Watson, A. B., & Kreslake, L. (2001). Measurement of visual impairment scales for digital video. *Proceedings of the SPIE, vol. 4299,* 79-89.

Measurement of visual impairment scales for digital video

Andrew B. Watson[a] and Lindsay Kreslake[b]

[a]NASA Ames Research Center, Moffett Field, CA 94035-1000, abwatson@mail.arc.nasa.gov

[b]Foothill College, Los Altos, CA

ABSTRACT

The study of subjective visual quality, and the development of computed quality metrics, require accurate and meaningful measurement of visual impairment. A natural unit for impairment is the JND (just-noticeable-difference). In many cases, what is required is a measure of an impairment scale, that is, the growth of the subjective impairment, in JNDs, as some physical parameter (such as amount of artifact) is increased.

Measurement of sensory scales is a classical problem in psychophysics. In the method of pair comparison, each trial consists of a pair of samples and the observer selects the one perceived to be greater on the relevant scale. This may be regarded as an extension of the method of forced-choice: from measurement of threshold (one JND), to measurement of the larger sensory scale (multiple JNDs). While simple for the observer, pair comparison is inefficient because if all samples are compared, many comparisons will be uninformative. In general, samples separated by about 1 JND are most informative. We have developed an efficient adaptive method for selection of sample pairs. As with the QUEST adaptive threshold procedure[1], the method is based on Bayesian estimation of the sensory scale after each trial. We call the method Efficient Adaptive Scale Estimation, or EASE ("to make less painful").

We have used the EASE method to measure impairment scales for digital video. Each video was derived from an original source (SRC) by the addition of a particular artifact, produced by a particular codec at a specific bit rate, called a hypothetical reference circuit (HRC). Different amounts of artifact were produced by linear combination of the source and compressed videos. On each pair-comparison trial the observer selected which of two sequences, containing different amounts of artifact, appeared more impaired. The scale is estimated from the pair comparison data using a maximum likelihood method. At the top of the scale, when all of the artifact is present, the scale value is the total number of JNDs corresponding to that SRC/HRC condition.

We have measured impairment scales for 25 video sequences, derived from five SRCs combined with each of five HRCs. We find that EASE is a reliable method for measuring impairment scales and JNDs for processed video sequences. We have compared our JND measurements with mean opinion scores for the same sequences obtained at one viewing distance using the DSCQS method by the Video Quality Experts Group (VQEG), and we find that the two measures are highly correlated. The advantages of the JND measurements are that they are in absolute and meaningful units and are unlikely to be subject to context effects. We note that JND measurements offer a means of creating calibrated artifact samples, and of testing and calibrating video quality models.

1. BACKGROUND

1.1. Need for accurate subjective measures of video quality

The design and use of digital video systems entail difficult tradeoffs amongst various quantities, of which the two most important are cost and visual quality. While there is no difficulty in measuring cost, beauty remains locked in the eye of the beholder. However in recent years a number of computational metrics have been developed which purport to measure video quality or video impairment. Metrics of this sort would be very valuable in providing a means for automatically specifying, monitoring, and optimizing the visual quality of digital video.

One impediment to the development of impairment metrics is that they must be designed to mimic, and must be tested against, a set of human subjective data. Until recently, no public source of such data existed.

1.2. VQEG Project

The Video Quality Experts Group (VQEG) has recently completed a large study comparing subjective data and predictions from a set of models[2, 3]. The data consisted of observers rating 20 source videos (SRCs) as processed by 16

hypothetical reference circuits (HRCs). An HRC is a particular set of processing operations, such as compression at a particular bit-rate. About 300 observers took part in the VQEG study. The ratings were obtained using the Double Stimulus Continuous Quality Scale (DSCQS) method of ITU-R BT.500-8[4].

There are several problems with rating data of this sort. First, they are quite variable. Second, they are subject to criterion and context effects. For example, the ratings given will depend upon the range of quality used in the experiment. Third, the scale on which they are rated has no inherent meaning, since different experiments may use different scales and different ranges of quality. An alternative approach is to use discriminability data to construct a measure of the perceptual difference between a reference and a processed video[5].

2. MEASUREMENT OF JND SCALES

2.1. The concept of JND scales

In the approach described here , rather than asking the observer to rate a given video, we ask the observer which of two videos is more impaired. This is called "pair comparison." From the responses to that simple question, we hope to measure the observer's internal "perceptual scale" for visual impairment. The idea is that each video gives rise to a mental estimate of impairment. This perceptual impairment, as a function of increasing physical impairment, is what we mean by the perceptual scale. This scale would be measured in units of JND (just-noticeable-differences).

2.2. Thurstone Scaling

We derive the scale from the pair comparisons by means of Thurstone's "Law of Comparative judgement"[6]. Thurstone proposed that physical sensory stimuli (such as sounds) might give rise to sensory magnitudes arranged along a one-dimensional internal sensory scale (such as loudness), as pictured in Figure 1. In this figure, two particular intensities x_1 and x_2 give rise to perceptual intensities $\Psi(x_1)$ and $\Psi(x_2)$.

The sensory magnitude varies from presentation to presentation, due to the inevitable variability of neural systems. In one particular case (Thurstone's "Case Five"), the distributions are assumed to be Normal, with a standard deviation of 1. In that case, the probability of a correct judgment in a pair comparison is a function only of the distance between the sensory magnitudes induced by the two intensities of the pair. We can therefore estimate these distances by finding which values would most likely give rise to the data in hand.

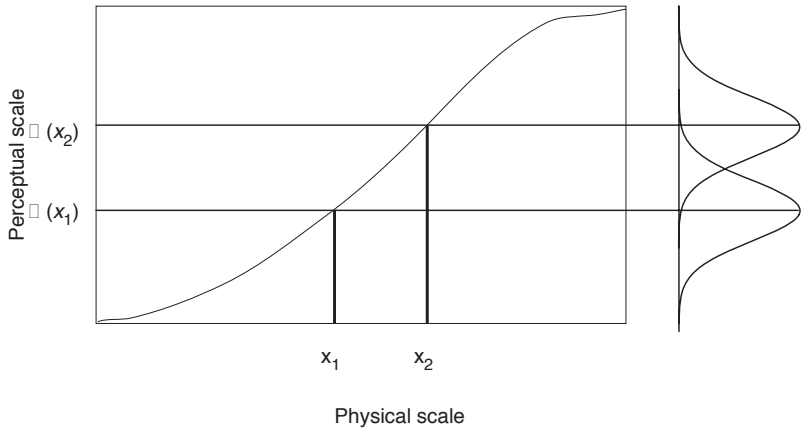

Figure 1 Thurstone Model.

If we specify trail number i by the record $\{x_{i,1}, x_{i,2}, R_i\}$ where $x_{i,1}$ and $x_{i,2}$ are the physical intensities and R_i is the response (R=1 if correct, 0 if incorrect), then the probability of being correct on trial number i is

$$P_i = C\left[\frac{\psi(x_{i,2}) - \psi(x_{i,1})}{\sqrt{2}}\right] \quad (1$$

where C is the cumulative Normal probability density function. After a sufficient number of trials, we can then estimate the perceptual scale function Ψ by maximizing the likelihood L,

$$L = \left(\prod_{i:R_i=0} 1 - P_i \right) \left(\prod_{i:R_i=1} P_i \right) \quad (2$$

with respect to some parameters that define the function Ψ. We consider two definitions of the scale function Ψ, and two resulting estimation methods: *sampled* and *functional*.

2.3. Sampled estimate of the scale function

Here we consider a scale function that is defined piecewise by its sampled values at each of the physical intensities used in an experiment. The parameters of the function are simply the values of the function at those points (more precisely, their differences). In practice, the parameters are estimated by minimizing the negative log of the likelihood function. An example of the results of this procedure for one session is shown in Figure 2. Each point in the figure corresponds to a weight used in the session. The vertical coordinate is the estimated number of JNDs (from x = 0) at that point. One point on this function will be of particular interest below: the value at x=1. We call this parameter J.

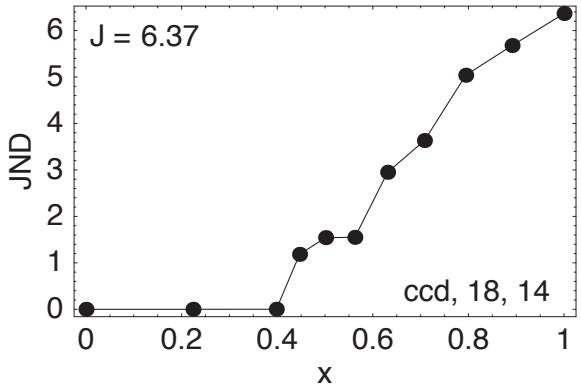

Figure 2. Sampled estimate of the scale function. The label in the lower right indicates observer and condition.

2.4. Functional estimate of the scale function

In this method we assume a particular mathematical parameterized form for the scale function, and fit it to the data, again using a maximum likelihood method. The function we have used is a thresholded power function:

$$\psi(x: M, T, Q) = M(1-T)^{-Q} \operatorname{Max}(0, x-T)^{Q} \quad (3$$

where x is the intensity, M is the total JNDs, T is the threshold below which intensity the scale remains at zero, and Q is the exponent of the power function. This function was in part inspired by results such as those in Figure 2, and also because this function is widely used in psychophysical scaling theory[7, 8]. The first term is present to ensure that the parameter M is the value of the function at x=1.

The results of applying this method to one session are shown in Figure 3, along with the previous piecewise estimates from Figure 2. The estimated parameters are also shown in the figure. In this case, we see very good agreement between the two methods, both in the overall course of the functions, and in the asymptotes (J and M). Agreement between the two methods is not always as good as that in (see Section 5.1) but on average the function adopted (Eqn 2) appears to be very reasonable.

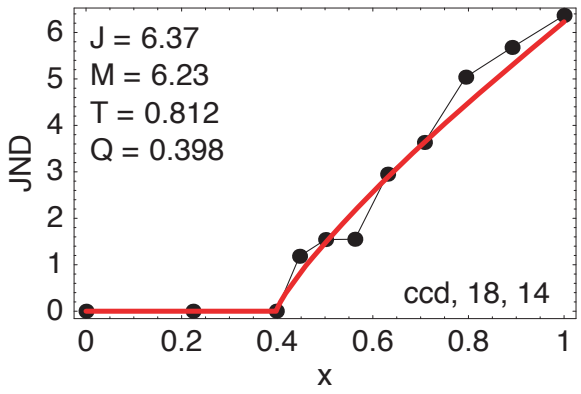

Figure 3. Functional estimate of the scale function.

3. EFFICIENT ADAPTIVE SCALE ESTIMATION: THE EASE METHOD

The preceding sections describe how to use the Thurstone model and maximum likelihood estimation to derive the scale function from a set of pair-comparision data. However, it should be clear that not all pairs yield useful results: if they are too far apart, the observer will always be correct, while if they are too close, the observer will be at chance. Roughly speaking, we would like the pairs spaced about one JND apart, and we would like them to span the entire physical range of interest. Unfortunately, to arrange such placement we need the scale function, which is what we are trying to estimate. We adopt an adaptive, iterative approach to this problem. We have given this new method the name EASE (Efficient Adaptive Scale Estimation). Here we give a brief description of this method. A longer report is in preparation.

EASE proceeds as pictured in Figure 4. A mathematical form is assumed for the scale function (we have typically used Eqn. 2 with Q fixed at 1). Initial parameters are guessed. Based on the function, a set of pairs of intensities is selected that are approximately a specified number of JNDs apart (jnd_step) and that span the intensity range of interest. One trial is then conducted at each of those pairs. All of the data are then used to estimate the function parameters. If sufficient data have been collected, the procedure stops, and the data may then be used to estimate the JNDs for the condition. If more trials are needed, a new set of pairs is derived from the newly estimated function, and the procedure continues. The termination rule we have used is to test whether the ratio trials/M >= 32. Simulations indicate that this is a reasonable compromise between accuracy and duration of the experiment.

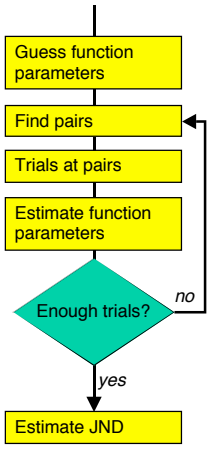

Figure 4. Flow diagram for the EASE method.

4. JND IMPAIRMENT SCALES FOR DIGITAL VIDEO

We turn now to the problem of measuring JND scales for impairment of digital video. Our goal here is to estimate the perceptual distance, in units of JND, between two video samples: one the *reference*, the other, that same sequence after it has been impaired by some form of processing (typically compression). These two sequences will form the end-points of a JND scale that we will measure using the EASE method.

4.1. Video Stimuli

We used a total of 25 conditions. These consisted of the complete matrix obtained by combining 5 SRCs with 5 HRCs. All conditions employed 60 Hz 525 line video. The selected SRCs are shown in Figure 5, along with the identifying number used in the VQEG study[3].

Figure 5. SRCs used in the experiment.

The HRCs are indicated in Table 1. This table also shows the mean DMOS (differential mean opinion score) obtained by each HRC in the VQEG experiment. These values were used to select a set of HRCs that spanned the quality range employed in the VQEG experiments.

HRC	Mean DMOS	Mbps	CODEC	Details
7	5.82	6	mp@ml	
5	13.83	8 & 4.5	mp@ml	Two codecs concatenated
9	21.18	3	mp@ml	
14	33.35	2	mp@ml	Horizontal resolution reduction
15	45.75	0.768	H.263	CIF, Full Screen

Table 1. Hypothetical Reference Circuits used in the experiment.

For each condition, we created a set of 21 videos that we call *blends*.A blend is a video that is the linear combination between two videos, a source video and that same video modified by a particular HRC. Thus if we write s for SRC number and h for HRC number (with h=0 for the special case of the reference), and write $v(s,h)$ for the video corresponding to a condition $\{s,h\}$, then a blend with intensity x is given by

$$\text{blend}(s,h,x) = (1-x)v(s,0) + x\ v(s,h) \quad (4$$

Note that the artifact due to a particular HRC can be written

$$\Delta(s,h) = v(s,h) - v(s,0) \quad (5$$

and thus the blend can also be written in the following way, which emphasizes that x is the proportion of the artifact present in the blend,

$$\text{blend}(s,h,x) = v(s,0) + x\Delta(s,h)\ . \quad (6$$

The arithmetic above should be understood as being applied to the raw numbers within the ITU-601 file that specify the values of Y and down-sampled Cb and Cr. Blends are a simple way of controlling, in a quantitative way, the amount of a particular artifact that is added to a source video[9]. We have used a series of 21 intensities spaced logarithmically from 0.1 to 1. These intensities are the samples along the physical artifact dimension whose corresponding perceived impairments we wish to measure.

4.2. Viewing Conditions, Apparratus, and Methods

Each trial consisted of two seven second video presentations, separated by a 1 second pause. After the presentations, the observer indicated which of the two videos was more impaired by pressing a button on a joy-stick computer input device. The computer then provided feedback ("right" or "wrong") using recorded voice. Trials were conducted using the EASE procedure in blocks of 32 trials. After each block (about 10 minutes) the observer was instructed to rest for five minutes. Ease parameters were trials/M = 32, and jnd_step = 1.273.

Video sequences were presented under computer control and displayed on a calibrated studio quality television monitor capable of displaying ITU-601 digital video streams. The display apparatus consists of an SGI Octane computer with SDI serial digital video input/output board, a Ciprico FibreChannel Disk Array, and a SONY BVM 20E1U monitor. Viewing conditions conformed to ITU-R BT.500-8[4]. The viewing distance was five picture heights.

Each presentation was seven seconds in duration, comprising the first seven seconds of each VQEG sequence. Observers were non-experts drawn from the local student and NASA population who were paid for their services and usually participated in a single 1-2 hour session. They were checked for normal color vision, and normal or corrected-to-normal acuity. We also recorded their age and gender.

5. RESULTS

Here we report data for 100 observers. For each condition, we collected at least three estimates of M, each from a different observer (except for SRC 21 at HRC 7 which has only two observers). For each estimate, we collected at least 20 trials per estimated value of M (trials/M >= 20). After data were collected, we used the methods described above to obtain a maximum likelihood estimates of M and J.

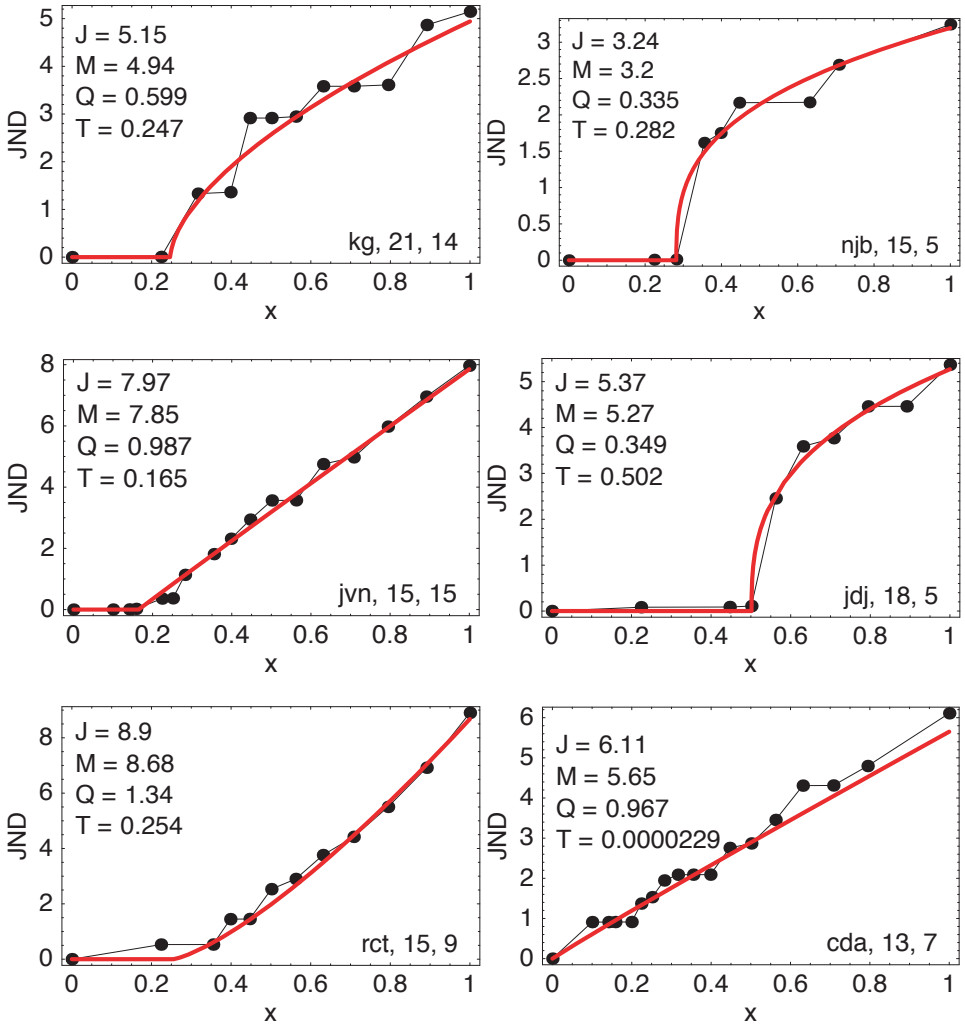

Figure 6. Six examples of estimated scale functions obtained using the EASE method. Each panel is for one observer and one condition. The observers initials, the HRC, and the SRC are indicated in the lower right of each panel. The smooth curve is the maximum likelihood fit of Eqn 3. The points are estimates of the sampled scale function.

Figure 6 shows six examples of fits to individual data sets. These six were chosen to show something of the variety and regularity of the results. The figure shows first of all that we are able to use the EASE method to estimate JND scale

functions for video impairment. Comparison of the smooth curve and sampled curve also suggests that our choice of fitting function (Eqn 3) was reasonable. The curves show instances in which there is a substantial threshold (T) and ones in which there is not; cases in which the exponent Q is greater than 1, less than one, and equal to one. The behavior and interaction of the parameters T and Q is beyond the scope of this report, however.

5.1. Comparing *J* and *M*

As noted above, we have considered two methods of estimating the total JND measure for a particular condition. The first measure J is obtained by estimation of the values of the JND function at each intensity used, while the second measure M is a parameter of a fitted function. Although *J* is a more model-free measure, we have noted that when too few trials/JND are collected, the value of *J* is biased upwards. This is shown in

Figure 7. Apart from this upward bias in *J*, which is more evident at larger values of *M*, the two measures give remarkably close results. In subsequent analyses we use the estimate *M* to avoid the potential bias in *J*.

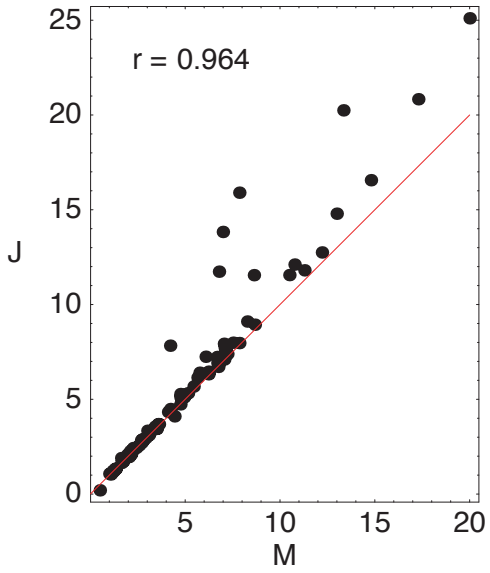

Figure 7. Comparison of estimates *M* and *J* of the JND for each measurement of each condition The Pearson correlation is indicated. The line is *J*=*M*.

5.2. Variability

The repeatability of our JND measurements is an obvious concern. An assessment of this can be obtained by looking at the standard deviation of our measurements. These are shown in Figure 8. Because the variance appears to increase with M, we also compute the coefficient of variation (sd/mean), which has a mean value of 0.35. It should be noted that this figure includes any between-observer variability, since the replications are from different observers.

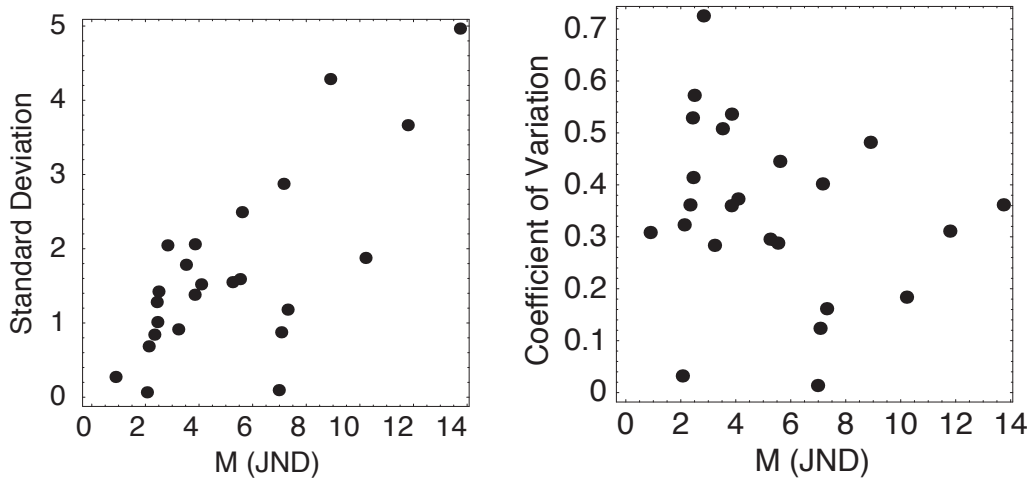

Figure 8. Standard Deviation and Coefficient of Variation of M as a function of M.

5.3. Effects of HRC and SRC

The means are shown in Figure 9. Lines connect conditions that share a common SRC, and the horizontal axis indicates the HRC. The figure shows that we were able to measure values of M as low as 1 JND and as large as 14 JND. In qualitative terms, the JND values vary in expected ways with HRC and SRC: lower bit rates and more "critical" sequences (e.g. 15, the notorious "mobile and calendar") yield larger JNDs.

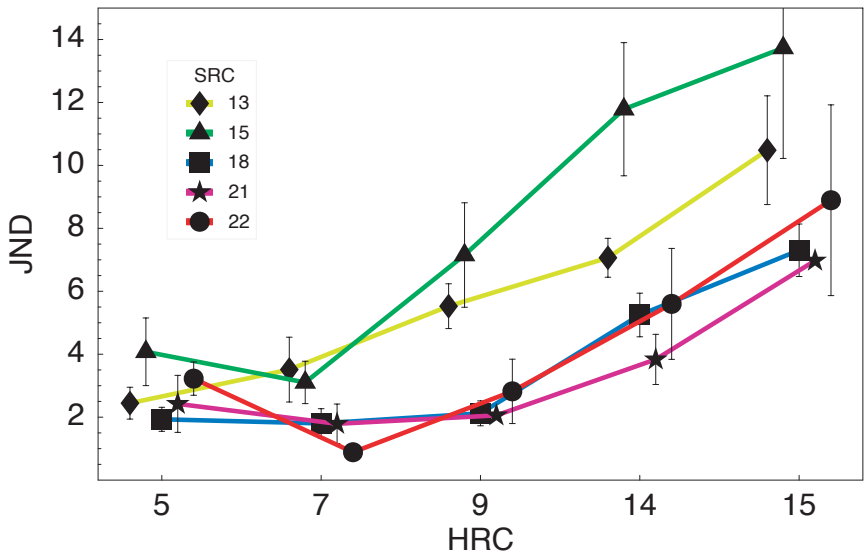

Figure 9. JND estimates (M) for 25 conditions based on five SRCs and five HRCs..Error bars indicate plus and minus one standard error. Each point is the mean of at least 3 observers, except for SRC 21 at HRC 7 which has only two. The curves for the various SRCs have been staggered horizontally for clarity.

5.4. JND vs DMOS

It is of interest to compare our JND measurements with the DMOS (differential mean opinion scores) measurements obtained by VQEG. This comparison is shown in Figure 10. The Pearson correlation between the two measures is 0.905. This agreement between the two varieties of measurement is reassuring, and suggests that JND measurements are at least as valid as DSCQS measurements of video quality. The JND measurements, of course, have the advantage of being context-free and on an absolute and meaningful scale.

The straight line shows the best fitting linear relation, constrained to pass through zero, which has a slope of about 0.231. The curved line is the best-fitting quadratic, with the form

$$JND = 1.917 + 0.125\,DMOS + 0.0012\,DMOS^2 . \quad (7$$

The non-zero intercept suggests that the JND measurement method described here is a more sensitive technique than the DSCQS method. Note that the precise relation between JND and DMOS will depend upon the context effects present in the DSCQS experiment.

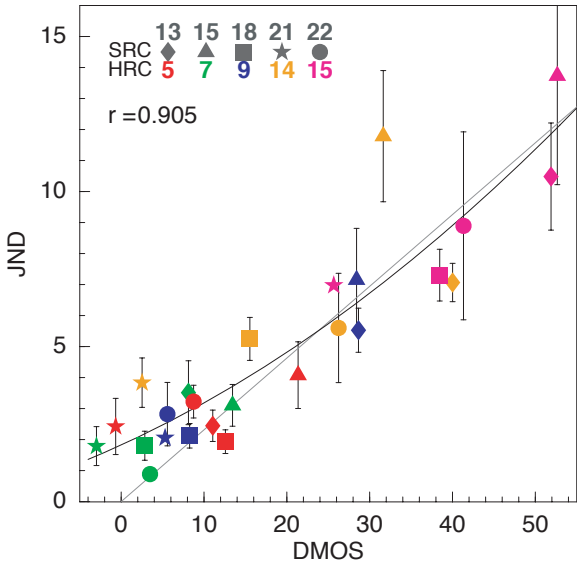

Figure 10. JND versus DMOS. The legend indicates the HRC and SRC for each condition. Only JND error was considered in performing this fit.

5.5. Calibration of Impairment metrics

One application of JND measurements is to calibrate computational impairment metrics. Here we illustrate this process with the DVQ model developed at NASA Ames Research Center[10]. In Figure 11 we show the relationship between DVQ impairment estimates and JND values for the same sequences. The Pearson correlation between the two quantities is 0.85.

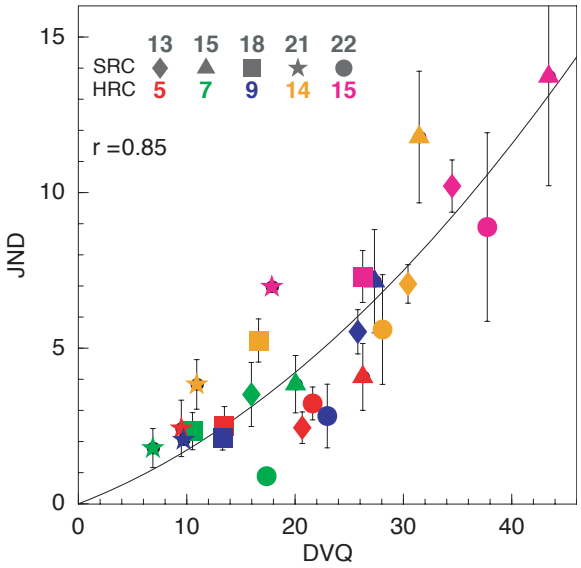

Figure 11. Relation between JND and DVQ estimates. Each point corresponds to the mean for one condition. The legend indicates the HRC and SRC for each condition. The curve is the best fitting quadratic constrained to go through the origin. The Pearson correlation between the two quantities is also indicated.

To describe further the relation between the two measures we have fit a quadratic function with an intercept of zero (so that both measures will be zero when the physical artifact is zero). The curve in the figure is the best fitting function of this sort, and is given by the relation

$$JND = 0.133\,DVQ + 0.00389\,DVQ^2 . \quad (8$$

This formula allows a simple recalibration of the DVQ values into units of JND. This would have considerable value for the users of such metrics, as it would allow them to express their results in standard units, and in units that have an intuitive meaning. This re-calibration is one of the primary values of the JND measurements.

6. APPLICATIONS OF JND MEASUREMENTS

Here we provide a brief description of three potential applications for the JND measurements reported here.

6.1. Calibration of Automatic Impairment Metrics

The first application is calibration of existing and future video impairment/quality metrics. At present, these metrics yield arbitrary units which have only relative meaning. Through correlation with JND measures, they can be converted to absolute JND measures. We have illustrated this calibration process above in Section 5.5.

6.2. Design of Impairment Metrics

A second application is the design of impairment/quality metrics. The detailed information about the growth of the scale function with increasing artifact provides a basis for design of metrics with similar properties, and presumably better prediction accuracy. Many metrics incorporate thresholds and other non-linearities, and these should be designed so as to render the scale functions shown here. For this purpose, one would need a more complete analysis of the regular features of these functions, as well as their variability. We hope to provide this in a future paper.

6.3. Standardized Samples of Impaired Video

A third application of these data is the creation of standard samples of video with specified JND levels. The JND values for each of the 25 sequences used here are know known. These sequences, along with their JND values, could be distributed on DVD or other media. These could serve as a basis for visual comparison in a variety of video production and distribution applications. They would also be a means of educating users of automated metrics as to the meaning of the numerical JND outputs.

6.4. Testing of Impairment Metrics

A final purpose is as a dataset for testing of video quality metrics. An advantage of this dataset is that it specifies subjective artifacts in absolute measures (JNDs), and the test could be on absolute, rather than relative terms (as in VQEG). A disadvantage of this dataset is that the sequences are known, and metrics could be tuned for these specific sequences. Nonetheless, testing of this sort would provide at least a minimal form of certification for proposed quality metrics.

7. CONCLUSIONS

1. We have proposed and implemented a new method for estimating subjective visual impairment of digital video in terms of a scale of perceived impairment measured in units of JND. The method is based on collection of pair-comparisons of samples of video with different amounts of impairment.
2. We have designed and implemented an efficient method (EASE) of collecting the pair comparison data. The EASE method appears to be a reliable method for measuring JNDs for processed video sequences.
3. We have collected JND measurements for 25 conditions consisting of 5 SRCs processed by 5 HRCs. These conditions were a subset of the VQEG conditions.
4. We have performed preliminary analyses of the JND data, with respect to variability, correlation with VQEG DMOS scores, and effects of SRC and HRC.
5. We have shown how the JND measurements may be used to calibrate automatic impairment metrics.
6. We have described a number of future applications of the JND data, including design, calibration and testing of metrics, as well as the creation samples of processed video with specified JND values.

ACKNOWLEDGEMENTS

The authors wish to thank the members of the IEEE Broadcast Technology Society, G-2.1.6 Compression And Processing Subcommittee, for their support and encouragement, especially John Libert, Alan Godber and Leon Stanger. We also wish to thank Amnon Silverstein for providing a paper[11] that was an early inspiration in our work.

REFERENCES

1. Watson AB, Pelli DG: **QUEST: a Bayesian adaptive psychometric method**. *Perception & Psychophysics* 1983, **33**:113-120.

2. Corriveau P, Webster A, Rohaly AM, Libert JM: **Video Quality Experts Group: The quest for valid objective methods**. *Proceedings of the SPIE* 2000, **3644**:129-139.

3. VideoQualityExpertsGroup: **Final report from the video quality experts group on the validation of objective models of video quality assessment**. In: *Final report from the video quality experts group on the validation of objective models of video quality assessment* (Editor Corriveau P, Webster A, Rohaly AM, Libert JM.). 2000.

4. ITU-R: **Recommendation BT.500-8: Methodology for the subjective assessment of the quality of television pictures**. In: *Book Recommendation BT.500-8: Methodology for the subjective assessment of the quality of television pictures* (.).: International Telecommunications Union; 1998.

5. Libert JM, Watson AB, Rohaly AM: **Toward developing a unit of measure and scale of digital video quality: IEEE Braodcast Technology Society Subcommittee on Video Compression Measurements**. *Proceedings of the SPIE* 2000, **3959**:160-167.

6. Thurstone LL: *The Measurement of Values*. Chicago: University of Chicago Press; 1959.

7. Engen T: **Psychophysics II. Scaling Methods**. In: *Woodworth and Schlosberg's Experimental Psychology* Edited by Kling JW, Riggs LA, vol. I, 3rd ed. pp. 47-86. New York: Holt, Rinehart and Winston; 1972: 47-86.

8. Falmagne JC: **Psychophysical measurement and theory**. In: *Handbook of Perception and Human Performance* Edited by Boff K, Kaufman L, Thomas J. pp. 1.3-1.66. New York: Wiley; 1986: 1.3-1.66.

9. Brunnstroem K, Eriksson R, Ahumada AJ: **Spatio-temporal discrimination model predicting IR target detection**. *Human Vision and Electronic Imaging IV* 2000, **3644**:403-410.

10. Watson AB, Hu J, McGowan JF, III: **Digital video quality metric based on human vision**. *Journal of Electronic Imaging* 2001, **10**:20-29.

11. Silverstein DA, Farrell JE: **Quantifying perceptual image quality**. In: *Image processing quality and capture*; 1998; Portland Oregon. 242-246.