AAW

COMMITTEE T1
CONTRIBUTION

Document Number: T1A1.5/94-131

****

STANDARDS PROJECT: Analog Interface Performance Specifications for Digital Video Teleconferencing/Video Telephony Service

****

TITLE: T1A1.5 Video Quality Project: Preliminary Results

****

ISSUE ADDRESSED: Correlation of objective and subjective measures of video quality

****

SOURCE: GTE Laboratories, Incorporated
(Gregory W. Cermak and David A. Fay)

****

DATE: 19 July 1994

****

DISTRIBUTION TO: T1A1.5

****

KEYWORDS: Video Quality, Objective Quality, Subjective Quality, Statistical Measures, Correlation

****

****

Notice

****

T1A1.5 Video Quality Project: Preliminary Results

Introduction

Need for measurement standards for video quality

Wolf (1990, p.2) 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." Measures of image quality do exist; the photographic, computer, television, and motion picture industries have faced this measurement issue for decades (e.g., Mezrich, Carlson, and Cohen, 1977). However, measuring the quality of images that have been *digitally coded and transmitted* presents a new set of problems, in the opinion of those familiar with video errors and artifacts (see Wolf, 1990).

Previous work

A substantial amount of work has measured the quality of analog video (see CCIR 654, 1986; CCIR 959-1, 1986). Various sorts of impairment have been calibrated with regard to their effect on *subjectively judged* picture quality, e.g., signal to noise ratio for continuous random noise, differential gain across the luminance scale, and echo. The keys to such calibration are (a) that the impairment be reproducible and (b) that human observers react to a given impairment in a predictable way. Reproducible test procedures had evolved that fairly represented the sorts of distortions to analog television signals one would meet in practice. Standard procedures for testing human viewers had also become accepted (CCIR 500 series). With the advent of digitally coded and compressed video, neither of the two keys to calibration could be taken for granted, and a new round of testing was begun, following the general strategy of the earlier work (e.g., Voran, 1991; Wolf, 1990).

This preliminary report

The video quality project has already existed for six years (see T1Q1.5 Working Group, 1988). Although we now have data on hand, it's anyone's guess how long it will take to issue a final report. The present report is meant to update the video quality measurement community in a timely way, so that we do not have to wait for a final report to get the benefits of the research that has been done. Readers should understand, however, that this report is not final and does not represent the position of the T1A1.5 committee.

Method

The test method is documented thoroughly in T1A1 document T1A1.5/94-118 entitled Subjective Test Plan (Morton, 1994). Highlights of the method appear below.

General strategy

Objective & subjective data

To obtain the "new objective measures of video transmission quality" mentioned at the outset, the strategy was to generate many candidate objective measures of video quality, then to choose those measures that best predict subjective video quality. The objective measures would come from considerations of the basic nature of moving video images and from previous experience with analog systems. The subjective measures would come from consumers. Any measurement instrument is ultimately calibrated against standard units; in the case of video quality, the standard is consumer judgment.

Standard, "official" techniques
Committee sentiment was that if consumer judgments were to be used, then at least the consumer judgments should be collected in the most respectable way possible. The most respectable measurement methods for the telecommunications industry are those sanctioned by the International Telecommunication Union, in particular those in Recommendation CCIR 500-5, so that is what we used.

Three labs
Subjective data were to be collected at three labs (NTIA, GTE, and Delta Information Systems). This improved the test program in four ways.

- The population of consumers is better represented by samples at three different locations (Boulder, Boston, and Washington DC, respectively).
- Although any single lab might somehow bias the subjective results, biases from three labs would be more likely to "average out."
- Inter-lab differences could be quantified. This is a very important point, because up until the current study the committee was plagued by vague fears of non-comparable data across labs, but there were no data to substantiate or allay the fears.
- More data could be collected. When one has noisy data, such as subjective data, the more data that get averaged, the more precise the estimate.

Sequences to be measured
Video systems
The committee, encouraged by the NTIA group (National Telecommunications and Information Administration), opted for testing a broad range of quality in video signals.[1] Although other methods of achieving a broad range of quality are possible -- such as deliberately modifying test signals with either analog or digital equipment -- the committee chose to use a broad range of actual video systems (Hypothetical Reference Circuits, or HRC's). The current authors' involvement with the T1A1.5 committee began just as the list of HRC's was approved for testing, so we do not know what other considerations led to the choice of the particular HRC's, other than to span a broad range of video quality. Twenty-five HRC's were chosen to represent the following categories (Appendix A shows the Test Plan's description of the individual HRC's). These systems cover a range of (nominal) bit rates from 112 kbps to 70 Mbps. Categories:

- High quality
- Vector quantization, medium rate
- Proprietary, low to medium rate
- Proprietary, medium to high rate
- QCIF, low rate
- QCIF, medium rate
- CIF, low rate
- CIF, medium rate
- CIF, high rate

[1] From a statistical point of view, it is always much easier to get good results if there is a large range in the variables one measures.

Scenes
NTIA has argued that testing should use a range of actual video material because the performance of digital systems varies with the kind of material transmitted (see Seitz, Wolf, Voran, and Bloomfield, 1994). In fact the word in the corridors has been that most of the interesting information has to do with differences among HRC's at the same bit rate in the way they handle a given scene (the statistical HRC-Scene "interaction"). Note that using typical video material is not a universal practice; the official MPEG test scenes are quite contrived (see Cermak, Teare, Tweedy and Stoddard, 1994).

Twenty-five scenes were chosen to represent five categories that are typical of video telephony, video teleconferencing, and entertainment video. These categories cover a wide range of movement and detail. Members of the committee donated the video footage. The categories:

- One person, mainly head and shoulders
- One person with graphics and/or more detail
- More than one person
- Graphics with pointing
- High object and/or camera motion

Each scene was edited to a length of nine seconds (plus three seconds of leader to allow the HRC to stabilize).

HRC-scene combinations
There are 625 possible combinations of 25 HRC's and 25 scenes. Most experimenters would not have used all possible combinations; efficient statistical designs for experiments allow good estimates of the effects of the variables HRC, Scene, and the HRC-Scene interaction with a fraction of the total possible combinations. However, the committee felt a need to see data on every last one of the HRC-scene combinations.[2] Therefore all scenes had to be run through all HRC's, and tapes had to be made of the 625 combinations. And, we wanted no editing errors. And, we required several orderings of the video scenes on the tapes to avoid any biases due to order effects.

Luckily, ABC was able to do all the editing in D2 (digital) format, using a randomization scheme supplied by NTIA and Delta Information Systems. The D2 version was then dubbed to BetacamSP for presentation to subjects at the individual labs. A complete set of stimuli runs to twelve tapes.[3]

Subjective measures

Presentation equipment
The stimulus tapes were played back on BetacamSP equipment through a Sony 1910 self-calibrating monitor. This equipment was the standard across the labs. A presentation room was set up with illumination and wall reflectance in compliance with CCIR 500-5. The room setup and light meter readings for the GTE Lab are shown in Appendix B.

---

[2] One can argue that the relationship between the objective and subjective measures across the HRC-scene combinations is all that matters. Data about the particular HRC's and scenes are ancillary and need not be complete. The committee, however, has taken the view there should be complete data at the level of the HRC-scene combination.

[3] The randomization called for 12 blocks of video sequences. Each subject was shown a different ordering of four blocks. The rather involved randomization scheme is described in the Test Plan.

Response scale

We used the method that CCIR 500-5 calls "the double-stimulus impairment scale method," in which the original of a given scene is presented, then the corresponding scene as processed through a particular HRC. The processed version is judged with respect to the original; we ask the subject how *impaired* the processed version is compared to its original state. The subject responds on a five-point scale whose points are labelled

- Imperceptible
- Perceptible, but not annoying
- Slightly annoying
- Annoying
- Very annoying.

Subjects

This preliminary report presents data from 79 subjects from the first two labs to have reported their data, GTE and NTIA. The criterion in selecting subjects was that they be either current users of video or video teleconferencing, or be likely users of such systems in the future. At least 50% of the sample was to be current users; in fact, 39% were recent users and another 10% had used video telephony or video conferencing at least some time in the past.[4] The "likely future users" were chosen to be similar to the actual users in age and type of job, namely technical and professional for the most part.[5] Females constituted 30% of the sample. The GTE Labs sample was recruited from the Route 128 high-tech community, but from outside GTE; these subjects were paid. The NTIA sample was mainly unpaid volunteers from within the NTIA/NIST facility in Boulder. Subjects' visual acuity and color vision were tested after the video quality testing, so we have data from both "normal" and "impaired" subjects for purposes of comparison.[6]

Procedure

Any given subject saw and judged four of the 12 tapes.[7] Each tape ran for slightly over half an hour. Subjects judged two tapes per day, separated by a break. Each subject had been randomly assigned to a tape group, to an ordering of the tapes within a group, and to an observing seat (left, center, or right). Instructions had been recorded on the audio track of the tape; we got the subject(s) settled then played the instructions. The instructions described the purpose of the study, gave an overview of the task, then led into six practice trials on scenes that were similar to (but distinct from) those appearing in the main experiment. The experimenter answered any questions, then stressed that if the subject should lose concentration for a video sequence, we'd prefer to go back and replay it rather than have the subject just guess.[8]

On each trial of the test the subject saw the original scene and a processed version. A voice on the audio track requested that the subject rate the scene, and identified the number of the scene. The number corresponded to numbers on the (paper) rating sheets so that subjects

---

[4] Each lab was to provide 30 datasets from "qualified" subjects, i.e., at least 15 from experienced users. The data from the two labs represent 31 recent users. We are using all the available data for analyses that appear in this report.

[5] The job categories of all subjects are available on request.

[6] Naturally, the population at large contains both visually-normal and visually-impaired persons.

[7] The tapes had been designed in three groups of four, so that in each group there was a similar range of HRC's, and for each HRC all 25 scenes appeared. Within each of the labs, the plan was to get at least 10 subjects per tape group. Over the three-lab experiment that would yield at least 30 subjects per HRC-scene pair, and the subjects would be spread across the US.

[8] Having gone through the test themselves, the experimenters knew how easy it is to lose concentration briefly during the test.

could keep track of where they were. The rating sheets had the five-point scale for each trial. The pace of the experiment was controlled by the tape; the experimenter mainly monitored, pushed the start and stop buttons, replayed trials as needed, and tried to jolly the subjects along if they looked bored.

After the final tape session we performed the vision tests then paid the subjects. We had collected demographic information at the beginning of the first session.

Objective measures from NTIA

Two general classes of objective measures have been proposed for the HRC's in the test, (a) measures that apply to an HRC independent of any particular scene, and (b) measures that apply to an HRC as it treats a particular scene. In the former category are measures like signal to noise ratio, spatial frequency response, and response time to a scene-cut. The present analyses used only the latter type of measure.

End-to-end system measures

A main element of the NTIA measurement philosophy is that measures should apply to the whole transmission system end-to-end (see Seitz, Wolf, Voran, and Bloomfield, 1994). They have developed a class of objective measures that are based on *comparisons*, just as the subjective measures are based on comparisons. A measurement is made of the signal as it enters the transmission system, and another as it arrives at the consumer's CPE. A family of measures is defined comparing these two measures in various ways.

More particularly, each frame of each scene is captured and analyzed. The analyses can be of information in the time domain or the spatial domain. A number of different measures are computed. Corresponding measures are made of the scene after it has been through a given HRC. The two sets of measures are combined according to explicit formulas (see Webster, 1993; Wolf, Pinson, Jones, and Webster, 1993). We paraphrase the NTIA group's descriptions of these measures below.

Temporal domain measures

Each frame is digitized, pixel-by-pixel. Consider two adjacent frames. Take the difference between those frames, pixel-by-pixel. Take the absolute values of those differences. Compute the mean and standard deviation of the absolute values across all the pixels. The basic building block measure of temporal information is computed as the following composite of the mean and standard deviation of the absolute differences between two frames:

$$TI = SQRT [Mean(diffs)**2 + Stdev(diffs)**2].$$

If there are 270 frames, then there will be 269 TI values.[9] This characterizes a single HRC-scene combination, and there are 625 such combinations. This is too much information to handle realistically for predicting subjective responses, so a composite measure is created out of the 269 basic TI values for a given HRC-scene combination. Similarly, a corresponding composite measure is computed for the original version of that scene.[10]

In computing the composite measures, one must imagine the two time series of frames (for the coded scene and the original scene) being aligned frame-by-frame. When a difference between frames 44 and 45 is computed in the coded scene, the corresponding difference between frames 44 and 45 is computed for the original scene. The composite measures use these two parallel streams of TI measures (see Fig. 1 below).

---

[9] NTIA reports that they actually analyzed 271 frames per scene.

[10] Strictly speaking, the composite measure is calculated using the TI values of the original and coded versions simultaneously.

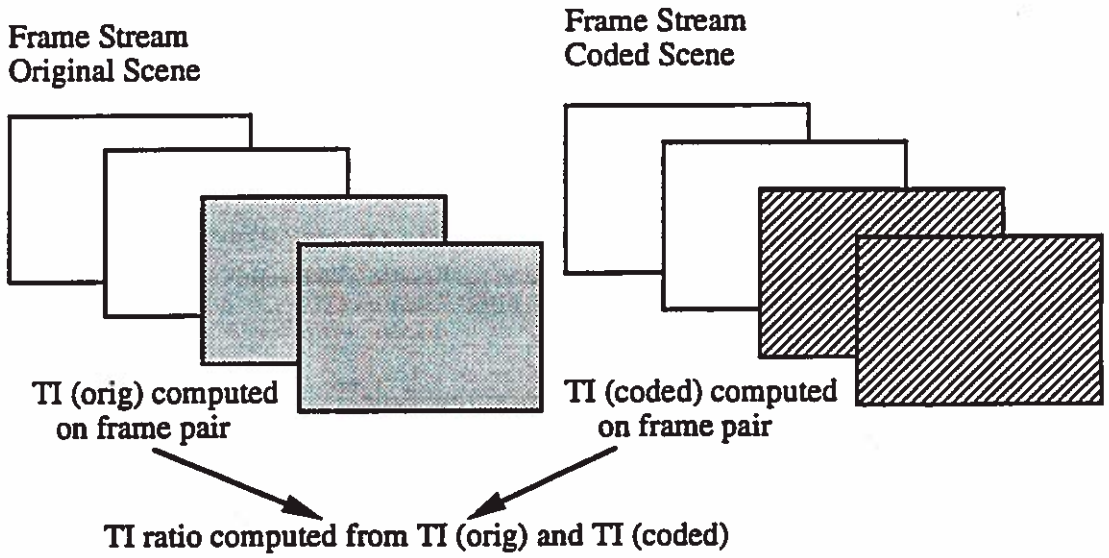


Figure 1. General scheme for defining Temporal Information measures.

The various measures below represent different approaches to creating a composite TI measure across the parallel frames in the original and coded scenes. All of these measures can be viewed as indicating added or lost motion in the coded scene compared to the original. Additions come from jerkiness and error blocks. Deletions come from frame repetition.

- P1 is the maximum of the TI ratio over the two streams of 269 frame pairs:
  MAX [log10{TI(coded) / TI(orig.)}]

- P2 is the RMS of the TI ratio over the two streams of 269 frame pairs:
  RMS [log10{TI(coded) / TI(orig.)}]

- P3 is the maximum minus the minimum of the TI ratio over the two streams of 269 frame pairs:
  MAX [log10{TI(coded) / TI(orig.)}] - MIN [log10{TI(coded) / TI(orig.)}]

- P4 is the mean of the positive values of the TI ratio over the two streams of 269 frame pairs minus the mean of the negative values of the TI ratio over the frame pairs.

P5 and P6 use a different ratio of the coded and original scenes than P1-P4, called the TI error ratio: {TI(orig.) - TI(coded)} / TI(orig.)

- P5 is the RMS of the TI error ratio over the two streams of 269 frame pairs:
  RMS[{TI(orig.) - TI(coded)} / TI(orig.)]

- P6 is the RMS of the stream of positive values (only) of the TI error ratio. This measure is sensitive to lost information, such as during frame repetition.

- P10 is a measure of frame repeat rate. In many HRC's the frame repeat rate is adaptive and changes over time. P10 is a measure of the upper end of the distribution of frame repeat rates for a scene-HRC combination. The definition

is algorithmic -- a series of operations to be performed on the original and coded frame streams; see Wolf et al., 1993.

- P11 is a measure of abrupt motion that is not due to scene cuts. The abrupt motion in the original is compared to the abrupt motion in the coded version. This is another algorithmic definition; see Wolf et al., 1993.

Spatial domain measures

Each frame is digitized, pixel-by-pixel. The digitized image is filtered by a process that emphasizes horizontal and vertical edges (a "Sobel filter.") The basic spatial information statistic, SI, is computed as the standard deviation of the pixels in the Sobel filtered image. SI is computed frame by frame for both the original scene and the coded scene. From these basic statistics a family of spatial measures is computed.

- P7 is the maximum absolute value of the SI error ratio. The SI error ratio is defined as

  $$\{SI(orig.) - SI(coded)\} / SI(orig.)$$

  This measure is computed on a frame-by-frame basis. This results in 270 values for P7 for each HRC-scene combination. That amount of information (270*625 values for just one parameter, P7) is not useful for later comparison with the subjective data -- it must be simplified to be workable. Therefore P7, as well as the measures P8 & P9, are composites defined across the 270 frame values. In the case of P7, the composite measure used is the maximum absolute value. When the SI error ratio is positive, P7 measures blurring; when negative, P7 measures added spatial information such as noise or blocks.

- P8 is the RMS of the SI error ratio (above) taken over the 270 frame values, i.e.

  $$RMS[\{SI(orig.) - SI(coded)\} / SI(orig.)]$$

- P9 is another version of P8 in which the RMS operation is applied inside the brackets to the primitive SI terms.

Fourier transform measures

The Fourier transform measures are also measures of spatial distortion in the processed scene as compared to the original scene. Imagine a two-dimensional Fourier transform of a single frame. Imagine also that the transform is centered in its rectangle. At the center of the rectangle is the amount of spatial information at very low frequencies; as one moves radially out from the center, information at higher and higher spatial frequencies is represented. (Typically, the amount of information or power falls off at higher frequencies, so the plot might be like an uneven hill centered in the rectangle.) Now imagine a thin ring about the center of the 2D spatial frequency spectrum. Integrate everything covered by that ring. Suppose we have this same operation for corresponding frames of the original and coded versions of the same scene. Now define the following two quantities:

$$PD = \{Amount\ (orig.) - Amount\ (coded)\} / Amount\ (orig.)\ if\ PD>0,$$

$$ND = \{Amount\ (orig.) - Amount\ (coded)\} / Amount\ (orig.)\ if\ ND<0.$$ [11]

PD represents distortion in which information is lost in a given spatial frequency band. ND represents distortion in which information is added in a given spatial frequency band.

---

[11] We use the term "Amount" rather than a term like "power" because we are not quite sure what the units are, but the idea is to compare the amount of signal in a given frequency band in the original with corresponding frequency band in the coded version.

For a given frame in the original and coded scene, the PD values are summed over all spatial frequency bands as a set of larger and larger rings are traced out over the Fourier transform rectangles. The comparable operation is performed for the ND values.[12] Two measures are defined from the PD and ND statistics:

- P12 = MAX (PD) over the 270 frames
- P13 = MAX |(ND)| over the 270 frames.

Results

Subjective measures

In the analyses that follow, we focus on the following questions:

- How accurate are the subjective data?
  How precise are the ratings of the HRC's and HRC-scene combinations?
  How credible are the subjective data?
  How big are the effects due to Lab and to individual differences among subjects?
- How sensitive are the ratings to HRC and scene?
- What do the subjective data tell us about what to expect from the objective-subjective correlations to follow?

The major variables we consider in the following analyses of the subjective data are Rating, HRC, Scene, and Subject.[13] We also consider other variables that apply to the Subject variable, namely, Lab, Team (the subset of four of the 12 tapes to which a subject was assigned), Gender, video teleconferencing experience (Video), performance on the vision tests (Vision), the chair location the subject viewed from (Seat), and whether the subject's data qualified according to the rules set down in the Test Plan (Qualify).

The analyses that follow are determined by the design of the experiment, and the design was "unbalanced" in the statistical sense of the term. Within any lab all 625 HRC-scene combinations were judged, so that part of the design was (nearly) balanced. However, each subject only judged about a third of the HRC's (but all 25 scenes for each HRC). Also, some HRC's appeared on the tapes of multiple "teams" of judges (for the sake of comparison), some HRC-scene combinations appeared more than once for each team and varied from team to team (as a consistency check), and an extra HRC (the null or original) appeared on each tape. Thus, the design was quite unbalanced with respect to the subjects, and was unbalanced due to the addition of the various consistency checks.

Each subject judged 250 HRC-scene combinations plus eight consistency check stimuli. For most of the analyses the consistency checks are treated separately from the rest of the data. The remaining 250 ratings for each subject encompass seven HRC's that were unique to that team, plus three HRC's that were common to at least two teams, for a total of 10 HRC's. Again, each HRC appeared with all 25 scenes.

---

[12] There is another complication which we have not described -- the operations described are actually performed within six subregions of each frame and then summed.

[13] We use the convention of capitalizing the *name* of a variable, but do not capitalize when referring to the item in general, e.g., Subject and Scene for the statistical variables used in the analyses, but subject and scene for the people in the experiment and the material they saw.

ANOVA

The main tool we use for answering the questions above is *analysis of variance* (ANOVA). The variance we refer to is variance or variability in the subjective ratings: The ratings vary from observation to observation. Using ANOVA we can decompose the variance in the ratings into effects of known factors, such as which HRC and scene were presented for a particular observation, and into unknown factors or error. In these analyses we used the 250 data points per subject that excluded extra catch trials. Also, we used all 79 subjects' data, irrespective of whether the subject passed all the screening criteria (see the section below on individual differences for justification).

Unfortunately, the size of the data set and the complexity of the experiment made it computationally infeasible to directly estimate the model specified by Crow of NTIA (Crow, 1994). The main results in Table 1 below should be viewed as a close approximation to the ANOVA for the data from the GTE and NTIA labs:

Table 1. An analysis of variance of the subjective data from GTE and NTIA.

Source	DF	Sum of Squares[14]	Mean Square[15]	F[16]	
HRC	24	16209	671.60	479.72	
Scene	24	4119	142.24	225.78	
Lab	1	27	25.98 (0.002)	0.86	(0.02)
Subject	77[17]	2423	30.35 (0.13)	72.26	(0.31)
HRC*Scene	576	2401	4.17 (0.12)	11.58	(-----)
HRC*Lab	24	34	1.40 (0.003)	1.07	(0.08)
Scene*Lab	24	15	0.63 (0.001)	0.79	(0.02)
Error	18974	7945	0.42 (0.42)		

The analysis shown in Table 1 accounted for 0.76 of the variance in the raw subjective ratings. An analysis that uses more "degrees of freedom" by including interactions of the variable Subject with other variables would account for approximately 0.84 of the variance.[18] We use the results in Table 1 in the discussions that follow.

Reliability

We have two independent estimates of the reliability of the subjective data. We also have at least two ways of looking at the reliability issue; we argue that one of them makes sense in the present context.

---

[14] "Type I" sum of squares is reported from the GLM procedure of SAS. SAS does not allow separate estimates of the nested variables Lab and Subjects, but we show the Lab main effect from a different run.

[15] For the "fixed effects" HRC and Scene we present the "Type III" mean square, which is corrected for the unbalanced design. For the "random effects," Lab and Subject, we present first the mean square calculated as if the effects were fixed (from GLM), and then, in parentheses, the mean square calculated based on a random effects model from SAS's VARCOMP procedure. VARCOMP, like GLM, is computationally intensive, so we were not able to calculate various interactions with the variable Subject.

[16] The F tests reported are the ones given in Crow (1994). Subject is tested against Error. The Tests for HRC*Lab and Scene*Lab use mean square estimates for HRC*Subject and Scene*Subject from the SAS GLM procedure applied to the individual teams' data; computing interactions with Subject in the full data set is too large a problem for GLM. Similarly, VARCOMP was run on individual teams' data to get estimates of HRC and Scene interactions with Subject for the tests of HRC*Lab and Scene*Lab for the random effects model -- in parentheses.

[17] One degree of freedom for Subject is subsumed in the Lab variable.

[18] We cannot actually calculate this model with either GLM or VARCOMP, but we were able to run both ANOVA and GLM on the GTE data alone and ANOVA on the combined GTE & NTIA data to get an idea of what we might have observed had we run GLM on the combined data.

1. Repeated trials. In each tape two trials were repeated, giving eight per subject or 632 for the 79 subjects. We can measure the amount of difference between the two occurrences of the same stimulus in a couple of ways:

- The intuitive, but misleading, way is simply to catalog the absolute differences between comparable ratings by the same subject. We get a datum for each *pair* of stimuli, or 316 data. Of these, 85 were 0.0, 133 were 1.0, 71 were 2.0, and 26 were more than 2.0.[19] The mean was 1.13. This value is about twice what the Test Plan aimed for and had been achieved in previous studies (e.g., Cermak et al., 1994; Voran and Wolf, 1992). We believe this error figure is inflated because of a *quantization* effect of the response scale. Whatever the subject's true uncertainty about their response, the scale forces a response that is an integer, and this intuitive measure of the error preserves the quantization.

- An indirect but more accurate measure of average error is via ANOVA of the repeat-trials data. The model is Rating = Order, Subject*HRC*Scene for each "team" separately (since each team viewed different stimuli for these trials). The RMS errors for the three teams were 0.51, 0.50, and 0.50 ("red," "green," and "orange," respectively). This error level beats the target error specified in the Test Plan because it achieves the specified raw RMS error (0.50), but with 79 rather than 90 subjects. And, this excellent level of error is achieved despite including nominally disqualified subjects.

  Note that on each trial, the ANOVA model was predicting an average rating based on the particular subject, HRC, scene, and order of presentation. This average was not an integer, and so the difference between predicted and the observed data also were not integers. These non-integers were about half the size of the integer values presented above. Of the two approaches to measuring error, this second one, based on averages, seems to us more reasonable. As we argue below, the basic unit of analysis should be an average rating across many subjects. A single subject's trial-by-trial behavior does not constitute a measuring instrument.[20]

2. ANOVA for all rating data. We have nearly 20,000 data points on which we can base an indirect estimate of error in the ratings. The estimate is actually of a combination of error and a non-error component that is likely to be very small, namely the Subject-by-HRC-by-Scene interaction. This composite is the *residual* after all the other systematic effects are accounted for. A practical problem in computing this indirect error estimate is that the SAS GLM procedure takes many hours of computing time for this model for the full data set. Therefore, we split the data by team before analyzing; this makes three much smaller problems, but also gives a balanced design for each of the sub-problems so that the more efficient ANOVA program can be used.[21] The model was

Rating = Subject, HRC, Scene, Subject * HRC, Subject * Scene, HRC * Scene.

---

[19] The sum is 315; one observation was missing. Recall that no subjects were excluded from this analysis, although the Test Plan calls for any subjects who were off in their repeat ratings by more than 2.0 to be dropped from the study.

[20] The people who build video systems may rely on their individual judgment of quality during the design process; that does not mean that individual judgment is a good measurement tool.

[21] GLM and ANOVA gave the same results for this problem on a team by team basis. GLM ran in 2 hours, two minutes; ANOVA ran in 22 seconds.

(SAS's approach to analysis of variance does not allow effects of nested variables to be estimated simultaneously, so the Lab variable was left out of this analysis. The Subject variable includes all the information that the Lab variable contains anyway.)

Results were RMS errors of 0.56, 0.58, and 0.60 for the red, green, and orange teams, respectively. This level of error is consistent with the estimates above from the repeated trials: Since the present estimates include the three-way interaction (however small it may be), they must be at least as large as the estimates based on the repeated trials.

Inter-lab comparability

Table 1 shows that, for either a fixed effects model or a random effects model of the Laboratory variable, Laboratory does not have a significant effect, either alone as a main effect, or as an interaction with HRC or Scene. That is, any difference between the two laboratories is the same as would be expected if the same laboratory ran the same experiment twice sampling about 40 subjects each time.

An apparent "team" effect also is caused by sampling -- both subjects and HRC's in this case. However, at least two teams judged the same HRC-scene combinations on 5900 occasions (out of 19725). Analyzing those data, differences between teams were just what one would expect on the basis of sampling different subjects (using the same kind of mixed ANOVA model as above); that is, there was no statistical effect of Team.

Errors

One time on each tape an original version of a scene appeared in place of an HRC as a quality check (the "null check"). The Test Plan requires that a subject be disqualified if they rated any null less than 4 on the five-point scale. Ten errors were recorded on nulls. Even these "errors" turn out to be largely attributable to variables in the experiment. Five of the ten were for the circuit scene which has a mean rating across all HRC's of 2.23. Two of the remaining five were for the famous flower garden, a "codec-buster" with a mean rating of 2.60.

Bounds on objective-subjective correlation and effects of averaging

The amount of error in the subjective data limits how well any set of objective measures can correlate with the subjective data (see Cermak, 1994). We can think of "error" in at least two ways. The most intuitive sort of error is of the repeated-trials type considered in the section on reliability. For subjective data with only this kind of error, in the amounts indicated by the repeat-trial data, 0.86 of the variance in the subjective data could possibly be predicted by an optimal objective measure (a correlation of 0.93).

However, this bound on goodness of fit depends on how the data are aggregated. The connection between goodness of fit and aggregation involves the variable Subject. A goodness of fit of 0.86 depends on being able to remove all variance that depends on differences between subjects (the error measurement we used above is for differences that occur within individual subjects but across trials). Removing variance that depends on between-subject differences is something we can do two ways, (a) analytically, and (b) by aggregating. We have been presenting analytic results. These results are equivalent to averaging data across all subjects. That is, a bound on goodness of fit of 0.86 is what one would expect if the data for each of the 625 HRC-scene combinations were first averaged across a very large number of subjects.[22]

[22] If we actually do average the ratings within each of the 625 stimuli, then do an ANOVA, the main effects for HRC and Scene account for 89.2% of the variance in the averaged ratings.

If one were using the raw individual ratings in a regression with objective measures of the stimuli, then the bound on goodness of fit would be affected by another source of "error." This source is between-subject differences. Objective measures of the sort listed in the section on Objective Measures from NTIA apply to HRC's and to scenes, not to subjects. All variance attributable to subjects is variance that is beyond the reach of these variables, no matter how correct they are physically. In Table 1, the amount of variance[23] accounted for by HRC, Scene, and HRC*Scene is 0.686. In disaggregate raw rating data, all the subject effects are present, so in fitting objective measures to the raw ratings, 0.686 would be the upper bound on goodness of fit (correlation = 0.828).

Relative importance of HRC, Scene and other variables

Table 1 gives measures of the relative importance of variables in influencing subjective judgments.

- The most important variable is the HRC -- which makes sense since the collection of HRC's spanned nearly three orders of magnitude in bit rate.
- Next most important was the particular scene. In retrospect, this also makes sense since the scenes were handpicked to match the range in capability of the various HRC's.
- Next most important was the particular subject. That is, individual subjects have characteristic mean ratings, independent of the particular HRC or scene. These individual differences are entirely irrelevant to measuring video quality. They are at best a nuisance. The statistician Crow of NTIA has advocated removing each subject's mean from the data, as have the present authors, but the Test Plan calls for the unprocessed rating data (which preserves these irrelevant individual differences) to be reported. We examine the issue of individual subjects further in the section *Individual differences*.
- Next most important was the HRC-Scene interaction. That is, either the *ordering* or the *spacing* of HRC's in terms of judged quality varied from scene to scene. Viewed the other way 'round: The *ordering* or *spacing* of scenes in terms of judged quality varied from HRC to HRC. This effect is not large compared to the main effects of HRC and Scene (see "mean squares" in Table 1).[24] In the section below on HRC-Scene interaction, we raise the possibility that much of this effect may be an artifact of the five-point rating scale used.
- Lab was not important. The committee can breathe a sigh of relief.

HRC and scene ratings

Mean ratings for the individual HRC's and scenes are given below. As can be seen in Table 2, for the HRC's Bit Rate is closely related to Rating (also see Fig. 9, p. 27), and for the scenes, scene Type is related to Rating. These means represent what in Table 1 are the "main effects" for HRC and Scene. One hears the comment that certain HRC's do not handle particular scenes very well, and the way this is often said leads one to think that HRC's are very idiosyncratic. The size of the main effects for HRC and Scene indicate, however, that certain HRC's always do well on all scenes, certain scenes are bad for all HRC's, and there is actually very little quirkiness that depends on a particular combination of HRC and scene. Also, there is no evidence for a "Susie effect," namely a distortion in ratings for HRC's because the scene is so wonderful it always gets the top rating (Susie is scene j; also see Figs. 4-6, p. 17-21).

---

[23] Neglecting the sum of squares due to Lab, which comes from a different analysis.

[24] The sum of squares for the interaction is substantial. In the AMMI analysis of the interaction which follows, the number of degrees of freedom used to account for the interaction is much less than the 576 in Table 1. Hence the size of the interaction, as measured by mean square, is effectively larger for the AMMI analysis.

Table 2. Mean ratings for each HRC and each scene; data from 79 subjects from the GTE and NTIA labs (also see Fig. 9, p. 28).

HRC	Bit rate, kbps	Rating	Scene	Type	Rating
1	Null	4.90	f	A	3.61
2	VHS	4.42	j	A	3.22
3	45,000	4.80	k	A	3.63
4	128	2.37	l	A	3.54
5	336	2.89	a	B	3.62
6	112	2.16	b	B	3.15
7	384	3.23	e	B	2.76
8	768	3.49	m	B	2.51
9	768	3.21	n	B	2.45
10	1536	3.74	w	B	3.03
11	128	1.76	d	C	2.99
12	128	1.81	g	C	3.30
13	168	1.85	o	C	2.78
14	384	1.96	p	C	3.10
15	112	1.74	q	C	2.26
16	128	1.81	r	C	2.71
17	128	2.22	c	D	2.62
18	168	2.47	s	D	2.23
19	256	2.74	t	D	2.54
20	384	3.26	u	D	3.50
21	384	2.70	v	D	2.68
22	768	3.71	x	D	3.04
23	768	3.25	h	E	2.60
24	1536	3.92	i	E	2.02
25	1536	3.70	y	E	2.61

HRC-scene interaction (AMMI analysis)

In a traditional analysis of the interaction between HRC's and scenes, we would note how much variance it accounted for in an ANOVA, graph it, and try to interpret it. This approach will work for simple 2 by 2 factorial designs (two variables each with two levels), but it is often inadequate when the variables in an interaction have many levels. As an example, consider the graph of the HRC by scene interaction in Figure 2. Here the mean rating for each scene is plotted for each of the 25 HRC's, arranged alphabetically. It's clear that there is structure to the data, but not at all apparent what that structure might be.

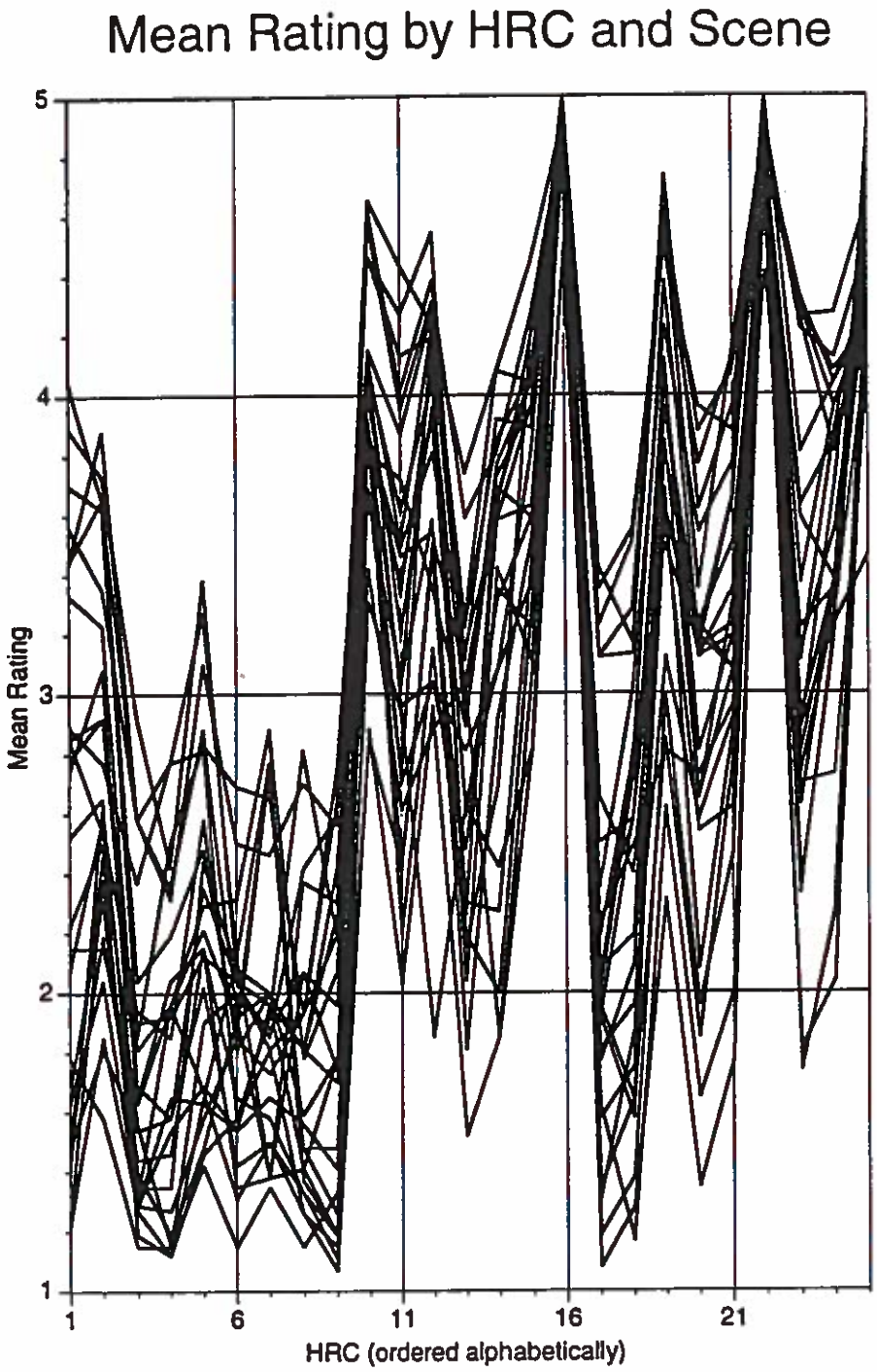


Figure 2: HRC by Scene Interaction (HRCs ordered alphabetically)

The situation improves somewhat if we organize HRC's by mean rating, as shown in Figure 3.

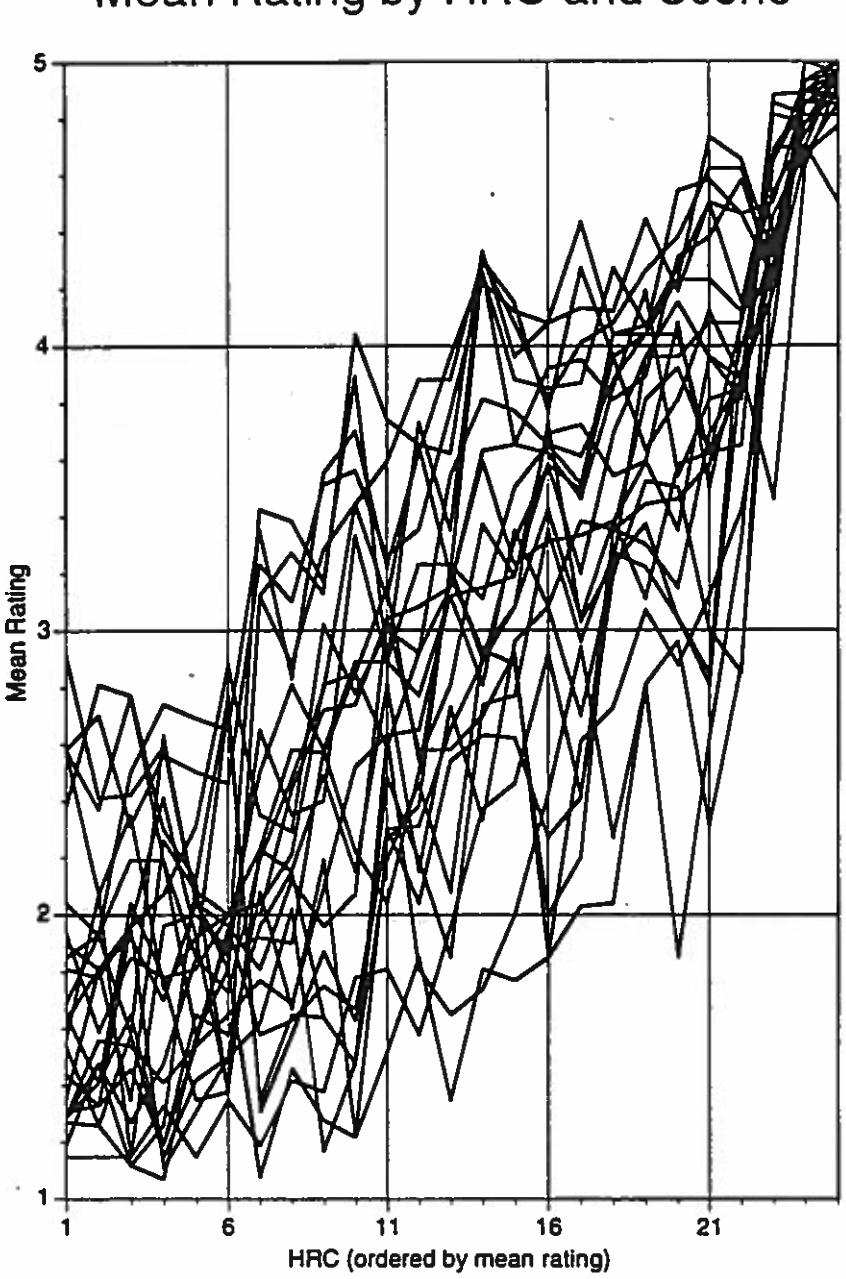


Figure 3: HRC by Scene Interaction (HRCs ordered by mean rating)

Here one prominent pattern in the data reveals itself: HRC's that do very well or very poorly, are more uniform in their ratings across scenes (also see Cermak et al., 1994). That is, the range of ratings across scenes decreases for HRC's at either end (this is much more apparent for HRC's at the high end). While one aspect of the structure of the data is brought out in Figure 3, other patterns are obscured.[25]

[25]Actually, one other structural feature is apparent in Figure 3. Scenes seem to divide into two groups within low-end HRC's: a large group of scenes that are rated about the same, and a small group of four or five scenes which are rated higher than the rest. If we ignore this smaller group, the data in the figure appears much more symmetrical, with low-end HRC's showing almost as narrow a range of ratings as high-end coders.

A different approach to analyzing two-way interactions was suggested in 1923 by Fisher and Mackenzie, who advocated modeling main effects with an additive Analysis of Variance and two-way interactions with a multiplicative Principal Components Analysis (PCA). Since PCA is extremely laborious to carry out without a computer, this approach lay dormant until the early 1950's when it was revived and extended by Williams (1952) and Pike and Silverberg (1952). Subsequent developments by Bradu and Gabriel (1974, 1978) and many others are discussed in Gauch (1992).

In AMMI (Additive Main Effects and Multiplicative Interactions) analysis, main effects in a two-way factorial experiment are modeled as usual with an additive Analysis of Variance. However, the residual from this analysis, which includes the two-way interaction, is modeled with a multiplicative PCA. PCA gives scores for each level of each variable usually scaled such that multiplying the scores together gives the predicted value of the residual. Addition of the predicted residual to the predicted main effects gives the prediction of the original scores. Thus the complete AMMI model is:

$$Y_{ijr} = \mu + \alpha_i + \beta_j + \Sigma_n(\lambda^{0.5}{}_n\gamma_{in})(\lambda^{0.5}{}_n\delta_{jn}) + \rho_{ij} + \varepsilon_{ijr}$$

where Y_{ijr} is the value of the dependent variable for row i, column j, and replication r, μ is the grand mean, α_i is the main effect for row i, β_j is the main effect for column j, λ_n is the singular value for PCA axis n, γ_{in} is the PCA score for row i on axis n, δ_{jn} is the PCA score for column j on axis n, ρ_{ij} is the residual for row i and column j if not all PCA axes are used and ε_{ijr} is the error.

AMMI is especially effective when used with a biplot (Bradu and Gabriel, 1978). A biplot combines two scatterplots in the same graph. In AMMI a biplot is used to plot main effects against PCA scores for both variables -- Scene and HRC in this case. This display is remarkably effective in revealing patterns in the data brought out by AMMI.

Figure 4 shows a biplot of the first PCA axis for the HRC x Scene interaction.

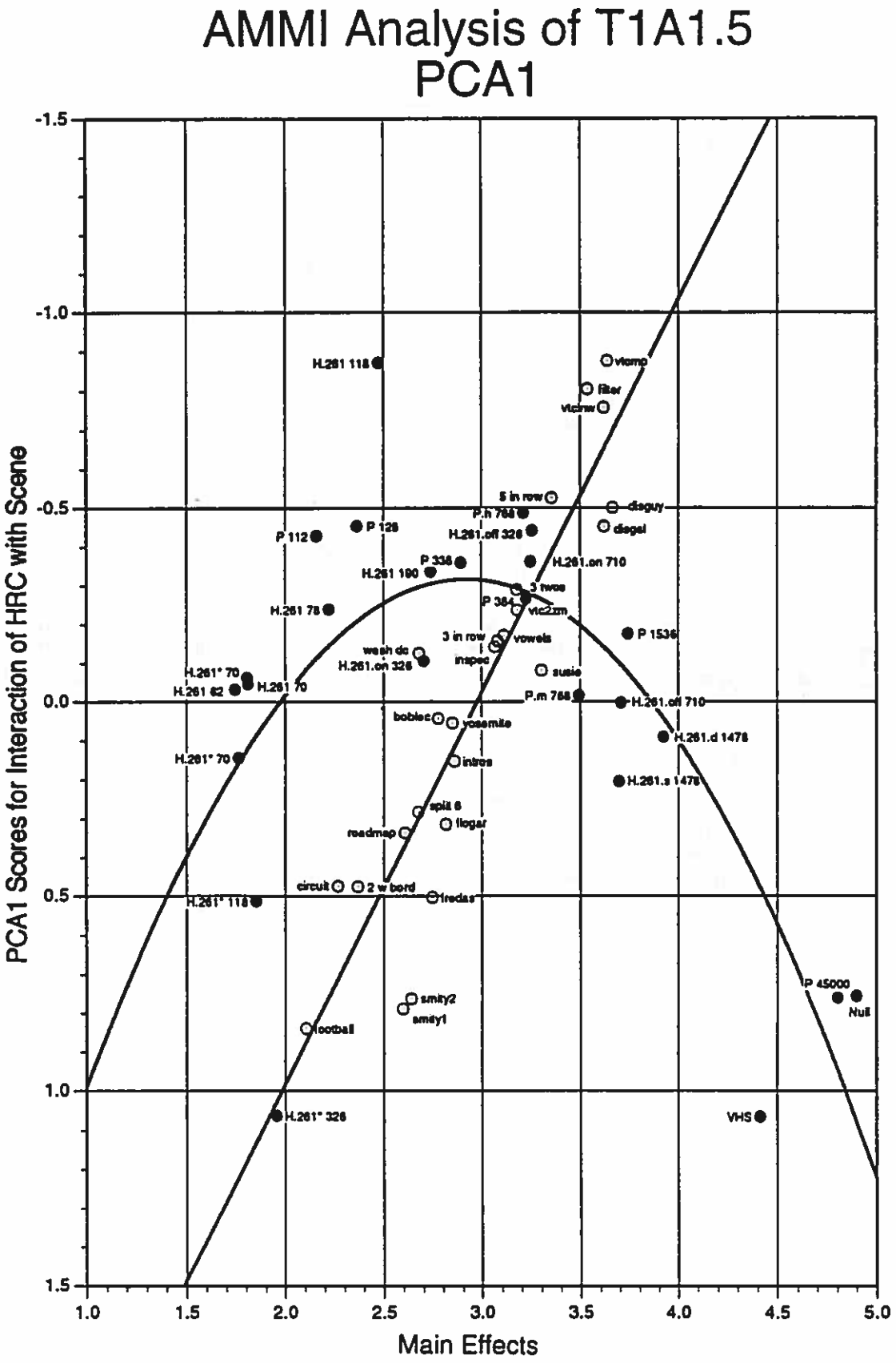


Figure 4: First Principal Component in AMMI Analysis

Main effects are represented on the abscissa and PCA scores on the ordinate. Points are labelled with the names of scenes or HRC's. HRC names generally indicate the type of algorithm used and the bit rate. For example, H.261 710 is an HRC from the H.261 family running at 710 kilobits per second. H.261 HRC's using QCIF are marked with an asterisk.

Two rules are needed to interpret the graph. First, points of the same type that are close together produce similar interactions. Thus the HRC's *P 45000* and *Null*, since they are close together, interact with scenes in similar ways. Likewise, scenes *smity1* and *smity2*, since they are close together, interact with HRC's in similar ways.

The second interpretive principle is that interactions for points of different types are given by multiplying their scores. If both scores are positive or both negative, their product will be positive, which when added to the main effect gives a higher rating than expected. However, if one is positive and the other negative, their product will be negative, which when added to the main effect gives a lower rating than expected.

Applying these interpretive rules in Figure 4, we see that good HRC's like *P 45000* give higher than expected scores for hard scenes like *football* and lower than expected scores for easy scenes like *vtcmp*. Poor HRC's like *H.261* 118* behave in the same way. This pattern has been emphasized in the figure by fitting curves to the two sets of PCA scores. Scenes are represented by a straight line, indicating that their participation in (the first principal component of) the interaction is a linear function of their main effects. HRC's, on the other hand, are fit with a quadratic curve to highlight the fact that good and bad HRC's behave similarly, and differently than intermediate HRC's.

What does this pattern mean? PCA, by itself, doesn't provide an interpretation of the pattern, but, often, by making the pattern apparent, an interpretation suggests itself. In this case, the first principal component of the interaction is clearly related to the pattern that showed up in Figure 3 -- HRC's with very high or very low means show a restricted range of ratings compared with HRC's in the middle. An obvious interpretation is that subjects have difficulty calibrating their use of the rating scale so that they run out of room at the top and the bottom of the scale. Because of these ceiling and floor effects, the ratings that subjects give to the HRC's that do particularly well or poorly do not reflect well their performance on different scenes, particularly compared with mid-range HRC's. (An identical pattern showed up as the first principal component of the AMMI analysis of the MPEG1 study reported in Cermak et al., 1994, which used the same rating scale; so there appears to be a general problem with this type of scale.)

A biplot for the second principal component is shown in Figure 5.

AMMI Analysis of T1A1.5
PCA2

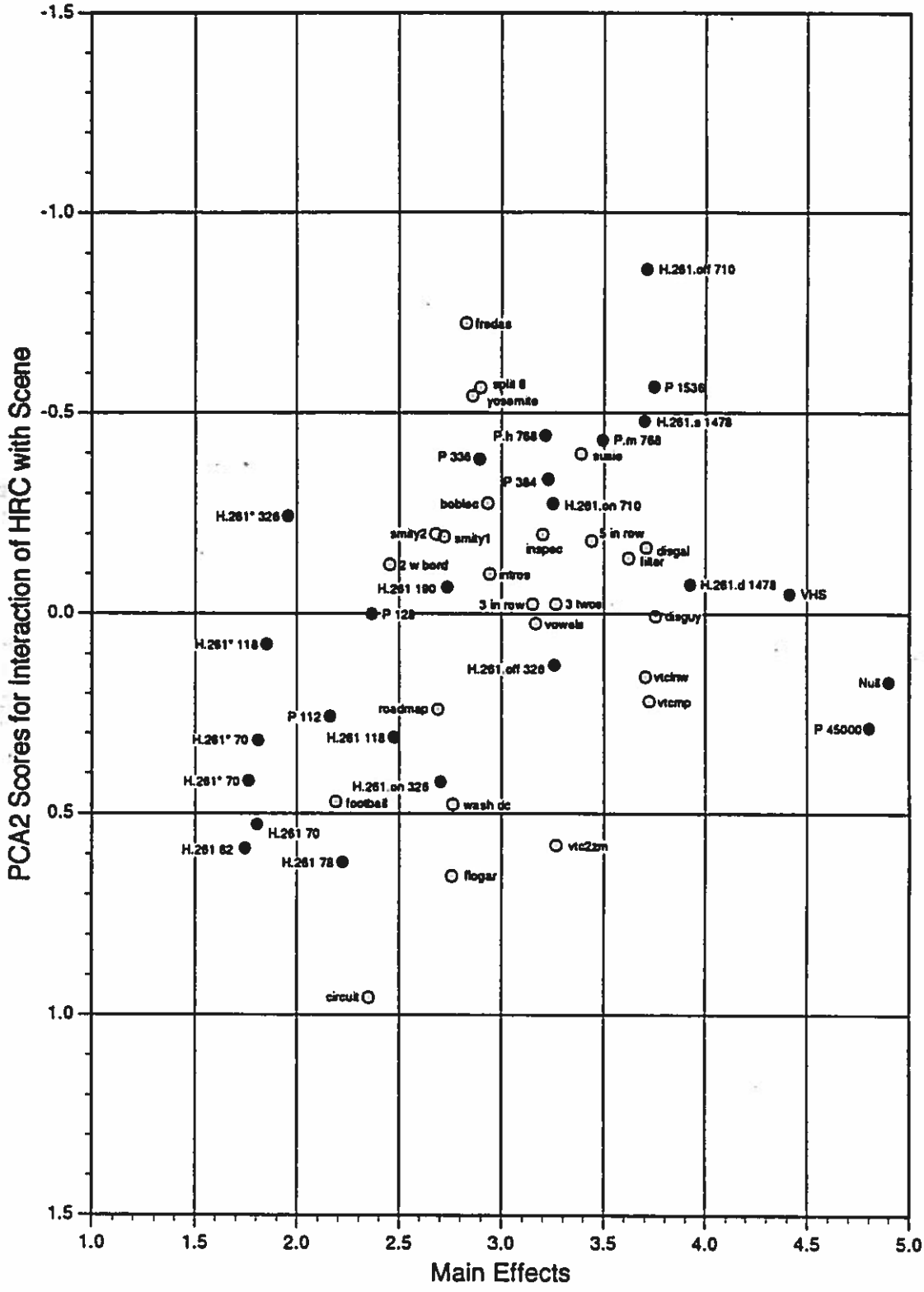


Figure 5: Second Principal Component in AMMI Analysis

Here there are no geometric configurations on which to base our interpretation so we must look in detail at the characteristics of scenes or HRC's that spread them out on the vertical (principal component) dimension in the plot. Close inspection of the scenes suggests an interpretation, albeit one that is far from secure. It appears that the dimension relates to movement in the scene. At one end of the dimension are scenes in which there is global movement created by pans and zooms: *circuit* zooms into a printed circuit, *flogar* pans across a flower garden, *football* follows a running football player and *wash dc* pans a finger tracing a route across a map. At the other end of the dimension are scenes containing local movement -- that is, the subject of the shot moves but the background doesn't: *fredas* shows Fred Astaire dancing, *split 6* is a composite image of six talking heads, *susie* shows a woman using a telephone, *boblec* shows a person lecturing at a blackboard, and so on.

If PCA2 represents movement of different types, then coders like H.261 710 and the higher bit rate coders like P 1536 and H.261 1478 do relatively better on local movement than global movement, while lower bit rate coders like H.261 62, 70, and 78 do relatively better on global movement than local. Very high bit rate HRCs like P 45000 and VHS, as well as some lower rate coders like P 128, are more balanced in their treatment of different types of movement.

The biplot for the third principal component is shown in Figure 6. Once again there are no clear geometric structures. However, there is a relatively clear interpretation of the PCA dimension: it represents the amount of perceptually important small detail in the scene. One indication of this is that all the scenes with text in them are found at one end of the dimension. These are highlighted in Figure 6 by underlining the name of the scene. *yosemite*, *roadmap*, vtc2zm and *wash dc* are all maps in which detail, including text, is perceptually prominent. *Circuit* and *filter*, both technical drawings, are similar. *boblec*, *vowels*, and *2 w bord* show text being written on a board. At the other end of the dimension are scenes in which detail is less important, typically people. These include *smity2*, *smity1*, *disgal*, *susie*, *3 in row*, and so on.

An odd consequence of this interpretation is that the H. 261 QCIF HRC's do *worse* on detail as their bit rate is increased. For example, a high detail scene like *yosemite* is coded relatively better by H.261* 70 than by H.261* 326, even though the latter has a higher bit rate. Contrast this with the opposite behavior for movement. These opposing effects may explain why QCIF HRC's don't get better overall as their bit rate increases.

While the variance associated with the HRC-Scene interaction is not large compared to the main effects, AMMI analysis is able to extract significant patterns related both to the rating scale and to idiosyncrasies of the coding algorithms. In particular, AMMI reveals in a particularly clear way the ceiling and floor effects often found with rating scales. As well, it shows that even after the overall effects of bit rate are removed, HRC's differ in how well they represent information distributed in time (movement) and space (detail).

AMMI Analysis of T1A1.5
PCA3

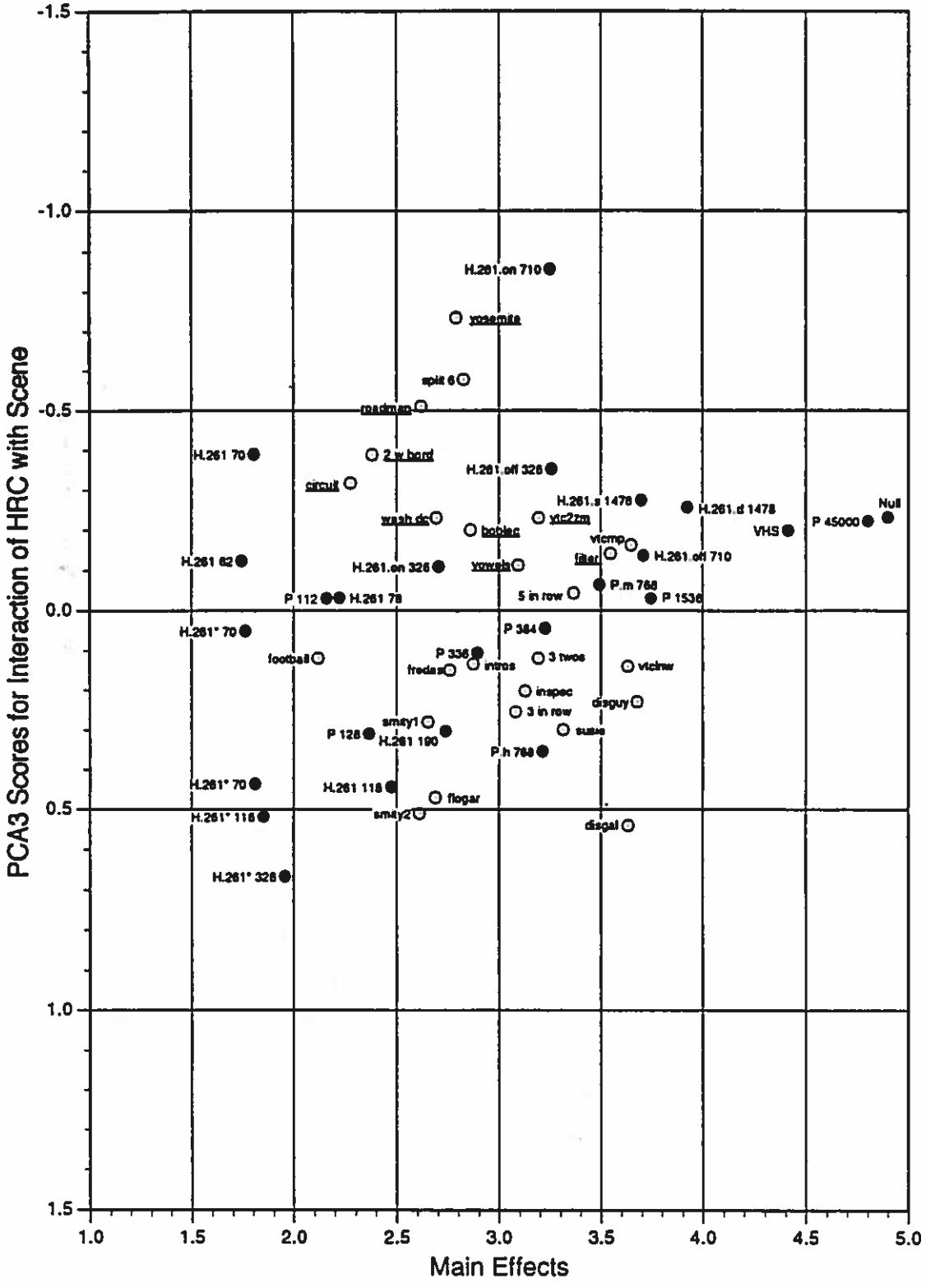


Figure 6: Third Principal Component in AMMI Analysis

Individual differences

Individual subjects use rating scales differently. Some subjects center their ratings about a fulcrum point of 2.60 on a 1 to 5 scale; others center their ratings about a fulcrum of 3.15, and so on. Or, one could say that some subjects have higher standards for video quality than others; it comes to the same thing for the present data set. One often hears complaints about this lack of standardization among subjects. Although the complaints are seldom explicit, they probably intend that differences among subjects are meaningful, important, and hard to deal with. We address these issues in turn, in the context of the present experiment.

- Differences are meaningful. That is, the differences among subjects are predictable from some sort of objective measure such as age, education, income, gender, and so on. In principle, if we knew enough about each subject we might be able to predict or account for subject differences in how they rate video quality. But, we certainly do not know enough about the subjects in this study. We substituted the variables Gender, Video (experience with video teleconferencing), Age, Acuity, Color, and interactions of HRC with each of these for the variable Subject in an analysis of the ratings. None of the main effects were significant when tested against the mean square for Subject. Interactions of HRC with Age and Video were significant, but accounted for 0.006 of the variance in the data. From a practical point of view, *subject differences are not meaningful in predicting ratings of video quality.*

  In particular, note that the results of the vision tests did not significantly predict either a main effect or an interaction with HRC. This point is important from a practical point of view: Quite a lot of money or time was spent getting extra subjects because of those who had failed either the acuity test or the color vision test. That money and time were apparently misspent.

- Differences are important. Even when there is a statistical effect, such as an interaction between experience in video teleconferencing and rating of HRC's, we have found such effects to be of very minor practical importance. For example, in Figure 7 below we see that the way the inexperienced and experienced subject groups responded to HRC's are very similar. Most of the disagreement seems concentrated in the two lowest-rated HRC's (low bit rate QCIF systems), and even then the disagreement is quite small. Of course, when there is no statistical effect, as in the case of Gender or Vision, the disagreement between groups is even smaller.

- Differences between subjects are hard to deal with. Two simple solutions to between-subject differences are: (1) Remove each subject's mean from their data. (2) Collect data from more subjects and average across the subjects. Removing the mean rating from each subject's data is based on the idea that a subject's mean score carries information about the subject rather than about individual HRC's and scenes, and that this information is of no practical use in video quality testing. A corollary is that the rating data we collect are not meaningful in any absolute sense -- just as temperature data are not meaningful in an absolute sense, since any temperature scale can be transformed into another that is also meaningful by adding constants and multiplying by constants.

  Averaging data across subjects neutralizes the effects of individual differences in mean ratings. Averaging also accomplishes something that subtracting means does not do: Averaging across subjects neutralizes *interactions* between subjects

and variables of interest, such as HRC and Scene. Averaging becomes more and more effective as the number of subjects grows, say, from 10 to 30 to 90.

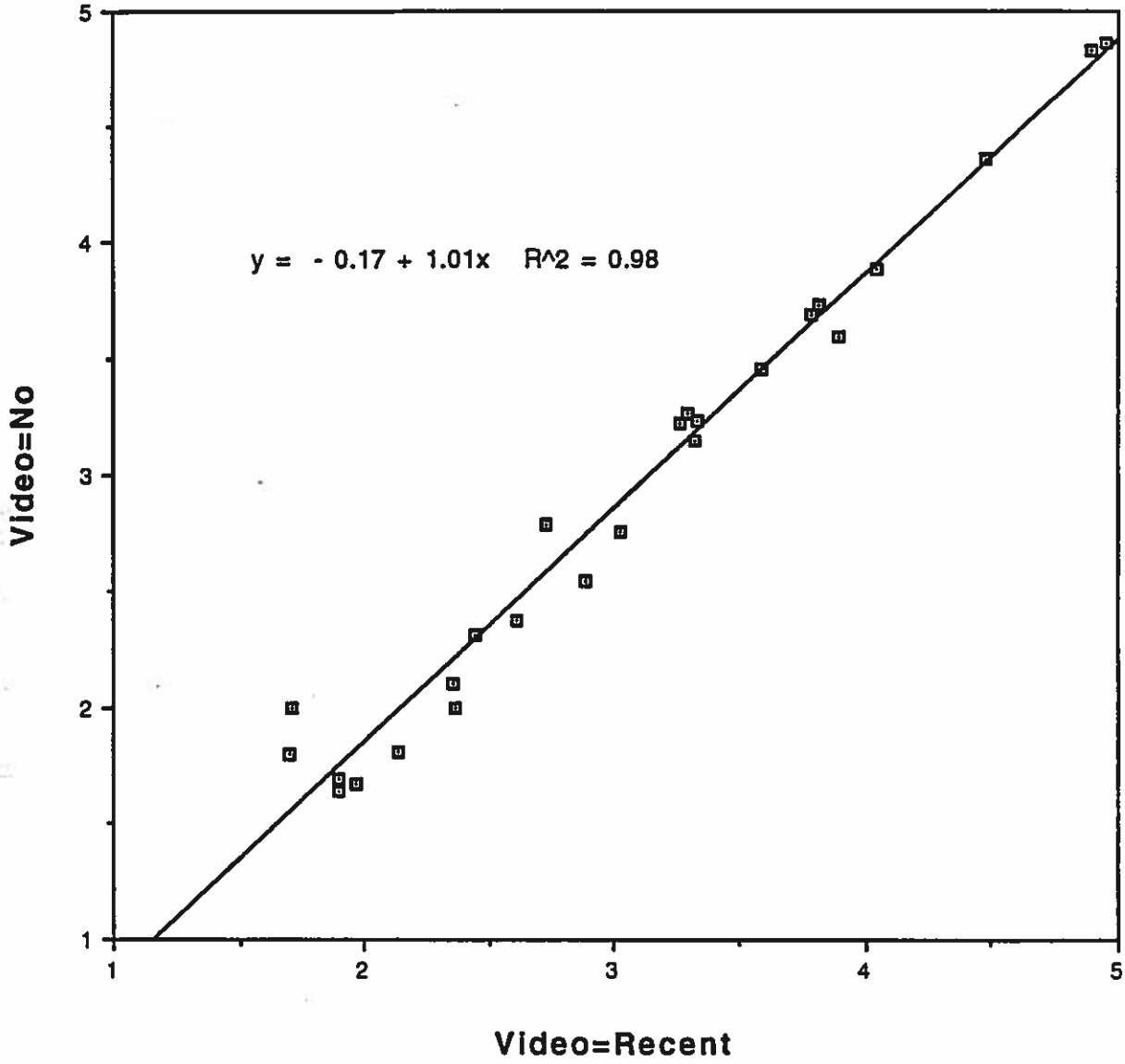


Figure 7. Average ratings of the 25 HRC's for subjects with recent experience in video teleconferencing compared to subjects with no such experience.

Objective measures alone

NTIA supplied two sets of objective measures, a laboratory version and a real-time version. We analyzed these sets of measures separately. For each set of measures we can ask (a) how many independent pieces of information there are, and (b) how related these measures are to the main experimental variables, HRC and Scene. In the section following this one we examine the issue of how well the objective measures predict the subjective ratings. Note that the laboratory and real-time data sets are of different sizes; the laboratory data contains all 13 measures, but the real-time data set contains only the first 11 measures.

Factor analysis

A Principal Components analysis of the laboratory measures showed two factors with eigenvalues greater than 1.0, the usual rule of thumb for selecting factors. We also examined a third factor; together the three factors accounted for 88.6% of the variance in the 13 measures.

- The first factor accounted for 64.5% of the variance in all the measures. As we shall see, this factor has an obvious interpretation as bit rate. The only measure that did not correlate highly with the first factor was P13, the Fourier transform measure of added information. Although all measures correlated highly with the first factor, the temporal domain measures P2-P6 correlated highest; of these, P6 had the highest correlation, 0.95.

- The second factor accounted for 16.8% of the variance in the objective measures. This factor was correlated most highly with the spatial domain measures P7-P9, with P9 having the highest correlation.[26]

- The third factor looked reasonable, even though the eigenvalue was 0.92 and the variance accounted for was only 7.3%. The third factor was P13, a measure of added spatial frequency information: P13 correlated 0.86 with this factor, no other measure correlated above 0.29, and P13 correlated with nothing else, even bit rate.

Results from the real-time measures were similar: The first factor accounted for 73.4% of the variance in the 11 measures and correlated most highly with P2-P6. The second factor accounted for 16.7% of the variance, and correlated most highly with P7-P9. There was no third factor with an eigenvalue even approaching 1.0; there also was no measure P13, the one that formed the third factor in the lab data set.

ANOVA

Each of the objective measures is likely to be sensitive both to particular HRC's and to scenes. We can gauge the relative sensitivity of the measures to HRC and Scene by means of ANOVA in which the independent variables are HRC and Scene, and the dependent variable is a given objective measure. Recall that the "main effect" of HRC alone is the variation in a given measure across HRC's, averaged over the scenes -- essentially a scene-independent effect due to the HRC. Similarly, the Scene main effect is independent of HRC. (It is not clear to us whether one should prefer a measure that is more sensitive to HRC's or to scenes.) Table 3 shows that the size of the HRC-Scene interaction is not a constant property across measures -- some measures are more sensitive to the peculiarities of specific HRC-scene combinations that cannot be predicted from the HRC and Scene main effects.[27] We are inclined to view this as a disadvantage in a measure -- sensitivity taken to the extreme of instability.

In Table 3, note the measures for which agreement between laboratory and real-time versions is poor: P3, P4, P10, and P11. Also note the unusual behavior of P13, a greater sensitivity to the particular scene than to the HRC.

---

[26] P1 correlated fairly highly with both of the first two factors, but did not correlate above 0.7 with either.

[27] What we are calling the HRC-Scene interaction also contains any measurement error; Pinson and Jones (1994) show this error to be less than 1% of the variance.

Table 3. Percent of variance in a measure accounted for by differences in HRC and Scene.

	Laboratory Measures			Real-Time Measures		
Measure	HRC	Scene	HRC*Scene	HRC	Scene	HRC*Scene
P1	36.5	34.2	29.3	32.0	36.8	31.2
P2	48.3	38.8	12.8	45.2	34.4	20.4
P3	53.8	32.8	13.4	41.2	41.5	17.3
P4	52.5	30.0	17.5	46.4	35.6	18.0
P5	66.2	20.2	13.6	62.4	24.0	13.6
P6	70.1	20.4	9.5	69.3	23.0	7.7
P7	54.5	31.0	14.5	54.7	30.7	14.6
P8	66.3	20.9	12.8	66.3	20.9	12.8
P9	68.6	18.5	12.9	68.7	18.4	12.9
P10	41.2	8.9	49.9	51.8	12.2	36.0
P11	62.4	22.9	14.7	49.1	33.6	17.3
P12	68.0	19.7	12.3	--	--	--
P13	12.6	41.9	45.5	--	--	--

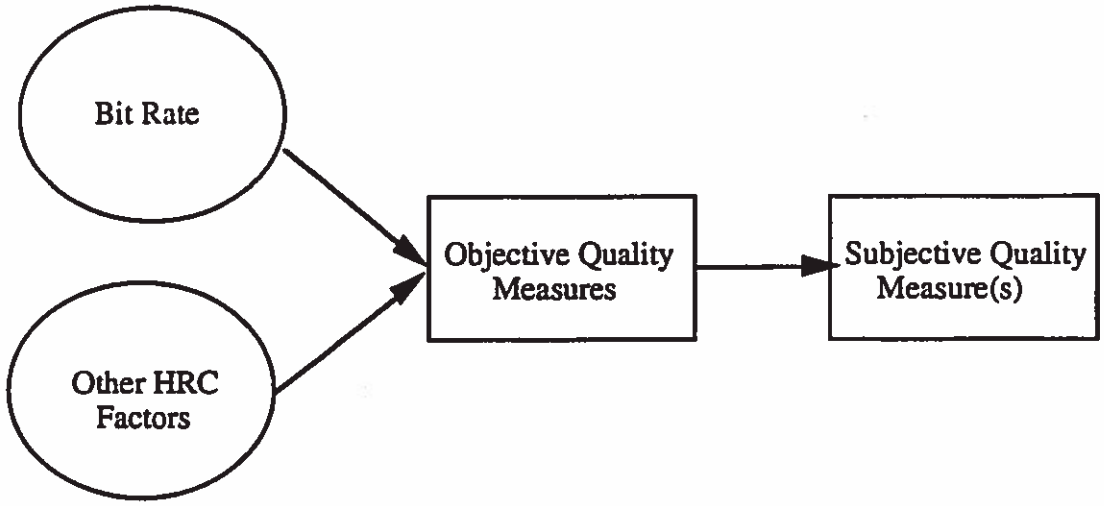
'Correlation' of objective and subjective measures

The whole point of the T1A1.5 video quality measurement program is to determine which objective measures best predict the subjective ratings. The obvious statistical tool is multiple linear regression ('regression' for short). We know from the factor analysis that we should use no more than three objective measures in a regression because there are no more than three independent pieces of information in the objective data set. We also know that the factors divide nicely into temporal domain measures, spatial domain measures, and Fourier measures. Therefore, we have decent *a priori* reasons for using a measure from each factor in the regression analysis. We will start with the measures that best exemplify their respective factors, P6, P9, and P13. But, first we take a small detour and examine the logically prior variable, Bit Rate.

Log(kbps)

Figure 8 diagrams two causal models of the data in the present experiment. Such diagrams often help clarify discussion. In both models bit rate is related to both the objective and subjective quality measures. In Model I, the objective video quality measures are logically prior to the subjective measures. Bit rate drives objective quality directly and subjective quality indirectly. In Model II, bit rate drives general video quality, which is *measured by* objective and subjective measures. In Model II there is no causal relationship between objective and subjective measures; they are related by virtue of their common relationship to Video Quality. The dashed arrow in Model II indicates that we model the objective and subjective measures *as if* there were a causal relationship between the two.

I. Objective Quality Drives Subjective Quality



II. Quality Drives Objective and Subjective Measures

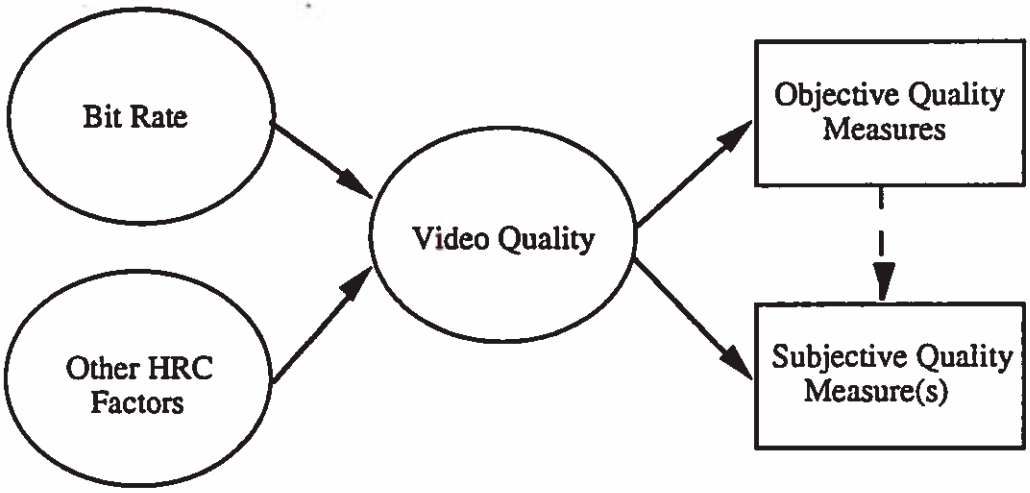


Figure 8. Two causal models of video quality.

(One cannot distinguish between Models I and II very well on the basis of this study's data. Model I predicts that the correlation between Bit Rate and the objective measures should be greater than the correlation between Bit Rate and Rating. In fact, the correlation of Bit Rate with Rating should be exactly the product of the correlation between Bit Rate and the objective measures and the correlation between the objective measures and Rating. Model II predicts that (a) the correlation between Bit Rate and objective measures could be either larger or smaller than (b) the correlation of Bit Rate and Rating. Therefore, only in one situation would the data distinguish between the two models: If the correlation between Bit Rate and Rating were greater than the correlation between Bit Rate and the objective measures, then Model I could be rejected in favor of Model II. In fact, the correlation between Bit Rate and Rating was 0.90, while the highest correlation between Bit Rate and a single objective measure was -0.87, for P6. Although this is not a strong test, it does make Model I look less attractive.)

The linear correlation of Rating with Bit Rate was 0.90. However, Figure 9 shows that more than a simple linear relationship is going on in the data. First, there is a definite curve in the data as the upper end of the rating scale is approached. Secondly, there are really two subsets of data points that exhibit different relationships between Rating and Bit Rate.[28] Within each of the subsets of points, the statistical relationship between Rating and Bit Rate is very large (the correlation is the square root of the R^2's reported in the figure, or 0.98 and 0.95). The lower group of points (one of which is hidden because it overlaps another) shows a smaller increase in subjective quality as bit rate increases. As Appendix A shows, these are mainly QCIF systems, but two CIF systems are also included. (The two leftmost points on the lower curve are the CIF systems; an alternative possibility for curve-fitting would be to leave the CIF points with the upper curve and isolate just the QCIF points on the lower curve.)

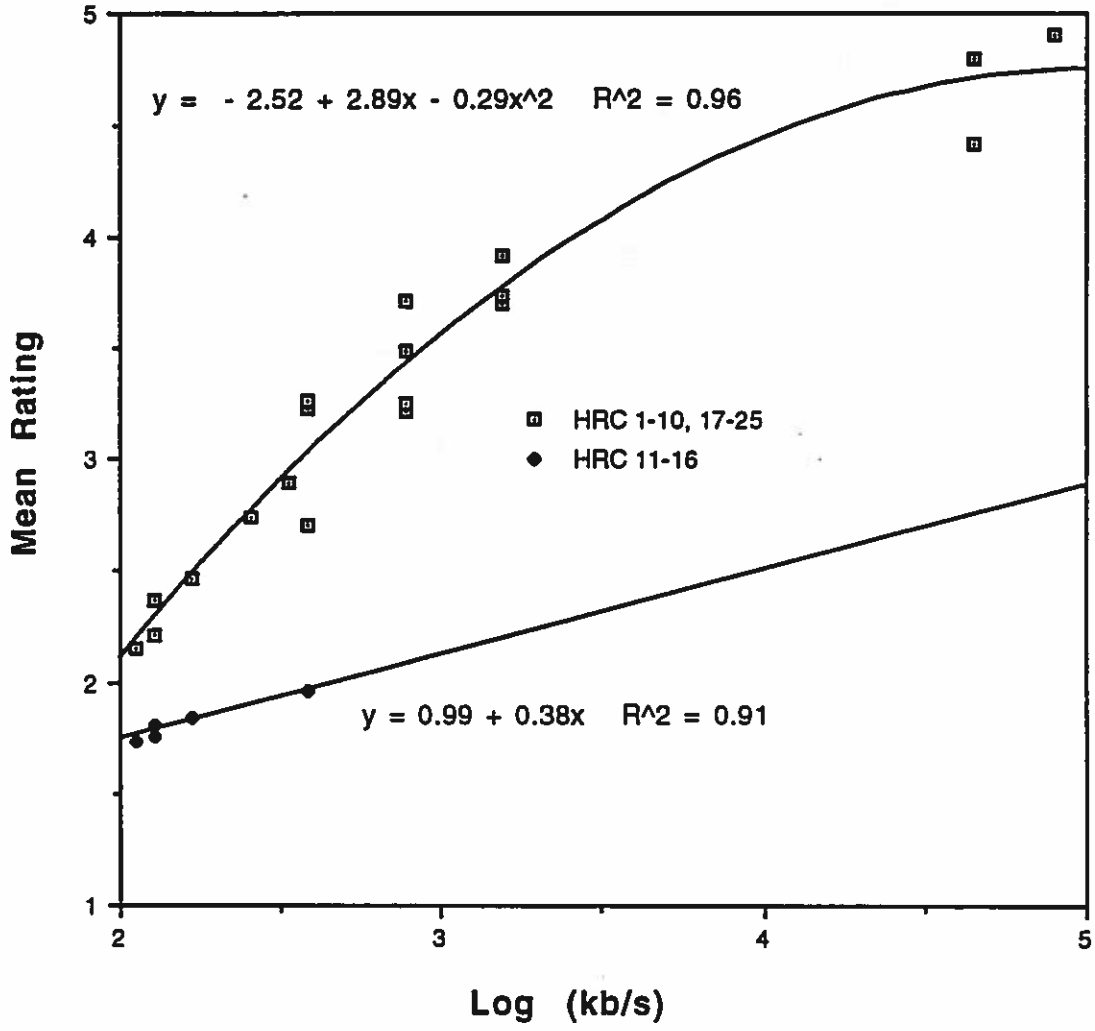


Figure 9. Mean ratings of 25 HRC's plotted against nominal bit rate.

[28] The equations given are strictly curve-fitting; they have no theoretical content and do not necessarily apply outside the range of the present data.

We observe:

- Bit Rate accounts for quite a lot of the variance in the subjective data; for a given bit rate on one of the two curves, the range among the various HRC's is limited to about half a rating point.

- The broad range of HRC's originally chosen in the Test Plan has an important influence on the results. Had the HRC's been limited to bit rates of, say, 384 kbps to 1.5 Mbps (2.6 to 3.2 on the x-axis of Fig. 9), the effect of bit rate might not have appeared so pronounced.

- The fact that the subjective data are fine enough to unmask the special nature of HRC's 11-16 should reassure us of the quality of subjective data.

Regression using raw data

The base case model used the 19725 raw ratings and the top three candidate objective measures P6, P9, and P13 for each of the 625 HRC-scene pairs. Thus, for each HRC-scene pair we have one objective prediction and more than 30 subjective responses to be predicted. Those subjective responses carry information about the individual subjects and their idiosyncratic responses to the particular HRC-scene combinations, which, of course, cannot be predicted by our objective measures. The results:

- P6, P9, and P13 were statistically reliable predictors of the raw ratings.

- The standardized parameter estimates for P6, P9, and P13 were -0.50, -0.26, and -0.05, respectively. These parameters are a measure of the relative importance of the objective measures in accounting for the subjective ratings.[29]

- This simple 3-parameter model accounted for 48.25% of the variance (correlation of predicted to observed of 0.695); RMS error was 0.93 rating unit.

- The corresponding ANOVA model that did not correct for subject effects accounted for 68.6% of the variance; this is the best that any model can possibly do that uses only information about HRC and Scene to predict raw ratings. A minimalist model using only the bit rate measure (which neglects scene differences) accounted for 40.0% of the variance in the ratings.

Is this a good result? We cannot say because no objective measure should be asked to predict individual subjective ratings. The main reason for analyzing the raw ratings is that the regression estimates are likely to be more stable (Tukey, 1994).

Results using averaged data

The subjective ratings for a given HRC-scene combination were averaged, then used in a regression with P6, P9, and P13. The results:

- P6, P9, and P13 were statistically reliable predictors of the averaged ratings.

- The standardized parameter estimates were -0.60, -0.31, and -0.07, respectively.

---

[29] The *standardized* estimates eliminate differences in scale of the the various objective measures. The standardization consists, in effect, of subtracting the mean of a measure from each raw observation, then dividing by the standard deviation of that measure. The interpretation of the resulting regression estimates is that by moving one standard deviation up, say, the P6 scale, we observe an average move of .50 standard deviation down the rating scale.

- The model accounted for 69.8% of the variance in the average ratings (correlation of predicted and observed of 0.835); RMS error was 0.61 rating unit.
- The corresponding ANOVA that accounts for subject effects would account for about 84% to 89% of the variance. Bit Rate accounted for 60.8% of the variance in the averaged ratings.

The results of the regression for the 625 stimuli are shown in Figure 10. Although visually the plot seems somewhat curved, a second degree polynomial does not fit this plot any better than a straight line. Are these good results? The objective measures are doing about 83% of maximum (0.698/0.840) in accounting for the average subjective ratings.[30]

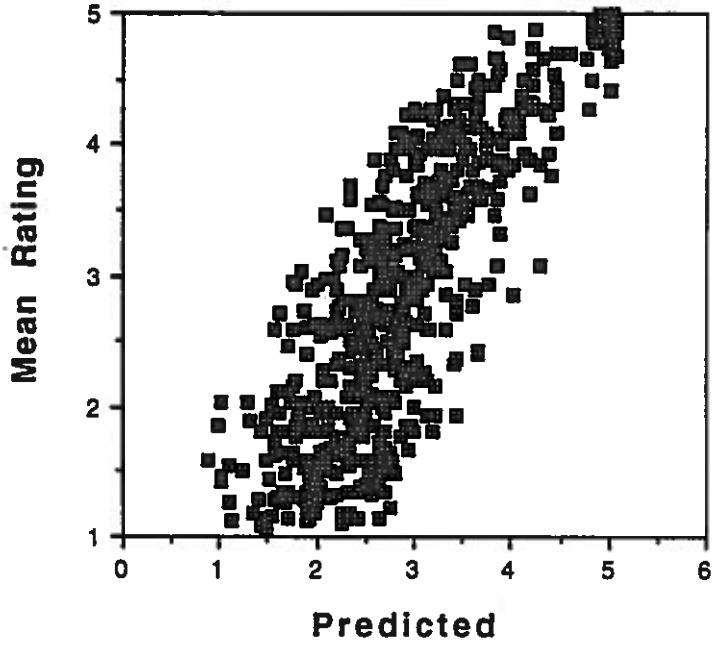


Figure 10. Mean ratings for 625 HRC-scene combinations vs. ratings predicted from a regression using objective measures P6, P9, and P13.

Figure 11 shows that the regression is not doing a complete job because there is still quite a lot of structure in the residuals. (If the regression were doing a very good job, the residuals would look entirely random, scattered about a straight line through 0.0 on the y-axis.) This structure exists in two forms. The first is the general curvilinear relationship between the residuals and the mean ratings. As the AMMI analysis showed, this structure may be due to end effects in the rating scale, something which no objective measure of video quality could be expected to handle. In this sense, the results may be about as good as one could hope for.[31]

---

[30] The ANOVA of 625 averaged ratings with respect to the variables HRC and Scene accounted for 89% of the variance in the ratings. If one uses that figure as the denominator, then the regression performed at about 78% of maximum.

[31] Transforming the data scales using obvious functions like powers and logs did not get rid of the curvilinearity.

The second kind of structure is simply the amount of scatter in the points in Figure 11. Each point represents an HRC-scene combination. The amount of scatter in those points that is due to error in the ratings is measured by a standard deviation of just 0.13.[32] The total amount of scatter actually present in those 625 residuals is measured by a standard deviation of 0.465 about the curved line. That means that peculiarities of HRC-scene combinations are not being picked up and accounted for by the objective measures, even though the measures actually apply individually to each of those 625 combinations.

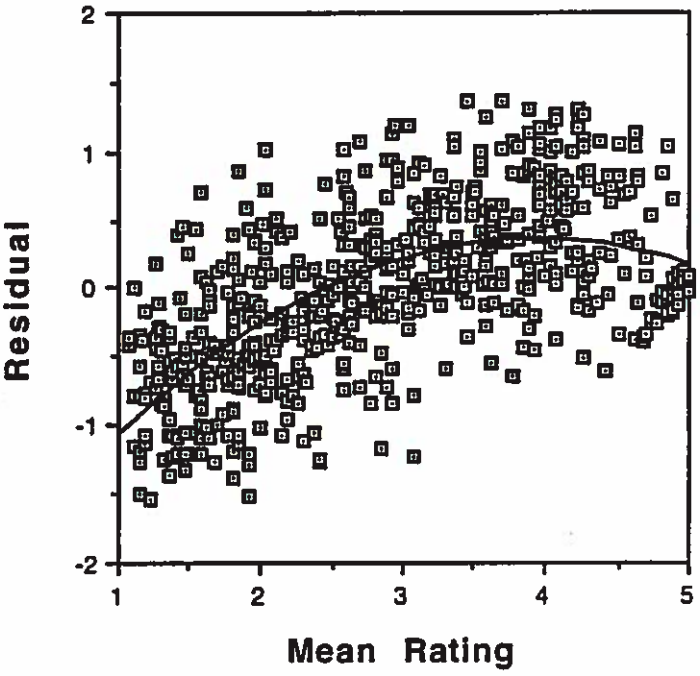


Figure 11. Residuals from the regression using P6, P9, P13 plotted against mean rating for 625 HRC-scene combinations.

If peculiarities of HRC-scene combinations are difficult for the objective measures to capture, one might simply average out those peculiarities by aggregating data across scenes. We did that. We averaged the subjective data and the objective measures across scenes for each of the 25 HRC's. We then did the corresponding regression using P6, P9, and P13. The objective measures accounted for 79.3 % of the variance in the averaged data (a correlation of 0.89 -- recall that the correlation of Bit Rate with the ratings was 0.90 at a comparable level of aggregation). Figure 12 shows the predicted and observed ratings. Again there is a slight curve at the upper end of the ratings.

We then examined the residuals from this linear regression, which are shown in Figure 13. As before, the residuals should be a random scatter about a flat line passing through 0.0 on the y-axis, but they are not. Again there is a marked curve, which may be an artifact of the rating scale. The amount of scatter about the curved line in Figure 13 is measured by a standard deviation of 0.25. But, the amount of error for each data point in Figure 13 is only 0.03 (that is, the standard error of the mean ratings for an HRC is, on average, 0.03; note, again, that this small standard error is based on the full data set of 79 subjects.)

---

[32] The standard error of each of those 625 points was calculated from the ratings. The average of the standard errors was 0.130.

Thus, even with the effects of different scenes removed from the data, the objective measures still do not completely capture the subjective data.

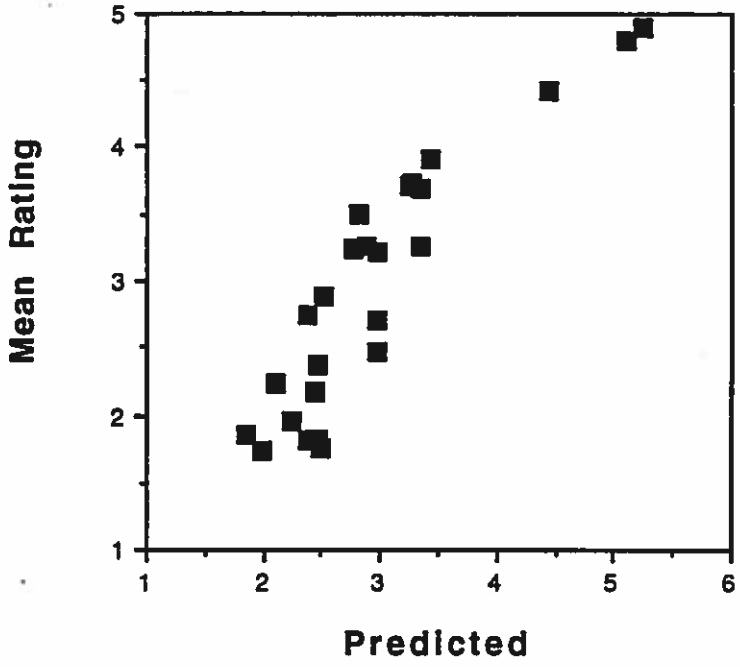


Figure 12. Mean ratings for the 25 HRC's plotted against ratings predicted from the objective measures P6, P9, P13.

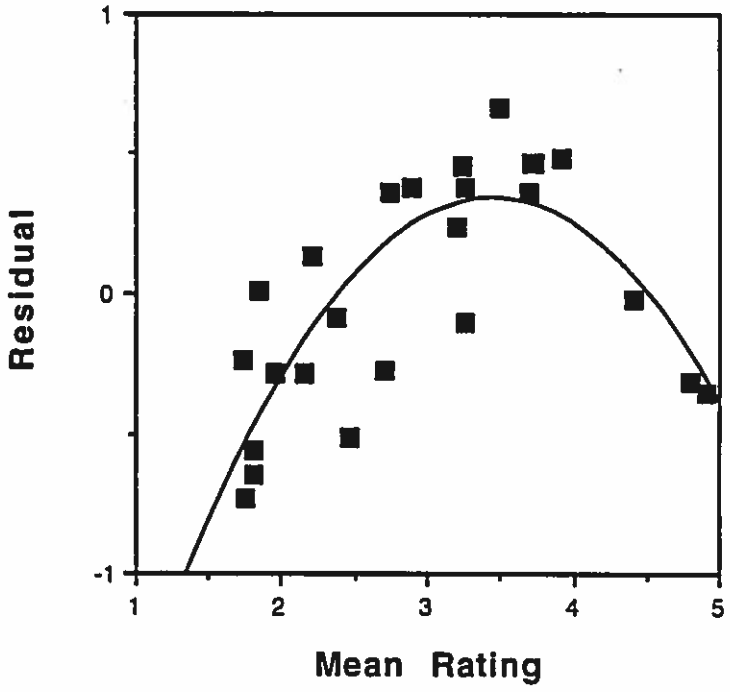


Figure 13. Residuals from the regression using P6, P9, P13 plotted against mean rating for 25 HRC-scene combinations.

Results using simplest RMS measures (P2, P8) and P13

In the analyses above we used objective measures that we had chosen on the basis of statistical analyses of the whole set of objective measures. We might have used a different strategy for choosing objective measures. One would be to use the simplest measure from each of the families of measures (temporal, spatial, and Fourier). On this strategy, P2, P8, and P13 could be chosen. A regression using the raw rating data and P2, P8, and P13 yielded an R-squared of 0.452 (compared to 0.483 for P6, P9, and P13). As in the previous analyses, the relative importance of the measures was *temporal*, *spatial*, and *Fourier*.

Results using Bit Rate as a covariate

Bit rate is usually a given constraint of a system, not a design factor. The coding algorithm, tradeoffs made between parameters within a type of algorithm, and preprocessing strategies for material of different kinds are the sorts of issues one hears described as interesting. Bit rate is like cubic inches in a car engine -- "sure bigger engines produce more power; there's nothing interesting in that." Given this view of bit rate, and given the powerful relationship it has with both objective and subjective video quality, it is important to see how well the various objective measures predict subjective quality if the influence of Bit Rate is factored out of the data.

P6, P9, P13

We used Bit Rate to predict the raw rating data in a regression, and output the residuals (i.e., the difference between each rating and a single best-fitting straight line in Fig. 9). The residuals then became the dependent variable (in place of Rating) in a regression using P6, P9, and P13 as predictors. Results:

- P6, P9, and P13 were statistically reliable predictors of the residuals. Thus, not only is there information in the subjective ratings that is not captured by Bit Rate, but that information is statistically related to the objective measures.
- The standardized parameter estimates were -0.02, -0.29, and -0.15, respectively. The regression weight for P6, the measure most highly related to Bit Rate, was affected the most (declined from -0.50 to -0.02).
- The model accounted for 10.4% of the variance in the residuals. RMS error was 0.95 rating unit, about the same as for the regression in which P6, P9, and P13 were used to predict the ratings. Thus, although the set of predictors {P6, P9, P13} is picking up something in the ratings beyond what Bit Rate picks up, the increase in predictive accuracy is marginal.
- Bit Rate accounted for 40.0% of the variance in the ratings. The set of predictors Bit Rate, P6, P9, P13 accounted for 51.1% of the variance in the ratings (in a separate regression).

P1, P9, P13

Perhaps another group of predictors would account for more of the residual if the predictors were chosen to be less correlated with Bit Rate. P1, P9, and P13 are the representatives of their respective classes of measures that correlate least highly with Bit Rate. Again we used the residuals from a Rating vs. Bit Rate regression as the dependent variable; this time P1, P9, and P13 were the predictors in a regression. Results:

- P1, P9, and P13 were statistically reliable predictors of the residuals.

- The standardized parameter estimates were -0.06, -0.30, and -0.14, respectively. Note that the regression weight for P1 was smallest. P1 is the measure most highly related to Bit Rate of the three, despite the fact that P1 was least related to Bit Rate among the class of temporal domain measures.

- The model accounted for 10.7% of the variance in the residuals. RMS error again was 0.95 rating unit.

- As before, Bit Rate accounted for 40.0% of the variance in the ratings. The set of predictors Bit Rate, P1, P9, P13 accounted for 50.5% of the variance of the ratings in a separate regression, slightly less than when P6 replaced P1, above.

Thus, the strategy of selecting objective measures that do not correlate highly with Bit Rate did not improve on the set P6, P9, and P13 (two of which were already the low correlators within their own classes of measures).

Summary and Conclusions

Objective measures of video quality do predict subjective ratings
Using data from 79 observers at two labs, and objective measures of 625 video recordings, we found a strong relationship between repeatable objective measures and the average subjective ratings of video quality. The best objective measures were representatives of (a) the temporal domain, (b) the spatial domain, and (c) the spatial frequency (Fourier) domain. The objective measures captured about 80% of the extractable information in the subjective data. The objective measures were not able to account for all effects in the subjective data.

Bit rate has an overwhelming influence on both the objective measures and the subjective ratings
An effective strategy in experimental design is to cover a wide range of the variables that are being investigated. We followed that reasonable strategy. But, in doing so we also allowed one variable, Bit Rate, to dominate our data. An area for future development might be objective measures that predict subjective quality when bit rate is held constant. Efforts in this direction have been successful in the past (Wolf and Webster, 1993).

The subjective data are of good quality
The random scatter in the data is not excessive. Lack of repeatability of a given rating for a single observer is 0.5 rating scale point, which was the target set by the Test Plan. The standard error for the average HRC-scene combination is just 0.13, substantially below the target in the Test Plan. The subjective measures were sensitive enough to identify a set of HRC's (11-16) that are systematically different from the rest of the HRC's. A component of the subjective data that is not accounted for by the objective measures may be due to an artifact of the response scale inherited from CCIR 500-5.

Methodological considerations
We found, as in previous work, that subject differences do not spoil the data and are unrelated to simple demographic measures. We intend to keep that fact in mind before spending time and money selecting special populations of subjects in future studies. We also found that lab differences do not spoil the data. We had expected to find lab differences, but not to be able to explain the differences uniquely. Instead we found no differences. The whole issue of subject effects can be avoided in the first place by removing subject means from the data. Besides helping reduce the variance in the data, removing subject means also inhibits one from thinking of the response scale as having some absolute meaning. Rating scales, in general, do not have an absolute meaning.

References

Bradu, D. & Gabriel, K. R. (1978). The biplot as a diagnostic tool for models of two-way tables. Technometrics, 20, 47-68.

Cermak, G.W. (1994). Accuracy in predicting subjective video quality. T1A1.5/94-127.

Cermak, G.W., Teare, S.K., Tweedy, E.P., & Stoddard, J. C. (1994). Consumer judgments of MPEG1 video. T1A1.5/94-303.

Crow, E. L. (1994). Methods for analysis of interlaboratory video performance standard subjective test data. T1A1.5/94-128.

Fisher, R. A., & Mackenzie, W. A. (1923). Studies in crop variation. II. The manurial response of different potato varieties. Journal of Agricultural Science, 23, 311-320.

Gauch, H. G. Jr. (1992). Statistical Analysis of Regional Yield Trials: AMMI Analysis of Factorial Designs. Amsterdam: Elsevier.

International Telecommunications Union (1986). 1986 - CCIR Recommendations and Reports of the CCIR, 1986 , Vol. XI - Part 1. Geneva: ITU.

International Telecommunications Union (1992). 1992 - CCIR Recommendations. Geneva: ITU.

Mezrich, J.J., Carlson, C.R., & Cohen, R.W. (1977). Image Descriptors for Display. Princeton: RCA Laboratories.

Morton, A.C. Subjective test plan (ninth draft). T1A1.5/94-118.

Pike, E. W. & Silverberg, T. R. (1952). Designing mechanical computers. Machine Design, 24, 131-137, 159-163.

Pinson, M. & Jones, C. (1994). NTIA/ITS objective test results. T1A1.5/94-120.

Seitz, N.B., Wolf, S., Voran, S., & Bloomfield, R. (1994). User-oriented measures of telecommunication quality. IEEE Communications Magazine, Jan., 56-66.

T1Q1.5 Working Group (1988). Project proposal titled "Analog interface performance specifications for digital video teleconferencing / video telephony service." T1Q1.5/88-061 REV.

Tukey, P.A. (1994). Whether to pre-average replications? T1A1.5/94-114.

Voran, S. (1991). Correlation between ITS objective measures and subjective video quality: Preliminary results on a set of 15 scenes. T1Q1.5/91-124.

Voran, S. & Wolf, S. (1992). Objective measures of video impairment: Analysis of 128 scenes. T1A1.5/92-112.

Webster, A. (1993). Methods of measurement for two objective video quality parameters based on their Fourier Transform. T1A1.5/93-153.

Wolf, S. (1990). Features for automated quality assessment of digitally transmitted video. NTIA Report 90-264.

Wolf, S., Pinson, M., Jones, C., & Webster, A. (1993). A summary of methods of measurement for objective video quality parameters based on the Sobel filtered image and the motion difference image. T1A1.5/93-152.

Wolf, S. & Webster, A. (1993). Objective performance parameters for NTSC video at the DS3 rate. T1A1.5/93-60.

Williams, E. J. (1952). The interpretation of interactions in factorial experiments. Biometrika, 39, 65-81.

Appendix A: The Individual HRC's
(reproduced from Subjective Test Plan, T1A1.5/94-118)

HYPOTHETICAL REFERENCE CIRCUITS

HRC	Algorithm (vendor)	Resolution	Total, Kbps	Audio, Kbps	Video, Kbps	Coding Mode	Frame Rate	FEC	Burst Errors
1	Null	-	-	-	-	-	-	-	Off
2	VHS	-	-	-	-	-	-	-	Off
3	Proprietary	V.High	45,000	-	-	-	-	-	Off
4	Proprietary	Med.	128	-	-	VQ	-	-	Off
5	Proprietary	High	336	-	-	VQ	-	-	Off
6	Proprietary	Med.	112	-	-	-	-	-	Off
7	Proprietary	Med.	384	-	-	-	-	-	Off
8	Proprietary	Med.	768	-	-	-	-	-	Off
9	Proprietary	High	768	-	-	-	-	-	Off
10	Proprietary	High	1536	-	-	-	-	-	Off
11	H.261(diff)	QCIF	128	56	70.4	INTER+MC	-	On	Off
12	H.261(same)	QCIF	128	56	70.4	INTER	10*	On	Off
13	H.261(same)	QCIF	168	48	118.4	INTER+MC	-	On	Off
14	H.261(diff)	QCIF	384	56	326.4	INTER+MC	-	On	Off
15	H.261(same)	CIF	112	48	62.4	INTER+MC	-	On	Off
16	H.261(same)	CIF	128	56	70.4	INTER+MC	-	On	Off
17	H.261(diff)	CIF	128	48	78.4	INTER+MC	-	On	Off
18	H.261(same)	CIF	168	48	118.4	INTER+MC	-	On	Off
19	H.261(same)	CIF	256	56	190.4	INTER+MC	15*	On	On
20	H.261(same)	CIF	384	56	326.4	INTER+MC	-	On	Off
21	H.261(same)	CIF	384	56	326.4	INTER+MC	-	On	On
22	H.261(diff)	CIF	768	56	710.4	INTER+MC	-	On	Off
23	H.261(same)	CIF	768	56	710.4	INTER+MC	-	On	On
24	H.261(diff)	CIF	1536	56	1478.4	INTER+MC	-	On	Off
25	H.261(same)	CIF	1536	56	1478.4	INTER+MC	-	On	Off

* Specified value. Actual frame rate may be determined through measurement.

Appendix B: Experimental Room Setup

Experimental Setup, Top View