High Resolution and Image Compression Using the Discrete Cosine Transform

Stanley A. Klein
School of Optometry, U.C. Berkeley, Berkeley, CA 94720

ABSTRACT

The information gathering capacity of the visual system can be specified in units of bits/min^2. The fall-off in sensitivity of the human visual system at high spatial frequencies allows a reduction in the bits/min^2 needed to specify an image. A variety of compression schemes attempt to achieve a further reduction in the number of bit/min^2 while maintaining perceptual losslessness. This paper makes the point that whenever one reports the results of an image compression study, two numbers should be provided. The first is the number of bits/min^2 that can be achieved using properties of the human visual system, but ignoring the redundancy of the image (entropy coding). The second number is the bits/min^2 including the effects of entropy coding. The first number depends mainly on the properties of the visual system, the second number includes, in addition, the properties of the image. The Discrete Cosine Transform (DCT) compression method is used to determine the first number. It is shown that the DCT requires between 16 and 24 bits/min^2 for perceptually lossless encoding of images, depending on the size of the blocks into which the image is subdivided. In addition, the efficiency of DCT compression is found to be limited by its susceptibility to interference from adjacent maskers. The present analysis suggests that the visual system requires many more bits/min^2 than the results of other researchers who find that .5 bits/min^2 are sufficient to represent an image without perceptible loss.

1. INTRODUCTION

A perceptually lossless reproduction of a visual scene is one that can not be discriminated from the original by a human observer. The amount of information in a display can be quantified by the number of bits/min^2 needed to specify the image. This paper asks what is the minimum number of bits per min^2 required for a perceptually lossless reproduction of a visual scene[1]? We use bits/min^2 rather than bits per pixel to make the analysis independent of pixel size or viewing distance.

The calculation of the number of bits/min^2 is not merely an abstract intellectual exercise. It is intimately connected with research on practical image compression, display technology, and metrics for assessing image quality. In each of these areas, knowledge of human visual performance is of central importance since the goal is to reduce the number of transmitted and displayed bits while maintaining high image quality by eliminating information in the signal that the human is presumably unable to see. Research in this area is becoming increasingly popular due to the severe bandwidth requirements of high definition television.

The purpose of the present article is to show how characteristics of human vision place limits on the amount of compression possible without perceptibly degrading an image. It is commonly claimed[1,2] that image compression is able to reduce the encoded information to less than 1 bit/min^2 without losing information that can be seen by the visual system. The findings to be reported in this paper, on the other hand, are that more than 15 bits/min^2 are needed. Why are our results so different from others? It is not simply that our criterion for image quality is more conservative than others. The two-alternative forced choice methods used by Watson[1] and Stein[3] are just as sensitive in revealing loss of image quality as our methods. We believe that our approach differs from that of others in the way stimuli are chosen. The initial images used by others are often degraded both in resolution (1.5 min/pixel) and in grey scale (8 bits/pixel), well below the discrimination capabilities of the human visual system (see Section 2). Another factor is that most image compressors use natural scenes. Natural scenes can result in lower values of bits/min^2 for two reasons: redundancy and non-optimal conditions for peak visual performance. It is the latter problem that is the focus of the present research. Rather than using natural scenes we examine simpler stimuli, chosen to reveal the full capability of the human visual system. Simple dart-board spoke patterns and frequency modulated square wave patterns[4,5] as well as shallow ramps are examples of patterns that we feel should be used in comparing compression algorithms.

There are three basic steps in image compression: 1) a transformation (e.g. Fourier), 2) application of the human visual system to eliminate invisible information by choosing appropriate contrast quantization steps, 3) entropy coding to efficiently transmit the information. Previous researchers[6,7] have lumped all three steps together and report a single value for bits/min^2 for a given image. This lumping, however, loses the distinction between how much of the savings of bits/min^2 is due to the characteristics of the visual model, and how much is due to the redundancy of the particular image that is chosen.

The compression literature would be clearer if two different meanings of "bits/min^2" were distinguished and if future researchers would report both numbers. First is the bits/min^2 due to the characteristics of the visual system without entropy coding. This is the value corresponding to the maximum number of bits/min^2 that can be processed by the visual system. Second is the value after the redundancy of the image is included. This second number has relevance to the practical issues of how to best transmit natural images. We believe, however, that the first number is also useful for estimating the "worst case" scenario in which, for example, high density text is embedded in the image to be transmitted. This first number is also important for comparing the variety of assumptions that different researchers make about the human observer. For this comparison it would be useful to know how many bits/min^2 are needed before the entropy coding stage that is so dependent on the particular image. That is the goal of this paper.

Our goals differ from those of others doing research in image compression. Research using natural scenes is relevant for high definition television compression, but may have misleading implications concerning the limitations of human vision. On the other hand, our results may not be directly applicable to compressing television images, but they clarify the role of the human observer. If the visual system is able to discriminate more than 15 bits/min^2, as we claim, then scenes with fewer bits/min^2 do not adequately test the full capacity of vision. Having an accurate estimate of the number of bits/min^2 is also useful for comparing psychophysics to anatomy and physiology. Anatomical evidence shows that in the central fovea there are approximately 4 cones/min^2[8] and between 8[9] and 12[10] ganglion cells/min^2. Thus our psychophysical finding of 16 bits/min^2 implies that cones must transmit about 4 bits/min^2 and ganglion cells must transmit between 1 and 2 bits/min^2. The proper discussion of this issue which includes temporal factors of how many bits/min^2 per sec are being transmitted will be examined in future research.

2. REQUIREMENTS OF A "PERFECT" DISPLAY

Before getting to the question of how many bits/min^2 are needed after compression we first ask how many bits/min^2 are needed before compression. We will argue that approximately 100 bits/min^2 are required for a perceptually lossless display. This result follows from the number of bits/min^2 being the product of two factors: the number of bits/pixel and the number of pixels/min^2. These two factors are now examined briefly. A more complete analysis is presented elsewhere[11]

2.1. The minimum number of pixels/min^2 and super-Nyquist sampling.

The maximum pixel size is governed by the resolution threshold of human vision. In order for a very thin line not to appear blurred, the pixel size must be smaller than the resolution threshold. This is because a thin line is sometimes represented by a single pixel and sometimes (when the location of the line falls between pixels) by a pair of half-intensity pixels. This procedure avoids the "jaggies" that could be detected by a vernier cue if a thin line with a slight tilt away from horizontal was always one pixel in width. Recent data[12] show that the resolution threshold is .33 min (3 pixels/min). Therefore the display must have at least 9 pixels/min^2 in order to be able to represent thin lines without perceptible loss. Note that 3 pixels/min is a denser sampling than would be expected from the Nyquist limit based on the 60 c/deg cutoff of human vision. This discrepancy is explained elsewhere[11].

2.2 The minimum number of bits/pixel

The trick in setting a limit on the minimum number of grey levels is to choose the proper image to examine. For a small disk, only about 100 grey levels are needed and therefore 7 bits would be adequate. An improvement over circular disks would be to use a grating between 2 and 6 c/deg, near the peak of the contrast sensitivity function (CSF), where the contrast threshold is .2%[13] corresponding to a luminance change of .4 %. One can do better yet. By using a square wave rather than a sinusoid the threshold is reduced by a factor of $4/\pi$. A further reduction by about a factor of 2 is achieved by using a drifting square wave grating. An even larger reduction, by a factor of 3, is obtained by placing the test grating on a pedestal grating that is slightly above threshold. Nachmias and Sansbury[14] and Stromeyer and Klein[15] showed that contrast discrimination is much better than contrast detection. This facilitation effect is especially strong for moving gratings[16]. Under these optimal conditions, it should be possible to discriminate a .1% change in luminance (.05%

change in contrast). A somewhat conservative estimate of the number of needed grey levels would be to require the display to have a luminance range from 2 to 200 cd/m^2 in .2% steps. The number of grey levels, N, needed to cover the 100-fold luminance range would then be:

$$1.002^N = 100 \quad (1)$$

or

$$N = \ln(100)/\ln(1.002) = 2{,}302. \quad (2)$$

A logarithmic attenuator with slightly more than 11 bits (11.2) would be needed to produce the 2,302 levels.

2.3 The total number of bits/min^2

By combining the resolution constraint that implied a minimum of 9 pixels/min^2 with the contrast discrimination constraint of 11.2 bits/pixel, one concludes that approximately 100 bits/min^2 are needed to display a general image without losing information that is visible to the human observer. The number of bits/pixel can be reduced from 11 to about 8 by preprocessing the image using a "dithering" technique based on clustering the image into 3x3 groups of 9 pixels and using local averaging to interpolate the luminances. The number of bits/min^2 would be reduced from 100 to 72 (9 pixels/min^2 x 8 bits/pixel), which is still a surprisingly high number. Further reductions are possible by making the pixel clusters larger than 3x3 pixels. The limiting reduction is the topic of image compression to be taken up next. The dithering process is not detected because the human visual system is insensitive to small luminance changes at high spatial frequencies. Dithering is excellent for the receiver since it doesn't require any processing by the receiver. It does, however, require substantial resources by the sender and slows the process of generating images. For the display to present an arbitrary image without perceptual loss, 100 bits/min^2 are still needed.

3. IMAGE COMPRESSION AND THE DCT ALGORITHM

The preceding section pointed out that a display must be capable of showing 100 bits/min^2 in order for the human observer not to notice degradation. We now show that compression techniques allow far fewer bits to fully specify the image for transmission. Image compression generally involves three steps.

1. In the first step, the image is transformed by a wide variety of methods (all using high-pass filters) to reduce the correlation between neighboring pixels. The simplest transformation (called Differential Pulse Code Modulation) takes the difference between neighboring pixels. The difference signal presumably has a smaller dynamic range than the original luminance signal and can thus be encoded with fewer bits.

2. The second step is to quantize the transformed image based on properties of the human visual system. The main factor that allows image information to be compressed is the visual system's poor sensitivity to high spatial frequencies. High spatial frequencies require the densest sampling so it is of crucial importance to reduce the bits needed for their encoding. We will also discuss possible savings that come from using Weber's Law which allows fine quantization steps in the range of small contrasts and coarse steps for large contrasts. This second step is the **only** step in which information is lost.

3. The third step is to make use of redundancy in the image to encode the quantized coefficients (entropy coding). No information is lost by this step.

Most schemes which make full use of the CSF falloff are based on the Fourier transform of the image. The Fourier approach is an improvement over Differential Pulse Code Modulation since, in the latter, the difference between neighboring pixels still contains many relatively low spatial frequencies to which the visual system is quite sensitive. Fourier transforms have a much narrower bandwidth, so the high frequency channels with the densest sampling do not contain low spatial frequency information and the quantization can be coarse.

3.1 The Discrete Cosine Transform (DCT)

The purpose of this section is to discuss properties of the human visual system that are relevant to the Discrete Cosine Transform (DCT), presently the most popular compression technique[17,18]. In fact, the DCT algorithm has been selected to become the Still Image Compression Standard[19]. The DCT is similar to the DFT (Discrete Fourier Transform) the method we were taught in grade school. DCT compression seems to work better than the DFT method, possibly because it allows smoother transitions between adjacent blocks. As will be seen, the DCT uses lower spatial frequencies than does the DFT.

The DCT transform is generated by dividing the pattern into square blocks and then reflecting each block about the x and y axes. There are two methods for reflecting. Consider a 3 pixel one-dimensional pattern with luminances a, b, and c. The even-DCT is based on the reflection: c, b, a, a, b, c; the odd-DCT is based on c, b, a, b, c. Symmetry implies that only cosine terms are needed in the expansion[7]. The even DCT transform matrix is given by:

$$t(f,x) = \cos(2\pi(x+.5)\, f/2N) \tag{3a}$$

and the odd-DCT matrix is:

$$t(f,x) = \cos(2\pi x f/(2N-1)) \tag{3b}$$

where N is the number of pixels across a block (pixels/min times min/block). The position, x, takes on integral values from 0 to N–1.

For the even-DCT f takes on values from 0 to (N-1)/2 c/block in steps of .5 c/block. The integer values of f correspond to the cosine terms, $\cos(2\pi xf/N)$, of the usual Fourier transform (DFT) since both are symmetric around the midpoint, x = (N-1)/2. The half integer values of f in the DCT correspond to the sine terms ($\sin(2\pi xf/N)$) of the DFT, where f is an integer in the DFT. Both the DCT for half-integer f and the DFT for odd-integer f are antisymmetric around the midpoint. For a 2 min block, the frequencies are multiples of .5c/block = (.5 cycle)/(2 min) = 15 c/deg. The highest frequency would be N-1 times the fundamental frequency. For 2 pixels/min, N would equal 4 and the highest frequency for the even-DCT would be f_{max} =45 c/deg. For 3 pixels/min, N=6 and f_{max} = 75 c/deg.

For the odd-DCT all frequencies are N/(N-.5) times the even-frequency case. Thus for N = 4 and 6, the fundamental frequencies are 17.14 c/deg and 16.36 c/deg respectively. The peak frequencies are 51.4 and 81.8 c/deg respectively. These values are to be compared to the values for the standard Fourier transform (DFT) where the fundamental frequency is 30 c/deg and the peak frequency is 60 and 90 c/deg for 2 and 3 pixels/min respectively with the 2 min block size. Since the sampling frequencies differ, the differences between the various compression algorithms is not surprising.

The unnormalized DCT coefficients are given by:

$$a(f) = \Sigma\, P(x)\, t(f,x) \tag{4}$$

where the summation over x ranges from 0 to N-1. The DCT coefficients can be normalized to contrast units by dividing each coefficient by the zero frequency coefficient (the mean luminance):

$$c(f) = 2\, a(f) / a(0). \tag{5}$$

The factor of 2 is needed so that the stimulus $P(x) = A\,(1 + \cos(2\pi fx/N))$ has a contrast of 100%. The N-1 contrast values for f>0 together with the mean luminance, a(0), preserve the full information about the original one dimensional stimulus. For two dimensions, there are N^2-1 contrasts, plus the mean luminance.

3.2 The contrast range extends from -200% to +200%

The contrast range is normally taken to go from -100 to +100%. Here we show that the range should be doubled to fully account for the capability of the human visual system. If the display is subdivided into blocks that are 8 min wide and the pixel separation is .5 min, then for the even-DCT, f takes on values from 0 to 56.25 c/deg in 3.75 c/deg steps. The quantization steps are based on the contrast sensitivity function. Suppose, at f = 45 c/deg the sensitivity is 4, implying the Michelson contrast at the detection threshold is 25%. The usual quantization procedure would be to divide the contrast range from 0 to 100% into four bins. In that case two bits would be sufficient to describe the full range of discriminable contrasts. An extra bit should be added to account for negative contrasts. In this section, we point out that a fourth bit is necessary assuming uniform quantization, since the contrast range should be ±200% rather than ±100%. In the next section we calculate the number of bits needed for the ±200% range for arbitrary levels of sensitivity and nonuniform quantization based on Weber's Law.

To show that the DCT values can go from -200% to +200%, consider the stimulus given by the following 16 pixel block:

$$P(x) = (0\ 0\ 0\ 0\ 0\ 0\ 0\ L\ L\ 0\ 0\ 0\ 0\ 0\ 0\ 0) \tag{6}$$

where L is the luminance of a pixel. From Eqs. 3, 4 and 5, the contrast values (in %) of the 16 DCT coefficients are:

$$c(f) = (200\ 0\ -196\ 0\ 185\ 0\ -166\ 0\ 141\ 0\ -111\ 0\ 76\ 0\ -39\ 0) \tag{7}$$

for frequencies from f = 0 to 7.5 c/block (for the even-DCT) in .5 c/block increments where for our example, with .5 min pixels, a block is 8 min. It can be seen that the contrast for f = 1 c/block is -196% and for f = 2 c/block it is 185%. This result is the same as the statement that a grating consisting of thin white lines on a black background has components of 200% contrast. Thus, the DCT coefficients must have a ±200% contrast range. The contrasts fall below 200% even at low spatial frequencies because the stimulus line (Eq. 6) is not infinitely thin. Only the even harmonics contribute because the stimulus is symmetric. The falloff at high spatial frequencies is a characteristic of the DCT which deemphasizes high spatial frequencies as compared to the DFT.

3.3 Quantization, including the effects of facilitation and Weber's Law

Before the quantization step, no information is lost since the transformation step (e.g. the DCT) of the compression process can be inverted. The quantization step is not reversible, and the relevance of research on human vision is to provide insight on how to quantize so that the lost information is not visible. The reduced sensitivity of the visual system at high spatial frequencies provides a major reduction in the needed bits. It is commonly believed by image compressors that further savings can be achieved by invoking Weber's Law, whereby the just noticeable difference (jnd) in contrast is proportional to the background contrast. In this section, the Weber's Law savings is calculated using an accurate representation of the transducer function. It is shown that a facilitation effect, which works in the opposite direction from Weber's Law, increases the number of bits needed for high fidelity transmission of an image.

The facilitation effect[14,15] is based on the finding that when a low contrast pedestal is present, a change in contrast can be detected even when the change is less than half its detection threshold. This facilitation can be understood in terms of an accelerated transducer function relating the detectability, d', to the contrast. The signal detection parameter, d', is the number of jnd's separating the signal from the blank. For contrasts below twice the detection threshold, the transducer function is approximately:

$$d' \approx (c/c_o)^2 \tag{8}$$

where the detection threshold, c_o, is defined such that d' = 1 when $c = c_o$. For $c \gg c_o$, Eq. 8 is no longer valid and contrast discrimination is governed by Weber's Law where the threshold becomes an increasing function of contrast. The contrast discrimination function is often called a "dipper function"[20] because of the decrease in threshold followed by an increase.

Weber's Law, which is close to being valid for large contrasts, can be written as:

$$c_{th} / c = W \tag{9}$$

where c_{th} is the contrast threshold and the Weber fraction,W, is usually between .1 and .2. A transducer function equivalent to Weber's Law is:

$$d' = (1/W) \ln(c/c_o) \tag{10}$$

A small increment of contrast, Δc, on a pedestal, c, leads to:

$$\Delta d' = (1/W)\, \Delta c / c \tag{11}$$

For a threshold increment, Δd' = 1, and Eq. 11 becomes the same as Eq. 9.

A single function that combines the low contrast facilitation (Eq. 8) with the high contrast Weber's behavior (Eq. 10) is given by[21]:

$$d' = \ln(1 + (c/c_o)^2(2W)) / \ln(1 + 2W) \tag{12}$$

This function is constructed so that, at very low contrast, d' is quadratic in c, and has a value of unity at $c = c_o$. At very high contrast, d' is approximately the same as in Eq. 10 since the numerator is:

$$\ln(1 + (c/c_o)^2(2W)) \approx 2 \ln(c/c_o) \tag{13}$$

for large c/c_o, and the denominator is:

$$\ln(1 + 2W) \approx 2W \tag{14}$$

for small W. This function was discussed by Klein and Levi[21], who pointed out that an exact expression for the Δc that produces $\Delta d' = 1$ could be derived from Eq. 12:

$$\Delta c = ((1 + 2W) c^2 + c_o^2)^{1/2} - c \tag{15}$$

The beauty of Eq. 12 is that it provides a closed expression for the number of jnd's from threshold to the highest contrast. This connection between d' and the number of jnds is based on the definition of one jnd corresponding to $\Delta d' = 1$.

If the contrast range from threshold to the maximum contrast is large, then Eq. 12 is approximately:

$$d' \approx (2 \ln(c/c_o) + \ln(2W)) / \ln(1 + 2W) \tag{16}$$

or

$$d' \approx 1/W \ln(c/c_o) + \ln(2W)/(2W) \tag{17}$$

if the Weber fraction W is much less than 1. The first term of Eq. 17 is the same as Eq. 10, and the second term is the correction for threshold effects. For example, if $W = .1$, $c_o = .0025$ (near the peak of the CSF), and $c = 2$ (see Section 3.2 for a discussion of why the maximum contrast, c, is 200%), the first term of Eq. 17 is 66.8 and the second term is -8.0, so the total number of jnd's is 58.8. The exact number, given by Eq. 12, is 64.5. The discrepancy between the two values is due to the true denominator being $\ln(1.2) = .182$ rather than .2.

Near the peak of the CSF, the classical formula for calculating the number of jnd's, as given by Eq. 10, is quite accurate. The problem is that the low spatial frequency regime near the peak of the CSF makes only a small contribution to the total number of bits/min^2 since low spatial frequencies are sparsely sampled. It is the high spatial frequency region that is most important, as is shown in Section 3.5. In the high spatial frequency regime, the quantity c/c_o is not very large and the correction terms discussed above become as important at the Weber's term.

3.4 Number of bits per coefficient and entropy coding

The next step in calculating the number of bits/min^2 is to determine how many bits are needed to specify the contrast level. The example following Eq. 17 showed that, for frequencies near the peak of the CSF about 64 contrast levels could be discriminated (opposite polarity sinusoids excluded). One's first thought is that it should take 6 bits ($\log_2(64)$) to specify the 64 quantized contrasts where each level has a different 6 bit address. If, however, the levels have different probabilities of occurrence then clever coding schemes can be used, in which case the number of bits needed to specify the levels is given by:

$$N = -\Sigma P(i) \log_2(P(i)) \tag{18}$$

where the summation over *'i'* goes all the d' contrast levels (d' =64 in the above example). P(i) is the probability that the *i*th contrast is present in the block. For example, if all 64 levels are equally probable, then $P = 1/64$ and $-\log_2(P) = 6$, so the number of bits in Eq. 18 equals 6 bits. On the other hand, if low contrast levels have greater probability, it is possible to use variable length Huffman coding (entropy coding, in general) whereby the most common levels are assigned shorter codes. There is no loss of information due to the coding step, information is lost only in the quantization step. In particular cases Huffman coding can lead to significant compression, but it is more common to find savings of only about 1 bit/pixel[6]. Furthermore, if large contrasts are present throughout the image or if the probability distribution of contrast levels changes from region to region then variable length coding does not offer much savings.

The assertion that there is not much benefit from entropy coding might be seen as the most controversial assumption of this paper. This is not so. As discussed earlier, the size of the reduction in bits/min^2 that can be achieved from entropy coding is totally dependent on the characteristics of the particular image. For a simple image that is mostly a uniform screen with only a small region devoted to the stimulus (as is common in our psychophysics experiments) a tremendous reduction in the average bits/min^2 can be achieved while, for a complex image with a lot of tiny embedded text, very little reduction would be expected. However, the goal of this paper is to isolate the number of bits/min^2 that are needed by the human visual system from factors that are dependent on the choice of image. In order to achieve this goal, we must ignore the possible bit reductions that can be achieved by doing clever things with the redundancy in the image. Thus our neglect of entropy coding is not really an assumption, but rather a consequence of our belief that two values of bits/min^2 should be reported, one that takes entropy coding into account and one that doesn't.

3.5 Totalling up bits/min^2

Consider a square image 4 min in width with .5 min pixels. The 4 min block size is not chosen arbitrarily. The human cortex seems to be organized in units called hypercolumns[22] that are about 4 x 4 min in extent in the fovea. We have conjectured[23] that within a hypercolumn the image is represented by size-tuned mechanisms rather than by continuing the retinotopic representation. Different hypercolumns thus correspond to the different blocks of the DCT.

The total number of bits required by the DCT compression can now be calculated:

$$\text{Total bits} = \sum_{f_x,\, f_y} N(f) \tag{19}$$

where N(f) is the number of bits at each spatial frequency and where $f = (f_x^2 + f_y^2)^{.5}$. The summation for the even-DCT is over the horizontal and vertical spatial frequencies f_x and f_y ranging from 0 to 52.5 c/deg in Δf = 7.5 c/deg steps. The steps are $\Delta f = 60/8 = 7.5$ c/deg because the block period is 4 min and in the DCT algorithm the even and odd transforms alternate in steps of half the base spatial frequency. The values of N(f) will be shown in Table 2.

Table 1 shows the contribution of each frequency range (frequency in c/deg is in first column) to the total number of bits/min^2. The second column is a representative CSF[21] with a peak sensitivity of about 300 and a cutoff near 60 c/deg. The analytic form of the CSF is a Cauchy function [21] given by:

$$\text{CSF} = 250\, f^{.5} \exp(-.13f). \tag{20}$$

The third column is the total number of discriminable contrasts up to 200%. This value is obtained by multiplying the d' value of Eq. 12 by a factor of 2 to account for DCT coefficients with a negative value. The fourth column is the number of bits needed to represent the contrast range of the third column

$$N(f) = \log_2(2d'). \tag{21}$$

In deriving this equation from Eq. 18, an assumption was made that the different quantization levels are equally likely to occur (no Huffman coding as discussed in Section 3.4).

spat		$c_{max} = 2$			$c_{max} = 1$			**equal quantization**		
freq	**CSF**	range	bits	total	range	bits	total	range	bits	total
0.00	0		11.	0.69		11.	0.69		11.	0.69
7.50	258.25	119.40	6.90	0.68	104.19	6.70	0.66	2065.97	11.01	1.08
15.00	137.76	105.61	6.72	1.32	90.41	6.50	1.28	1102.05	10.11	1.98
22.50	63.64	88.67	6.47	1.91	73.48	6.20	1.83	509.11	8.99	2.65
30.00	27.72	70.45	6.14	2.41	55.30	5.79	2.27	221.74	7.79	3.06
37.50	11.69	51.59	5.69	2.79	36.68	5.20	2.55	93.51	6.55	3.21
45.00	4.83	32.67	5.03	2.96	19.03	4.25	2.50	38.64	5.27	3.11
52.50	1.97	15.47	3.95	2.72	6.29	2.65	1.82	15.74	3.98	2.73
60.00	0.79	4.47	2.16	1.70	1.30	0.38	0.30	6.35	2.67)	2.09
TOTAL (Polar summation)				17.2			13.9			20.6
TOTAL (Cartesian summation)				19.0			16.0			27.3

Table 1. Total number of bits/min^2 for DCT compression with 8 pixels per block
The top row represents the mean intensity of the 4' x 4' patch. It is assumed that 11 bits are needed to represent the mean luminance, which results in a contribution of 11/16 = .69 bit/min^2 to the total. The total from polar summation is the sum of all values in the total columns. The Cartesian summation is discussed in the text and the individual contributions are shown in Table 2.

In crder to clarify how many bits come from each frequency range it is useful to transform Eq. 19 from Cartesian to polar coordinates:

$$\text{Total bits} = \pi/2 \sum_{f=0}^{56.75} (f/7.5) N(f)\, \Delta f \tag{22}$$

The summation is in frequency steps of 7.5 c/deg. The factor of $(\pi/2)f/7.5$ is approximately the number of components in the upper right quadrant of radius f/7.5. The number of bits/min^2 from each frequency range, T(f), are listed in the fifth column, where:

$$T(f) = \pi/2\ (f/7.5)\ N(f) / 16. \quad (23)$$

The factor of 16 is used here because there are 4 x 4 min in the stimulus patch, so in order to get bits/min^2 the total number of bits must be divided by 16. The main contributions come from between 30 and 52.5 c/deg. The frequencies below 20 c/deg, which are the only frequencies considered in most compression schemes, involve less than 20% of the bits/min^2 that are needed for perfect image quality. Frequencies above 30 c/deg require 70% of the total bits. These numbers should be kept in mind when one hears of compression schemes in which the highest sampling frequency is 16 c/deg. Finally at the bottom of the fifth column are two values giving the total number of bits/min^2. The first value is 17.2 bits/min^2 that comes from summing the values in the fifth column corresponding to the polar coordinate summation. The bottom value is 19.0, corresponding to the more accurate Cartesian summation of Eq. 19. The individual components of the Cartesian summation are presented in Table 2.

The 6th through 8th columns show similar calculations for the case in which the maximum contrast is taken to be 100% rather than 200%. The total number of bits/min^2 is 13.9 for this case. This analysis shows that 3 bits/min^2 are ignored by calculations based on 100% maximum contrast. This increase in contrast range has only a minor effect on spatial frequencies below 45 c/deg. In a 7.5 c/deg span around 45 c/deg, for example, the number of extra bits/min^2 due to the extended range is 2.96 - 2.50 =.46 bits/min^2. In the same frequency span around 52.5 c/deg, however, there are an extra 2.72 - 1.82 = .90 bit/min^2. The reason that the extended contrast range has more impact at high frequencies is that at these high frequencies, the extended range lies in the facilitation region of the dipper function whereas at lower frequencies the extended contrast range lies in the Weber regime where the quantization steps are coarser so that relatively fewer steps are added by going to 200%.

Table 1 showed the polar coordinate representation of the number of bits/min^2 for each range of spatial frequencies summed over all orientations. Table 2 is the Cartesian coordinate version of the number of bits/min^2. The row and column headings are the 8 spatial frequencies from 0 to 52.5 that are present in the DCT for .5 min sampling. Only values for which the CSF is above .5 (threshold=200%) are displayed and are included in the sum. That is, components that can not get above threshold are not included. The sum of these values is tabulated in the bottom row of Table 1 for the case of $c_{max}=2$ (the data from Table 2), $c_{max}=1$, and for equal quantization. A column and row at 60 c/deg has been added to include the extra contributions when the sampling is increased from 2 pixels/min to 3 pixels/min. The spatial frequencies above 60 c/deg for 3pix/min are not shown since they are below threshold. The point made in Section 2.1 should be remembered, however, that even though there are no contributions above 60 c/deg, 3 pixels/min are needed for a lossless display[11].

	0.0	**7.5**	**15.**	**22.5**	**30.**	**37.5**	**45.**	**52.5**	**(60.0)**
0.00	**.69**	.43	.42	.40	.38	.36	.31	.25	(.14)
7.50	.43	**.43**	.42	.40	.38	.35	.31	.24	(.13)
15.00	.42	.42	**.41**	.39	.37	.34	.30	.22	(.10)
22.50	.40	.40	.39	**.38**	.36	.32	.27	.18	
30.00	.38	.38	.37	.36	**.33**	.29	.23	.13	
37.50	.36	.35	.34	.32	.29	**.24**	.16		
45.00	.31	.31	.30	.27	.23	.16	**.07**		
52.50	.25	.24	.22	.18	.13				
(60.0)	(.14)	(.13)	(.10)						

TABLE 2. Cartesian representation of bits/min^2 for a 4 x 4 min block with 2 pixels/min sampling and $c_{max}=2$. The numbers in parentheses are the additional contributions from increasing the sampling to 3 pixels/min. The sum is 19.0 and 19.7 for the 2 and 3 pixels/min cases.

In order to demonstrate the effect of even vs. odd DCT, sampling at 2 vs. 3 pixels/min and different block sizes, we present the results of the number of bits/min^2 for the exact Cartesian summation in Table 3. It is seen that, for a block size of 60 min, there are negligible differences between even and odd DCT or between 2 and 3 pixels/min. For small block sizes the differences are large. The even-DCT requires more bits/min^2 than the odd version since the spatial frequency sampling grain is smaller. Based on this result it might be expected that the odd-DCT would be more popular than the even DCT. The opposite is true because a fast-Fourier version is available for the even but not the odd DCT.

pixels/min	block size	Even-DCT (bits/min^2) c=2	c=1	uniform	Odd-DCT (bits/min^2) c=2	c=1	uniform
2 pixels	1 min	28.6	27.2	35.3	25.1	22.3	29.5
	2 min	22.1	19.5	30.2	19.3	16.2	26.7
	4 min	19.0	16.0	27.3	17.4	14.4	25.4
	8 min	17.6	14.5	25.7	16.8	13.7	24.8
	60 min	16.5	13.3	24.3	16.3	13.2	24.2
3 pixels	1 min	32.9	27.2	47.5	26.8	24.7	34.9
	2 min	24.0	19.5	35.9	20.3	17.1	31.2
	4 min	19.7	16.0	31.0	18.3	15.0	28.7
	8 min	18.0	14.5	28.6	17.3	14.0	27.4
	60 min	16.7	13.3	26.5	16.6	13.2	26.4

Table 3. The total number of bits/min^2 for the even and odd DCT for 2 and 3 pixels/min, and for 1, 2, 4, 8 and 60 min block sizes. As in Table 1, the calculation is made for quantization based on d' with a maximum contrast of 200% and 100%, and also for uniform quantization.

3.6 Problems with Weber's Law Quantization

The number of bits/min^2 that we have just calculated may be an underestimate because the effect of nonlocal masking has been ignored so far. To see how this works, suppose the image we would like to compress is a thin line at the left side of the 4 min block. The pattern has an intensity profile of (assuming .5 min/pixel):

$$P(x) = (26\ 25\ 25\ 25\ 25\ 25\ 25\ 25) \tag{24}$$

where 0 represents the lowest intensity and 255 is the highest. This line was chosen because its contrast of 4% ((26-25)/25) and width of .5 min gives a line strength of 2%min which is slightly above the line detection threshold[24,25]. Suppose, however, that a high intensity pixel is present at the other side of the block:

$$P_s(x) = (26\ 25\ 25\ 25\ 25\ 25\ 25\ 255) \tag{25}$$

where the subscript, s, stands for signal. The observer's task is that of discriminating the pattern in Eq. 25 from

$$P_b(x) = (25\ 25\ 25\ 25\ 25\ 25\ 25\ 255) \tag{26}$$

where the subscript, b, stands for blank. Since the pixel spacing is .5 min, in the example of Eq. 25 the low intensity line is separated by 3.5 min from the high intensity line on the right of the block. Experiments on the detection of lines near edges[26] show that the range of the masking on the dark side of the edge is less than 5 min. We would thus expect very little threshold elevation for the stimulus of Eq. 25 as compared to the stimulus of Eq. 24.

The seven coefficients of the DCT of P_s and P_b for nonzero frequencies, given by Eqs. 3, 4 and 5, are expressed in percent contrast:

$$c(f) = 2\Sigma\, t(f,x)\, P(x) \,/\, \Sigma\, P(x)$$

$$c_s(f) = (-104.9, 98.8, -89.0, 75.6, -59.4, 40.9, -20.9) \text{ percent.} \tag{27}$$

$$c_b(f) = (-104.2, 99.0, -88.4, 75.8, -59.0, 41.0, -20.7) \text{ percent} \tag{28}$$

The difference between the coefficients for the blank and the stimulus are:

$$c_b(f) - c_s(f) = (\ .7,\ \ .2,\ \ .6,\ \ .2,\ \ .4,\ \ .1,\ \ .2) \text{ percent} \tag{29}$$

The seven spatial frequencies represented in Eqs. 27-29 range from 7.5 to 52.5 c/deg in steps of 7.5 c/deg (60/8). The zero frequency coefficient is not shown since it contains no information when the coefficients are expressed in contrast units. It is clear that we have a problem. In order to detect the faint line at the left of the block, information must be preserved to at least .2% out of a range of about 200% (going from -105% to +99%). The problem is even worse. One must do more than simply detect this difference in one coefficient, since the noise can be variable if the strength of the masking line changes. One must look at <u>correlations</u> between the coefficients in order to discriminate the signal from the test line on the left in the presence of the much stronger noise from the line on the right. Quantization schemes based on Weber's law would run into trouble because quantization based on a 10% Weber fraction would totally obliterate all information about the threshold line on the left of the block. The conclusion is that Weber's Law does not provide reliable constraints for quantizing the Fourier coefficients when nonlocal maskers are present. In order to detect the weak line in the presence of the nearby mask it is necessary to know several coefficients to an accuracy of the unmasked threshold. This requires uniform quantization as is reported in the last three columns of Table 1. The resulting larger number of bits/min^2 is a consequence of Weber's quantization being faulty for high fidelity transmission of low intensity lines when nearby noise is present in the same block.

The last three columns of Table 1 and also two of the columns of Table 3 provide an estimate for the number of bits/min^2 that are needed in order to avoid the effects of nonlocal masking. From the examples presented in Eqs 24 and 25, in order to preserve low contrast information in the presence of a high contrast distant stimulus, the quantization must be in fine steps throughout the entire contrast range. The contrast range is -2 to +2 and the smallest increment needing to be preserved could be as low as half of the contrast threshold, due to the facilitation effect discussed in Section 1.2. The total number of quantized steps is displayed in the 9th column of Table 1. The value in this column is 8 x CSF. Three factors of 2 are present: (1) The smallest contrast that is discriminable is about 1/(2 CSF), where the factor of 2 is needed to account for the facilitation effect. (2) The largest contrast that can be generated is 200% so the number of contrast steps is 2 x (2 x CSF). (3) Another factor of 2 is needed for negative contrasts. This approach leads to 27.3 bits/min^2 using Cartesian summation and the even-DCT, which is significantly higher than the 19 bits/min^2 that was based on Weber's law.

5. SUMMARY

The calculation of bits/min^2 began with an analysis of the requirements of a video display. The human resolution threshold of .33 min implied that a .33 min pixel size was needed (smaller than the Nyquist limit[11]), and the human's ability to discriminate .2% contrast changes implied that an 11 bit logarithmic attenuator was needed to allow for all discriminable luminances for a luminance range of 100. These considerations led to an estimate of 100 bits/min^2 needed by a display to present an image with no loss.

Compression algorithms allow images to be transmitted with far fewer bits than are needed for the display. By using the Discrete Cosine Transform with a 4 min block size and nonuniform quantization based on Weber's law, an image can be encoded in 17 to 19 bits/min^2, depending on whether the odd or even DCT is used. In the discussion of ways to quantize the DCT coefficients, several novel points were made. (1) The contrast range of the coefficients is from -200% to +200%, a larger range than is usually used. For spatial frequencies above 30 c/deg, where most of the information resides, doubling the contrast range increases the number of bits/min^2 by 3. (2) The transducer function given by Eq. 12 (a closed form expression for d' as a function of contrast) gives the number of discriminable contrast levels from 0 to the contrast specified. The Weber's Law behavior provides the main limit on the number of discriminable levels for spatial frequencies below about 30 c/deg. For higher spatial frequencies, the Weber's regime spans a relatively small extent and does not provide a large savings of bits. The facilitation effect[14,15] which is also incorporated into Eq. 12 extends the number of discriminable contrasts in the near threshold regime. This extension is especially important at high spatial frequencies.

The 4 minute block size for the DCT may be appropriate for several reasons: (1) masking effects are strong within 4 min and fall off outside this range[27,28]. (2) The visual cortex is organized in units of hypercolumns which correspond to about 4 min in extent. (3) We have speculated[23] that size tuned filters (the DCT transform) operate within a hypercolumn and position coding (local sign) operates between hypercolumns.

A distinguishing feature of our analysis is that the images used to test the compression schemes are carefully chosen to reveal flaws produced by the compression. For the DCT compression, a near threshold thin line flanked by a high contrast line at a distance of 3.5 min was used to argue that the Weber's Law quantization may have underestimated the needed number of bits/min^2. We presume that text embedded in a natural scene would provide similarly stringent tests of compression. Other images that are excellent for testing compression are high contrast square wave spoke patterns and frequency modulated gratings[4].

The usual tests of compression use complex scenes. Complex images are useful for evaluating how compression degrades the type of images normally transmitted on television. The problem with complex images is that it is difficult to quantify the magnitude of the loss. The naive observer may not know where to look for the degraded information, since the amount of degradation is scene dependent. If, for example, one wanted to transmit detailed text on a weather map (a combination of graphics and text) then the full resolving power and hyperacuity power of the human observer may be needed. Our approach is to calculate the number of bits/min^2 needed to see selected images that isolate particular visual characteristics. The human observer is allowed to use all visual cues to discriminate between the original image and the compressed image. This approach does not take into account the image dependent redundancy. We believe the compression due to redundancy should be reported separately.

Our approach using special stimuli to find the limiting bits/min^2 of the human visual system has several advantages: (1) The conditions are easy to specify. We do not need to make arbitrary assumptions about pixel size and bits/pixel of the original image. Our criterion is that the compression scheme should work for all images. (2) Since we are determining the total amount of information entering the visual system, our findings should be related to the number of ganglion cells per min^2 times the number of bits of information each ganglion cell is able to transmit in a given time interval. We hope to make this comparison in the future. In this extension to physiology we will need to examine temporal factors. The psychophysical limits on the number of bits/min^2 per sec must be measured.

Acknowledgements: We thank Thom Carney for helpful suggestions and Al Ahumada for his thoughts following the oral presentation. This research was supported by the Air Force Office of Sponsored Research, AFOSR 89-0238 and by the National Eye Institute EY0-4776.

References

1. A. B. Watson, "Efficiency of a model human image code," J. Opt. Soc. Am. A **4**, 2401-2417 (1987).

2. E. H. Adelson, E. Simoncelli and R. Hingorani, "Orthogonal pyramid transforms for image coding," SPIE- Visual Communications and Image Processing II, 50-58 (1987).

3. C. S. Stein, A. B. Watson and L. E. Hitchner, "Psychophysical rating of image compression techniques," SPIE - Human Vision, Visual Processing, and Digital Display, 198-208 (1989).

4. J. A. McCormick, R. Alter-Gartenberg and F. O. Huck, "Image gathering and restoration: information and visual quality," J. Opt. Soc. Am. A **6**, 987-1005 (1989).

5. F. O. Huck, S. John, J. A. McCormick and R. Narayanswamy, "Image gathering and digital restoration: End-to-end optimization for visual quality," Appl.Vision **16**, 85-88 (1989).

6. A. N. Netravali and B. G. Haskell, *Digital Pictures: representation and compression* (Plenum Press, New York, 1988).

7. W. K. Pratt, *Digital Image Processing* (John Wiley & Sons, New York, 1978).

8. J. Hirsch and C. A. Curcio, "The spatial resolution capacity of human foveal retina," Vision Res. **9**, 1095-1101 (1989).

9. S. J. Schein and F. M. de Monasterio, "Mapping of retinal and geniculate neurons onto striate cortex of macaque," J. Neuroscience **7**, 996-1009 (1987).

10. H. Wassle, U. Grunert, J. Rohrenbeck and B. B. Boycott, "Cortical magnification factor and the ganglion cell density of the primate retina," Nature **341**, 643-645 (1989).

11. S. A. Klein and T. Carney, "How many bits/min2 are needed for the perfect display?" SID - Society for Information Display, Annual Meeting (1990).

12. D. M. Levi and S. A. Klein, "Equivalent Intrinsic Blur in Spatial Vision," Vision Res. In Press (1990).

13. F. W. Campbell and J. G. Robson, "Application of Fourier analysis to the visibility of gratings," J. Physiol. (London) **197**, 551-566 (1968).

14. J. Nachmias and R. V. Sansbury, "Grating contrast: Discrimination may be better than detection," Vision Res. **14**, 1039-1042 (1974).

15. C. F. Stromeyer III and S. A. Klein, "Spatial frequency channels in human vision as asymmetric (edge) mechanisms," Vision Res. **14**, 1409-1420 (1974).

16. C. F. Stromeyer III, R. E. Kronauer, J. C. Madsen and S. A. Klein, "Opponent-movement mechanisms in human vision.," J. of Opt. Soc. Am. **1**, 876-884 (1984).

17. N. Ahmed, T. Natarajan and K. R. Rao, "Discrete Cosine Transform," IEEE Trans. on Comm. 90-93 (1974).

18. K. N. Ngan, K. S. Leong and H. Singh, "Cosine transform coding incorporating human visual system model," SPIE - Visual Comm. & Image Proc. **707**, 165-171 (1986).

19. G. Wallace, "Still image compression standard," SPIE/SPSE Symposium on Electronic Imaging Science & Technology, 18A (1990).

20. G. E. Legge and J. M. Foley, "Contrast masking in human vision," J. Opt. Soc. Am. **70** (1980).

21. S. A. Klein and D. M. Levi, "Hyperacuity thresholds of one second: Theoretical predictions and empirical validation," J. Opt. Soc. Am. A **2**, 1170-1190 (1985).

22. D. H. W. Hubel T. N., "Uniformity of monkey striate cortex: a parallel relationship between field size, scatter, and magnification factor," J. Comp. Neurol. **158**, 295-306 (1974).

23. S. A. Klein and D. M. Levi, "Position sense of the peripheral retina," J. Opt. Soc. Am A **4**, 1543-1553 (1987).

24. S. Hecht and E. U. Mintz, "The visibility of single lines of various illuminations and the retinal basis of visual resolution," J. Gen. Physiol. **22**, 593-612 (1939).

25. S. A. Klein, E. Casson and T. Carney, "Vernier acuity as line and dipole detection," ARVO 1988, p. 371, Vision Res., In Press (1990).

26. J. O. Limb and U. Tulunay-Keesey, "Spatiotemporal characteristics of thresholds adjacent to a luminance edge," J. Opt. Soc. Am 1209-1219 (1981).

27. G. Westheimer and G. Hauske, "Temporal and spatial interference with vernier acuity," Vision Res. **15**, 1137-1141 (1975).

28. D. M. Levi, S. A. Klein and A. P. Aitsebaomo, "Vernier acuity, crowding, and cortical magnification," Vision Res. **25**, 963-977 (1985).