distortion metric should be able to predict the observers' responses for both of these experiments.

3.1. Observers

A total of 11 individuals (the authors and others involved in the project) served as observers in these experiments. The number of observers that participated in each experiment is specified below. All had normal or corrected-to-normal visual acuity.

3.2. Stimuli

Apparatus. Stimulus presentations were controlled by a Sun 3/260 computer. Images were presented on a 19-inch color monitor (Pixar II, Sony GDM-1950). The largest images, 13.7 by 13.7 cm, were centered on the display screen and subtended 21.4 degrees of visual angle at a viewing distance of 35 cm. The luminance of the display was calibrated such that it varied linearly from 1.85 cd/m^2 (all pixels black) to 42.54 cd/m^2 (all pixels white). This range was divided into 256 equal gray-scale values. Observers viewed the display with their head held steady by a chin rest and forehead support. The display was viewed binocularly under flourescent room lights.

Images. Gray-scale images were digitized from black and white photographs at 8 bits/pixel (bpp). The content of images used for these experiments included portraits, still life, and outdoor scenery. The full-size images used in Experiment 1 were 512×512 pixels in size. The spatial

sampling rate at the viewing distance specified above was thus 23.1 pixels/degree. Image presentation was synchronized with the vertical retrace of the display and was completed within 16.67 ms (the refresh rate of the display was 60 Hz).

In addition to the original images described above, 40 different compressed images were computed for each original, at bit rates ranging from 0.1 bpp to 4.0 bpp, in 0.1 bpp increments. Images were compressed and de-compressed according to the JPEG draft standard [1] with sequential coding and Huffman coding options. The luminance quantization table and the Huffman code tables used were from the Examples section of [1]. These compressed/decompressed (i.e. distorted) images were compared with the original images in the two experiments described below.

In the second part of Experiment1, and in Experiment 2, sub-blocks of size 128×128 pixels were taken from the original and compressed images described above. These sub-blocks were used in order to mitigate the effects of attention shifting in the larger images.

3.3. Procedure - Experiment 1 (Threshold)

Two-alternative, temporal forced-choice trials in conjunction with a staircase procedure were used to measure distortion thresholds. Each trial consisted of two 500-msec observation intervals, separated by 500 msec. Trials were preceded by a 1-sec interval during which a fixation mark was displayed on the otherwise blank screen. One of the observation intervals contained the original image and the other contained a distorted version of the same image. The distorted image was presented randomly in the first or second interval with equal probability. The observer's task was to indicate, by pressing one of two buttons, the interval containing the distorted image. The observer received feedback in the form of coded tones to indicate correct/incorrect responses. Each response initiated the next trial.

The staircase procedure was used to determine the level at which the distortion produced by compression could be detected at the 70 percent correct level (see [13])*. (It is implicitly assumed that the observer's ability to detection distortion is monotonically decreasing with increasing bit rate, or equivalently, is monotonically increasing with increasing Q-factor.) Two consecutive correct responses produced a reduction in the bit rate of the distorted image (i.e. an increase in the distortion), and one incorrect response produced an increase in the bit rate. Initially, the distorted image was coded at 0.1 bpp, for which the distortion was clearly visible, and the staircase moved in relatively large steps, 1.6 bpp. Following the second, fourth, sixth, and eighth reversals, the step size was reduced by one half. Following the eighth reversal, the step size remained constant at 0.1 bpp. The staircase terminated after 11 reversals had occurred. Distortion threshold was defined as the mean bit rate associated with the last three reversals.

This staircase procedure was used for 8 full-size (512×512) images and for 12 smaller image sub-blocks (128×128). Each staircase typically consisted of 40-50 trials, and each daily testing session lasted between 45 and 60 minutes. Three staircases were performed for each full-sized image and the mean of the last two were used for the analyses presented below. A single staircase was completed for each of the image sub-blocks. For full-size images, the order within a session was randomized such that a staircase was completed for each image in the group before beginning the next staircase for a given image. Thresholds for each of the eight full-sized images were obtained from five observers and thresholds for each of the 12 image sub-blocks were obtained from nine observers.

A modification of the standard staircase procedure was also employed for each of the full-sized images. After the standard procedure was completed on a given image, informal debriefing of the observers provided an indication of which region of the image most clearly revealed the distortion. This information was then used in a second set of staircases to direct observers'

attention to the region of each image in which the distortion was most easily detectable. Distortion thresholds were obtained with this modified staircase procedure from the same five observers (the "Hints" condition.)

3.4. Procedure - Experiment 2 (Suprathreshold)

Four of the images used in Experiment 1 were subdivided into 16 128×128 sub-blocks. Each full-sized image was compressed/decompressed at 0.5 bpp before being sub-divided. A modified method of paired comparisons was then used to rank the image sub-blocks with respect to the amount of perceived distortion. Image sub-blocks were ranked in groups of 16, such that each block was compared only to other blocks taken from the same image.

In the standard method of paired comparisons, every item is compared with every other item in a group, resulting in a large number of comparisons (120 in our case). In our modified method, a smaller set of comparisons was performed, from which a complete ranking can be inferred. An implicit assumption here is that the comparisons are consistent, that is, if item A is preferred to item B, and B is preferred to C, then A will be preferred to C.

Each paired comparison was conducted in the following way. Four image sub-blocks were displayed, as shown in Figure 3.1. The four images were the original and distorted versions of sub-blocks i and j, as shown. The subjects were asked to determine subjectively which of the two blocks appeared more distorted when compared with the original.

The result of each comparison was used in conjunction with a sorting algorithm to arrive at a complete ranking of the 16 sub-blocks. We used the HEAPSORT algorithm [14], which can sort a list of N objects with $O(N \log N)$ comparisons. This algorithm was implemented as a computer program running on the Sun 3/260; each comparison of items i and j in this program would initiate a comparison of distorted sub-blocks as described above. The exact number of comparisons cannot be determined *a priori*, but in our experiment approximately 60-70 comparisons

were required to sort the 16 sub-blocks. The result for each observer would be a ranking of the the 16 sub-blocks, ordered from least distorted to most distorted. An additional step of allowing the subjects to modify their ranking by making adjacent pairwise permutations was included at the completion of sorting program. In this way it was assumed that the subjects were satisfied with their completed rankings.

4. Results

4.1. Experiment 1

Mean threshold bit rates for the full-sized images, as estimated with the standard staircase procedure and the staircase procedure with hints, are presented in Figure 4.1. The mean threshold bit rate across images for the standard staircase was 1.89 bpp, and the mean across images for the staircase with hints was 2.36 bpp. A two-way analysis of variance (see Table 4.1), image by procedure, indicates that the improved performance with hints is statistically significant, and that the difference between images is marginally significant. There was no significant interaction effect.

In order to evaluate the behavior of both the Type I and Type II metrics, data were analyzed for the 12 128×128 image sub-blocks (3 each from 4 of the larger images). For each of these sub-blocks, several quantities were computed: 1) the actual bit rate at the threshold of just-noticeable distortion, 2) the Q-factor for the JPEG algorithm, also at threshold, 3) the value of each of the Type I metrics, at threshold, and 4) the value of the Type II metrics for each of the original sub-blocks.

The variabilities of the Type I metrics at threshold were analyzed. Because the actual values of the metrics have no intrinsic meaning (they were scaled somewhat arbitrarily in software to yield numerical results in the 1-10 range) the coefficient of variance (standard deviation divided by mean) was used to to quantify the variability of each metric. These are shown in Table 4.2.

DCON achieved the lowest coefficient of variance (0.301).

In Table 4.3 the coefficients of variance for the ensemble of Type II metrics are shown. In this case, we expect a good metric to have a large coefficient of variance, since this would imply that the measure is bringing out significant differences in the qualities of the original image blocks. MI and SS/MI had coefficients of variance of 0.700 and 0.916, respectively.

Because we expect the ideal Type I metric to be a constant over all compressed images when evaluated at the threshold of just-noticeable distortion, the experimental error in the determination of such a metric should not be correlated with system variables such as threshold bit rate and threshold Q-factor. If this were not the case, it would imply that there was residual information in the Q-factor or bit rate which the metric itself was not conveying. For this reason, in our statistical analysis we examined correlations between Type I metrics, evaluated at threshold, and the threshold bit rates.

Table 4.4 shows the correlation analysis for Type I metrics versus threshold bit rate. All the correlation values are quite low. This is illustrated further in Figure 4.2, in which the DCON metric is plotted as a function of bit rate for each of the 12 sub-blocks. The circles indicate the threshold point of just-noticeable distortion for each of the curves; there is no apparent pattern in the collection of threshold points. Figure 4.3 gives similar curves for the DCON metric vs. Q-factor; again the threshold points are plotted. In this figure a pattern is evident for the threshold points: most of the points lie close to convex curve of decreasing slope. Similar patterns were evident for other metrics. From this pattern it can be concluded that the DCON metric by itself will not summarize everything about the visibility of the distortion, since knowledge of the Q-factor evidently would provide more information toward predicting the threshold point.

While one would hope that a good Type I metric would, at threshold, be uncorrelated with bit rate and Q-factor, a good Type II metric *should* be correlated with these latter quantities. This

is because the intrinsic qualities of the image should or may reveal something about the level of distortion that can be tolerated. In Table 4.5 the correlations between the Type II metrics and the threshold bit rate are presented, and in Table 4.6 the correlations between the metrics and the threshold Q-factors are presented. Note the particularly high correlation between MI and threshold Q-factor ($r = 0.836$). To illustrate this further, in Figure 4.4 the threshold Q-factor is plotted as a function of MI for the 12 image sub-blocks. A regression line is also plotted.

4.2. Experiment 2

For the analysis of the data from the block-sorting experiment nonparametric methods of rank correlation, as described in Kendall [15], we used. Only the ranks assigned to the 16 sub-blocks according to the distortion metrics were considered, rather than the numerical values of the metrics themselves. In this way each metric can be treated as an observer. The objective was to determine which metric gives rankings which are in agreement with the human observers.

The rank data were examined to determine whether or not there was agreement among the 9 observers. Kendall proposes a test statistic, called the "coefficient of concordance", which is a measure of this agreement. This statistic, appropriately scaled, is approximately subject to a χ^2 distribution, and can be used to reject the null hypothesis that the rankings from the 9 observers were chosen independently and uniformly from the set of 16! possible permutations. The coefficient W itself, which can take on values between 0 (no agreement) and 1 (identical rankings), was in the range 0.76 (PATH) to 0.89 (YOSEMITE) ($p < 0.01$). Thus it was concluded that there was significant agreement among the rankings of the 9 observers.

To compare the rankings given by the metrics to those given by human observers, the Spearman ρ statistic was used. This is a measure of the agreement between two rankings of N objects, normalized so that $-1 \leq \rho \leq +1$. A value of 1 indicates that the rankings are identical (complete agreement), while a value of -1 indicates that one ranking is exactly the reverse of the

other (complete disagreement). To evaluate a given distortion metric, the Spearman ρ statistic was computed for the observer/metric pair for each of 9 observers, then the average taken across observers.

Mean Spearman ρ statistics (rank versus metric) across observers for each image are plotted in Figure 4.5 for each of the metrics tested. Inspection of this figure shows the substantial variability obtained, not only between metrics, but particularly between the same metrics applied to different images. No single metric was consistently highly correlated with psychophysical ranking for all images; for some metrics, correlations were positive for some images and negative for other images. In general, Type I metrics tended to have lower, more variable correlations than Type II metrics. Overall, of the Type I metrics, DCON was most highly correlated with observer rankings (mean $\rho = 0.6814$) suggesting that the amount of perceived distortion increased with this particular pointwise distance measure. However, other metrics (MSENL and MSE in particular) were more highly correlated with some of the images.

Type II metrics, with the exception of Spectrum Slope, tended to be more highly and more consistently correlated with observer rankings than Type I metrics. Overall, SS/MI had the highest mean correlation (mean $\rho = 0.763$); MI was also highly correlated with observer ranking (mean $\rho = -0.719$).

4.3. Comparison of Experiment 1 and Experiment 2

The 12 128×128 blocks used in the threshold experiment were chosen based on an informal assessment of the visibility of the distortion at 0.5 bpp. From each of the 4 images, three sub-blocks were chosen to have low, medium, and high perceived levels of distortion, respectively. This informal assessment was borne out by the rankings assigned to the 3 sub-blocks within the group of 16 sub-blocks for the respective images. For example, a sub-block with low

perceived distortion had a low median rank, and a sub-block with high perceived distortion had a high median rank.

To see if the results of the two psychophysical experiments were consistent, median ranks of each of the 12 sub-blocks from the suprathreshold experiment were compared with the threshold bit rates and Q-factors for these same blocks. (Note that the rank of a particular sub-block is only meaningful in comparison to the ranks of other sub-blocks from the same image.) These results are summarized in Figure 4.6. The threshold bit rates did not exhibit a pattern which is consistent with the suprathreshold rankings. On the other hand, the threshold Q-factors within each group of 3 sub-blocks had the opposite ordering of the median ranks, in all 4 cases. This implies that those blocks for which the distortion is most objectionable at 0.5 bpp are the same blocks for which the Q-factor must be lowered in order to reach the distortion threshold.

5. Discussion of Results

1. The results of the threshold experiment for full 512×512 images indicate that high-quality images can be obtained using the JPEG algorithm at around 2 bpp. The naive observer will be unable to detect distortion due to compression at this level, at least for ordinary photographic images. Under very careful scrutiny, and perhaps with a little practice and prior knowledge, the observer can detect distortion at slightly higher bit rates.

2. Because of the issue of global versus local attention in the interpretation of images, smaller image blocks were used in those experiments involving the computed metrics. In so doing, one must remain cautious about attaching too much significance to the actual bit rates associated with these blocks. The JPEG compression algorithm, as implemented with the Huffman coding option, is a variable bit rate code. For a fixed Q-factor, or overall image bit rate, there can be considerable inter-block variability in bit rates. While of course the average bit rate is the single most important parameter in any data compression scheme, it is not safe to assume

that a rate of R bpp over some small region of the image is representative of image quality over that region. This variability in bit rate can explain in part the low correlations in Table 4.1. For small image blocks, the Q-factor is a more reliable measure of information (or information loss) since it determines the coefficient quantization bin width, and is a constant. Note, for example, the high correlations between the threshold Q-factor and Type II metrics, and the agreement between suprathreshold rankings and threshold Q-factors.

3. MSE is a poor predictor of observer response. Note the high variability of the measure at threshold, and the results of the suprathreshold experiment as shown in Figure 4.5.

Another curious fact about MSE is that, when the Q-factor or the average bit rate is held fixed, then MSE tends to correlate *positively* with bit rate over small image blocks. This may seem counter-intuitive at first; however, on further reflection, one could envision a scenario in which image blocks which are less probable according to the Huffman code tables could simultaeously suffer increased squared-error and increased bit rate. For example MSE and BITS exhibit similar behavior in Figure 4.5, although neither one is a good predictor of observer responses. Recall that BITS is the true block bit rate when the average bit rate over the 16 image blocks was 0.5 bpp.

4. Of the Type I metrics, DCON appears to have reasonable performance, both in terms of the variability at threshold, and in agreement with the suprathreshold rankings. What is most interesting about this result is not the particular metric, but the fact that it is a simple *pointwise* measure which does not take into account spatial properties or local contrast in the image. Because of the pattern in the relationship between DCON and threshold Q-factor, as shown in Figure 4.3, we cannot assume that this measure summarizes all of the relevant information about distortion thresholds. Indeed, Figure 4.3 leads us to believe that some combination of the DCON metric and the Q-factor could lead to excellent prediction of distortion thresholds. Still, we are

encouraged to search further for other pointwise measures. Although it is possible to find metrics which provide an optimal fit to the psychophysical data presented here, it is recommended instead that the search for the "best" pointwise measure be accompanied by a larger experiment involving more observers and more images.

5. Quite similar results were obtained for MSENL and MANNOS, both in the correlation with suprathreshold rankings, and in variability at threshold. The rank correlations were similar to those of DCON, whereas the threshold variabilities were larger by about a factor of 2. The difference between these two metrics is that MANNOS employs a spatial frequency weighting which adds significantly to the computational complexity. The present results suggest that there is no advantage to be gained by this frequency weighting. Also, proper use of this frequency weighting requires exact knowledge of the viewing distance and pixel size, which can vary depending on the viewing situation. Metrics which are especially sensitive to viewing conditions should probably be avoided.

6. The success of mean intensity in predicting observer ranking in the suprathreshold experiment, and the Q-factor in the threshold experiment, was remarkable. It was not especially surprising, however, in light of the physiological evidence that perceived differences in intensity are related to mean luminance, and also our informal debriefings with the observers which indicate that distortion is indeed much more visible in darker regions of the images. The engineering implications of this result may be significant, as this particular quantity is trivial to compute. Coding schemes which take into account the mean intensity of a block of pixels can easily be envisioned: for example, one could vary the Q-factor in the JPEG algorithm according to local mean luminance. The mean luminance for an 8×8 block of pixels is simply the DC coefficient from the Discrete Cosine Transform (DCT), so the information is already available for use.

7. There does not seem to be a masking effect associated with the distortion introduced by the JPEG compression algorithm. Note, for example, that when the LCON metric is large, the distortion is actually *more* visible, as evidenced by the lower threshold Q-factor. The reason for this is the fact that the JPEG algorithm suppresses high-frequency components within in each 8×8 block, and thus in the distorted image there is less contrast to create a masking effect. Observers will detect and respond to this change in local contrast. This phenomenon also forms the rationale behind the LIP metric, which measures pointwise differences in local contrast rather than intensity.

8. The SS/MI metric is a good predictor of the suprathreshold rankings. Still, we do not recommend that it be considered as a viable metric, for the following reasons. First, SS by itself has no value in predicting observer rankings, and second, the coefficient of variance of SS is very low (0.155). This suggests that the SS metric carries very little information about the image block, and that the correlation between SS/MI and observer rankings can be explained almost entirely by the MI factor. Also, SS is extremely time-consuming to compute, involving (as does MANNOS) a two-dimensional Fourier transform, and several additional multiplications and additions per pixel.

6. Summary and Conclusion

We have presented results of two psychophysical experiments designed to evaluate computable distortion metrics for photographic images. It was determined that mean-square-error is not a good distortion measure. Other simple pointwise distance measures (MSENL, DCON) correlated well with observer suprathreshold rankings; DCON in particular exhibited the lowest variability at the distortion threshold. Computationally intensive metrics such as MANNOS and LIP do not seem to be worth the extra effort, even though they can be justified on physiological

grounds. The use of the metrics based on the spatial frequency properties of the human visual system is especially discouraged, requiring as it does precise knowledge of the viewing conditions.

The importance of mean intensity in the assessment of image distortion introduced by the JPEG algorithm was most surprising. What is interesting about this is that, whereas much attention was paid to the spatial frequency response of the human visual system in the development of the standard, no consideration has been given to even simple pointwise nonlinear models. It is our conclusion that the nonlinear response to luminance is a dominant effect which should not be ignored if one wants to incorporate the human visual system into algorithm design. Our specific recommendation is that a simple pointwise measure such as DCON be used as a distortion measure in further image compression and rate-distortion studies, and that the search for better pointwise metrics continue.

7. Acknowledgements

The authors would like to express their grateful appreciation to students Steven Attwood, Ladan Hedayati, and Thomas Ku, for their technical assistance in maintaining the software for the psychophysical experiments, coordinating the experimental sessions with the subjects, and some manipulation of the data. Steven Attwood and Ladan Hedayati were supported by Southwestern Bell Technology Resources Inc. Thomas Ku was supported by the National Science Foundation under the Summer Undergraduate Research Assistantship program.

References

1. Joint Photographic Experts Group (JPEG) ISO CD 10918, "Digital Compression and Coding of Continuous-Tone Still Images", ISO/IEC JTC1/SC2/WG10, January 1991.
2. W. Pratt, *Digital Image Processing*, John Wiley and Sons, New York, 1978.
3. R. Clarke, *Transform Coding of Images*, Academic Press, London, 1985 (Chapter 6, The Human Visual Response).
4. T. Stockham, Jr., "Image Processing in the Context of a Visual Model", *Proc. IEEE*, vol. 60, no. 7, July 1972.
5. C. Hall and E. Hall, "A Nonlinear Model for the Spatial Characteristics of the Human Visual System", *IEEE Trans. Systems, Man, and Cybernetics*, vol. SMC-7, no. 3, March 1977.
6. D. Sakrison, "On the Role of the Observer and a Distortion Measure in Image Transmission", *IEEE Trans. Communications*, vol. COM-25, no. 11, November 1977.
7. J. Limb, "Distortion Criteria of the Human Viewer", *IEEE Trans. Systems, Man, and Cybernetics*, vol. SMC-9, no. 12, December 1979.
8. K.-H. Tzou, *A Physiologically Based Human Visual System Model for Threshold Vision and Image Processing*, D.Sc. dissertation, Washington University Sever Institute of Technology, August 1983.
9. J. Mannos and D. Sakrison, "The Effects of a Visual Fidelity Criterion on the Encoding of Images", *IEEE Trans. Information Theory*, vol. IT-20, no. 4, July 1974.
10. M. Jourlin and J. Pinoli, "A Model for Logarithmic Image Processing", *J. Microscopy*, vol. 149, pt. 1, pp. 21-35, January 1988.
11. J. Brailean, B. Sullivan, C. Chen, and M. Giger, "Evaluating the EM algorithm for image processing using a human visual fidelity criterion", *Proc. ICASSP 91*, pp. 2957-2960, May 1991.
12. E. Cargill, K. Donohoe, G. Kolodny, J. Parker, R. Zimmerman, "Analysis of Lung Scans Using Fractals".
13. G. Wetherhill and H. Levitt, "Sequential Estimation of Points on a Psychometric Function", *Brit. J. Mathematical and Statistical Psychology*, vol. 18, pp. 1-10, 1965.
14. A. Aho, J. Hopcroft, and J. Ullman, *The Design and Analysis of Computer Algorithms*, Addison-Wesley, Reading, Mass., 1974.
15. M. Kendall, *Rank Correlation Methods*, Charles Griffin and Co., London, 1970.

Tables

Factor	df	F-Ratio	Probability
Image	7,56	2.633	0.020
Hints	1,56	8.305	0.006
Image x Hints	7,56	0.393	0.902

Table 4.1. Analysis of variance for threshold bit rates for the full-sized images.

Metric	Coef. Var.
MSE	0.765
MSENL	0.648
MANNOS	0.661
LIP	0.388
DCON	0.301

Table 4.2. Variability of Type I metrics at threshold.

Metric	Coef. Var.
MI	0.700
SS	0.155
SS/MI	0.916
LCON	0.297

Table 4.3. Variability of Type II metrics.

Metric	r
MSE	-0.319
MSENL	0.011
MANNOS	0.280
LIP	0.044
DCON	0.137

Table 4.4. Correlation analysis for Type I metrics at threshold versus threshold bit rate.

Metric	r
MI	-0.208
SS	-0.189
SS/MI	-0.261
LCON	0.583

Table 4.5. Correlation analysis for Type II metrics versus threshold bit rate.

Metric	r
MI	0.836
SS	-0.248
SS/MI	-0.506
LCON	-0.766

Table 4.6. Correlation analysis for Type II image metrics versus threshold Q-factor.

Figure Legends

Figure 3.1. Display of image sub-blocks used in suprathreshold experiment.

Figure 4.1. Mean threshold bit rates for each of the full-sized images. Results of the standard staircase procedure are represented by white bars and the staircase procedure with hints by the shaded bars. Error bars represent + 1 standard error of the mean.

Figure 4.2. DCON metric as a function of bit rate for the 12 image sub-blocks. Circles indicate the distortion threshold averaged across the observers. For bit rates below the threshold the distortion is visible, and for bit rates above the threshold the distortion is not visible.

Figure 4.3. DCON metric as a function of Q-factor for the 12 image sub-blocks. Circles indicate the distortion threshold averaged across the observers. For Q-factors above the threshold the distortion is visible, and for Q-factors below the threshold the distortion is not visible.

Figure 4.4. Threshold Q-factor as a function of MI metric for 12 image sub-blocks. The dashed line is the regression line.

Figure 4.5 Mean Spearman ρ coefficients for each metric versus the observers' ranks. Each of the images is represented with a different shading pattern. The top panel contains results for Type I metrics and the bottom panel contains results for Type II metrics.

Figure 4.6. *Top Panel:* Mean threshold bit rate for each of the image blocks. *Middle Panel*: Mean threshold Q-factor for each of the image blocks. *Bottom Panel:* Median rank for each of the image blocks. A lower rank indicates a smaller amount of perceived distortion. The labels along the horizontal axis indicate the names of the images and the identifying number of the sub-block within that image.

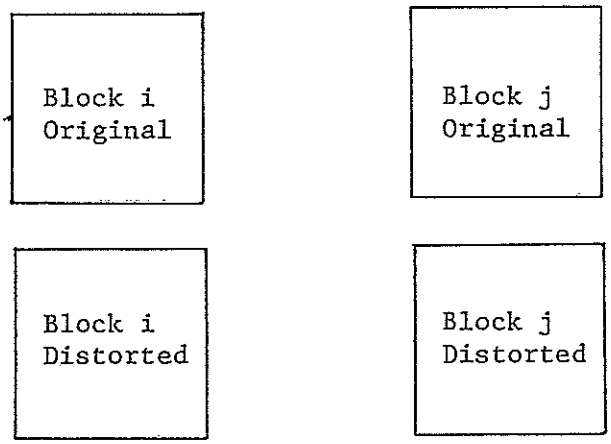

Figure 3.1. Display of image sub-blocks used in suprathreshold experiment.

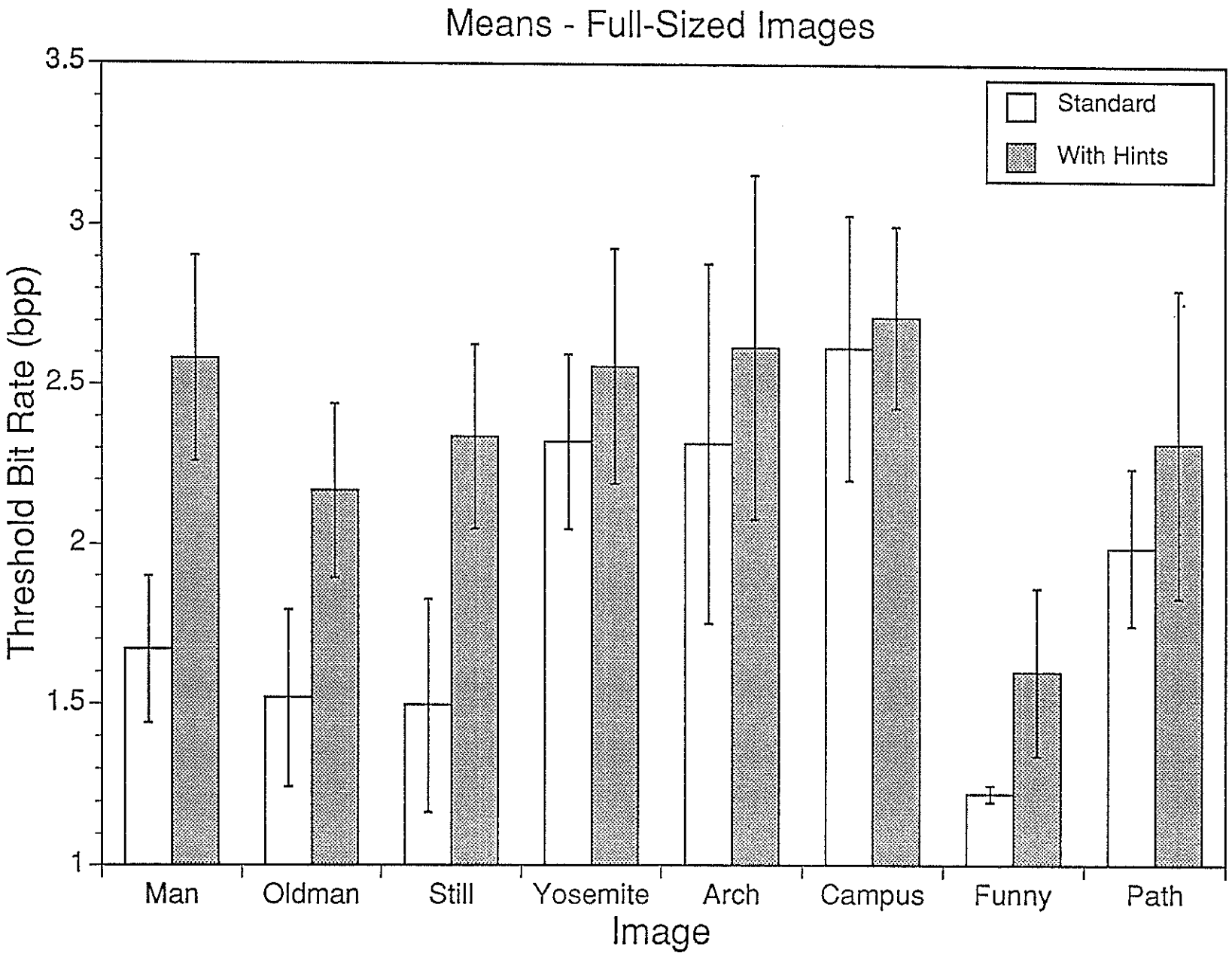
Means - Full-Sized Images
Standard
With Hints
Threshold Bit Rate (bpp)
3.5
3
2.5
2
1.5
1
Man
Oldman
Still
Yosemite
Arch
Campus
Funny
Path
Image

Figure 4.1

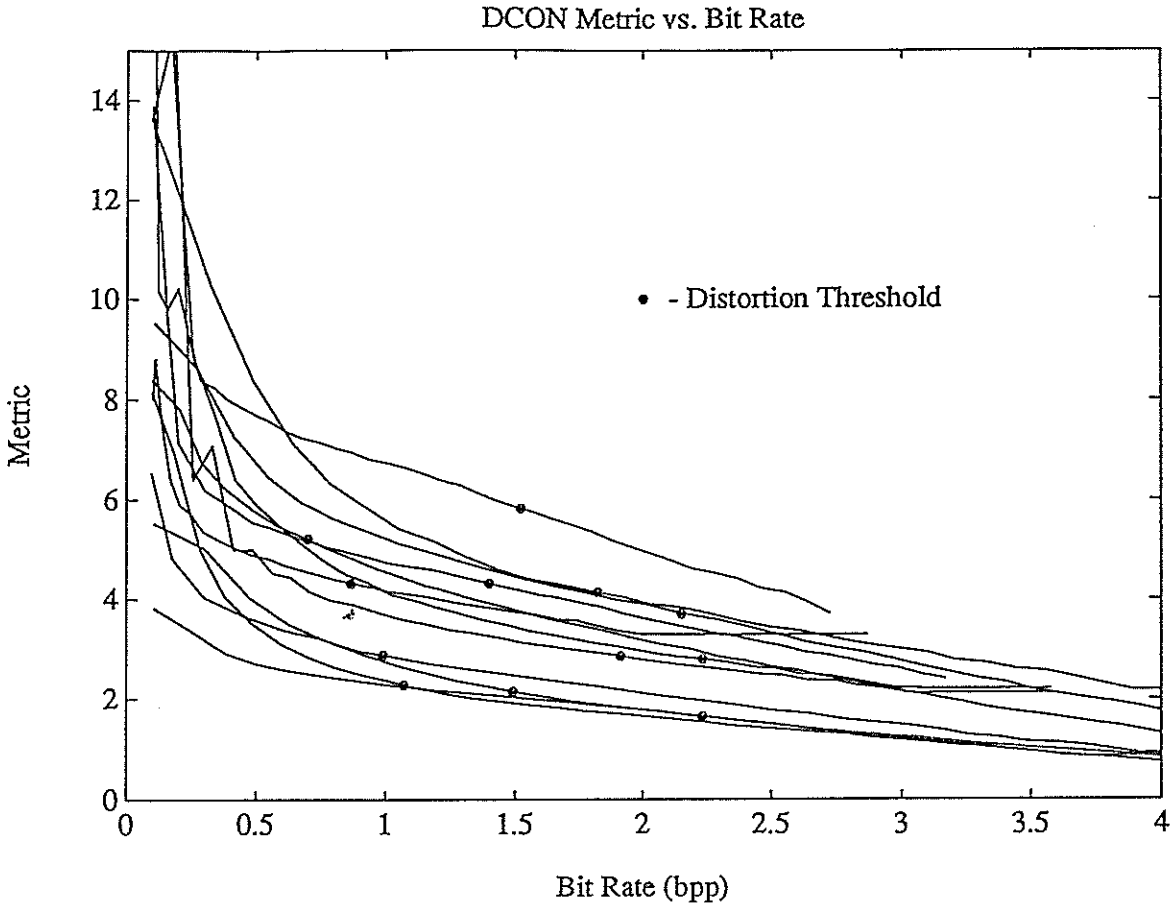
DCON Metric vs. Bit Rate
Metric
Bit Rate (bpp)
• - Distortion Threshold
0
2
4
6
8
10
12
14
0
0.5
1
1.5
2
2.5
3
3.5
4

Figure 4.2.

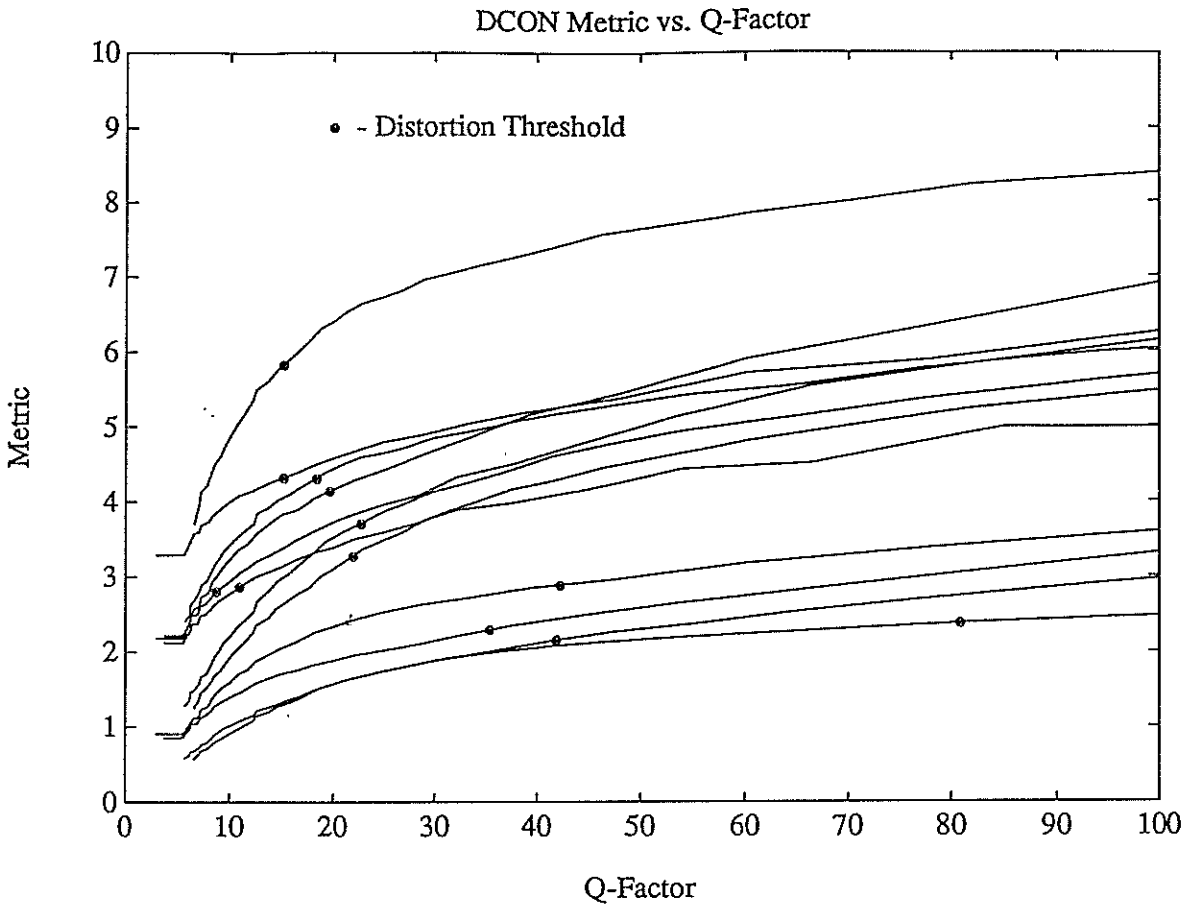
DCON Metric vs. Q-Factor
- Distortion Threshold
Metric
Q-Factor
0
1
2
3
4
5
6
7
8
9
10
0
10
20
30
40
50
60
70
80
90
100

Figure 4.3.

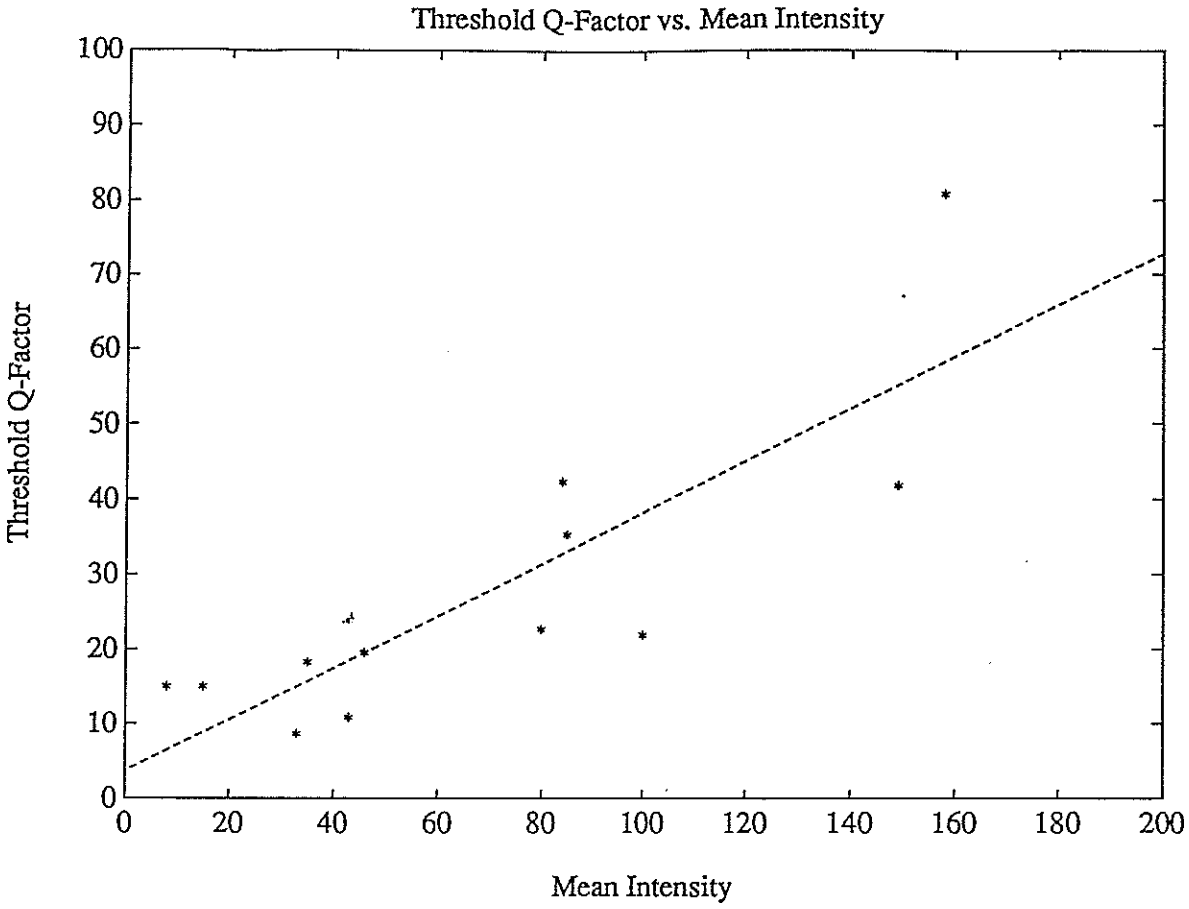
Threshold Q-Factor vs. Mean Intensity
Threshold Q-Factor
Mean Intensity
0
10
20
30
40
50
60
70
80
90
100
0
20
40
60
80
100
120
140
160
180
200

Figure 4.4.

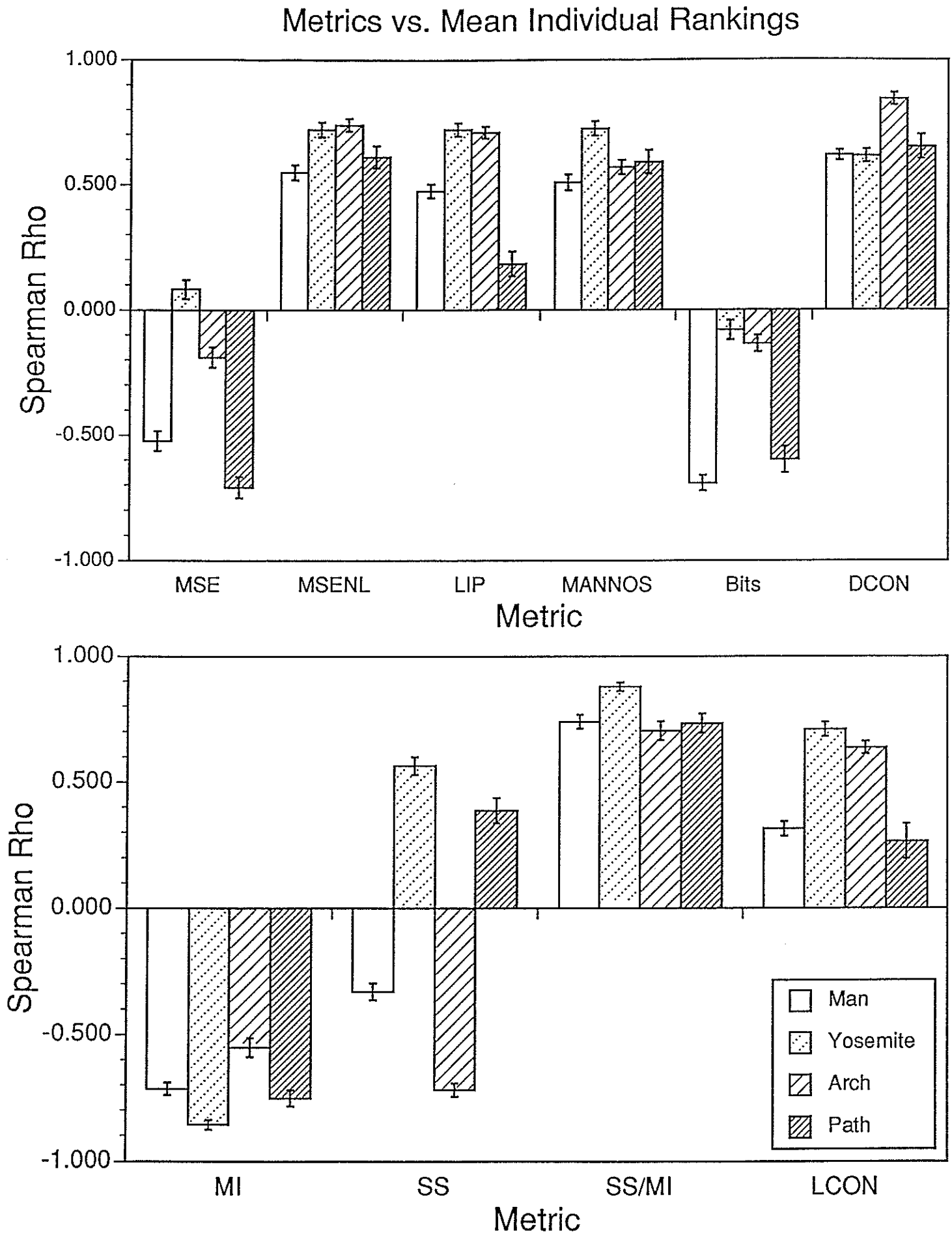
Metrics vs. Mean Individual Rankings
1.000
0.500
0.000
-0.500
-1.000
Spearman Rho
MSE
MSENL
LIP
MANNOS
Bits
DCON
Metric
1.000
0.500
0.000
-0.500
-1.000
Spearman Rho
MI
SS
SS/MI
LCON
Metric
Man
Yosemite
Arch
Path

Figure 4.5

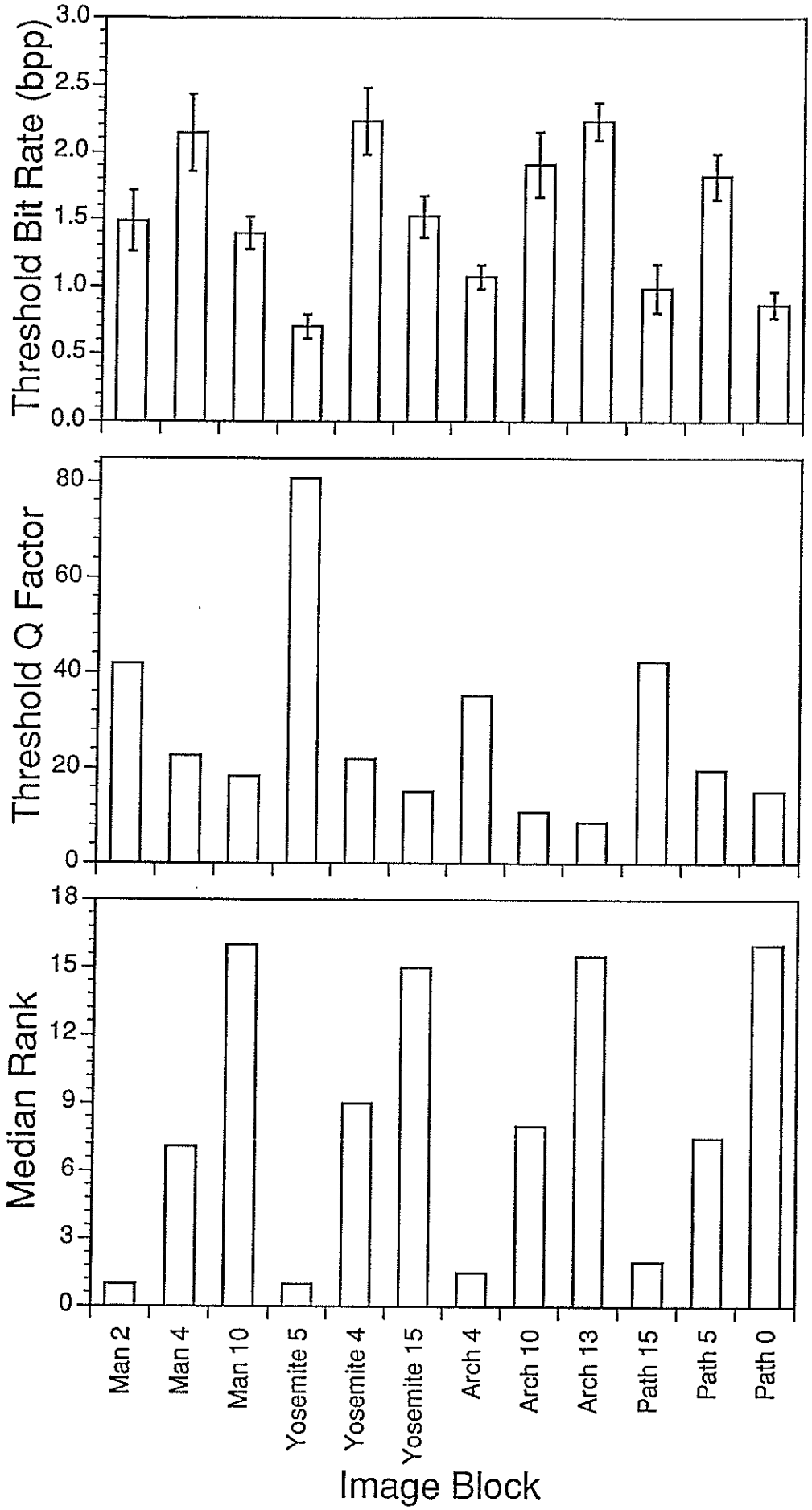
Threshold Bit Rate (bpp)
3.0
2.5
2.0
1.5
1.0
0.5
0.0
Threshold Q Factor
80
60
40
20
0
Median Rank
18
15
12
9
6
3
0
Man 2
Man 4
Man 10
Yosemite 5
Yosemite 4
Yosemite 15
Arch 4
Arch 10
Arch 13
Path 15
Path 5
Path 0
Image Block

Figure 4.6