Objective and Subjective Measures of MPEG Video Quality

Stephen Wolf,
Margaret H. Pinson, Arthur A. Webster

Institute for Telecommunication Sciences
Boulder, CO

Gregory W. Cermak,
E. Paterson Tweedy

GTE Laboratories Incorporated
Waltham, MA

Abstract

In 1996, the American National Standards Institute (ANSI) adopted ANSI T1.801.03, which presents a number of new objective video quality metrics for quantifying the effects of digital compression and transmission impairments. The measurements in ANSI T1.801.03 were selected based on an extensive multilaboratory quality assessment study that included video systems from bit rates of 64 kbit/sec to 45 Mbit/sec and video test scenes that spanned a wide range of spatial and temporal coding difficulties. The set of objective video quality measurements effectively accounted for subjective judgments by human viewers. While 25 video systems were tested, this multilaboratory study did not include MPEG video systems, and did not cover any bit rates between 1.6 and 10 Mbit/sec. This paper presents the results from two MPEG studies designed to fill in the bit-rate gap in the previous multilaboratory study. In these studies, we concentrated on bit rates from 1.5 - 8.3 Mbit/sec and examined the performance of MPEG 1 and MPEG 2 codecs (coder-decoders) specifically. The effectiveness of the ANSI standard objective video quality metrics was examined for these bit rates and coding technologies. Our analysis revealed that the objective video quality metrics primarily measure four principal components of video quality: added edges, lost edges, added motion, and lost motion; we found that parameters selected from these principal components can be used as effective predictors of subjective quality ratings for entertainment video systems.

1. Background

The development of *objective* measures of the quality of compressed digital video was motivated by a number of factors. For example, Wolf (pg.2) [1] writes: "New, objective measures of video transmission quality are needed by standards organizations, end users, and providers of advanced video services. Benefits would include impartial, reliable, repeatable, and cost effective measures of video and image transmission system performance and increased competition among providers as well as a better capability of procurers and standards organizations to specify and evaluate new systems." Ardito and Visca [2] write: "...we may expect that in the future there will be increasingly faster development of new coding systems, especially related to the matter of image compression, and it would be an unrealistic hypothesis to carry out subjective assessment for every different coder available on the market."

Several groups have been working on new objective video quality measures; e.g., Research Centre of RAI (Ardito & Visca) [2]; Sarnoff Laboratories (Lubin) [3]; National Telecommunications and Information Administration, Institute for Telecommunication Sciences (NTIA/ITS) and their associates (Wolf, Voran, Webster et al.) [1, 4, 5]; Bellcore (Cotton) [6], and KPN Research (Beerends).[7] In 1996, ANSI adopted ANSI T1.801.03 [8], which presents a number of new objective video quality metrics for quantifying the effects of digital compression and transmission impairments. (This standard has also been submitted to the ITU-T for consideration for international standardization.[9]) The measurements in ANSI T1.801.03 were selected after an extensive multilaboratory quality assessment study of video systems from bit rates of 64 kbit/sec to 45 Mbit/sec, and video test scenes that spanned a wide range of spatial and temporal coding difficulties. The set of objective video quality measurements effectively accounted for subjective judgments by human viewers (Cermak & Fay).[10] While 25 video systems were tested, this multilaboratory study did not include MPEG video systems, and did not cover any bit rates between 1.6 and 10 Mbit/sec.

The two MPEG studies described in this report were conducted to fill in the bit-rate gap in the previous multilaboratory study. In these studies, we concentrated on bit rates from 1.5 - 8.3 Mbit/sec and examined the performance of MPEG 1 and MPEG 2 codecs (coder-decoders) specifically. The effectiveness of the ANSI standard objective video quality metrics was examined for these bit rates and coding technologies. Several other metrics also were evaluated as part of this study. Two matrix versions of the ANSI spatial information (SI) distortion, named *Possob* (positive Sobel difference) and *Negsob* (negative Sobel difference), were used to compute input and output SI distortions on a pixel-by-pixel basis. The analysis revealed that the objective video quality metrics primarily measure four principal components of video quality: added edges, lost edges, added motion, and lost motion and that parameters selected from these principal components can be used as effective predictors of subjective quality ratings for entertainment video systems.

2. Overview of the Two MPEG Studies

The data and analyses reported here are from two previous data-collection efforts, one on MPEG 1 codecs [11] and one on MPEG 2 codecs.[12] In both studies, four analog Hypothetical Reference Circuits (HRC's) were compared with the MPEG HRC's. (The term Hypothetical Reference Circuit refers to a specific realization of a video transmission system. Such a video transmission system may include coders, digital transmission circuits, decoders, and analog processing of the video signal.) Objective video quality measures should be general enough to apply to both

digital codecs and analog systems. Also note that the MPEG 1+ (i.e., enhanced resolution MPEG 1) codec at 3.9 Mbit/sec was included in the second study for comparison purposes. By including a common set of HRC's in both studies, the subjective judgments from the two studies could be directly compared with one another.

The HRCs tested in Study 1 were:

1. MPEG 1 — Bit rate 1.5 Mbit/sec; Vertical resolution 240 lines
2. MPEG 1 — Bit rate 2.2 Mbit/sec; Vertical resolution 240 lines
3. MPEG 1+ — Bit rate 3.9 Mbit/sec; Vertical resolution 480 lines
4. MPEG 1+ — Bit rate 5.3 Mbit/sec; Resolution 330-400 pixels by 480 lines
5. MPEG 1+ — Bit rate 8.3 Mbit/sec; Resolution 330-400 pixels by 480 lines
6. Original scene with an SNR (signal-to-noise ratio) of 34 dB
7. Original scene with an SNR of 37 dB
8. Original scene with an SNR of 40 dB
9. Original scene recorded and played back from a VHS VCR
10. Original scene with no further processing

And, in Study 2, the HRCs were:

1. MPEG 2 — Bit rate 3.0 Mbit/sec; Resolution 352 pixels by 480 lines
2. MPEG 1+ — Bit rate 3.9 Mbit/sec; Resolution 352 pixels by 480 lines
3. MPEG 2 — Bit rate 3.9 Mbit/sec; Resolution 352 pixels by 480 lines
4. MPEG 2 — Bit rate 5.3 Mbit/sec; Resolution 704 pixels by 480 lines
5. MPEG 2 — Bit rate 8.3 Mbit/sec; Resolution 704 pixels by 480 lines
6. Original scene with an SNR of 34 dB
7. Original scene with an SNR of 37 dB
8. Original scene with an SNR of 40 dB
9. Original scene recorded and played back from a VHS VCR
10. Original scene with no further processing

The random noise for HRC's 6-8 in each study was added to the signals by attenuating a modulated version of the signals before inputting them to a demodulator. The SNR was measured with a video test instrument. To avoid introducing jitter when recording these signals, the noise on the synchronizing pulses was removed by regenerating the pulses in a processing amplifier. The VHS unit used for HRC 9 was a consumer model, rather than a laboratory model. Note that MPEG 1+ at 3.9 Mbit/sec and the comparison HRC's 6-10 were used in both studies.

The same set of scenes was used in both studies. The scenes spanned a range of coding difficulty, within the general domain of entertainment. They were not all chosen to stress the

codecs as much as possible. Each scene was 14 sec long. Four of the scenes were clips from movies and four of the scenes were clips from sporting events. Two of the movie clips were low motion scenes with subtle facial expressions. The rest of the clips contained considerable motion and spatial detail. The sources for the movie clips were commercial laser discs copied to ½ inch professional tape using a Y/C component connection. The sports event scenes were supplied by local broadcasters on ½ inch professional tape.

3. Objective Measures

3.1 Performance Measurement Issues for Digital Video Systems

A digital video transmission system that performs adequately for video teleconferencing might be inadequate for entertainment television. Specifying the performance of a digital video system as a function of the video scene coding difficulty yields a much more complete description of system performance. Recognizing the need to select appropriate input scenes for testing, algorithms have been developed for quantifying the expected coding difficulty of an input scene based on the amount of spatial detail and motion.[13] Other methods have been proposed for determining the picture-content failure characteristic for the system under consideration.[14] National and international standards have been developed that specify standardized video scenes for testing digital video systems.[15, 16] Use of these standards helps ensure that adequate care is taken when systems from different suppliers are evaluated.

3.2 Summary of the Objective Measurement Methodology

The objective performance measurement system used in this study digitizes the input and output video streams in accordance with ITU-R Recommendation BT.601 [17] (the objective parameters presented in this paper were applied to the luminance component only, which is sampled at 720 pixels by 486 lines) and extracts features from these digitized frames of video. *Features* are quantities of information that are associated with individual video frames. These features are used to quantify fundamental perceptual attributes of the video signal, such as spatial and temporal detail. Parameters are calculated using comparison functions that operate on two parallel sequences of these feature samples (one sequence from the output video frames and a corresponding sequence from the input video frames). The ANSI T1.801.03 standard contains parameters derived from three types of features that have been proven to be useful: (1) scalar features, where the information associated with a specified video frame is represented by a scalar; (2) vector features, where the information associated with a specified video frame is represented by a vector of related numbers; and (3) matrix features, where the information associated with a specified video frame is represented by a matrix of related numbers.

In general, the transmission and storage requirements for measuring an objective parameter based on scalar features are less than those required for an objective parameter based on vector features. These, in turn, are less than those required for an objective parameter based on matrix features. Significantly, scalar-based parameters have produced good correlations with subjective quality. This demonstrates that the amount of reference information that is required from the video input to perform meaningful quality measurements is much less than the entire video frame. This important new idea of compressing the reference information for performing video quality measurements has significant advantages, particularly for such applications as long-term maintenance and monitoring of network performance. Since a historical record of the output scalar features requires very little storage, these features may be archived efficiently for future

reference. Then, changes in the digital video system over time can be detected by simply comparing these past historical records with current output feature values.

The performance metrics in ANSI T1.801.03 can be used in-service or out-of-service for applications that detect the operational readiness of one-way, 525-line video systems that use digital transport facilities (e.g., maintenance, fault detection, and quality monitoring). The ultimate goal is to refine and extend this technology to produce objective methods that can replace subjective experiments for a wide range of applications.

3.3 Producing Frame-by-Frame Objective Parameter Values from Features

Frame-by-frame parameter values can be computed by applying mathematical comparison functions to each input and output feature value pair (the algorithms for temporally aligning output and input images are discussed below). Useful comparison functions include the log ratio (logarithm base 10 of the output feature value divided by the input feature value), and the error ratio (input feature value minus output feature value, all divided by the input feature value). These frame-by-frame objective parameter values give distortion measurements as a function of time.

Subjective tests conducted in accordance with ITU-T Recommendation P.910 [13] or CCIR Recommendation 500 [14] produce one subjective mean opinion score (MOS) for each HRC-scene combination. Since these video clips are normally about 10 sec in length, it is necessary to "time collapse" the frame-by-frame objective parameter values before they are correlated to subjective MOS. ANSI T1.801.03 specifies several useful time-collapsing functions such as maximum, minimum, and root mean square (rms). The maximum and minimum functions are useful for detecting the extremes of video quality while the rms function is a good indicator of the overall average.

3.4 Calculation of Gain, Level Offset, and Active Video Shift

Calibration is an important consideration when input and output video frames are compared directly. Neglecting calibration can produce large measurement errors in the parameter values. For example, both nonunity channel gains and nonzero level offsets can have a significant effect on the calculations of peak signal to noise ratio (PSNR) and other parameters defined in ANSI T1.801.03. Thus, robust methods for measuring gain, level offset, and active video shift (i.e., spatial registration of input and output video frames) are specified in the standard. These methods require the use of still video and, in the case of the gain and level offset calculations, that still video is a test pattern defined in the standard.

An alternative method for performing dynamic calibration measurements using the sampled input and output video had to be devised for the MPEG experiments because the standardized calibration frames were not included on the original source tapes.[18] This calibration analysis revealed that it is quite common for digital video systems to have substantial nonunity gains, level offsets, horizontal shifts, and vertical shifts of the output video. We also discovered that important calibration quantities can change dynamically depending upon the scene content. In light of this analysis, a separate gain g, level offset l, horizontal shift h_s, and vertical shift v_s were computed for each clip (i.e., each HRC-scene combination). We median-filtered the time histories of the calibration quantities for each clip and applied these filtered corrections to each output frame before computing the objective parameters. Note that within-scene variations from the calibration quantities are not removed by this approach. These within-scene variations could thus be detected as impairments by the objective parameters.

3.5 Temporal Alignment (i.e., Video Delay)

The output video frames should be temporally aligned, or registered, to the input video frames before the objective parameters are computed. Temporal misalignment of the input and output video streams results from accumulated video delays in the end-to-end transmission circuit (e.g., coder, digital transmission channel, and decoder). There are two fundamental methods that can be used to perform temporal alignment. The first method, called constant alignment [8], gives one time delay measurement for the entire output video stream. The second method, called variable alignment [19], gives a time delay measurement for each individual output video field. Objective parameters can be computed using either temporal alignment method. When constant alignment is used, frame by frame distortion metrics measure errors produced by both spatial impairments and repeated output frames. With variable alignment, frame-by-frame distortion metrics measure only those errors produced by spatial impairments; and the error caused by repeated output frames is quantified separately using variable frame delay statistics.

3.6 Temporal Alignment Observations

For higher quality transmission systems such as MPEG (i.e., systems that rarely drop frames), a field-accurate constant alignment method has proven to be a simple and excellent technique for measuring video delay.[18, 20] This technique has the added advantage of being an in-service method of measurement for video delay. For transmission systems that repeat frames, drop frames, or perform temporal warping (i.e., variable video delay), constant alignment produces a temporal alignment that reflects the average alignment of the ensemble of output video frames being examined. For the current studies, the constant alignment technique was used prior to computing the objective parameters.

3.7 Summary of Objective Parameters Used for the MPEG 1+ and MPEG 2 Tests

Table 1 presents a summary of the objective parameters that were computed for each HRC-scene combination in the MPEG 1+ and MPEG 2 studies. Parameter definitions and detailed methods of measurement are based on ANSI T1.801.03.[8, 9] Annex B of ANSI T1.801.03 is particularly informative as it provides pictorial representations of parameter responses to various spatial and temporal distortions (e.g., tiling, blurring, error blocks, and jerkiness, see ANSI T1.801.02-1996 [21] for definitions of these impairments).

Table 1 Summary of Objective Parameters

Parameter	Description	ANSI Method of Measurement
711	maximum added motion energy	Section 7.1.1 of ANSI T1.801.03-1996
712	maximum lost motion energy	Section 7.1.2
713	average motion energy difference	Section 7.1.3
714	average lost motion energy with noise removed	Section 7.1.4
715	percent repeated frames	Section 7.1.5

716	maximum added edge energy	Section 7.1.6
717	maximum lost edge energy	Section 7.1.7
718	average edge energy difference	Section 7.1.8
719	maximum HV to non-HV edge energy difference, threshold=20	Section 7.1.9
719_60	maximum HV to non-HV edge energy difference, threshold=60	Section 7.1.9 using an r_{min} of 60 instead of 20
719a	minimum HV to non-HV edge energy difference, threshold=20	Section 7.1.9 using the feature comparison function in section 6.5.1.5
719a_60	minimum HV to non-HV edge energy difference, threshold=60	Section 7.1.9 using an r_{min} of 60 instead of 20 and the feature comparison function in section 6.5.1.5
7110	added edge energy frequencies	Section 7.1.10
7110a	lost edge energy frequencies	Section 7.1.10 using modified feature comparison function to sum the lost frequencies (i.e. sum positive part instead of negative part)
721	maximum added spatial frequencies	Section 7.2.1
722	maximum lost spatial frequencies	Section 7.2.2
732	minimum peak signal to noise ratio	Section 7.3.2
733	average peak signal to noise ratio	Section 7.3.3
Negsob	negative Sobel difference	Mean of the negative part of the input minus output pixel by pixel differences of SI_r values (see section 6.1.1.1), mean $[\text{Sobel(input)}-\text{Sobel(output)}]_{np}$ ($[X]_{np}$ defined in section 6.5.1.9)
Possob	positive Sobel difference	Mean of the positive part of the input minus output pixel by pixel differences of SI_r values (see section 6.1.1.1), mean $[\text{Sobel(input)}-\text{Sobel(output)}]_{pp}$ ($[X]_{pp}$ defined in section 6.5.1.7)

The horizontal and vertical (HV) to non-HV edge energy difference parameters were computed using an r_{min} threshold of 60 in addition to the recommended r_{min} threshold of 20. An r_{min} threshold of 20 included nearly every pixel in the sampled video frames due to the amount of noise that was present in the source video. With an r_{min} threshold of 60, the noise was eliminated effectively from the calculation. To remove the effect of scene length, the added edge energy

frequencies and lost edge energy frequencies parameters were computed using a mean calculation rather than the sum calculation specified by ANSI T1.801.03.

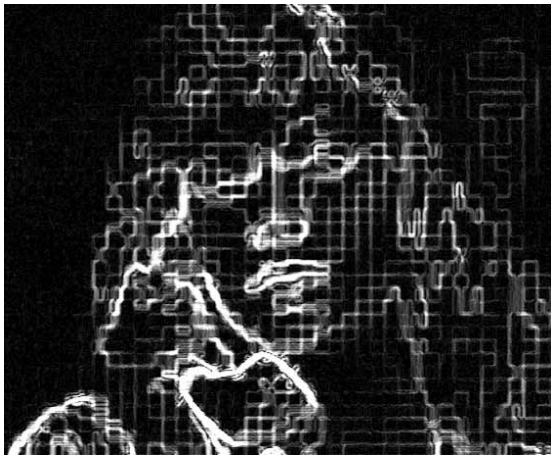
Two matrix versions of the ANSI spatial information (*SI*) parameters were included in the analysis. These two parameters (Negsob and Possob) are illustrated in Figure 1. Figure 1 (a) is the input image, Figure 1 (b) is the spatially registered output image, Figure 1 (c) is the spatial information of the input image (SI_r [input]), Figure 1 (d) is the spatial information of the output image (SI_r [output]), and Figure 1 (e) is the error between the two spatial information images (i.e., SI_r [error] = SI_r [input] - SI_r [output]). In Figure 1 (e), zero error has been scaled to be equal to mid-level gray (128 out of 255 for an 8-bit display). When false edges are present in the output image (e.g., blocks, edge busyness, etc.), the *SI* error is negative and appears darker than gray (Negsob parameter). When edges are lost in the output image (e.g., blurred), the *SI* error is positive and appears lighter than gray (Possob parameter). In this manner, the two types of error can be clearly separated on a pixel-by-pixel basis when both are present in the output image.

The ability to separate impairments on a pixel-by-pixel basis is one advantage of the *SI* matrix equivalents over the *SI* scalar features presented in ANSI T1.801.03. Since *SI* scalar features use summary statistics from the input and output *SI* images, impairments can be missed when two impairments with opposite responses are present (for instance, lost edges and added edges). However, it is possible to design scalar features that can separate certain impairments that have opposite responses (e.g., blocking can be separated from blurring by looking at the direction of the spatial gradient; see Annex B, section B.3 of ANSI T1.801.03). The primary disadvantages of using matrix features are that they require a tremendous amount of extra storage (or transmission bandwidth), and precise spatial registration of the input and output images is required.

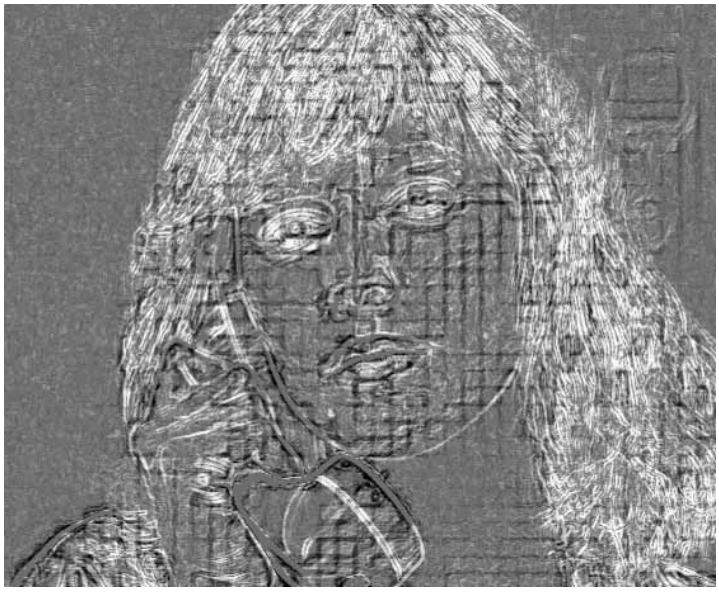
(a) Input

(b) Output

(c) SI of Input



(d) SI of Output



(e) SI of Input minus SI of Output

Figure 1 Illustration of Negsob and Possob Parameters

4. Subjective Data

4.1 Method Used to Collect Subjective Data

A variant of one of the methods specified in CCIR 500 [14] was used to collect the subjective data. Human observers watched recorded video segments on a single high-quality monitor in a room with controlled illumination. The video segments were presented in pairs, so that each judgment was a comparison of two video treatments. The observers made subjective judgments and recorded them on answer sheets.

The method for collecting subjective judgments of video quality also differed from the CCIR 500 method used in the 1994 multilaboratory study (see Cermak et al.[11] for rationale and details) in three ways:

- HRC's were compared to each other, not to the original, unprocessed clip. For a given number of "trials" (exposures to stimuli), this method provides a larger number of exposures to the HRC's being tested.
- The judgment that observers made was different from the impairment scale methods. Rather than rating on a five-point impairment scale, observers (a) chose the better HRC in each pair, then (b) estimated the difference between the value of the two HRC's in dollars per month. This method correlates highly with the impairment scale method; it also provides technical advantages over the impairment scale method.
- The video clips were recorded and played back on a video disc, rather than on a ½ inch professional tape recorder. The performance specifications for the video disc machine are marginally lower than for the tape machine (>45 dB video SNR, 450 pixels horizontal resolution). The video disc has the advantages of random access and computer control. The ordering of stimuli was randomized separately for each subject in real time. Also, the pairings of HRCs and scenes were randomized; over the course of the full experiment, each HRC was paired with each scene an approximately equal number of times, but on any specific trial the scene was selected randomly. This sampling procedure is based on the logic that the HRC's we tested are known, fixed, and limited in number, while the scenes are sampled from a potentially infinite pool.

In the MPEG 1+ study 30 observers provided subjective responses in dollars per month. The observers were not laboratory employees but were chosen to be cable television customers, familiar with the signal quality of cable television, and also accustomed to paying for it. Their demographics were unremarkable. The MPEG 2 study also used a sample of 30 consumers with the same overall description as the MPEG 1+ study. Some of the same subjects participated in both studies, but the studies were separated by nearly a year, more than enough time for subjects to forget fine details of visual stimuli.

4.2 Summary of Subjective Data

The basic subjective data for this study are the mean dollar ratings for each HRC-scene combination, averaged across 30 observers. Each rating represents the average difference between a given HRC and the other HRC's with which it was compared. Thus, negative values are possible. Table 2 and Table 3 show the mean ratings (in dollars per month) for the MPEG 1+ and MPEG 2 studies, respectively. The standard errors of the table values for MPEG 1+ and MPEG 2 on the order of 0.7 and 1.0, respectively (there being half as many trials per subject in the MPEG 2 study than in the MPEG 1+ study).

Table 2 Mean subjective ratings of HRC-scene combinations for MPEG 1+ study

Scene	1.5 Mbit/ sec	2.2 Mbit/ sec	3.9 Mbit/ sec	5.3 Mbit/ sec	8.3 Mbit/ sec	34 dB	37 dB	40 dB	VHS	Original
2001	0.86	-0.57	2.79	1.33	2.53	-7.92	-3.93	-2.12	2.35	3.85
Graduate	-4.37	-6.06	0.84	0.22	1.97	-7.88	-4.98	-1.39	-0.11	3.09
Godfather	0.46	-0.19	0.80	1.70	2.18	-8.44	-2.22	-3.34	1.79	4.04
Being There	1.23	0.68	2.29	2.36	2.97	-9.14	-4.76	-0.65	1.81	2.91
Basketball	-4.26	-1.04	0.31	2.46	3.50	-6.84	-1.88	0.47	2.71	3.17
Baseball	-2.37	-0.41	3.56	2.30	2.00	-8.05	-5.57	-3.15	5.21	4.38
Hockey 1	-5.65	-5.53	-0.29	0.89	2.52	-3.94	1.97	2.39	3.79	4.16
Hockey 2	-4.61	-3.92	2.39	2.11	0.58	-5.12	-0.36	2.75	2.74	3.94

Table 3 Mean subjective ratings of HRC-scene combinations for MPEG 2 study

Scene	3.0 Mbit/ sec	3.9 Mbit/ sec (1+)	3.9 Mbit/ sec	5.3 Mbit/ sec	8.3 Mbit/ sec	34 dB	37 dB	40 dB	VHS	Original
2001	3.40	1.17	2.57	3.29	2.56	-10.47	-6.29	0.24	2.00	2.90
Graduate	-0.13	1.68	1.11	1.94	1.16	-10.09	-4.78	-2.65	0.23	3.38
Godfather	0.20	-0.72	2.80	3.17	1.13	-9.45	-6.75	-4.50	3.54	3.26
Being There	2.00	1.64	3.70	1.89	3.95	-9.50	-5.43	-2.13	1.30	2.35
Basketball	0.15	-0.68	0.22	1.36	3.42	-6.33	-2.73	-0.60	5.40	3.60
Baseball	-1.00	3.35	1.44	2.50	4.20	-7.29	-6.69	-1.37	4.20	4.22
Hockey 1	2.38	-0.13	0.23	1.69	3.85	-6.06	-4.06	-0.10	1.36	2.38
Hockey 2	-0.24	-3.60	3.69	0.86	3.17	-8.89	-1.91	-0.26	1.25	4.15

Other papers have presented analyses of these subjective data in some detail.[11, 12] In both data sets the ratings were statistically related to the variables: HRC, scene, and the specific HRC-scene combinations, as would be expected. Other analyses demonstrate that the subjective data are not excessively noisy and show systematic differences between the way observers react to analog vs. digital HRC's. We do not present further analyses of the subjective data by themselves here. Instead, we concentrate on analyses of the objective data as predictors of the subjective data.

5. Statistical Analyses

5.1 Methods

5.1.1 Strategy

The theoretical goals of the data analysis were to

- find the “best” set of objective measures for predicting the subjective judgments, and
- determine how close to optimal these predictors are.

Two characteristics of most data sets complicate the problem of finding the "best" set of predictors and force one to use compensating data analysis strategies: (a) noise and (b) redundancy. Two consequences of noise are (a) that a different set of predictors will best fit in different, but comparable, data sets; and (b) the best fit will never be 1.0. Two consequences of redundancy in a set of variables are (a) different subsets of variables will fit a data set (essentially) equally well; and (b) if too many redundant variables are used as predictors, results can be very unstable from one analysis to the next, especially in the presence of noise (a phenomenon known as “over-fitting”).

Because of the realities of the characteristics of data,

- the actual goals of the analysis were to find a generalizable and meaningful set of predictor measures;
- several sets of predictors may be essentially equally good; and
- the fit of these good sets of predictors will be less than 1.0.

Strategies for handling data with noise and redundancy were:

- measure the redundancy in the set of predictor variables;
- pre-specify the maximum number of variables to be used in any analysis on the basis of the measure of redundancy;
- use variables that are known *a priori* to be causally related to the dependent variable whenever possible; and
- verify that a candidate set of predictor variables can be generalized to another data set or sample.

5.1.2 Redundancy

The set of 20 objective measures is based on a few fundamental quantities such as spatial and temporal differences in pixel brightness. The measures fall into families of closely-related measures (see above). A statistical measure of the amount of redundancy in the set of 20 measures is the number of orthogonal (i.e., uncorrelated) variables needed to account for most of the variance in the set of measures. The analysis that computes this measure is “principal components analysis.” In this analysis the original data set of measurements of the signals is represented as a linear combination of the original measurements. The particular linear combination is given by the eigenvectors of the correlation matrix of the measures. The corresponding eigenvalues represent the amount of variation in the original correlation matrix that is attributable to particular linear combinations of the original variables. The original data matrix can always be reproduced if the number of principal components is equal to the number of original variables. However, the data matrix can often be closely approximated with fewer

principal components than there are variables, especially when the original variables are correlated. Generally, one considers the number of principal components for a data set to be the number whose eigenvalues are greater than 1.0. In practice, an analysis is considered successful if it accounts for about 70 or 80% of the variance in a set of measures with a number of components equal to about a third or a fourth of the number of original variables.

5.1.3 Reliability

The reliability issue is important because it limits the statistical fit of even a perfect objective measure.[22, 23] That is, if the subjective judgments have noise in them (as we know they certainly will), then even perfect objective measures will not be able to predict the subjective judgments perfectly. The reliability of a variable is defined as the ratio {the variance in the variable if it were measured perfectly} / {the variance in the variable if it were measured perfectly, plus error}. This definition is theoretical because one never observes "the variance in the variable if it were measured perfectly." However, the ratio still can be estimated using observable quantities, as follows.[23]

- The denominator is just the variance in the variable as actually observed. This variance is, by hypothesis, composed of both the true value and error. The estimator for the denominator is the mean square (variance) pooled across the two subsamples, i.e., the MPEG 1+ and MPEG 2 studies.
- The numerator is estimated by the covariance of the observed variable across the two studies. This simple estimator is based on the assumption that the error in the two studies is independent and uncorrelated with the variable itself. In this case, the covariance of the observed variable with itself is the same as the variance of the variable if it were measured perfectly.

The term "reliability" is somewhat misleading when applied to objective measures of video quality. If a measure receives a low reliability score, one might think of the measure as defective, while in fact the measure may be responding accurately to real differences in the video streams between the two studies. Despite this incorrect connotation, the term "reliability" is the one that the statistics literature recognizes. We analyzed repeated measurements to compute estimates of the statistical reliability of the objective measures and of the subjective measure. Five of the HRC's and all eight of the scenes were nominally the same across the two experiments. The repeated HRC's were MPEG 1+ at 3.9 Mbit/sec, the cable simulations at 34, 37, and 40 dB SNR; and VHS. We say "nominally the same" because the two tapes of the HRC's and scenes were not identical frame-by-frame and pixel-by-pixel. In this sense, when we speak of a measurement in the present study we refer to the end-to-end process of obtaining the video signal and preparing it for measurement, as well as the digitization and parameter computation.

5.1.4 Regression

We used a standard regression program for most of the analyses in which the objective measures were used to predict the subjective judgments. We also used a "stepwise" regression as a secondary analysis. Stepwise regression is an exploratory data analysis technique that seeks a best-fitting set of predictor variables via an automated algorithm. Stepwise regression is an exploratory technique in the sense that it can suggest hypotheses on the basis of one data set for testing in another data set. (The "best" set of variables found through stepwise regression is rarely the set that is most generalizable.) Prior to the regression analyses, the MPEG 2 rating data were re-scaled according to the formula: New Rating = 0.721 + 0.833 * Rating, so that the

ratings from the two experiments would be comparable.

5.2 Results

5.2.1 Redundancy in Objective Measures

Redundancy was separately analyzed for the MPEG 1+ and MPEG 2 data sets, as well as for the two data sets combined. The separate analyses agreed qualitatively, so only the analysis of the two sets combined is reported here. A principal components analysis showed four principal components with eigenvalues greater than 1.0, and these principal components jointly accounted for 80% of the variance in the 20 objective measures. The following describes the components:

1. The first component, as in the two data sets separately, correlated highest with measures from the 719 series, 721, and Negsob (added edges and added spatial frequencies). The first component accounted for 34% of the variance.
2. The second component, again similar to the second component for the two data sets separately, accounted for 26% of the variance and correlated most highly with measures 717, 722, and Possob (lost edges and lost spatial frequencies).
3. The third component accounted for 12% of the variance and correlated highest with 7110a (lost edge energy frequencies) and 714, 715 (lost motion).
4. The fourth component, accounting for 7% of the variance, correlated highest with measure 7110 (added edge energy frequencies; 7110 and 7110a were slightly negatively correlated with each other).

5.2.2 Regression

Any one regression analysis, on any one data set is unlikely to produce a generalizable result. However, multiple analyses on multiple data sets that produce similar answers form the basis for credible and potentially generalizable results. The following regression analyses were performed:

1. MPEG 1+ data alone, using measures from principal components analysis. We used only a single variable from each of the four principal components that passed the eigenvalue test as predictors in the regression analysis.
2. MPEG 1+ data alone, using Sobel image measures. The first two principal components of both data sets correlate nearly maximally with the two complementary Sobel image measures. Because these measure are of *a priori* interest, we performed a regression analyses using the Sobel measures as representatives of the first two components.
3. MPEG 1+ data alone, exploratory stepwise analysis. Stepwise regression enters variables sequentially, choosing the next variable that maximizes the square of the correlation coefficient R^2 given the preceding variables. Typically, results of a stepwise analysis are sensitive to noise in the data, and thus may not be reliable when used in isolation. However, when used in combination with other analyses, stepwise analysis can be informative.

From the analyses of the MPEG 1+ data set, we hypothesized (a) that a variable from each of the first three principal components of the objective data set is worth trying; (b) the most likely variable from the first principal component is Negsob; (c) an R^2 above 0.7 is achievable.

4. MPEG 2 data alone, using measures from principal components analysis.
5. MPEG 2 alone using the best MPEG 1+ measures. A set of candidate "best" predictors from the MPEG 1+ analysis was Negsob, 713 (motion difference), and 717 (lost edges). The adjusted R^2 fit of this model to the MPEG 2 data was 0.815; this was quite an improvement over the variables derived from the principal components, and also an improvement over the T1A1.5 multilaboratory data set.
6. MPEG 2 alone, exploratory stepwise analysis.

The three hypotheses from the MPEG 1+ data set were supported in the MPEG 2 data set. Thus, we considered these hypotheses in the analysis of the joint data set.

7. MPEG 1+ and 2 using measures from principal components. The measures that best correlated with the first four principal components, respectively, of the combined data set were Negsob (added edges), 722 (lost spatial frequencies), 714 (lost motion), and 7110 (added edge energy frequencies). The adjusted R^2 for this set of predictors was 0.704. The variables 7110 and 722 were not significantly correlated, as was the case in the analysis of the MPEG 1+ data set. Again, Negsob had by far the largest effect.
8. MPEG 1+ and 2 using variables from MPEG 2 analyses. A slightly different set of variables had been identified in the analysis of the MPEG 2 data namely, Negsob and 714, as above, as well as Possob and 711. The adjusted R^2 for this set of variables was a more respectable 0.769, and all variables were significant (Possob marginally).
9. MPEG 1+ and 2 using exploratory stepwise analysis. The first three variables selected, and the only three that appreciably improved the fit of the model, were Negsob, 711, and 714, respectively. These three parameters measure added edges, added motion, and lost motion, respectively. The adjusted R^2 fit of the three-variable model was 0.763 (correlation coefficient R = 0.87). Figure 2 shows the predicted ratings from this model plotted against the actual subjective ratings. The (standardized) parameters of this model are 0.555, -0.347, and -0.220 for the variables Negsob, 711, and 714, respectively. The unstandardized parameters of the model are 0.224, -8.662, and -7.547, respectively, with an intercept constant of 4.327. (Unstandardized parameters for a four-parameter model that included a parameter for measuring lost edges; i.e., the variables Negsob, 711, 714, and 717 are 0.209, -9.374, -6.127, and -3.241, respectively, with an intercept of 4.380.)
10. Peak signal to noise ratio (PSNR) has been used as a measure of video quality for years. We report its ability to predict subjective judgments in the present joint data set: $R^2 = 0.181$ for average PSNR (parameter 733); and $R^2 = 0.095$ for minimum peak SNR (parameter 732). By contrast, the R^2 for Negsob for the joint data set was 0.657.

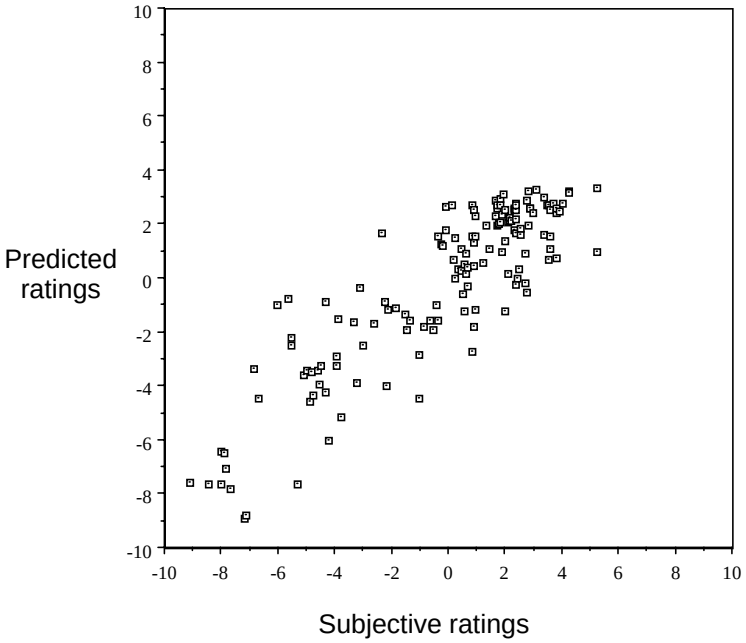


Figure 2 Predicted ratings vs. subjective ratings for the three-parameter model

5.3 Interpretation of Results

5.3.1 Which Measures Work Best

For the current data sets, the best single predictor of subjective video quality is Negsob (the mean of the negative portion of the differences in pairs of Sobel images). Recall that Negsob becomes large in absolute value when the coded video has false edges added to it, as in blocking. This variable is both consistent across data sets and powerful in its ability to predict.

After Negsob, the ability to predict increases with the addition of another two or three variables. Exactly which ones are chosen is not terribly crucial as long as representatives from the following families of measures are included:

- Possob or the family of measures of lost edge information (e.g., 717 for lost edge energy, or 722 for lost spatial frequencies).
- 714 or the family of measures of lost motion (e.g., 715 for repeated frames).
- 711 or measures of added motion (e.g., 713 for average motion difference), or measures of edge energy difference (e.g., the 719 family).

The inclusion of matrix versions of spatial information (*SI*) distortion (i.e., Negsob, Possob) increased the amount of subjective variance that was explained by the objective metrics by about 5 - 8%. Thus, for the current studies, the price paid for compressing the *SI* information into a set of scalar quality features appears to be about a 5 - 8% reduction in prediction efficiency.

The particular package of measures that best predicts subjective judgments may depend somewhat on the particular domain of HRC's and scenes for which one wants to make predictions. For example, if one is interested in comparing only MPEG HRC's running at different bit rates, then one package of measures could be slightly better, while if one were comparing MPEG systems to VHS video recorders and cable television, then another package might predict slightly better. If one were interested in determining acceptable bit rates for one

kind of content (e.g., sports), then one package of measures might be slightly better, but if one were interested in another kind of content (e.g., news and weather) then another package of measures might be slightly better.

5.3.2 How Good Is the Statistical Fit?

In the combined data set, the objective measures were able to account for 0.763 - 0.769 of the variance, depending on whether three or four predictor variables were used. By way of comparison, in the T1A1.5 multilaboratory study (Cermak and Fay [10], pg. 28), the fit was not quite as good: R^2=0.706.

Another relevant comparison is with the maximum R^2 that could have been achieved, given the level of error in the data. More than a quarter century ago the statistician Cochran [23] (pg. 22) dealt with the problem of estimating R^2 in the presence of error: "This paper deals mainly with the relation between R^2, the squared multiple correlation coefficient between *y* and the *X*'s when these are correctly measured, and R'^2, the corresponding value when errors of measurement are present." We use Cochran's equation 3.6 (pg. 24):

R'^2 = R^2 * (reliability of *y*, subjective data) * (weighted average of reliabilities of *X*'s, objective data).

Suppose R^2 were 1.00 in the case of no error of measurement, then R'^2 = 1.00 * 0.890 * 0.949 = 0.845, where 0.949 is a weighted sum of the reliabilities of the best predictors, 711, 714, Negsob. (The weights are the absolute values of the standardized regression coefficients for 711, 714, and Negsob, scaled to sum to 1.00.) See Cermak et al.[18] for the reliabilities of the objective measures.

R'^2 = 0.845 is the upper bound for prediction of subjective ratings by objective measures when error of measurement is present in the amounts we observed in this study. Compared to 0.845, the observed 0.763 is 90% of maximum. As in the case of the T1A1.5 study, the ability of the objective measures to predict subjective responses is good but shows some room for improvement.

6. Conclusions

The current generation of objective video quality measures has achieved good prediction of subjective ratings for entertainment-level HRCs. The objective measures captured about 90% of the subjective information that could be captured considering the level of measurement error present in the subjective and objective data. We have not attempted to tune this set of objective measures for a specific testing situation. Further work is required to evaluate the potential of fine-tuning the measures for specific applications in testing equipment. However, the current objective measures can be considered as reasonable candidates for testing applications.

The kinds of objective variables that effectively predict subjective responses well for MPEG video systems are

(a) measures of the addition of false edges, in particular the matrix measure Negsob,

(b) measures of lost sharpness of edges, and

(c) measures of change in motion (i.e., lost motion, and added motion).

A traditional objective variable that does not effectively predict subjective responses for MPEG video systems is PSNR. PSNR captured only about 21% of the subjective information that could be captured considering the level of measurement error present in the subjective and objective data.

In conclusion, by using a set of three or four objective measures as indicated above, a correlation coefficient of 0.87 is achieved. Since the covariance/variance analysis indicates that 0.92 is the best possible for this study, this result is very good.

7. References

Note: Copies of ANSI contributions referenced below can be obtained from the T1 Secretariat, Alliance for Telecommunications Industry Solutions, 1200 G Street, NW, Suite 500, Washington, DC 20005.

---

1. Wolf, S., “Features for automated quality assessment of digitally transmitted video,” NTIA Report 90-264, June 1990.
2. Ardito, M., and Visca, M., "Correlation between objective and subjective measurements for video compressed systems," SMPTE Journal, V 105, n 12, pg. 768-773, December 1996.
3. Lubin, J., "A visual discrimination model for imaging system design and evaluation," Report from the David Sarnoff Research Center, February 1995.
4. Webster, J., Jones, C., Pinson, M., Voran, S., Wolf, S., “An objective video quality assessment system based on human perception,” SPIE Human Vision, Visual Processing, and Digital Display IV, vol. 1913, February 1993.
5. Voran, S., “The development of objective video quality measures that emulate human perception," IEEE Global Telecommunications Conference (GLOBECOM), December 1991.
6. Cotton, B., “An objective model for video quality performance,” ANSI T1A1 contribution number T1A1.5/96-105, March 1996.
7. ITU-T Contribution to Question 22/12, COM 12-7 (Netherlands), “Objective measurement of video quality,” February 1997.
8. ANSI T1.801.03-1996, “American National Standard for Telecommunications - Digital Transport of One-Way Video Telephony Signals - Parameters for Objective Performance Assessment,” Alliance for Telecommunications Industry Solutions, 1200 G Street, NW, Suite 500, Washington DC 20005.
9. ITU-T Contribution to Question 22/12, COM 12-66-E (USA), “Selections from the Draft American National Standard: Digital Transport of One-Way Video Signals - Parameters for Objective Performance Assessment,” January 1996.
10. Cermak, G. W., and Fay, D. A., “T1A1.5 video quality project: GTE Labs analysis,” ANSI T1A1 contribution number T1A1.5/94-148, September 1994.
11. Cermak, G. W., Tweedy, E.P., Ottens, D. W., and Teare, S.K., “Consumer acceptance of MPEG1 video at 1.5 to 8.3 Mb/s.” ANSI T1A1 contribution number T1A1.5/96-108, May 1996.
12. Cermak, G. W., Teare, S. K., Tweedy, E. P., and Stoddard, J.C., “Consumer acceptance of MPEG2 video at 3.0 to 8.3 Mb/s,” Broadband Access System, W.S. Lai, S.T. Jewell, C.A.

Siller, I. Widjaja, & D. Karvelas (eds.) Proc. SPIE 2917, pg. 53-62, 1996.

13. ITU-T Recommendation P.910, “Subjective video quality assessment methods for multimedia applications,” Recommendations of the ITU, Telecommunication Standardization Sector.

14. CCIR Recommendation 500-5, “Method for the subjective assessment of the quality of television pictures,” Recommendations and Reports of the CCIR, 1992.

15. ITU-R Recommendation BT.802-1, “Test pictures and sequences for subjective assessments of digital codecs conveying signals produced according to Recommendation ITU-R BT.601,” Recommendations of the ITU, Radiocommunication Sector.

16. ANSI T1.801.01-1995, “American National Standard for Telecommunications - Digital Transport of Video Teleconferencing/Video Telephony Signals - Video Test Scenes for Subjective and Objective Performance Assessment,” Alliance for Telecommunications Industry Solutions, 1200 G Street, NW, Suite 500, Washington DC 20005.

17. ITU-R Recommendation BT.601-4, “Encoding Parameters of Digital Television for Studios,” Recommendations of the ITU, Radiocommunication Sector.

18. Cermak, G., Tweedy, P., Wolf, S., Webster, A., Pinson, M., “Objective and subjective measures of MPEG video quality,” ANSI T1A1 contribution number T1A1.5/96-121, October 1996.

19. ANSI T1.801.04-1997, “American National Standard for Telecommunications - Multimedia Communications Delay, Synchronization, and Frame Rate Measurement,” Alliance for Telecommunications Industry Solutions, 1200 G Street, NW, Suite 500, Washington DC 20005.

20. Wolf, S., Webster, A., “Objective and subjective video performance testing of DS3 rate transmission channels,” ANSI T1A1 contribution number T1A1.5/93-60, April 1993.

21. ANSI T1.801.02-1996, “American National Standard for Telecommunications - Digital Transport of Video Teleconferencing/Video Telephony Signals - Performance Terms, Definitions, and Examples,” Alliance for Telecommunications Industry Solutions, 1200 G Street, NW, Suite 500, Washington DC 20005.

22. Bollen, K.A., Structural Equations With Latent Variables, New York, Wiley, 1989.

23. Cochran, W.G., “Some effects of errors of measurement on multiple correlation,” Journal of the American Statistical Association, No. 65, pg. 22-34, 1970.