Watson, A. B., Hu, J., & McGowan, J. F., III. (2001). Digital video quality metric based on human vision. *Journal of Electronic Imaging, 10*(1), 20-29.

DVQ: A digital video quality metric based on human vision

Andrew B. Watson, James Hu, John F McGowan III

NASA Ames Research Center, Moffett Field, CA 94035

ABSTRACT

The growth of digital video has given rise to a need for computational methods for evaluating the visual quality of digital video. We have developed a new digital video quality metric, which we call DVQ (Digital Video Quality)[1]. Here we provide a brief description of the metric, and give a preliminary report on its performance. DVQ accepts a pair of digital video sequences, and computes a measure of the magnitude of the visible difference between them. The metric is based on the Discrete Cosine Transform. It incorporates aspects of early visual processing, including light adaptation, luminance and chromatic channels, spatial and temporal filtering, spatial frequency channels, contrast masking, and probability summation. It also includes primitive dynamics of light adaptation and contrast masking. We have applied the metric to digital video sequences corrupted by various typical compression artifacts, and compared the results to quality ratings made by human observers.

Key Words: digital video quality fidelity hvs metric perception vision quantization

1 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[2, 3, 4, 5]. These metrics incorporate simplified models of human visual sensitivity to spatial and chromatic visual signals. Here we describe a new video quality metric, which we call DVQ (Digital Video Quality)[1], 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. Here we provide a brief description of the metric, and give a preliminary report on its performance. DVQ accepts a pair of digital video sequences, and computes a measure of the magnitude of the visible difference between them. The metric is based on the Discrete Cosine Transform. It incorporates aspects of early visual processing, including light adaptation, luminance and chromatic channels, spatial and temporal filtering, spatial frequency channels, contrast masking, and probability summation. It also includes primitive temporal dynamics of both light adaptation and contrast masking. 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. We have applied the metric to digital video sequences corrupted by various typical compression artifacts, and compared the results to quality ratings made by human observers.

2 INTRODUCTION

The emerging infrastructure for digital video lacks a critical component: a reliable means for automatically measuring visual quality. Such a means is essential for evaluation of codecs, for monitoring broadcast transmissions, and for ensuring the most efficient compression of sources and utilization of communication bandwidths. In previous papers[1, 6], we gave a preliminary description of a new video quality metric, which we called DVQ. Here we provide a brief review of the design of the metric, and then describe application of this metric to a set of video materials for which human subjective ratings are available. This provides a test of whether the metric can accurately predict human subjective ratings.

Recently a number of video quality metrics have been proposed[7, 8, 9, 10, 11, 12]. 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.

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 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.

3 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.

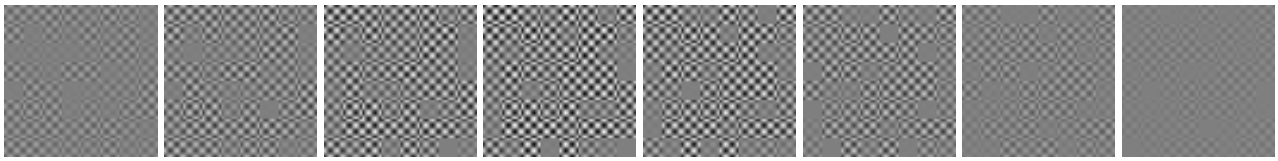


Figure 1. Dynamic DCT noise. In this example the DCT frequency was {3,3}, the number of blocks was 8 x 8, the frame rate was 60 Hz, the Gaussian time scale was 5 frames, and the temporal frequency was 2 Hz.

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[13], using a total of 32 trials/threshold. DCT frequencies tested were {0,0}, {0,1}, {0,2}, {0,3}, {0,5}, {0,7}, {1,1}, {2,2}, {3,3}, {5,5}, {7,7}. Temporal frequencies tested 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.

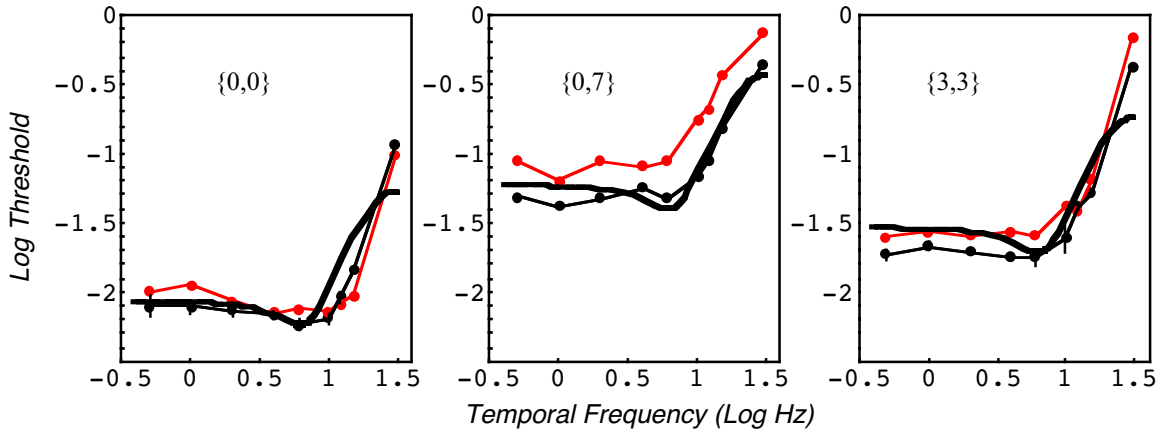


Figure 2. Selected contrast thresholds for dynamic DCT quantization noise. Points are data of two observers (JQH and HKK); error bars show ± 1 standard deviation. The numbers in braces in each panel show the horizontal and vertical DCT frequencies. The thicker curve is the model.

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.

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

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. Note that p can also be derived from the viewing distance specified in pixels as $p \approx$ *distance_in_pixels*/57.3[3].

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

The orientation function (Figure 3c) accounts for two effects: the higher threshold for oblique frequencies, and the imperfect visual summation between two component frequencies[14]. It is given by

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

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.

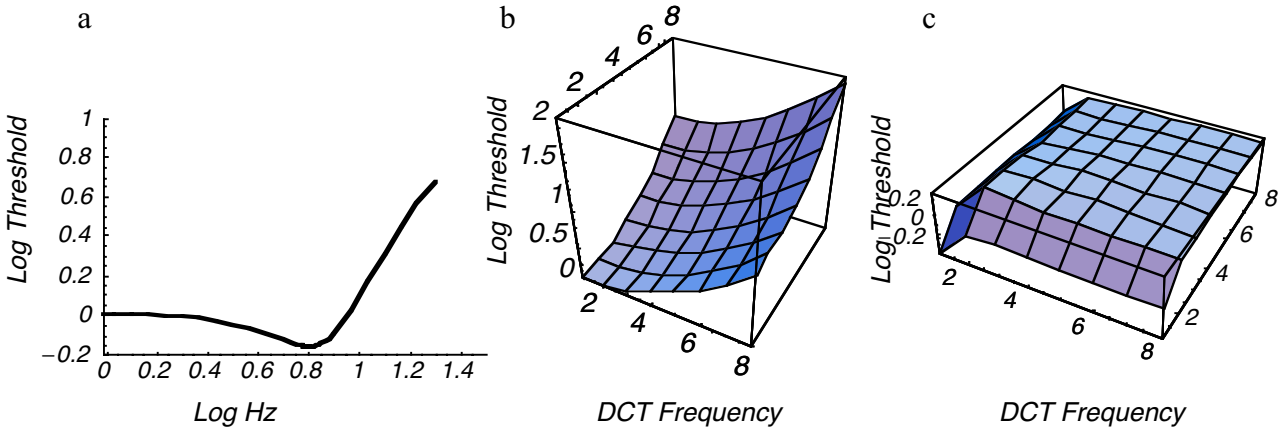


Figure 3. Temporal, spatial and orientation components of the dynamic DCT threshold model.

Because the spatial display resolution in these experiments was 32 pixels/degree, the highest horizontal or vertical frequency tested is a 16 cycles/degree. Examination of Figure 3 shows that we have not tested the complete usable range of spatial frequencies, particularly high frequencies, which might be especially relevant to high-resolution imagery such as HDTV. However, our mathematical model will extrapolate to these higher frequencies, and we suspect that because sensitivity to these frequencies is low, errors in our extrapolation will not have a major effect.

4 DVQ

All video quality metrics are inherently models of human vision. For example, if root-mean-squared-error (RMSE) is used as a quality metric, this amounts to the assumption that the human observer is sensitive to the summed squared deviations between reference and test sequences, and is insensitive to aspects such as the spatial frequency of the deviations, their temporal frequency, or their color. The DVQ (Digital Video Quality) metric is an attempt to incorporate many aspects of human visual sensitivity in a simple image processing algorithm. Simplicity is an important goal, since one would like the metric to run in real-time and require only modest computational resources. One of the most complex and time consuming elements of other proposed metrics[7, 11, 12, 15, 16] are the spatial filtering operations employed to implement the multiple, bandpass spatial filters that are characteristic of human vision. We accelerate this step by using the Discrete Cosine Transform (DCT) for this decomposition into spatial channels. This provides a powerful advantage since efficient hardware and software are available for this transformation, and because in many applications the transform may have already been done as part of the compression process.

Figure 4 is an overview of the processing steps of the DVQ metric. These steps are described in greater detail elsewhere[1], here we provide only a brief review. The input to the metric is a pair of color image sequences: reference (R), and test (T). The first step consists of various sampling, cropping, and color transformations that serve to restrict processing to a region of interest and to express the sequences in a perceptual color space. This stage also deals with de-interlacing and de-gamma-correcting the input video. The sequences are then subjected to a blocking (BLK) and a Discrete Cosine Transform (DCT), and the results are then transformed to local contrast (LC). Local contrast is the ratio of DCT amplitude to DC amplitude for the corresponding block. The next step is a temporal filtering operation (TF) which implements the temporal part of the contrast sensitivity function. This is accomplished through a suitable recursive discrete second order filter. The results are then converted to just-noticeable differences by dividing each DCT coefficient by its respective visual threshold . This implements the spatial part of the contrast sensitivity function (CSF). At the next stage the two sequences are subtracted. The difference sequence is then subjected to a contrast masking operation (CM), which also depends upon the reference sequence. Finally the masked differences may be pooled in various ways to illustrate the perceptual error over various dimensions (POOL), and the pooled error may be converted to visual quality (VQ).

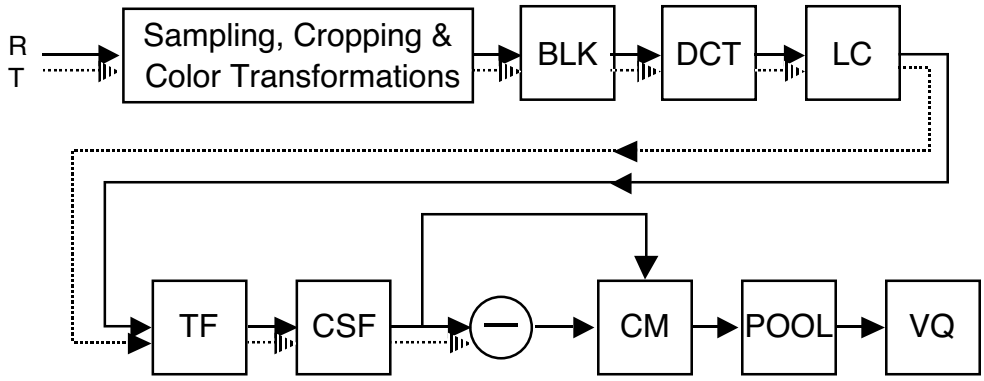


Figure 4. Overview of DVQ processing steps.

The parameters of the metric have been estimated from psychophysical data, both from the existing literature and from measurements of visibility of dynamic DCT quantization error described above.

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.

Color Transformations

The first step in the process is the conversion of both image sequences to the YOZ color space[14]. 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.

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.

Local Contrast

The DCT coefficients are converted to units of local contrast in the following way. First we extract the DC coefficients from 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_l. 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.

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.

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*. Use of the jnd as a measure of visibility in image quality metrics dates to the work of Carlson and Cohen[Carlson, 1980 #191].

Contrast Masking

Contrast masking is accomplished by first constructing a *masking sequence.* This begins as the reference sequence, after jnd 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* (typically 0.9), 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[17].

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*).

5 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.

6 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[18]. Foley's data also show a facilitation effect at sub-threshold mask contrasts, which we do not attempt to model.

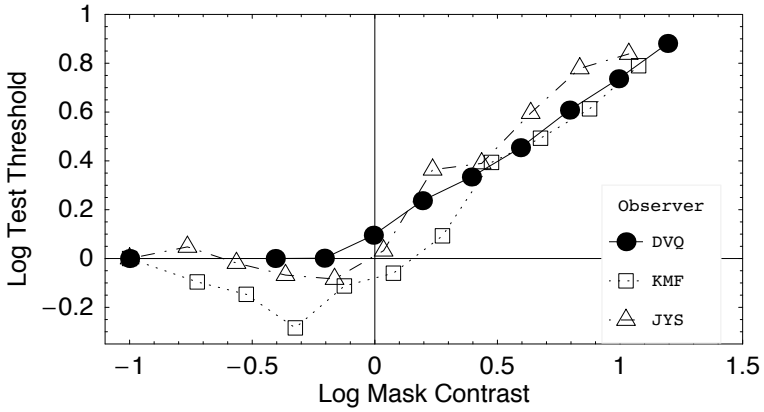


Figure 5. Test threshold versus mask contrast. Test was a 2 cycle/deg Gabor function, and mask a 2 cycle/deg sinusoidal grating, both with a duration of 2 frames at 60 Hz. Test and mask contrasts are expressed as Log[jnd], where 1 jnd is the unmasked threshold. Data points are from Foley[18] for similar 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[19], 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.

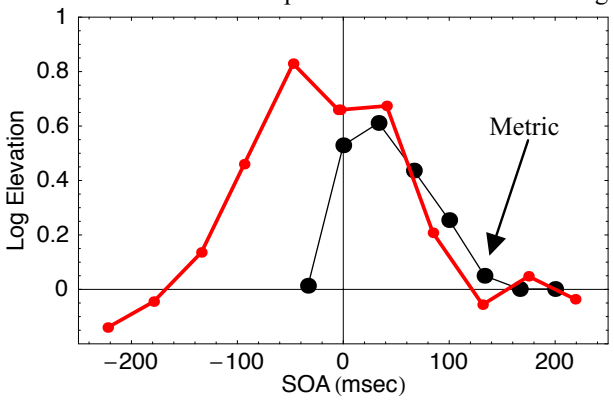


Figure 6. Contrast masking vs delay from mask to test (SOA). Test and mask were 1 cycle/deg sinusoidal gratings 2x2 deg in size with a duration 1/30 sec. Mask contrast was 0.23 (one log unit above threshold).

7 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 reference and test sequences, and their difference added to a uniform gray background, are shown in Figure 7.

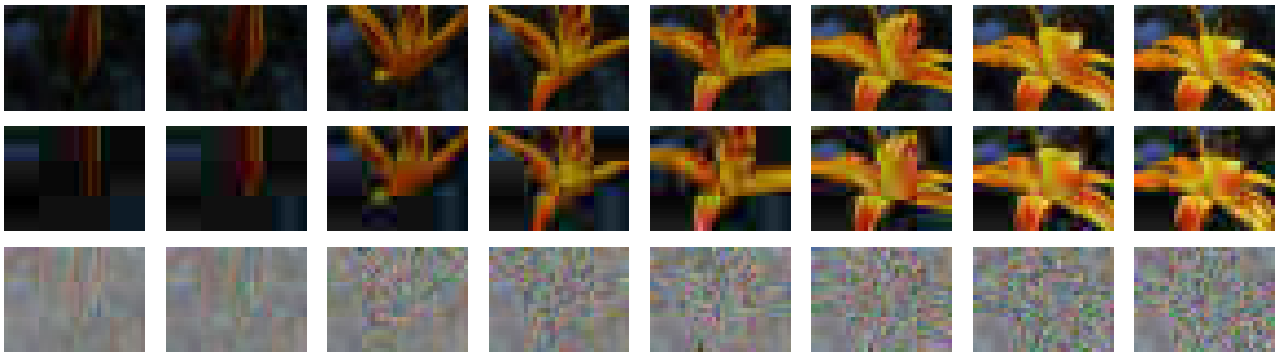


Figure 7. Reference (top) and test (middle) video sequences, and their difference (bottom).

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.

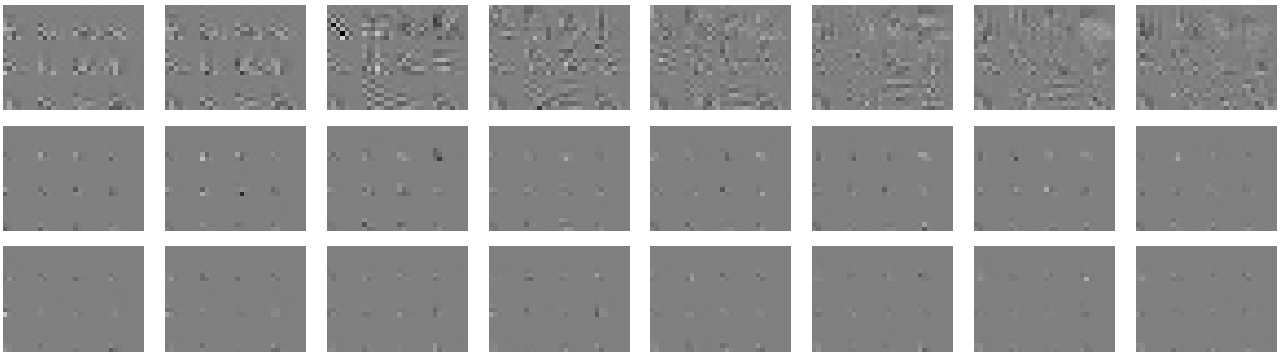


Figure 8. Perceptual errors computed by the DVQ metric for the test and reference sequence of Figure 7. From top to bottom, the three rows show perceptual errors in the Y, O, and Z color channels.

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.

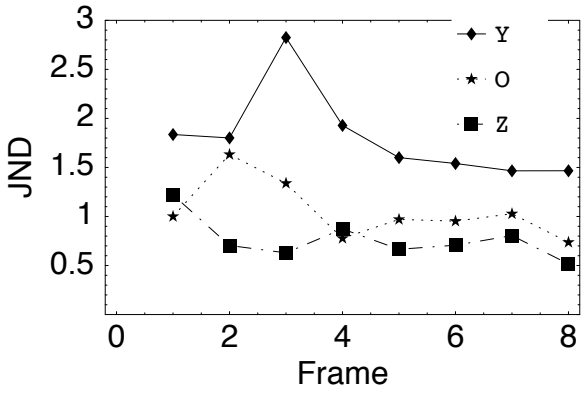


Figure 9. JND errors for each color channel and frame, pooled over space and frequency.

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.

8 TESTING THE METRIC

To test the DVQ metric we need to compare its results for specific video sequences to the quality estimates provided by human observers. There are few, if any, publicly available data sets of this kind. Most such materials are developed in commercial settings and subject to proprietary concerns. We were able to secure one data set, consisting of digital image sequences and associated subjective ratings. The data set is described more extensively elsewhere[20, 21]; here we provide only summary details regarding the video sequences and subjective data.

Video Sequences

The video materials consisted of a total of 65 sequences, of which 5 were reference sequences and 60 were processed sequences, obtained by passing the 5 reference sequences through 12 hypothetical reference circuits (HRCs) that will be described below. Each sequence was in ITU-601 PAL format[22] (576x720, interlaced, 4:2:2 sampling), and was 9 seconds (225 frames) in duration.

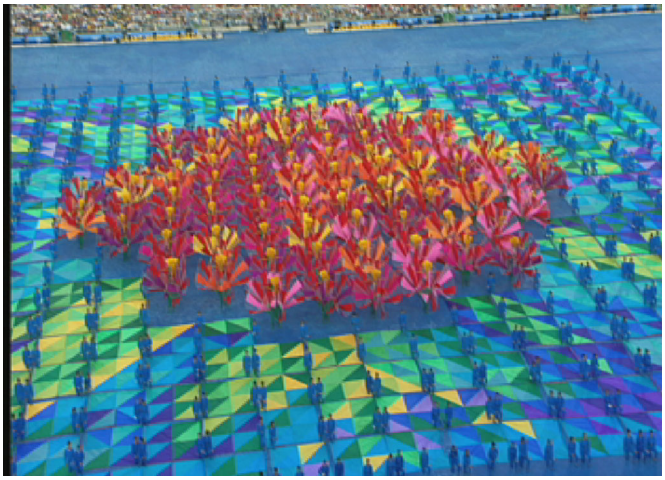
Reference Sequences

The five reference sequences were selected to span a wide spectrum of typical video, and to emphasize various challenges to video compression, such as saturated color, panning, rapid movement, fine detail, etc. The first frame of each of the five reference sequences is shown in Figure 11, along with one frame from a processed sequence.

football

wall

flower



barcelona

mobile

wall (HRC 1)

Figure 11. The first five images show the first frames of the reference sequences. The last image shows one frame from HRC 1 applied to “wall”.

Hypothetical Reference Circuits

Processed sequences were produced by passing the five reference sequences through each of 13 hypothetical reference circuits (HRCs). Each HRC consisted of a particular combination of MPEG codec, bit-rate, and possibly conversion from digital to analog (PAL) and back again to digital prior to encoding. The identities of the particular codecs are not important for present purposes, so we assign them the arbitrary labels A and B. An HRC consisting of no processing is also included. The thirteen HRC's are defined in Table 1.

HRC	Codec	Bit-rate (Mbits/sec)	PAL processing
1	A	2	
2	A	3	
3	A	4.5	
4	A	7	
5	A	10	
6	B	2	
7	B	3	
8	B	4.5	
9	B	7	
10	B	10	
11	A	3	yes
12	B	3	yes
13	none		

Table 1. HRC definitions.

Subjective data.

The subjective data were obtained from 25 observers using the Double Stimulus Continuous Quality Scale (DSCQS)[23]. In this method on each trial the observer views in succession a processed (test) sequence and the corresponding reference sequence, and assigns to each a quality rating between 0 and 100. Here we examine the difference in the scores for reference and test, which we will call the *impairment* score.

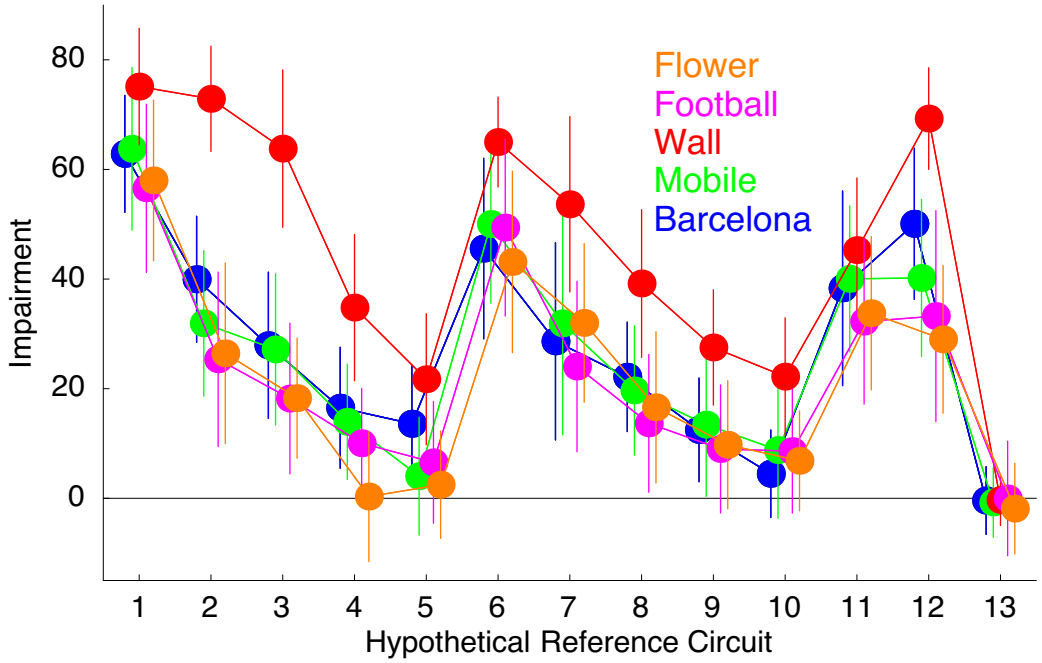


Figure 12. Impairment scores for five reference sequences processed by thirteen HRC's. Means and ± one standard deviation are shown

In Figure 12 we plot the mean impairment scores for each condition (combination of source and HRC). Each point is the mean of a single trial from each of 25 observers. The general pattern of results is similar for all five sources, though the "wall" sequence yields generally higher levels of impairment than the others. As expected, impairment declines with bit-rate, approaching zero at a bit-rate of 10 Mbits/sec for both codecs. Note the considerable variability among observers, which places a limit on the predictability of the results.

9 DVQ PREDICTIONS

The DVQ metric was prototyped in the Mathematica programming language, and subsequently implemented in c. Predictions shown here were computed on a Silicon Graphics Octane computer. To accelerate computations, we have also sometimes made use of software to distribute calculations for different sequences to an array of computers.

The DVQ metric computes elementary perceptual errors indexed by DCT frequency, block, color, and field of the video sequence, and allows selective pooling over subsets of these dimensions. In Figure 13, we show the perceptual error pooled over all dimensions except time (video field). The results are shown for the "flower" source and HRC's 1-5. Particularly for the lowest bit rate, the perceptual error is clearly bursty and periodic, the periodicity presumably due to the GOP structure.

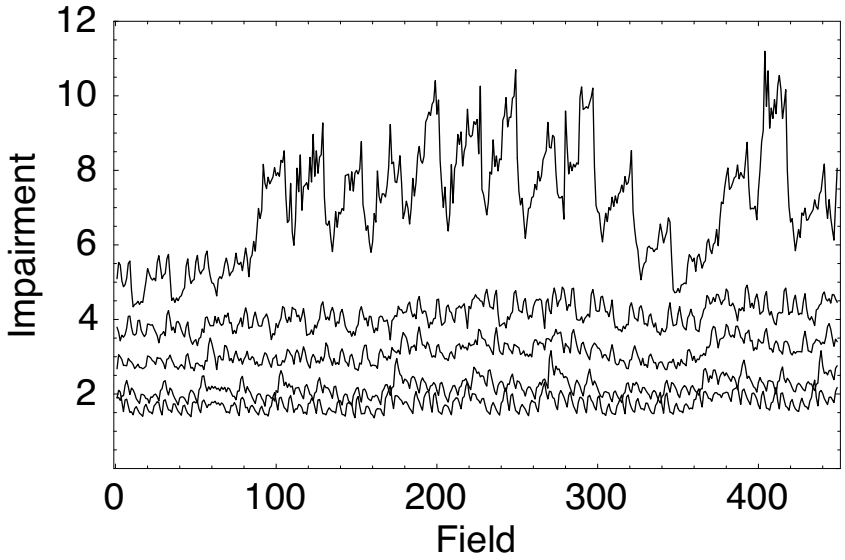


Figure 13. Impairment in each field for the source "flower" and HRC's 1-5 (codec A at bit-rates of 2-10 Mbit/sec).

Time-varying predictions of this sort would be useful for predicting time-varying subjective assessments of video quality[24], but regrettably the requisite sequences and ratings are not publicly available. The total impairment score for a sequence is obtained by pooling over fields, using a Minkowski metric with an exponent of 4[1]. The five traces above, pooled in this way, yield scores of {34.7, 19.14, 14.56, 10.3, 8.21}.

10 COMPARING DATA AND METRIC

A simple method of comparing the impairment predictions generated by DVQ and the subjective data is to compute the root-mean-squared (rms) error between predictions and data. However, it has frequently been observed that when plotted against simple objective measures such as rms error or bit-rate, subjective impairment scores show a sigmoidal form[25]. This is attributed to two aspects of human judgements: the flattening at the low end is attributed to a threshold, while the flattening at the high end is attributed to a saturation. In simple terms, impairments below a certain level are invisible, and impairments above a certain level are all considered about equally bad. Although such thresholds and saturation phenomena should be a part of an accurate model, they may be absent from current models, and thus in comparing model and data the possibility of a further non-linear transformation between model and data is often entertained. The non-linear transformation we have considered is a cubic polynomial. Thus we have first found the best-fitting cubic polynomial, and then computed the residual rms error. This fit is shown in Figure 14A. Finally, we present the results by plotting the data against the transformed predictions, as shown in Figure 14B. Following transformation, a unit slope straight line (as shown) is the best fitting polynomial. For this fit, the parameters of the polynomial were $\{a_0 = -1.635, a_1 = 0.573, a_2 = 0.0603, a_3 = -0.00085\}$.

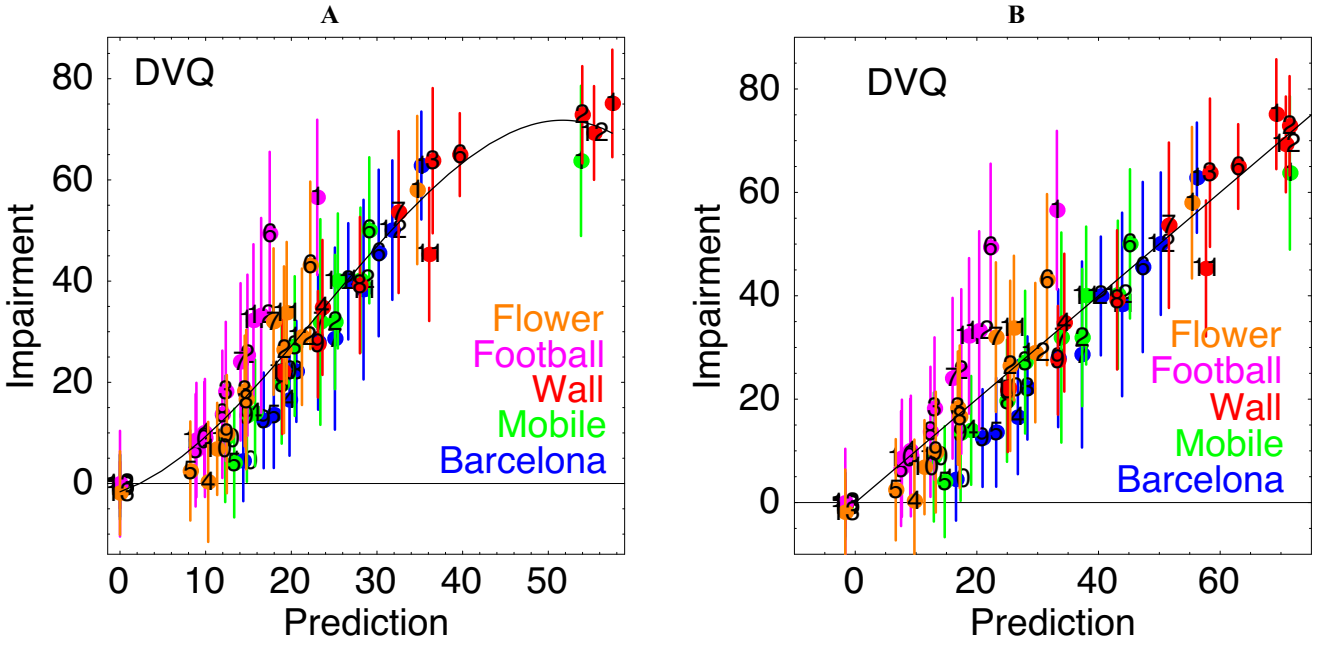


Figure 14. Comparison of DVQ prediction and subjective data. A: DVQ predictions vs data with fitted polynomial; B: transformed predictions. Numbers within points indicate the HRC.

The rms error of the fit of the DVQ metric is 14.61. For comparison, we have also computed the predictions of a model consisting of the rms error between the digital values of the reference and processed sequences (RMSE), and a model consisting simply of the inverse bit-rate of each sequence (MBPS). We have also assessed the predictions of the Sarnoff model[9] for these data, as reported in [21]. All of these may be compared to the rms error of the individual observers relative to their mean for a given condition (source and HRC), which is a measure of the underlying variability of the data. The rms errors for all of these models are noted in Table 2.

While more detailed statistical conclusions must await further analysis, some preliminary observations may be warranted. First, there is unlikely to be a statistical difference between DVQ and Sarnoff models for this data set. A choice between the two metrics must be based on other considerations, such as speed, simplicity, availability, or cost. Second, the variability in the data puts a bound on how well any metric can fit the data, and on how much effort we should expend on improvements to the models. Third, both Sarnoff and DVQ metrics appear to perform significantly better than the RMSE model. Furthermore, this data set does not exercise one feature of these models that should be expected to set them clearly above the RMSE metric, namely variations in display resolution and viewing distance.

Metric	**Rms Error**
DVQ	14.61
Sarnoff	14.40
RMSE	16.80
Mbits/sec	16.70
Condition Means	12.65

Table 2. Rms error of fit of various metrics to the subjective data.

Although the overall DVQ fit is reasonable, the “football” source appears to depart systematically from the model predictions. We have not yet discovered the explanation for this discrepancy. Interestingly, it is also evident in the fit of the

RMSE and Sarnoff models. One possibility is that the scales used by observers are different for each source sequence. In this case, for example, the observers may be more demanding in their quality expectations for the “football” sequence.

11 CONCLUSIONS

We have evaluated the performance of the DVQ video quality metric by comparing its predictions to judgements of impairment made by 25 human observers viewing 5 reference sequences as processed by 12 HRCs. The DVQ metric performs about as well as the de facto standard (the Sarnoff model), and considerably better than models based on simple bit-rate or rms error. The quality of the predictions suggests the metric may be useful in practical applications. However the metric shows what appear to be systematic failures of prediction (the “football” sequence at low bit-rates) whose explanation awaits further research.

12 ACKNOWLEDGEMENTS

We thank Alexander Schertz of the Institut fuer Rundfunktecknik GmbH, and Tektronix, Inc. for providing the digital video sequences and subjective data used in this report. This work was supported by NASA Grant 199-06-12-39 from the Office of Life Science and Microgravity Research.

13 REFERENCES

1. A.B. Watson,"Toward a perceptual video quality metric," *Human Vision, Visual Processing, and Digital Display VIII*, 3299, 139-147 (Year).

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

3. 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).

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

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

6. A.B. Watson, J. Hu, J.F. McGowan, III and J.B. Mulligan,"Design and performance of a digital video quality metric," *Human Vision, Visual Processing, and Digital Display IX*, Proc. SPIE, 3644, 168-174 (Year).

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

8. 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 (Year).

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

10. 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 (Year).

11. 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 (Year).

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

13. A.B. Watson and D.G. Pelli,"QUEST: a Bayesian adaptive psychometric method," *Perception & Psychophysics*, 33(2), 113-20 (1983).

14. H. Peterson, A.J. Ahumada, Jr. and A. Watson,"An Improved Detection Model for DCT Coefficient Quantization," *Human Vision and Electronic Imaging*, Proc. SPIE, 1913, 191-201 (Year).

15. A.B. Watson,"Multidimensional pyramids in vision and video," (1991).

16. A.B. Watson,"Perceptual-components architecture for digital video," *Journal of the Optical Society of America A*, 7(10), 1943-1954 (1990).

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

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

19. 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).

20. M. Ravel, J. Lubin and A. Schertz, "Pruefung eines Systems fuer die objektive Messung der Bildqualitaet durch den Vergleich von objektiven und subjektiven Testergebnissen," **FKT vol.52 nr. 10**, (1998).

21. A. Schertz, "IRT/Tektronix Investigation of Subjective and Objective Picture Quality for 2-10 Mbit/s MPEG-2 Video," Technischer Bericht des IRT **nr. B 159/97**, (1997).

22. ITU-R,"Recommendation ITU-R BT.601-5, Studio encoding parameters of digital television for standard 4:3 and wide screen 16:9 aspect ratios,", (1995).

23. ITU-R, "Methodology for the subjective assessment of the quality of television pictures," International Telecommunications Union **BT.500-7**, (1995).

24. R. Hamberg and H.d. Ridder,"Continuous assessment of perceptual image quality," *Journal of the Optical Society A*, 12(12), 2573-2577 (1995).

25. J. Allnatt,*Transmitted picture assessment*, John Wiley & Sons, New York (1983).