Toward a perceptual video quality metric

Andrew B. Watson[1]

NASA Ames Research Center, Moffett Field, CA 94035-1000

ABSTRACT

The advent of widespread distribution of digital video creates a need for automated methods for evaluating the visual quality of digital video. This is particularly so since most digital video is compressed using lossy methods, which involve the controlled introduction of potentially visible artifacts. Compounding the problem is the bursty nature of digital video, which requires adaptive bit allocation based on visual quality metrics, and the economic need to reduce bit-rate to the lowest level that yields acceptable quality.

In previous work, we have developed visual quality metrics for evaluating, controlling, and optimizing the quality of compressed still images[1, 2, 3, 4]. These metrics incorporate simplified models of human visual sensitivity to spatial and chromatic visual signals. Here I describe a new video quality metric that is an extension of these still image metrics into the time domain. Like the still image metrics, it is based on the Discrete Cosine Transform. An effort has been made to minimize the amount of memory and computation required by the metric, in order that might be applied in the widest range of applications. To calibrate the basic sensitivity of this metric to spatial and temporal signals we have made measurements of visual thresholds for temporally varying samples of DCT quantization noise.

Keywords: vision, digital video, perception, quality, fidelity

1. INTRODUCTION

Recent years have seen the introduction and widespread acceptance of several varieties of digital video. These include digital television broadcasts from satellites (DBS-TV), the US Advanced Television System (ATV), digital movies on a compact disk (DVD), and digital video cassette recorders (DV). In the near future we can expect to see widespread terrestrial broadcast and cable distribution of digital television. All of these systems, and any we can foresee, depend upon lossy compression of the video stream. Lossy compression may introduce visible artifacts, and indeed there is an economic incentive to reduce bit-rate to the point were artifacts are almost visible. For this reason, there is an urgent need for reliable means for automatically evaluating the visibility of compression artifacts, and more generally, the visual quality of digital video.

Recently a number of video quality metrics have been proposed[5, 6, 7, 8, 9, 10]. Possible disadvantages of these metrics are that they may either not be based closely enough upon human perception, in which case they may not accurately measure visual quality, or that they may require amounts of memory or computation that restrict the contexts in which they may be applied. The goal of this project has been to construct a metric that is reasonably accurate but computationally efficient.

Here I describe a new objective fidelity metric for digital video. A novel feature of this metric is that it is based on the Discrete Cosine Transform (DCT). Because all of the coding standards mentioned above make use of this transform, it is expected that efficientmeans for implementing the DCT will be readily available, or indeed, that the necessary transforms will already have been done. In its use of the DCT, this metric resembles the DCTune metric that we have developed for optimization of still image compression[3]. We have given this metric the name DVQ, for Digital Video Quality.

We begin this paper with a report of new data on the visibility of dynamic DCT quantization noise. We fit these data with a simple mathematical model that subsequently forms a part of the DVQ metric. We then describe the individual processing

[1] abwatson@mail.arc.nasa.gov • http://vision.arc.nasa.gov/

Part of the IS&T/SPIE Conference on Human Vision and Electronic Imaging III
San Jose, California • January 1998 • SPIE Vol. 3299 • 0277-786X/98/$10.00

steps of the DVQ metric. We then compare metric outputs to some published psychophysical data, and describe application of the metric to an example video sequence.

2. VISIBILITY OF DYNAMIC DCT QUANTIZATION NOISE

The DVQ metric computes the visibility of artifacts expressed in the DCT domain. Therefore we have first made measurements of human visual thresholds for a novel visual stimulus which we call *dynamic DCT noise*. This is produced by first computing an image composed of a square array of 8x8 pixel blocks, within each of which is placed a DCT basis function of the same frequency. Over a sequence of frames, each basis function is then modulated by a Gabor function in time (the product of a Gaussian and a sinusoid) of a particular temporal frequency and phase. From block to block, the phase is randomly distributed over the interval [0, 2 Pi]. This signal resembles in some ways the quantization error to be expected from a single DCT coefficient, but it is confined to a narrow band of temporal frequency. An example sequence is shown in Figure 1.

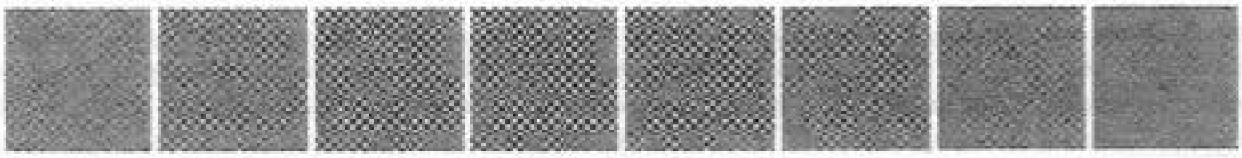

Figure 1. Dynamic DCT noise. In this example the DCT frequency was {3,3}, the number of blocks was 8 x 8, the frame rate was 60 Hz, the Gaussian time scale was 5 frames, and the temporal frequency was 2 Hz.

Other methods were as follows. Display resolution was 32 pixels/degree, and display frame rate was 120 Hz. Mean luminance was 50 cd/m^2. Viewing was binocular with natural pupils from a distance of 91.5 cm. Thresholds were collected using a QUEST staircase[11], using a total of 32 trials/threshold. DCT frequenciestested were {0,0}, {0,1}, {0,2}, {0,3}, {0,5}, {0,7}, {1,1}, {2,2}, {3,3}, {5,5}, {7,7}. Temporal frequenciestested were 0, 1, 2, 4, 6, 10, 12, 15, 30 Hz. Three observers participated: RNG, a 25 year old male, JQH, a 35 year old male, and HKK, a 25 year old female. All used their usual spectacle correction.

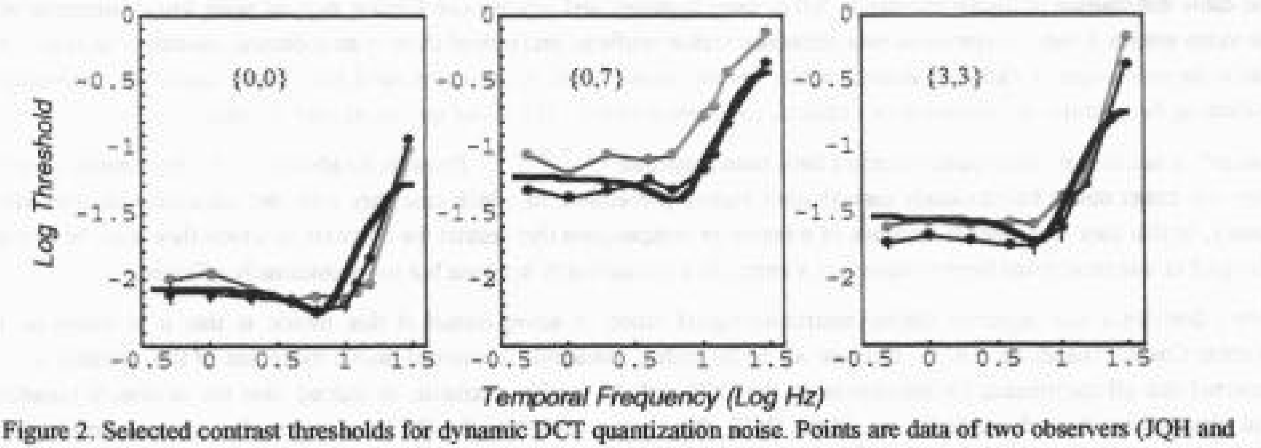

Figure 2. Selected contrast thresholds for dynamic DCT quantization noise. Points are data of two observers (JQH and HKK); error bars show ± 1 standard deviation. The numbers in braces in each panel show the horizontal and vertical DCT frequencies. The thicker curve is the model.

A subset of the data is shown in Figure 2. The data show an expected increase in threshold at high spatial and temporal frequencies. In both spatial and temporal frequency domains, the data are roughly low-pass in form. For this reason we have considered a simple, separable model that is the product of a temporal function, a spatial function, and an orientation function:

$$T(u,v,w) = T_0\ T_w(w)\ T_f(u,v)\ T_a(u,v) \qquad (1)$$

The factor T_0 is a global or minimum threshold. The remaining functions are defined in such a way that they have unit peak gain, so that the minimum threshold is given directly by T_0. The temporal function (Figure 3a) is the inverse of the magnitude response of a first-order discrete IIR low-pass filter with a sample rate of w_s Hz and a time constant of τ_0 seconds.

$$T_w(w) = \left| \frac{-1 + e^{\frac{1+i2\pi\tau_0 w}{\tau_0 w_s}}}{-1 + e^{\frac{1}{\tau_0 w_s}}} \right| \qquad (2)$$

The spatial function (Figure 3b) is the inverse of a Gaussian, with a parameter of f_0, corresponding to the radial frequency at which threshold is elevated by a factor of e^{π}. The factor of $p/16$, where p is the display resolution in pixels/degree, converts from DCT frequencies to cycles/degree.

$$T_f(u,v) = Exp\left(\pi \frac{u^2+v^2}{f_0^2} \left(\frac{p}{16} \right)^2 \right) \qquad (3)$$

The orientation function (Figure 3c) accounts for two effects: the higher threshold for oblique frequencies, and the imperfect visual summation between two component frequencies[12]. It is given by

$$T_a(u,v) = 2^{\frac{\beta-1}{\beta}} \Big/ \left(1 - \frac{4 r u^2 v^2}{\left(u^2+v^2\right)^2} \right) \qquad (4)$$

where r and β are parameters. This model has been fit to the data of the two observers, and the results are shown by the red curves in Figure 2. Despite the simplicity of the separable model, a reasonable fit to the data is obtained. This model will be used below to calculate visibility of differences between two video sequences.

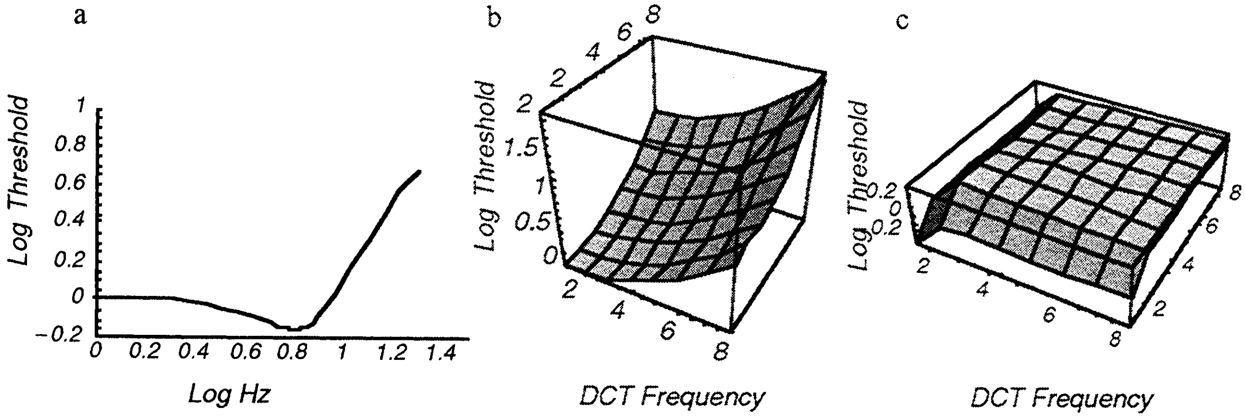

Figure 3. Temporal, spatial and orientation components of the dynamic DCT threshold model.

3. DVQ OVERVIEW

Figure 4 is an overview of the processing steps of the DVQ metric. These steps will be described in greater detail in subsequent sections. The first step is a possible cropping, to exclude regions whose quality is not of interest. This step may also include registration of the two sequences, if that is required. The next step is a transformation from the input video color format, such as RGB or YCbCr, to the color space YOZ. The next step is transformation of each video frame to its blocked DCT. The results are then transformed to local contrast. The next step is a temporal filtering operation. The temporally filtered coefficients are then converted to just-noticeable differences (jnds) by multiplying each DCT coefficient by its corresponding entry in the spatial contrast sensitivity function (SCSF). At the next stage the two sequences are subtracted. The difference sequence is then subjected to a contrast masking operation. Finally the masked differencesmay be pooled in various ways to illustrate the perceptual error over various dimensions.

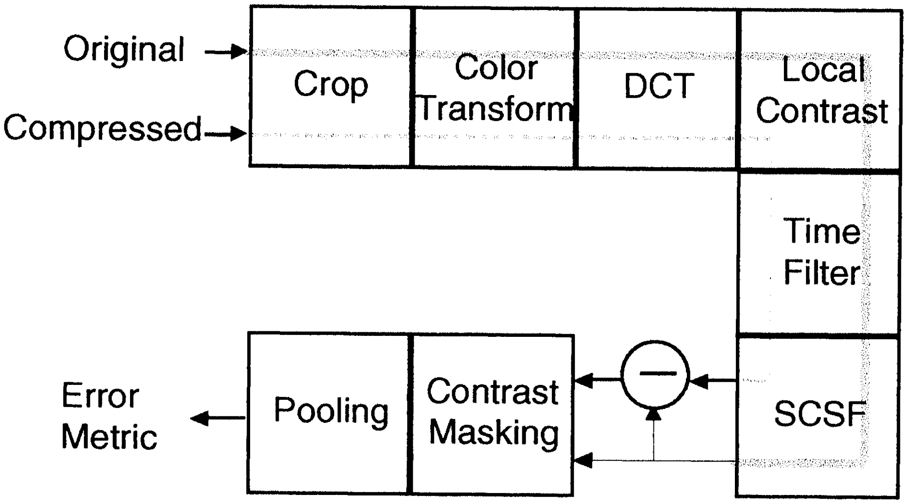

Figure 4. Overview of the DVQ video quality metric.

3.1. Input

The input to the metric is a pair of color image sequences. The dimensions of this input are {s,f,c,y,x}, where s = sequence (2), f = frames, c = color (3), y = rows, and x= columns. The first of the two sequences is the reference, the second is the test. Typically the test will differ from the reference in the presence of compression artifacts. The input color space must be defined in sufficient detail that it can be transformed into CIE coordinates, for example by specifying the gamma and chromaticity coordinates of each primary. Two common examples used in this paper are a linear (gamma=1) RGB space, and YCbCr with gamma=2.2.

3.2. Color Transformations

The first step in the process is the conversion of both image sequences to the YOZ color space. This is a color space we have previously used in modeling perceptual errors in still image compression. The three components of this space are Y (CIE luminance in candelas/m^2), O, a color-opponent channel given by O = .47 X -.37 Y -.1 Z, and a blue channel given by the CIE Z coordinate. Transformation to the YOZ space typically involves 1) a gamma transformation, followed by 2) a linear color transformation. These operations do not alter the dimensionality of the input.

3.3. Blocked DCT

At this point a blocked DCT is applied to each frame in each color channel. The dimensions of the result are {s, f, c, by, bx, v, u}, where by and bx are the number of blocks in vertical and horizontal directions, and where now v=u=8.

3.4. Local Contrast

The DCT coefficients are converted to units of local contrast in the following way. First we extract the DC coefficientsfrom all blocks. These are then time filtered, using a first-order, low-pass, IIR filter with a gain of 1 and a time constant of τ_1. The

DCT coefficients are then divided by the filtered DC coefficients on a block by block basis. The Y and Z blocks are divided by Y and Z DC coefficients;the O is divided by the Y DC. In each case, a very small constant is added to the divisor to prevent division by zero. Finally, the quotients are adjusted by the relative magnitudes of their coefficients corresponding to a unit contrast basis function. These operations convert each DCT coefficient to a number between -1 and 1, that expresses the amplitude of the corresponding basis function as a fraction of the average luminance in that block.

The DC coefficients themselves are converted in a similar fashion: the mean DC over the entire frame is subtracted, and the result is divided by that mean.

3.5. Temporal Filtering

Both sequences are then subjected to temporal filtering. The temporal filter is a second-order IIR filter, as described above in the fit of the dynamic DCT noise data. Use of an IIR filter minimizes the number of frames of data that must be retained in memory. For even greater simplicity, a first order filter may be used.

3.6. JND Conversion

The DCT coefficients, now expressed in local contrast form, are now converted to just-noticeable-differences(jnds) by dividing by their respective spatial thresholds, as specified by the Equations (3) and (4). These thresholds are first multiplied by a spatial summation factor *s*, whose purpose and estimation are described below. The thresholds for the two color channels are either derived from the luminance thresholds[3] or based on additional chromatic thresholds. After conversion to jnds, the coefficients of the two sequences are subtracted to produce a *difference sequence*.

3.7. Contrast Masking

Contrast masking is accomplished by first constructing a *masking sequence*. This begins as the referencesequence, afterjnd conversion. This sequence is rectified, and then time-filtered by a first-order, low-pass, discrete IIR filter, with a gain of g_1 and a time constant of τ_2. These values are then raised to a power *m*, any values less than 1 are replaced by 1, and the result is used to divide the difference sequence. This process mimics the traditional contrast masking result in which contrasts below threshold have no masking effect, and that above threshold the effect rises as the mth power of mask contrast in jnds[13].

3.8. Minkowski Pooling

The dimensions of the result at this point are $\{f, c, by, bx, v, u\}$, where, to remind, *f* is frames, *c* is color channels, *by* and *bx* are the number of blocks in vertical and horizontal directions, and where $v=u$ are the vertical and horizontal frequencies. These elementary errors may then be combined over various dimensions, or all dimensions, to yield summary measures of visual error. This summation is done using a Minkowski metric,

$$J_x = M\left(j_{f,c,by,bx,y,x}, \beta\right) = \left(\sum_x \left|j_{f,c,by,bx,y,x}\right|^\beta\right)^{\frac{1}{\beta}} \qquad (5)$$

In this equation we have indicated summation over all six dimensions, but any subset of these dimensions may be considered as well. A virtue of the Minkowski formulation is that it may be nested. For example, we may first sum over only the color dimension *(c)*, and then these results may subsequently be summed over, for example, the block dimensions (*by* and *bx*).

4. DYNAMIC DCT NOISE SIMULATIONS

The metric incorporates the mathematical model fit to the thresholds for dynamic DCT noise, as given in Equations (1-4). This model is used to establish the threshold for elementary errors in individual DCT coefficients, that is, threshold for a single DCT basis function in a single 8 x 8 pixel block. However, the stimuli used in the psychophysical experiments were arrays of 8 x 8 blocks, and thus their sensitivity was enhanced by probability summation over sensitivity to a single block. Traditional probability summation calculations would predict this factor to be $(8^2)^{1/\beta}$, which for a beta of 3 would be 4. Through simulations of actual dynamic DCT noise stimuli, we have found a factor of s = 3.7 to provide a good fit.

5. CONTRAST MASKING SIMULATIONS

To validate and calibrate the masking component of the metric, we performed a simulation in which both sequences contained a mask consisting of two frames of a sinusoidal grating of 2 cycles/degree with a size of 2 degrees. The test sequence also contained a test grating of the same size and spatial frequency. We varied the test contrast to find the value that would yield unit output from the metric. The figure shows that as mask contrast exceeds the threshold contrast, threshold rises, and does so in a fashion similar to the comparison data from Foley[14]. Foley's data also show a facilitation effect at sub-threshold mask contrasts, which we do not attempt to model.

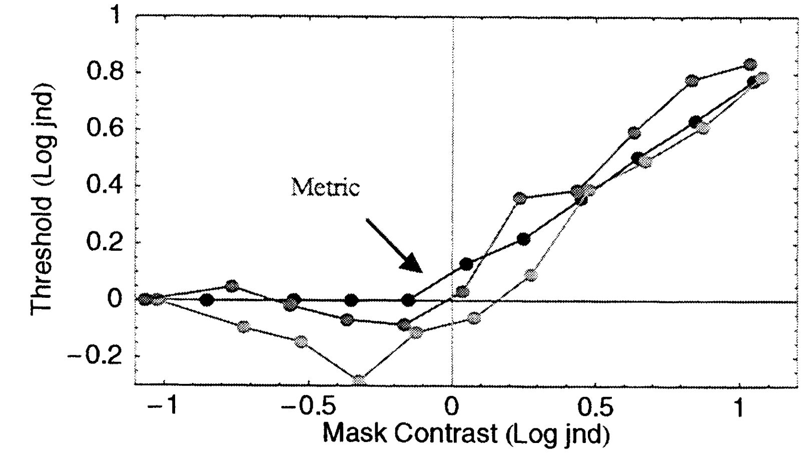

Figure 5. Test threshold versus mask contrast. Test was a 2 cycle/deg Gabor function, and mask a 2 cycle/deg sinusoidal grating, both with a duration of 2 frames at 60 Hz. Test and mask contrasts are expressed as Log[jnd], where 1 jnd is the unmasked threshold. Data points are from Foley[14] for similar conditions.

In a second simulation, we sought to estimate a time constant for the temporal filter that is applied to the contrast masking signal. We approximated the conditions of an experiment of Georgeson and Georgeson[15], in which threshold for a grating target was measured at various times relative to presentation of a grating mask of the same spatial frequency. As shown in Figure 6, a time constant of 0.04 seconds approximately reproduces the decay of masking following the masker (so-called "forward masking") but does not produce the substantial masking that occurs for test presentations that occur before the mask (so-called "backward masking"). At this time we do not attempt to model backward masking.

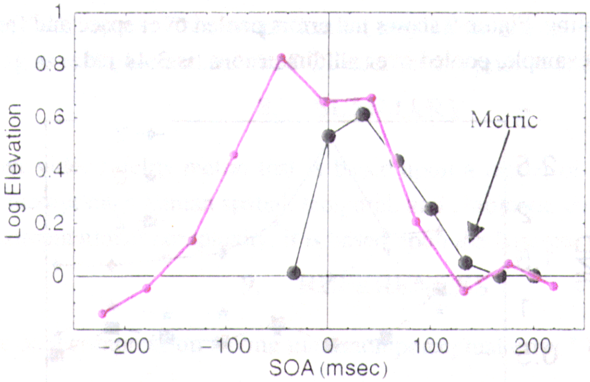

Figure 6. Contrast masking vs delay from mask to test (SOA). Test and mask were 1 cycle/deg sinusoidal gratings 2x2 deg in size with a duration 1/30 sec. Mask contrast was 0.23 (one log unit above threshold).

6. VIDEO SIMULATIONS

To illustrate the application of the DVQ metric to an image sequence, we created a small (24 x 32) short (8 frames) sequence, which we then corrupted via a DCT and quantization. These referenceand test sequences, and their differenceadded to a uniform gray background, are shown in Figure 7.

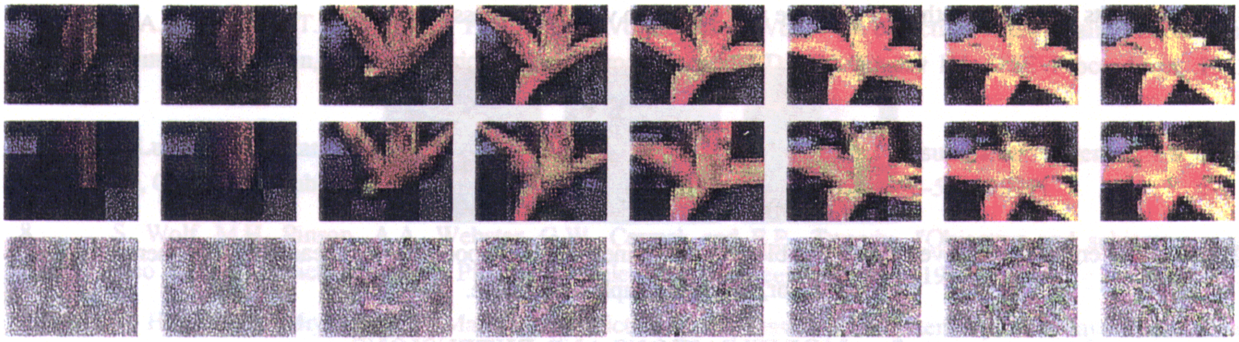

Figure 7. Reference (top) and test (middle) video sequences, and their difference (bottom).

The output of the DVQ metric is shown in Figure 8. Each panel is for one frame of one color channel, within which can be seen the individual 3 x 4 array of DCT coefficient blocks. The errors are most prominent in the latter half of the sequence, and in the Y channel. Errors in the color channels are predominantly at the lowest DCT frequencies.

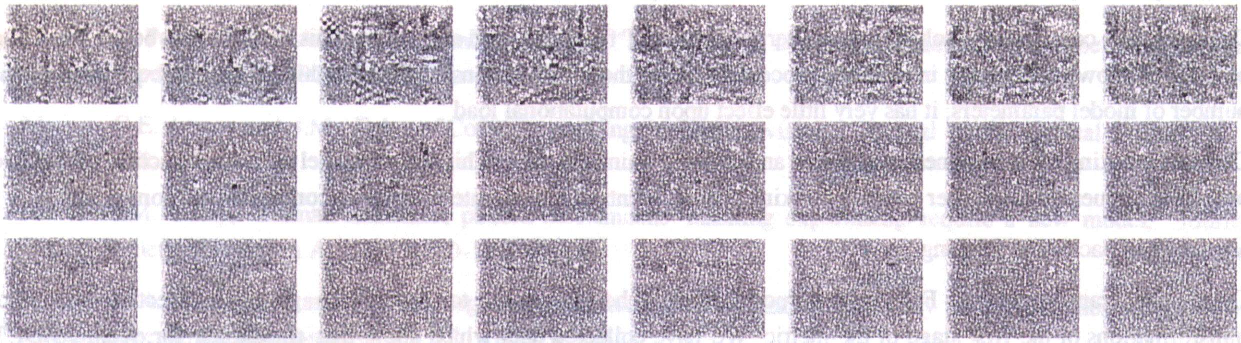

Figure 8. Perceptual errors computed by the DVQ metric for the test and reference sequence of Figure 7. From top to bottom, the three rows show perceptual errors in the Y, O, and Z color channels.

As an illustration of one form of pooling, Figure 9 shows jnd errors pooled over space and frequency, that is, over each panel in Figure 8. The total error for this example, pooled over all dimensions, is 3.44 jnd.

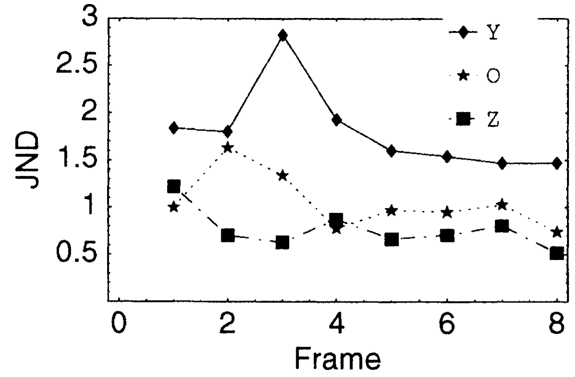

Figure 9. JND errors for each color channel and frame, pooled over space and frequency.

Another useful form of pooling is illustrated in Figure 10. Here the errors have been pooled over all dimensions except DCT frequency and color. The results show that the visible errors predominate at the low frequencies, and primarily in the Y channel. By means of this information, intelligent adjustments in the quantization matrices of the compression process can be made. This suggests a general method of adaptive rate control in video compression systems.

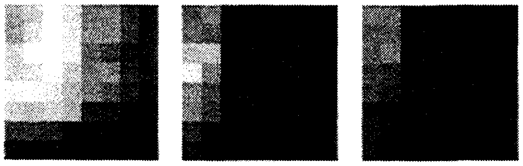

Figure 10. JND errors pooled over frames and blocks. Each panel shows the pooled error at each DCT frequency for each color, over the complete sequence.

7. MODIFICATIONS AND EXTENSIONS

Although a primary goal in this project has been to produce a metric with minimal computational or memory demands, a number of extensions and modifications to the basic scheme may be contemplated. Among these are the following.

- Higher order temporal filters for contrast sensitivity, light adaptation, and contrast masking. These would allow better modeling of temporal effects, at a substantial cost in computation and memory.
- Different time constants in each temporal filter for each DCT frequency and each color. This would allow better modeling of the well known differences in temporal processing along these dimensions. Although this is a great expansion in the number of model parameters, it has very little effect upon computational load
- Contrast masking via a nonlinear transducer and contrast-gain control[16]. This sort of model defers subtraction of the two processed sequences until after contrast masking occurs. It entails a moderate increase in computational complexity.
- Modeling of backward masking.
- Omit color transformations. For some purposes, it may be reasonable to bypass the gamma correction and color transformations of the first stage of the metric. We have collected data which show that thresholds for dynamic DCT noise on a display with a gamma of 2.2 are quite similar to those with a gamma of 1. Likewise, the YCbCr color space

can be used instead of the YOZ space. This simplification may allow one to apply the metric internal to the typical digital video encoder, using as reference and test the DCT coefficients prior to and following quantization.

8. SUMMARY

I have described a new objective video fidelity metric that is based upon a model of human spatial, temporal and chromatic visual processing. The metric incorporates human spatial, temporal, and chromatic contrast sensitivity, light adaptation, and contrast masking. To reduce computational complexity, it is based upon the Discrete Cosine Transform.

9. REFERENCES

1. A.B. Watson, "Image data compression having minimum perceptual error," US Patent 5,629,780, (1997).

2. A.B. Watson, G.Y. Yang, J.A. Solomon and J. Villasenor, "Visibility of wavelet quantization noise," IEEE Transactions on Image Processing, 6(8), 1164-1175 (1997).

3. A.B. Watson, "Perceptual optimization of DCT color quantization matrices," IEEE International Conference on Image Processing, 1, 100-104 (1994).

4. A.B. Watson, "Image data compression having minimum perceptual error," US Patent 5,426,512, (1995).

5. C.J.v.d.B. Lambrecht, "Color moving pictures quality metric," International Conference on Image Processing, I, 885-888 (1996).

6. A.A. Webster, C.T. Jones, M.H. Pinson, S.D. Voran and S. Wolf, "An objective video quality assessment system based on human perception," Human Vision, Visual Processing, and Digital Display IV, SPIE Proceedings, 1913, 15-26 (1993).

7. J. Lubin, "A Human Vision System Model for Objective Picture Quality Measurements," International Broadcasters' Convention, Conference Publication of the International Broadcasters' Convention, 498-503 (1997).

8. S. Wolf, M.H. Pinson, A.A. Webster, G.W. Cermak and E.P. Tweedy, "Objective and subjective measures of MPEG video quality," Society of Motion Picture and Television Engineers, 160-178 (1997).

9. T. Hamada, S. Miyaji and S. Matsumoto, "Picture quality assessment system by three-layered bottom-up noise weighting considering human visual perception," Society of Motion Picture and Television Engineers, 179-192 (1997).

10. K.T. Tan, M. Ghanbari and D.E. Pearson, "A video distortion meter," Picture Coding Symposium, 119-122 (1997).

11. A.B. Watson and D.G. Pelli, "QUEST: A Bayesian adaptive psychometric method," Perception and Psychophysics, 33(2), 113-120 (1983).

12. H. Peterson, A.J. Ahumada, Jr. and A. Watson, "An Improved Detection Model for DCT CoefficientQuantization," SPIE Proceedings, 1913, 191-201 (1993).

13. G.E. Legge and J.M. Foley, "Contrast masking in human vision," Journal of the Optical Society of America, 70(12), 1458-1471 (1980).

14. J.M. Foley, "Human luminance pattern mechanisms: masking experiments require a new model," Journal of the Optical Society of America A, 11(6), 1710-1719 (1994).

15. M.A. Georgeson and J.M. Georgeson, "Facilitation and masking of briefly presented gratings: time-course and contrast dependence," Vision Research, 27, 369-379 (1987).

16. A.B. Watson and J.A. Solomon, "A model of visual contrast gain control and pattern masking," Journal of the Optical Society A, 14, 2378 - 2390 (1997).