Committee T1 Performance
Standards Contribution

Document Number:	T1A1/ 97-010
TIBBS File:	
DATE:	Jan 24, 1997
STANDARDS PROJECT:	Analog Interface Performance Specifications for Digital Video Teleconferencing/Video Telephony Service (T1Q1-12)
SUBJECT:	Proposed US contribution to ITU-T SG12 Objective and subjective measures of MPEG video quality: Summary of experimental results.
SOURCE:	T1A1.5
CONTACT:	GTE Laboratories: Greg Cermak (phone: 617-466-4132, email: gwc0@gte.com), Pat Tweedy NTIA/ITS: Arthur Webster (303-497-3567), Steve Wolf
KEY WORDS:	video quality, MPEG, subjective, objective, correlation, performance assessment, multimedia.
DISTRIBUTION:	T1A1

ITU - Telecommunication Standardization Sector
STUDY GROUP 12

Delayed Document No.
Original: English

Geneva 7-18 April, 1997

Questions: 10/12, 11/12

Source: USA[1] (Proposed)

Title: OBJECTIVE AND SUBJECTIVE MEASURES OF MPEG VIDEO QUALITY: SUMMARY OF EXPERIMENTAL RESULTS

Abstract

The United States is interested in the development of useful and standard methods for the performance assessment of audiovisual services and systems. This contribution is an informational progress report on further developments and experimental results based on the techniques described in COM 12-66-E. The material contained in this contribution is intended to strengthen the value of the basic techniques, however the limitations listed in COM 12-66-E are still applicable. Further work in this area is planned including development and comparative analysis of other measurement methods such as algorithms based on a model of the human visual system.

The results of a video performance assessment test are summarized. MPEG video at several bit rates and coding methods was rated both subjectively and objectively.

The recently approved ANSI standard objective metrics for video quality (and others) were used in the test. The statistical analysis of the data is summarized. 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 The metrics are thus shown to work well for MPEG encoded video.

[1] Contact person: Arthur Webster
tel: +1 303 497 3567
fax: +1 303 497 5323
Email: webster@its.bldrdoc.gov

OBJECTIVE AND SUBJECTIVE MEASURES OF MPEG VIDEO QUALITY: SUMMARY OF EXPERIMENTAL RESULTS

Introduction

In this contribution the results of a video performance assessment test are summarized. MPEG video at several bit rates and coding methods was rated both subjectively and objectively. The subjective viewing test utilized a novel method for rating the video. Viewers were asked to rate (in dollars) the amount they would be willing to pay per month to receive the better of each pair of conditions.

The objective metrics for video quality defined in ITU-T COM 12-66-E (and others) were used in this test. Statistical analysis of the data is summarized. 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.

The conclusion is reached that by using a set of 3 or 4 objective measures as indicated a correlation coefficient of .87 is achieved. Since the covariance/variance analysis indicates that .92 is the best possible for this study, the result is significant. The metrics are thus shown to work well for MPEG encoded video.

The summary of the study is offered for information and to support the proposal that these measures (or a subset of them) be considered for inclusion in the draft new Recommendation on objective assessment of video quality.

Proposal

It is proposed that the twenty objective measures (or a subset of the twenty) presented in this contribution be considered for inclusion in the draft new Recommendation on objective measures of video quality.

It is proposed that the subjective method (with modifications for different currencies) utilized in this study be considered for inclusion in ITU-T P.910 when the Recommendation is revised.

Background

In ITU-T Contribution COM 12-66-E [15] a number of objective video quality measurements is presented. These measurements were selected after a multi-lab quality assessment study that included video systems from bit rates of about 100 kb/s to 45 Mb/s. This set of objective quality measurements performed well in accounting for subjective judgements by human viewers. While 25 video systems were included, this study did not cover MPEG video systems.

The two MPEG studies presented in this contribution were conducted to fill in the bit-rate gap in the previous multi-lab study. In particular, the current studies concentrate on bit rates from 1.5 to 8.3 Mb/s and they examine MPEG 1 and MPEG 2 codecs specifically. The effectiveness of the ANSI standard objective video quality metrics (defined in COM 12-66-E [15]) are examined for these bit rates and coding technologies. In addition, two other measures are included in this study. These metrics are two matrix versions of spatial information (SI) distortion named *Possob* (positive Sobel difference) and *Negsob* (negative Sobel difference). (See section 6.1.1.1 of COM 12-66-E [15] or P.910[4] for a definition of spatial information). Spatial and temporal registration (or alignment) of the input and output images is critical for successful implementation of matrix measures. For more information on image calibration and alignment, see [14].

1. Overview of the Two MPEG Studies

1.1 HRCs[2] (Hypothetical Reference Circuits) and Scenes

The data and analyses reported here come from two previous data-collection efforts, one on MPEG1 codecs (i.e., coder-decoders) and one on MPEG2 codecs [1, 2].

The HRCs tested in Study 1 were,:

1. MPEG 1 Bit rate 1.5 Mb/s Vertical Resolution 240 lines
2. MPEG 1 Bit rate 2.2 Mb/s Vertical resolution 240 lines
3. MPEG 1+ Bit rate 3.9 Mb/s Vertical resolution 480 lines
4. MPEG 1+ Bit rate 5.3 Mb/s Resolution 330-400 pixels X 480 lines
5. MPEG 1+ Bit rate 8.3 Mb/s Resolution 330-400 pixels X 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 Mb/s Resolution 352 (codec setup) X 480 lines

[2] The term Hypothetical Reference Circuit (HRC) refers to a specific realization of a video transmission system. Such a video transmission system may include coders, digital transmission circuits, decoders, and even analog processing (e.g., VHS) of the video signal.

2. MPEG 1+ Bit rate 3.9 Mb/s Resolution 352 (codec setup) X 480 lines
3. MPEG 2 Bit rate 3.9 Mb/s Resolution 352 (codec setup) X 480 lines
4. MPEG 2 Bit rate 5.3 Mb/s Resolution 704 (codec setup) X 480 lines
5. MPEG 2 Bit rate 8.3 Mb/s Resolution 704 (codec setup) X 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 HRCs 6-8 in each study was added to the signals by attenuating a modulated version of the signals before passing them on to a demodulator. The SNR was measured with a Tektronics VM700 video test instrument. To avoid introducing jitter when recording these signals, the noise on the synchronizing pulses was removed by regenerating them in a processing amplifier. The VHS unit used was a consumer model, rather than a laboratory model. Note that MPEG 1+ at 3.9 Mb/s and the comparison HRCs 6-10 were used in both studies.

The same set of scenes was used in both studies. The scenes were chosen to span 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 seconds long. Four of the scenes are clips from movies and four of the scenes are clips from sporting events. The sources for the movie clips were commercial laser discs copied to MII equipment using a Y/C component connection. The sports event scenes were supplied by local broadcasters on Betacam SP tape.

2. Objective Measures

2.1 Performance Measurement Issues for Digital Video Systems

2.1.1 Input Scene Dependencies

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 [3, Annex A of 4]. Other methods have been proposed for determining the picture-content failure characteristic for the system under consideration [Appendices 1 and 2 to Annex 1 of 5]. National and international standards have been developed that specify standard video scenes for testing digital video systems [4, 6, 7]. Use of these standards helps ensure that adequate care is taken when evaluating systems from different suppliers.

2.2 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.601-4 [10] and extracts features from these digitized frames of video. Features are quantities of information that are associated with individual video frames. They 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 standard contains parameters derived from three types of features that have proven 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.

Further refinement in objective video quality assessment promises to lead to "in-service" objective methods for measuring video quality that may be good enough to replace subjective experiments. Already it is possible to perform non-intrusive, in-service performance monitoring which is useful for applications such as fault detection, automatic quality monitoring, and dynamic optimization of limited network resources.

2.2.1 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 will be 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 CCIR Recommendation 500 [5] or ITU-T Recommendation P.910 produce one subjective mean opinion score (MOS) for each HRC-scene combination. Since these video clips are normally about 10 seconds in length, it is necessary to "time collapse" the frame-by-frame objective parameter values before they are correlated to subjective MOS. COM 12-66-E [15] specifies several useful time collapsing functions such as maximum, minimum, and root mean square (rms). The maximum and minimum are useful to catch the extremes of video quality while the rms is a good indicator of the overall average.

2.3 Description of the Video Processing System

A computer-controlled frame capture and storage system was used to sample and store the video clips from the two MPEG studies. Video is received on Betacam SP tape cassettes. An HP workstation controls both a Sony BVW-65 and a Truevision ATVista

frame grabber installed in a PC. For the results in this paper, only the luminance channel from the Betacam SP deck was used.

The ITU-R Recommendation BT.601 A/D sampling rate of 13.5 MHz results in a frame size of 720 x 486 pixels. Each pixel is sampled using 8 bits giving 256 discrete levels of luminance. The A/D is adjusted to sample black (normally 7.5 IRE) as 16 and white (normally 100 IRE) as 235.

Using the dynamic tracking and remote control capabilities of the BVW-65, NTSC fields 1 and 2 are grabbed and combined to produce an NTSC frame. The NTSC frame is stored in TIFF format on a video optical disc jukebox which allows storage of up to 1 hour of uncompressed video.

This data collection and storage system ensures the availability of each frame or field at any timecode during the processing by the HP workstation. The optical jukebox provides random access to input and output frames, which enables the objective video quality measurement system to implement matrix metrics (based on pixel by pixel comparisons of entire frames), as well as scalar and vector metrics.

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

Calibration is an important issue whenever input and output video frames are being directly compared. Neglecting calibration can produce large measurement errors in the parameter values. For example, both non-unity channel gains and non-zero level offsets can have a significant effect on the calculations of peak signal to noise ratio (PSNR) and other parameters defined in COM 12-66-E.

COM 12-66-E [15] specifies robust methods for measuring gain, level offset, and active video shift (i.e., spatial registration of input and output video frames). 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 these calibration measurements had to be devised for the MPEG experiments because the calibration frames were not included on the source tapes. For a description of this new method see [14].

The calibration analysis revealed that it is quite common for digital video systems to have substantial non-unity gains, level offsets, and horizontal and vertical shifts of the output video. For details see [14].

In light of the calibration analysis, a separate gain g, level offset l, horizontal shift h_s, and vertical shift v_s was computed for each clip (i.e., each HRC-scene combination). This was done by median filtering the calibration quantities for that clip. Each frame of the clip was then corrected using the median filtered calibration quantities for that clip before any objective parameters were computed. Note that within-scene variations from the calibration quantities are not removed by this approach. These within-scene variations will thus be detected as impairments by the objective parameters.

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

The output video frames must be temporally aligned, or registered, to the input video frames before the objective parameters can be 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, decoder). There are two fundamental methods that can be used to perform temporal alignment. The first method, called constant alignment, gives one time delay measurement for the entire output video stream. The second method, called variable alignment, gives a time delay measurement for each individual output video frame (See ITU-T COM-12-75-E [16]). 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.

2.5.1 Temporal Alignment Test Results

For higher quality[3] transmission systems like MPEG, the field-based constant alignment method presented in [15] has proven to be a simple and excellent technique for measuring video delay. They have the added advantage of being "in-service" methods of measurement for video delay. For transmission systems that repeat frames, drop frames, or perform temporal warping, 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 chosen as the one to use for computation of the objective parameters.

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

This section presents a tabular summary of the objective parameters that were computed for each HRC-scene combination in the MPEG 1+ and MPEG 2 studies.

Parameter	Description	Reference for Method of Measurement (COM 12-66-E)
711	maximum added motion energy	Section 7.1.1
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

[3] In this case, high quality refers to the temporal aspects of the video (i.e., systems that rarely drop frames) and includes analog video transmission systems as well as high bit-rate digital video systems.

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	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	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	missing edge energy frequencies	Section 7.1.10 using modified feature comparison function to sum the missing 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 $[\text{Sobel(input)-Sobel(output)}]_{np}$ ($[X]_{np}$ 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 $[\text{Sobel(input)-Sobel(output)}]_{pp}$ ($[X]_{pp}$ defined in section 6.5.1.7)

Notes:

1. The "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. It was observed that an r_{min} threshold of 20 included nearly every pixel in the sampled video frames due to the amount of noise which was present in the source video. With an r_{min} threshold of 60 the noise was effectively eliminated from the calculation

2. The "added edge energy frequencies" and "missing edge energy frequencies" parameters were actually computed using a mean calculation rather than a sum calculation in the comparison function in section 6.5.1.9 to remove the effect of scene length.

3. Subjective Data

3.1 Methods Used to Collect Subjective Data

The method used to collect subjective data was a variant of the CCIR 500 method. Recorded video segments were played back to human observers 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 multi-lab study (see [2], for rationale and details). Three main differences were

- HRCs 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 HRCs being tested. Rather than the original being presented, say, 80 times while all other HRCs are presented eight times, as in the "standard" method, in the current method the original is presented eight times as a comparison and the other 72 exposures are equally spread among the other HRCs.
- 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 HRCs in dollars per month. This method does correlate highly with the impairment scale method, but also provides other technical advantages (see [2]).
- The video clips were recorded and played back on a video disc, rather than on a Betacam SP tape recorder. The performance specifications for the video disc machine are marginally lower than for the tape machine (>45 dB video S/N, 450 pixels horizontal resolution). The video disc has the advantages of random access and computer control. The ordering of stimuli was separately randomized 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 approximately an equal number of times, but on any specific trial the scene was selected randomly. This sampling procedure is based on the logic that the HRCs we are testing 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 data in the dollar-rating task. The observers were not laboratory employees. They were chosen to be cable TV customers, familiar with the signal quality of cable TV, and also familiar with paying for TV service.

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.

3.2 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 HRCs with which it was compared. Table 1 shows the mean ratings for the MPEG 1+ study and shows the mean ratings for the MPEG 2 study. The standard errors of the values in Table 1 are on the order of 0.7, and in the standard errors are on the order of 1 (there being half as many trials per subject as in the MPEG 1+ study).

Other papers have presented analyses of these subjective data in some detail [1, 2]. In both data sets the ratings are statistically related to the variables: HRC, Scene, and the specific HRC-Scene combinations. This is what one would expect, and the subjective data are in accord with expectations. Other analyses demonstrate that the subjective data are not excessively noisy and show systematic differences between the way observers react to analog vs. digital HRCs. 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.

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

Scene	1.5 Mb/s	2.2 Mb/s	3.9 Mb/s	5.3 Mb/s	8.3 Mb/s	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 2 Mean subjective ratings of HRC-scene combinations, MPEG 2 study

Scene	3.0 Mb/s	3.9 Mb/s 1+	3.9 Mb/s	5.3 Mb/s	8.3 Mb/s	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

4. Statistical Analyses

4.1 Methods

4.1.1 Strategy

The theoretical goals of the analysis are to

- Find the "best" set of objective measures for predicting the subjective judgments, and
- Determine how close to optimal these predictors are.

Two features of most data sets complicate the problem of finding the "best" set of predictors and force one to use compensating data analysis strategies. The complicating features of data are (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.

Because of the realities of data,

- The actual goals of the analysis are 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 dealing with data with noise and redundancy are:

- Measure the redundancy in the set of predictor variables;
- On the basis of the measure of redundancy, pre-specify the maximum number of variables to be used in any analysis;

- Use variables that are known a priori to be causally related to the dependent variable whenever possible;
- Verify that a candidate set of predictor variables generalizes to another data set or sample.

4.1.2 Redundancy

The set of 20 objective measures are 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." 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 fourth the number of original variables.

4.1.3 Reliability

The reliability issue is important because it limits the statistical fit of even a perfect objective measure (see [12, 13]). 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 definition of reliability of a variable is: 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, one can still estimate the ratio using observable quantities, as follows (see [13]).

- 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.

We used the method of analyzing repeated measurements to compute estimates of the statistical reliability[4] of the objective measures and of the subjective measure. Five of the HRCs and all eight of the scenes were nominally the same across the two experiments. The repeated HRCs were MPEG1+ at 3.9 Mb/s, the cable simulations at 34, 37, and 40

---

[4] 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 accurately responding 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.

dB S/N, and VHS. We say "nominally the same" because the two tapes of the HRCs 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 digitizing and computing.

4.1.4 Regression

We use a standard regression program found in the SAS statistical software package for most of the analyses in which we use the objective measures to predict the subjective judgments. We also use a "stepwise" regression as a secondary analysis. Stepwise regression is an exploratory data analysis technique that looks for a best-fitting set of predictor variables via a mechanical algorithm. Stepwise 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 stepwise regression finds is rarely the set that is most generalizable.)

4.2 Results

4.2.1 Redundancy in objective measures

MPEG 1+ data set alone. The 20 objective measures, applied to the MPEG 1+ data set of 72 HRC-scene pairs, yielded four "factors" in a principal components analysis. The four factors accounted for 81% of the variance in the 20 measures. The factors are described:

1. The first component accounted for 33% of the variance in the data. The four measures with the largest correlation were 719 and 719_60 (two measures of edge energy difference), 721 (a measure of added spatial frequency), and Negsob (a measure of the difference between the Sobel transforms of the original and processed images).
2. The second principal component accounted for 28% of the variance, and the pattern of correlations was complementary to that of the first principal component (high where the first was low, and vice versa). The three measures with the largest correlations were 712 (lost motion), 722 (lost spatial frequency), and Possob (a second, complementary measure based on differences in Sobel images).
3. The third principal component accounted for 13% of the variance. The four measures that correlated highest with this component were 7110a (added edge energy), 713, 714, and 715 (types of motion difference, including repeated frames).
4. The fourth component accounted for 6% of the variance. It correlated highest with 7110 and 713 (types of motion difference).

MPEG 2 data set alone. The MPEG 2 data set also yielded four principal components with eigenvalues greater than 1.0; the four accounted for 83% of the variance in the data. Descriptions:

1. The first component accounted for 44% of the variance in the data set. It correlated equally well with six of the measures: the suite of four 719 variants (edge energy difference), 721 (added spatial frequency), and Negsob (difference in Sobel images). This principal component is very similar to the first principal component of the MPEG 1+ data set.
2. The second component accounted for 21% of the variance. Its four highest correlations were with measures 717 (lost edge energy), 732 and 733 (peak signal to noise ratio), and Possob (the other measure of differences in Sobel images). Again, the second component is similar across the two data sets.
3. The third principal component accounted for 9% of the variance. It correlated most highly with measures 7110 (added edge energy) and 713 (motion difference). This principal component is similar to the fourth component of the MPEG 1+ data.
4. The fourth principal component accounted for 8% of the variance. It correlated most highly with the measures 7110a (another measure of added edge energy) and 714 (another measure of motion difference). This principal component corresponds to the third component of the MPEG 1+ data.

Thus, the MPEG 2 data set replicates the pattern of results from the MPEG 1+ data set quite well. The total amount of redundancy in the measures was very similar, and the pattern of redundancy was similar across the two sets of HRCs.

MPEG 1+ and MPEG 2 data sets together. A principal components analysis of the two data sets together revealed a similar pattern of results (as one might expect). Four principal components had eigenvalues greater than 1.0, and jointly accounted for 80% of the variance. Descriptions of the components:

1. The first component, as in the two data sets separately, correlated highest with measures from the 719 series, 721, and Negsob. It 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.
3. The third component accounted for 12% of the variance and correlated highest with the added edge energy (7110a) and motion difference measures (714, 715).
4. The fourth component, accounting for 7% of the variance, correlated highest by far with measure 7110 (added edge energy; 7110 and 7110a are slightly negatively correlated with each other).

4.2.2 Regression

Any one regression run, on any one data set, is unlikely to produce a generalizable result. However, multiple runs on multiple data sets that produce similar answers form the basis for credible and potentially generalizable results. The following analyses form a sequence in which the details do not generalize from analysis to analysis, but the general pattern of results does generalize.

MPEG 1+ alone using measures from principal components analysis. The principal components analysis showed that the data do not support more than four orthogonal

variables. This fact does not absolutely require that the regression use four or fewer variables. However, practical experience shows that fewer rather than more variables actually generalize to other data sets. Therefore we used only a single variable from each of the four principal components that passed the eigenvalue test.

These variables were the measures 7110 (added motion energy), 713 (average motion difference), 719_60 (edge energy difference), and 722 (lost spatial frequency). The adjusted R^2 for this regression was 0.586. By comparison, the R^2 for the best model in the T1A1.5 three-lab study, using comparably averaged subjective data, was 0.706. Also, two of the four variables were not significant, viz., 7110 and 722. Variables only from the first and third principal components were significant as predictors of the subjective ratings. Thus, we might hope that we could do better with the MPEG data than we did just using variables from the principal components analysis.

MPEG 1+ alone, using Sobel image measures. The first two principal components of both data sets correlate nearly maximally with the two Sobel image measures. Because these measure are of a priori interest, we ran a regression using the Sobel measures as the representatives of the first two components. The remaining two measures were 713 and 7110. The adjusted R^2 for this regression was 0.689, quite a bit better. The most interesting outcome of this regression was that the Sobel measure Negsob was by far the most important variable. Again, the variable 713 from the third principal component was a significant predictor, and neither of the variables from the second and fourth principal components were significant (i.e., Possob and 7110, respectively).

MPEG 1+ alone, exploratory stepwise analysis. Stepwise regression enters variables sequentially, choosing the next variable that maximizes R^2 given the preceding variables. Typically, results of a stepwise analysis are sensitive to noise in the data, so are not to be trusted in isolation. However, when used in combination with other analyses, stepwise can be informative. In the present data set, the order of entry of significant predictors was: Negsob, 713, and 717. Negsob and 713 were significant predictors in the preceding analyses. The measure 717 (lost edge energy) is highly correlated with the other candidate measures from the second principal component (722 and Possob) that turned out not to be significant for this data set. The R^2 for this three-variable model was 0.737, which is respectable by comparison with the T1A1.5 results.

From the analyses of the MPEG 1+ data set, we can take forward the hypotheses (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.

MPEG 2 alone using measures from principal components analysis. As in the case of the MPEG 1+ data set, results of the principal components analysis suggested four or fewer measures should be used in the regression analysis. The measures that best fit the first four components, respectively, were 719a_60 (edge energy difference), 717 (lost edge energy), 7110 (added edge energy), and 714 (lost motion). The adjusted R^2 for the regression with these variables was a respectable 0.718. However, variable 7110 was not significant as a predictor (as was true of the MPEG 1+ data set).

MPEG 2 alone using best MPEG 1+ measures. A set of candidate "best" predictors from the MPEG 1+ analysis was Negsob, 713 (motion difference), and 717. The adjusted R^2 fit of this model to the MPEG2 data was 0.815, quite an improvement over the variables derived from the principal components, and also an improvement over the T1A1.5 multi-lab data set. However, even with this good fit, both 713 and 717 were only marginally significant, suggesting that an even better fit might be possible.

MPEG 2 alone, exploratory stepwise analysis. The order of entry of three clearly significant variables was Negsob, Possob, 714 and 711; 711 was marginally significant. The adjusted R^2 for this model was 0.849, another improvement. Again, Negsob was by far the most important variable, achieving a fit of 0.774 by itself.

The three hypotheses from the MPEG 1+ data set were supported in this data set. (Principal components three and four in the MPEG2 data set need to be switched for the first hypothesis to be exactly true.) We take these hypotheses into the analysis of the joint data set.

MPEG 1+ & 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, 722 (lost spatial frequency), 714 (lost motion), and 7110 (added edge energy). The adjusted R^2 for this set of predictors was 0.704. The variables 7110 and 722 were not significant, as was the case in the analysis of the MPEG1 data set. Again, Negsob had by far the largest effect.

MPEG 1+ & 2 using variables from MPEG 2 analyses. A slightly different set of variables had been identified in the analysis of the MPEG2 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).

MPEG 1+ & 2 using stepwise. The first three variables to enter the equation, and the only three that appreciably improved the fit of the model, were Negsob, 711, and 714, respectively. (Possob was marginal). The adjusted fit of the three-variable model was 0.763.

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) ; 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.

4.3 Interpretation of results

4.3.1 Which measures work best

For the current data sets (i.e. higher resolution formats with little or no frame repetition), 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 picked 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 (717 for lost edge energy, or 722 for lost spatial frequency, a measure of edge sharpness).
- 714 or the family of measures of lost motion (713 for average motion difference or 715 for repeated frames).
- 711 or measures of motion difference (713) or measures of edge energy difference (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 to 8 percent. 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 to 8 percent reduction in prediction efficiency.

The particular package of measures that predicts subjective judgments best may depend somewhat on the particular domain of HRCs and scenes for which one wants to make predictions. For example,

- If one is interested in comparing only MPEG HRCs running at different bit rates, then one package of measures could be slightly better, while if one were comparing MPEG to VHS and cable, 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.

4.3.2 How good the statistical fit really is

In the combined data set the objective measures were able to account for 0.763 to 0.769 of the variance, depending on whether one used three or four predictor variables. By way of comparison, in the T1A1.5 multi-lab study ([11], pg. 28), the fit was not quite as good: R^2=0.706.

Another comparison that is relevant 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 dealt with the problem of estimating R^2 in the presence of error ([13], pg. 22): "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) * (weighted average of reliabilities of X's).

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 beta coefficients for 711, 714, and Negsob, scaled to sum to 1.00.) See [14] 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 have seen in the present study. Compared to an R^2 of 0.845 (correlation coefficient of .92), the observed R^2 of 0.763 (correlation coefficient of .87) is 90% of maximum. As in the case of the T1A1.5 study, the ability to predict is good but shows some room for improvement.

5. 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 measures to apply to a specific testing situation, so we cannot say for certain whether this current set of measures has the potential to be fine-tuned for application in testing equipment. However, the current objective measures should be considered as reasonable candidates for testing applications.
- The kinds of objective variables that predict subjective responses well for MPEG video are
  (a) Measures of the addition of false edges, in particular the matrix measure Negsob,
  (b) Measures of lost sharpness of edges,
  (c) Measures of change in motion.
- A traditional objective variable that does not predict subjective responses well for MPEG video 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.
- The conclusion is reached that by using a set of 3 or 4 objective measures as indicated above, a correlation coefficient of .87 is achieved. Since the covariance/variance analysis indicates that .92 is the best possible for this study, this result is very good.

6. References and Bibliography

[1] Cermak, G. W., Teare, S. K., Tweedy, E. P., and Stoddard, J.C. (1996), "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, pp. 53-62.

[2] Cermak, G. W., Tweedy, E.P., Ottens, D. W., and Teare, S.K. (1996b). "Consumer acceptance of MPEG1 video at 1.5 to 8.3 Mb/s." ANSI T1A1 contribution number T1A1.5/96-108. [5]

[3] A. A. Webster, C. T. Jones, M. H. Pinson, S. D. Voran, S. Wolf, "An objective video quality assessment system based on human perception," SPIE Human Vision, Visual Processing, and Digital Display IV, vol. 1913, Feb 1993.

[4] ITU-T Recommendation P.910, "Subjective video quality assessment methods for multimedia applications," Recommendations of the ITU (Telecommunication Standardization Sector).

[5] CCIR Recommendation 500-5, "Method for the subjective assessment of the quality of television pictures," Recommendations and Reports of the CCIR, 1992.

[6] 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).

[7] 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.

[8] 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.

[9] 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, N. W., Suite 500, Washington DC 20005.

[10] ITU-R Recommendation BT.601-4, "Encoding Parameters of Digital Television for Studios," Recommendations of the ITU (Radiocommunication Sector).

---

[5] Copies of ANSI T1A1 contributions can be obtained from the T1 Secretariat, Alliance for Telecommunications Industry Solutions, 1200 G Street, NW, Suite 500, Washington, DC 20005.

[11] Cermak, G. W., and Fay, D. A., "T1A1.5 Video quality project: GTE Labs analysis," ANSI T1A1 contribution number T1A1.5/94-148, Sept, 1994.

[12] Bollen, K.A., Structural Equations With Latent Variables, New York, Wiley, 1989.

[13] Cochran, W.G., "Some effects of errors of measurement on multiple correlation," Journal of the American Statistical Association, No. 65, pg. 22-34, 1970.

[14] G. Cermak, P. Tweedy, S. Wolf, A. Webster, and M. Pinson, "Objective and subjective Measures of MPEG Video Quality," ANSI T1A1 contribution number T1A1.5/96-121, October 28, 1996.

[15] ITU-T Contribution to Question 22/12, COM 12-66-E "Selections from the Draft American National Standard: Digital Transport of One-Way Video Telephony Signals - Parameters for Objective Performance Assessment", (USA), January 1996.

[16] ITU-T Contribution to Question 22/12, COM 12-75-E "Visual channel delay and frame rate measurement - initial results with a prototype system", (AT&T),March 1996.