Committee T1 Performance
Standards Contribution

Document Number:	T1A1.5/96-121
TIBBS File:	
DATE:	Oct 28, 1996
STANDARDS PROJECT:	Analog Interface Performance Specifications for Digital Video Teleconferencing/Video Telephony Service (T1Q1-12)
SUBJECT:	Objective and Subjective Measures of MPEG Video Quality
SOURCE:	GTE Laboratories, NTIA/ITS
CONTACT:	GTE Laboratories: Greg Cermak (phone: 617-466-4132, email: gwc0@gte.com), Pat Tweedy NTIA/ITS: Stephen Wolf (phone: 303-497-3771, email: steve@its.bldrdoc.gov), Arthur Webster, Margaret Pinson
KEY WORDS:	video quality, MPEG, subjective, objective, correlation
DISTRIBUTION:	Working Group T1A1.5 (announced via t1a15@t1.org)

NOTICE: *Identification in this report of certain commercial equipment, instruments, protocols, or materials does not imply recommendation or endorsement by NTIA, ITS, or GTE Laboratories, nor does it imply that the material or equipment identified is necessarily the best available for the purpose.*

This contribution contains information that was prepared to assist Committee T1 and specifically Technical Subcommittee T1A1 in their work program. This document is submitted for discussion only, and is not to be construed as binding on GTE. Subsequent study may lead to revision of the details in the document, both in numerical value and/or form, and after continuing study and analysis GTE Telephone Operations specifically reserves the right to change the contents of this contribution.

1. Introduction

The T1A1.5 working group has been working toward a set of standards for the measurement of the quality of compressed digital video [e.g., 2, 6, 11, 12, 13, 17. 19, 20, 26]. The benefits of standards for the measurement of video quality have been cited by many (e.g., see [15], pg.2). New, objective measures of video transmission quality are needed by standards organizations, end users, and providers of advanced video services. Such measures will promote impartial, reliable, repeatable, and cost effective assessment 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.

The T1A1.5 working group has been approaching the issue of video quality standards by means of a research program. The general scientific method used has been to test digital codecs and take data of two types: (a) a set of objective measures, and (b) subjective judgments by human judges. Statistical analyses reveal which objective measures best predict the subjective judgments. A multi-lab collaborative study of this type [see 20, 21], mounted by T1A1.5 members, covered a wide range of digital video systems, from bit rates of about 100 kb/s to 45 Mb/s. A set of objective measures of video quality developed at NTIA/ITS performed well in accounting for subjective judgments by human observers on these same systems.

The T1A1.5 multi-lab study was large, well done, and successful. But, it was not conclusive in the sense of pre-empting future studies. Furthermore, this study did not cover high bit-rate entertainment video systems very thoroughly (by design): Only three systems were at or above 1.5 Mb/s, and of those one was VHS. No systems were tested with bit rates between 1.5 and 45 Mb/s.

The present studies were conducted to fill in the bit-rate gap in the previous T1A1.5 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 T1.801.03 objective video quality metrics are examined for these bit rates and coding technologies. In addition, the NTIA/ITS video quality laboratory has been recently upgraded to implement and test matrix metrics (e.g., metrics that perform pixel by pixel comparisons of the input and output images) on large video data sets. This added capability (which did not exist for the previous T1A1.5 multi-lab study) has made possible the evaluation of three matrix video quality metrics; peak signal to noise ratio (PSNR), and two previously introduced [15, 16] matrix versions of spatial information (*SI*) distortion (see section 6.1.1.1 of ANSI T1.801.03 for a definition of spatial information for a pixel). One matrix *SI* distortion metric measures the amount of false edges in the output image and the other measures the amount of missing edges. Since spatial registration of the input and output images is critical for successful implementation of matrix measures, a considerable effort has been made here to describe the image calibration algorithms that were used by the objective measurement system.

2. Overview of the Two Studies

2.1 HRCs[1] 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]. Both of these studies followed the strategy:

- Choose a set of HRCs for testing that includes as wide a range of video quality as possible within the usage domain (in this case, *entertainment*).
- Among the HRCs, include current products for comparison, e.g., VHS and cable.
- Test each of the HRCs with the same set of test sequences.
- Test each HRC end-to-end, i.e., using a full cycle of coding, transmission, and decoding.
- Use test sequences that are typical of the material that actual consumers would view using such HRCs.
- Use recordings of the HRC-scene pairs, rather than creating each sequence live during testing and analysis.

Following this strategy, the HRCs tested were, in Study 1:

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 Horizontal resolution 330-400 pixels, Vertical resolution 480 lines
5. MPEG 1+ Bit rate 8.3 Mb/s Horizontal resolution 330-400 pixels, Vertical resolution 480 lines
6. Original scene with a signal-to-noise ratio of 34 dB
7. Original scene with a signal-to-noise ratio of 37 dB
8. Original scene with a signal-to-noise ratio 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

[1] 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 a signal-to-noise ratio (SNR) of 34 dB
7. Original scene with a signal-to-noise ratio of 37 dB
8. Original scene with a signal-to-noise ratio 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. Two studies, rather than one larger study, were conducted because the MPEG 2 codecs were not available at the time of the first study.

The same set of scenes was used in both studies. The scenes were chosen to span a *range* of 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. The length of the scenes was chosen so that the sum of all combinations of the scenes and the HRCs plus an original of each scene would be less than 20 minutes, which is the limit for high quality record/playback from the Panasonic video disc machine we used. 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. The clips chosen for the two studies were as follows:

1. A clip from the movie "2001: A Space Odyssey". It shows a man running in a cylindrical track in a space ship. The runner remains stationary with respect to the camera. The circular walls apparently move from behind the camera (and viewer), rotating about an axis parallel to the plane of the picture. The walls have quite a lot of detail and sharp edges.
2. A clip from the movie "The Graduate". It shows a slow camera zoom towards a woman (Ann Bancroft) sitting on a chaise. Behind her is a background of leaves that are large enough so that many edges appear.
3. A clip from the movie "The Godfather". It shows two men talking in low ambient light, with very little apparent color (Al Pacino at a restaurant with an enemy of Don Corleone). The camera focus is soft. The important movements are subtle facial expressions.

4. A clip from the movie "Being There" showing two men talking (Peter Sellers and a government representative). Again there is very little color, and the only movements are subtle facial expressions.

5. Ice hockey clip #1 is dominated by a fight in which the camera remains stationary, but there is much movement among the players. The background is very high-contrast, consisting of bright ice with a highly detailed and colorful crowd above the ice.

6. Ice hockey clip #2 shows much movement up and down the ice with the camera following a skater or the puck, panning across the background. The clip is from the same game (and the same background of ice and crowd) as hockey clip #1.

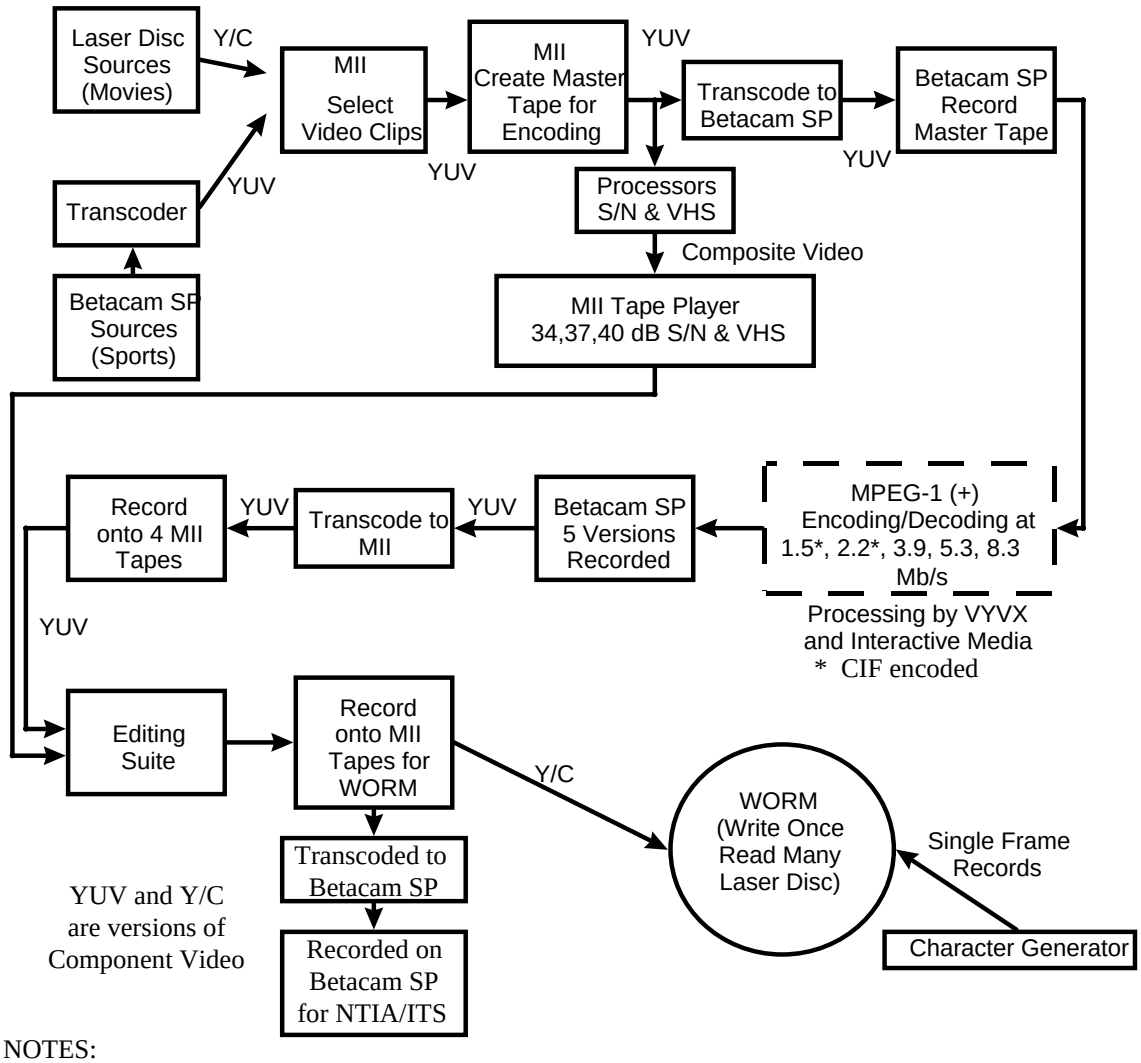
7. A basketball clip includes many scene cuts (from one camera to another). One main sequence shows a close-up of a player (Charles Barkley) running down the court, with the background crowd and other players a blur behind him. Another shows a close-up of the Bulls' coach moving slowly in front of the bench and crowd. The third main sequence (packed into the 14 sec scene) shows a long distance shot of half-court play in which there is a great amount of fine detail, but the total amount of movement on the screen is small.

8. A baseball clip also includes several scene cuts. The viewer sees two close-ups of the pitcher on different pitches, with a stationary and moderately detailed background. There are two shots of batters stationary against the background of stadium walls and crowd. There are also two shots of base-runners trotting against the background of the field markings after a walk. Finally, there is a long distance shot in which the camera tracks a long fly ball (barely visible in the original), with the field, stadium walls, and crowd as the background.

2.2 Production of Video Material

2.2.1 MPEG 1+ study

Figure 1 describes the steps in producing the video material (a) in the form it was shipped for objective analysis, and (b) in the form it was presented to consumers for ratings. The video processing for the objective analysis and the subjective testing followed the same series of steps until the final step. The reader will note that there are more tape generations than would be ideal. Whatever noise was added to the video signal during this processing became part of the end-to-end system performance that was evaluated both by the consumers and the objective measures. The added noise was certainly not of a magnitude to hide other processing artifacts, and it affected all of the HRCs equally.



NOTES:

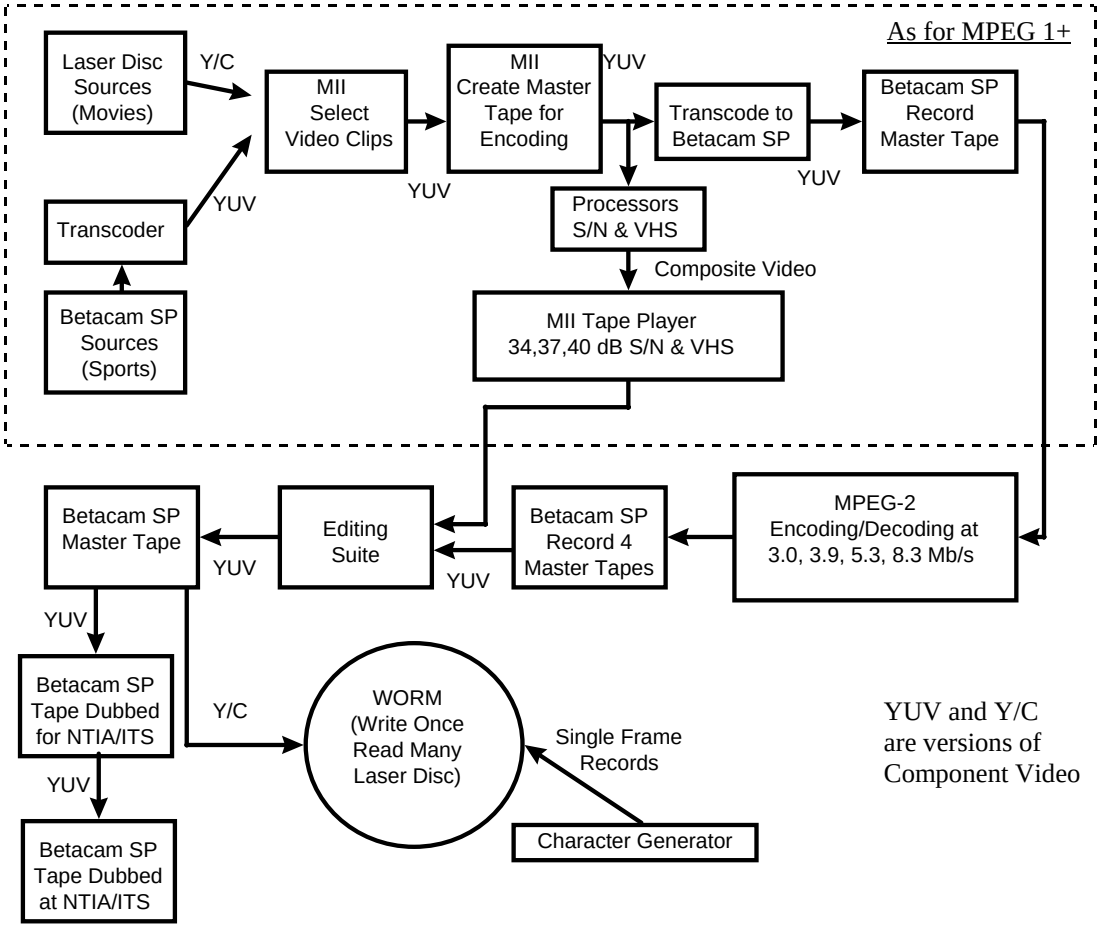
1. The use of a WORM disc was considered desirable to avoid the creation of sets of Betacam SP tapes each providing random orders of processed video clips. The WORM disc can be controlled from a PC to generate sets of random sequences.
2. At the time this work was carried out editing could only be carried out using a pair of MII recorder/players. Only one Betacam SP recorder/player was available.
3. Source material (sports) provided by broadcast stations was provided on Betacam SP tapes.
4. For MPEG-1 processing by outside organizations it was necessary to provide them with Betacam SP tapes - they did not own MII equipment.

Figure 1 Process used to create MPEG 1+ WORM disc for subjective testing and Betacam SP tape for objective testing

2.2.2 MPEG 2 study

Figure 2 describes the steps in producing the video material for the MPEG2 study (a) in the form it was shipped for objective analysis, and (b) in the form it was presented to consumers for ratings. The reader will note that there were fewer processing steps to produce the WORM disc in this study than in the preceding MPEG1+ study. This would

not affect the relationships among the HRCs within the MPEG 2 study, compared to the relationships of HRCs within the MPEG 1+ study. However, it might give the HRCs from the MPEG 2 study a slight advantage over the HRCs in the MPEG 1+ study. (We did not see such an effect, however, in our own observation; the analog tape equipment used is of very high quality.)



NOTES:

1. The use of a WORM disc was considered desirable to avoid the creation of sets of Betacam SP tapes each providing random orders of processed video clips. The WORM disc can be controlled from a PC to generate sets of random sequences.
2. For MPEG 2 Betacam SP editing was available, although the original MPEG 1+ MII source tape was used to allow valid subjective measure comparisons between MPEG 1+ and MPEG 2.

Figure 2 Process used to create MPEG 2 WORM disc for subjective testing and Betacam SP tape for objective testing

Note that one extra dub of the Betacam SP was required at NTIA/ITS to insert vertical interval time code (VITC), which was required for frame capture by the NTIA/ITS objective measurement system but which was inadvertently left off in the first Betacam SP dub. Side by side subjective comparisons of the video from the two Betacam SP tapes revealed that a slight amount of visible noise was introduced by this extra Betacam SP dub.

3. Objective Measures

3.1 Performance Measurement Issues for Digital Video Systems

3.1.1 Input Scene Dependencies

The advent of video compression, storage, and transmission systems has exposed fundamental limitations of techniques and methodologies that have traditionally been used to measure video performance. Traditional performance parameters have relied on the "constancy" of a video system's performance for different input scenes. Thus, one could inject a test pattern or test signal (e.g., a static multi-burst), measure some resulting system attribute (e.g., frequency response), and be relatively confident that the system would respond similarly for other video material (e.g., video with motion).[2] A great deal of research has been performed to relate the traditional analog video performance parameters (e.g., differential gain, differential phase, short time waveform distortion, etc.) to perceived changes in video quality [3, 4, 5]. While the recent advent of video compression, storage, and transmission systems has not invalidated these traditional parameters, it has certainly made their connection with perceived video quality much more tenuous. Digital video systems adapt and change their behavior depending upon the input scene. Therefore, attempts to use input scenes that are different from what is actually used "in-service" [3] can result in erroneous and misleading results. Variations in subjective performance ratings as large as 3 quality units on a subjective quality scale that runs from 1 to 5 (1=lowest rating, 5=highest rating) have been noted in tests of commercially available systems. While quality dependencies on the input scene tend to become much more prevalent at higher compression ratios, they also are observed at lower compression ratios. For example see [6], where subjective test results of 45-Mb/s contribution quality systems (i.e., systems now used by broadcasters to transmit over long-line digital networks) revealed one transmission system with multiple tandem codecs whose subjective performance varied from 2.16 to 4.64 quality units.

A digital video transmission system that works fine 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

---

[2] The subjective, or user-perceived, quality of analog video systems can also depend upon the scene content. For example, a fixed analog noise level may be less objectionable for some scenes than others.

[3] With "in-service" measurements, the transmission system is available for use by the end-user. With "out-of-service" measurements, the transmission system is not available for use by the end-user.

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 [7, Annex A of 8]. 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 9]. National and international standards have been developed that specify standard video scenes for testing digital video systems [8, 10, 11]. Use of these standards assures that users compare apples to apples when evaluating systems from different suppliers.

3.1.2 New Digital Video Impairments

Digital video systems produce fundamentally different kinds of impairments than analog video systems. Examples of these include tiling, error blocks, smearing, jerkiness, edge busyness, and object retention [12]. To fully quantify the performance characteristics of a digital video system, it is desirable to have a set of performance parameters, where each parameter is sensitive to some unique dimension of video quality or impairment type. This is similar to what was developed for analog impairments (e.g., a multi-burst test would measure the frequency response, and a signal-to-noise ratio test would measure the analog noise level). This discrimination property of performance parameters is useful to designers trying to optimize certain system attributes over others, and to network operators wanting to know not only when a system is failing but where and how it is failing.

Also of interest is how a user weighs the different performance attributes of a digital video system (e.g., spatial resolution, temporal resolution, or color reproduction accuracy) when subjectively rating the quality of the experience. The process of estimating these subjective quality ratings from objective performance parameter data is an important new area of work that will be discussed below.

3.1.3 The Need for Technology Independence

The constancy of analog video systems over the past 4 decades provided the necessary long term development cycle to produce today's accurate analog video test equipment. In contrast, the rapid evolution of digital video compression, storage, and transmission technology presents a much more difficult performance measurement task. To avoid immediate obsolescence, new performance measurement technology developed for digital video systems must be technology independent, or not dependent upon specific coding algorithms or transport architectures. One way to achieve technology independence is to have the test instrument perceive and measure video impairments like a human being. Fortunately, the computational resources needed to achieve these measurement operations are becoming available.

3.2 A New Objective Measurement Methodology

The above issues have necessitated the development of a new measurement methodology for testing the performance of digital video systems. Rather than being limited to artificial test signals, this methodology is one that can use natural video scenes. Figure 3 presents the reference model for measuring end-to-end video performance parameters and summarizes the principles of the new measurement methodology detailed in ANSI

T1.801.03, “American National Standard for Telecommunications - Digital Transport of One-Way Video Telephony Signals - Parameters for Objective Performance Assessment” [13]. This standard specifies a framework for measuring end-to-end performance parameters that are sensitive to distortions introduced by the coder, the digital channel, or the decoder shown in Figure 3.

Performance measurement systems digitize the input and output video streams in accordance with ITU-R Recommendation BT.601-4 [14] and extract features from these digitized frames of video. Features are quantities of information that are associated with individual video frames. They 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.

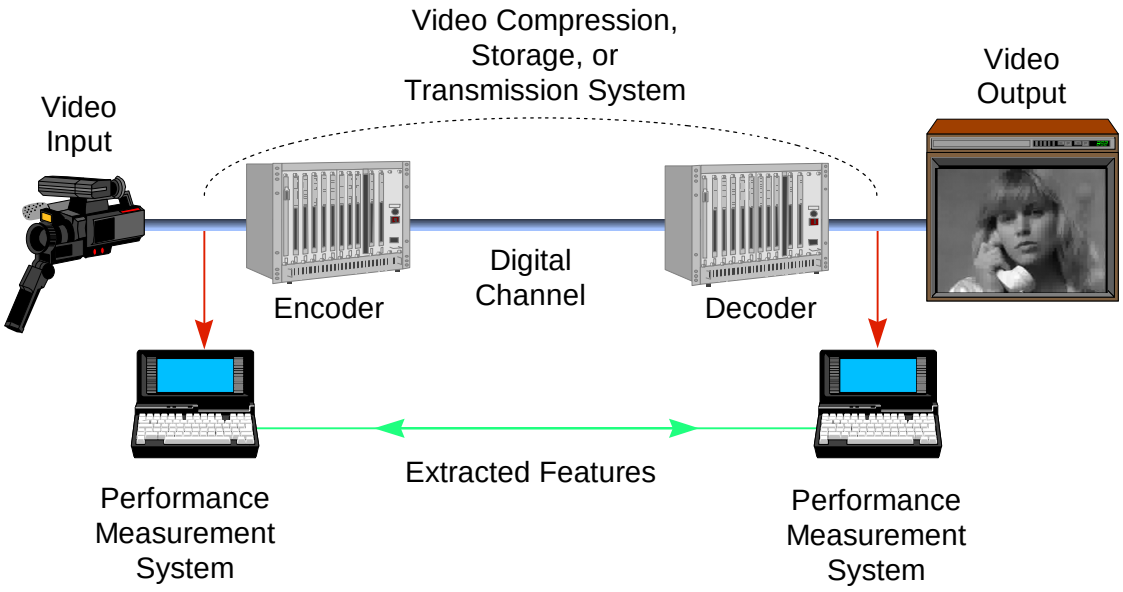


Figure 3. ANSI T1.801.03 reference model for measuring video performance.

In general, the transmission and storage requirements for measuring an objective parameter based on scalar features is less than that required for an objective parameter based on vector features. This, in turn, is less than that required for an objective parameter based on matrix features. Significantly, scalar-based parameters have produced good correlations to subjective quality. This demonstrates that the amount of reference information that is required from the video input to perform meaningful quality measurements is much less than the entire video frame. This important new idea of

compressing the reference information for performing video quality measurements has significant advantages, particularly for such applications as long-term maintenance and monitoring of network performance. Since a historical record of the output scalar features requires very little storage, they may be efficiently archived for future reference. Then, changes in the digital video system over time can be detected by simply comparing these past historical records with current output feature values.

Further refinements in the art of compressing video quality information holds out the promise of producing an "in-service" method for measuring video quality that will be good enough to replace subjective experiments in many cases. This extension would make it possible to perform non-intrusive, in-service performance monitoring, which would be useful for applications such as fault detection, automatic quality monitoring, and dynamic optimization of limited network resources.

3.2.1 Example Features

This section presents examples from each of the three classes of features (scalar, vector, matrix). The first example to be presented is scalar features based on statistics of spatial gradients in the vicinity of image pixels. These spatial statistics are indicators of the amount and type of spatial information, or edges, in the video scene. The second example is scalar features based on the statistics of temporal changes to the image pixels. These temporal statistics are indicators of the amount and type of temporal information, or motion, in the video scene from one frame to the next. Spatial and temporal gradients are useful because they produce measures of the amount of perceptual information, or change in the video scene. The third example is a vector feature that is based on the radial averaged frequency content of a video scene. Finally, several examples of matrix features are presented, included the commonly used peak signal to noise ratio (PSNR).

3.2.1.1 Spatial Information (SI) Features

Figure 4 demonstrates the process used to extract spatial information (*SI*) features from a sampled video frame. Gradient or edge enhancement algorithms (i.e., Sobel filters) are applied to the video frame. At each image pixel, two gradient operators are applied to enhance both vertical differences (i.e., horizontal edges) and horizontal differences (i.e., vertical edges). Thus, at each image pixel, one can obtain estimates of the magnitude and direction of the spatial gradient (the right-hand image in Figure 4 shows magnitude only, called SI_r in ANSI T1.801.03). A statistic is then calculated on a selected subregion of the spatial gradient image to produce a scalar quantity. Examples of useful scalar features that can be computed from spatial gradient images include total root mean square energy (this spatial information feature is denoted as SI_{rms} in ANSI T1.801.03) , and total energy that is of magnitude greater than r_{min} and within $\Delta\theta$ radians of the horizontal and vertical directions (denoted as $HV(\Delta\theta, r_{min})$ in ANSI T1.801.03). Parameters for detecting and quantifying digital video impairments such as blurring, tiling, and edge busyness are measured using time histories of *SI* features.

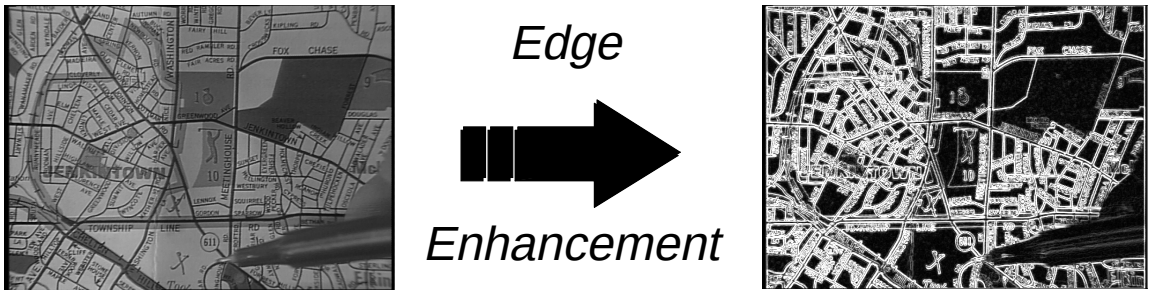


Figure 4. Example spatial information features.

3.2.1.2 Temporal Information (TI) Features

Figure 5 demonstrates the process used to extract temporal information (*TI*) features from a video frame sampled at time *n* (i.e., *frame n* in the figure). First, temporal gradients are calculated for each image pixel by subtracting, pixel by pixel, *frame n-1* (i.e., one frame earlier in time) from *frame n*. The right-hand image in Figure 5 shows the absolute magnitude of the temporal gradient and, in this case, the larger temporal gradients (white areas) are due to subject motion. A statistical process, calculated on a selected subregion of the temporal gradient image, is used to produce a scalar feature. An example of a useful scalar feature that can be computed from temporal gradient images is the total root mean square energy (this temporal information feature is denoted as TI_{rms} in ANSI T1.801.03). Parameters for detecting and quantifying digital video impairments such as jerkiness, quantization noise, and error blocks are measured using time histories of temporal information features.

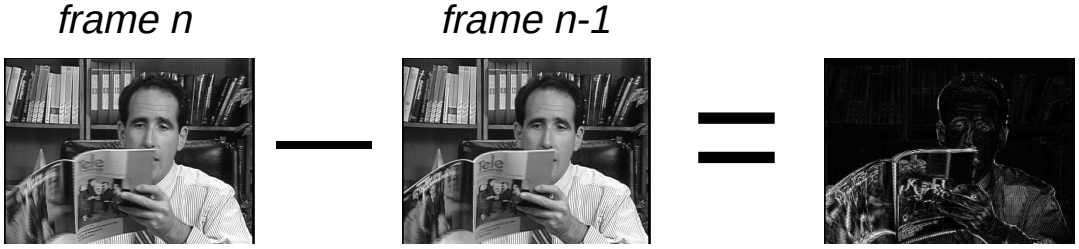


Figure 5. Example temporal information features.

3.2.1.3 Spatial Frequencies Feature

A vector feature can be computed from the Fourier transform of a square (*N* horizontal pixels by *N* vertical lines) sub-region of the sampled video frame. This vector feature, denoted by

$$\mathbf{f} = \begin{bmatrix} f(0) \\ f(1) \\ . \\ . \\ . \\ f\left(\frac{N}{2}-1\right) \end{bmatrix},$$

is computed from the magnitude of the two dimensional Fourier transform **F** shown in Figure 6 by radial averaging of the spatial frequency bins. The individual elements of the vector, computed as

$$f(k) = \frac{1}{N_k} \sum_{i,\,j} F(i,\,j) \quad \text{for all } i \text{ and } j \text{ such that } \quad k-1 < \sqrt{i^2 + j^2} \le k,$$

give the total amount of spatial frequency information at each spatial frequency *k*. Graphically, the Fourier magnitude points *F*(*i*, *j*) that are contained within the shaded ring of Figure 6 are averaged to produce a value for each frequency ring *k*.

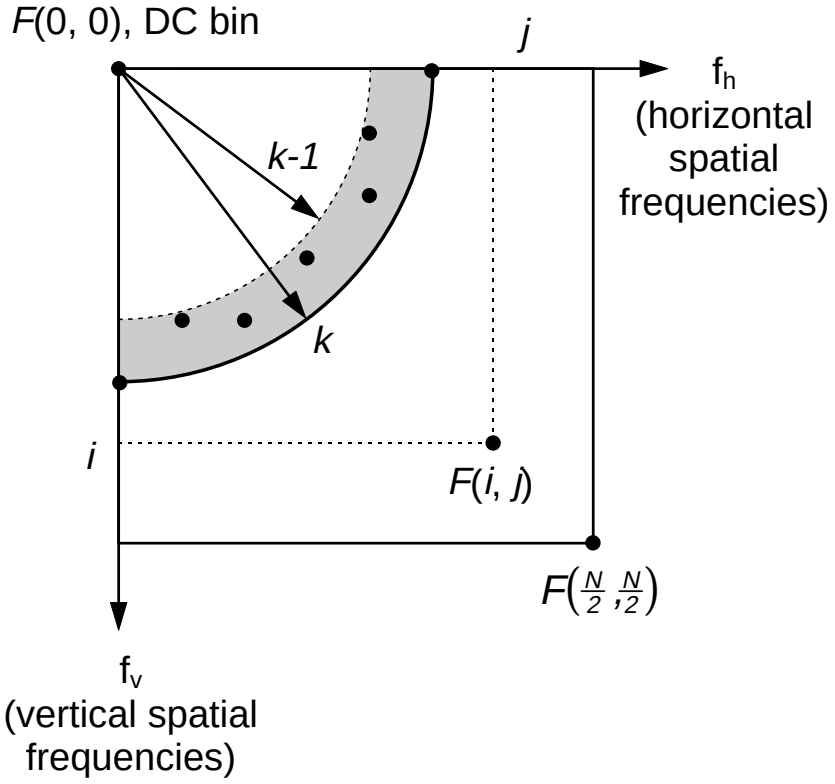


Figure 6 Radial averaging over the Fourier magnitude to produce a vector

Distortions in the output video are detected by comparing the radial averaged vector from the output image with the radial averaged vector from the corresponding input image.

Added noise in the output produces extra high frequency content. Blurring of the output image produces missing high frequency content. Unlike traditional multi-burst measurements, this new frequency response measurement technique can measure dynamic changes in system response as the input scene changes.

3.2.1.4 Example Matrix Features

The entire image can also be used as a reference feature. One well known parameter that is measured from the whole image feature is peak signal to noise ratio (PSNR). PSNR is computed from the error image which is obtained by subtracting the output image from the input image (a standardized method of measurement for PSNR is given in ANSI T1.801.03). Other matrix features and parameters are possible. The spatial information (*SI*) image, illustrated in Figure 4, can also be used as a matrix feature. Parameters based on this matrix feature were first introduced in [15] and applied to a subjectively rated data set in [16].

For the MPEG 1+ and MPEG 2 experiments, two *SI*-based matrix parameters were included in the analysis. These two parameters (Negsob, and Possob), are illustrated and compared with PSNR in Figure 7. The top left image is the input image, the top center image is the spatially registered output image, and the top right image is the error between the input and the output image (i.e., error = input - output). In this case, zero error has been scaled to be equal to mid-level gray (128 out of 255 for an 8 bit display). The bottom left image is the spatial information of the input image (SI_r[input]), the bottom center is the spatial information of the output image (SI_r[output]), and the bottom right image is the error between the two spatial information images (i.e., SI_r[error] = SI_r[input] - SI_r[output]). Once again, zero error has been scaled for mid-level gray. When false edges are present in the output image (e.g., blocks, edge busyness, etc.), the *SI* error is negative and appears darker than gray (Negsob parameter). When edges are missing in the output image (e.g., blurred), the *SI* error is positive and appears lighter than gray (Possob parameter). In this manner, the two types of error can be clearly separated on a pixel-by-pixel basis when both are present in the output image. Note that the enhancement of image artifacts is much greater in the *SI* error image (bottom right) than in the PSNR error image (top right). It will be shown below that these *SI* distortion metrics produce much higher correlations to subjective score than PSNR for the subjectively rated MPEG data sets.

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

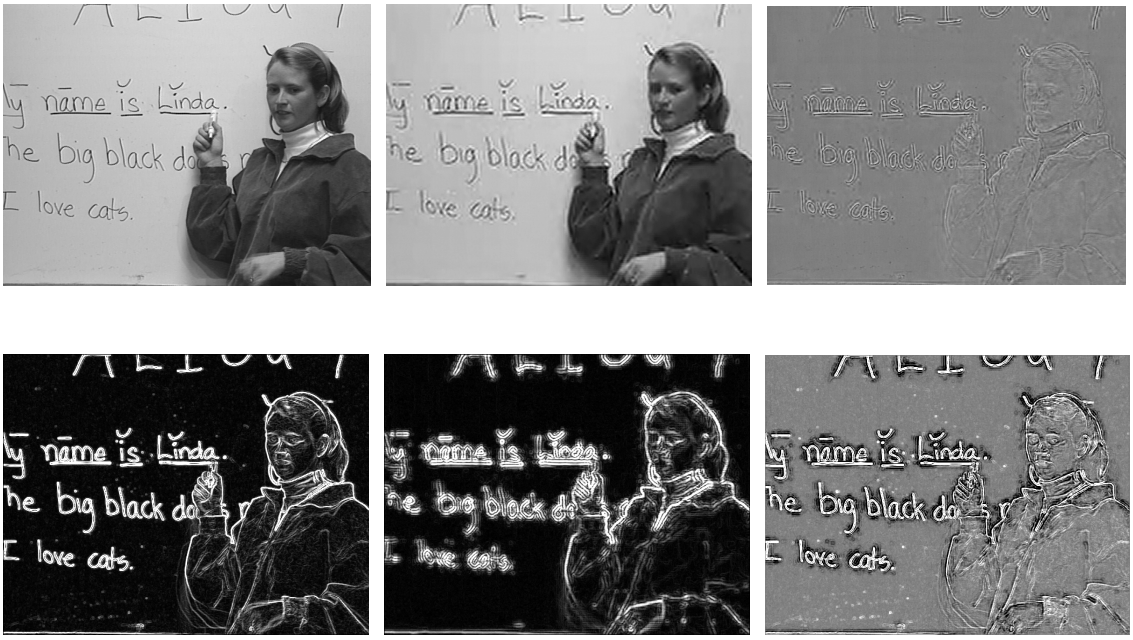


Figure 7 Comparison of *SI* error with PSNR error

3.2.2 Producing Frame-by-Frame Objective Parameters 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.

3.2.3 Temporal Reduction of the Frame-by-Frame Parameter Values

Subjective tests conducted in accordance with CCIR Recommendation 500 [9] 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. ANSI T1.801.03 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.

3.3 Description of NTIA/ITS 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. The system block diagram is shown in Figure 8. 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. In order to avoid clipping the data, 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.

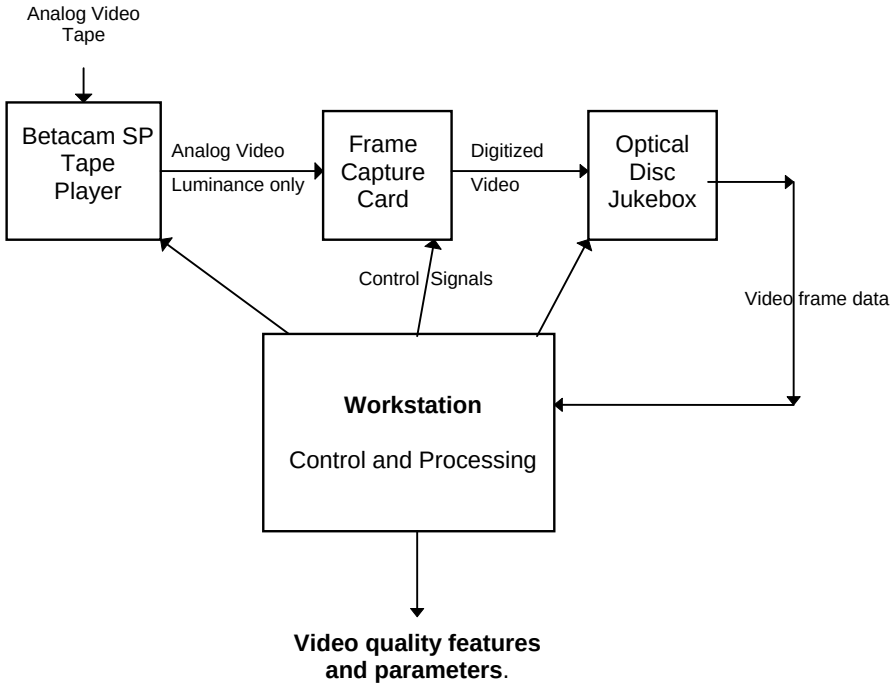


Figure 8 NTIA/ITS Video Processing System

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

This section is included for the benefit of those seeking to implement the image calibration procedures that were used in the current studies. The reader may choose to skip ahead to section 3.5 on page 24.

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 in the ANSI T1.801.03 standard.

ANSI T1.801.03-1996 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 ANSI calibration frames were not included on the source tapes. This section presents an adaptation of the methods in ANSI T1.801.03 for calculating gain, level offset, and active video shift using natural motion video. The method has the added advantage of being able to track dynamic changes in gain, level offset, and active video shift. The method has proven useful for channels that change their calibration characteristics on a scene by scene basis (e.g., an MPEG channel that is re-tuned for each scene to optimize quality).

3.4.1 Overview of Algorithm

The basic calibration algorithm is applied to a single field from the output video stream. For each selected output field, the following quantities are computed:

1. The closest matching field from the input video stream.

2. The estimated gain and level offset between the output field and the closest matching input field.

3. The estimated active video shift (horizontal and vertical spatial shift) between the output field and the closest matching input field.

The interdependence of the above listed quantities produces a “chicken or egg” measurement problem. Calculation of the closest matching input field requires that one know the gain, level offset, and active video shift. However, one cannot determine these quantities until the closest matching input field is found. If there are wide uncertainties in the above three quantities, a full exhaustive search would require a tremendous number of computations. The approach taken here is to reach the solution using an iterative

search algorithm. For robustness, the basic calibration algorithm can be independently applied to several output fields and the results averaged.

3.4.2 Description of Basic Calibration Algorithm for One NTSC Output Field

The basic calibration algorithm for one selected output field is described in this section. The next section discusses how multiple applications of this basic calibration algorithm can be used to track dynamic changes in the calibration quantities or to obtain robust estimates of static calibration quantities.

3.4.2.1 Inputs to the Algorithm

The following is a list of quantities that must be pre-specified in order for the search algorithm to work. The initial search limits should be generous enough to include the correct calibration point. *A priori* knowledge of the transmission channel behavior may be used to help define the initial search limits (e.g., minimum and maximum video delay may be used to specify the range of input fields to search).

1. o_m, the current output field on which to perform the calibration, sampled according to ITU-R Recommendation BT.601 (horizontal extent: 0 to 719 pixels, vertical extent: 0 to 242 active video lines). The image pixel at vertical and horizontal coordinates (v=*i*, h=*j*) will be denoted by $o_m(i, j)$, where (0,0) is the top-left pixel in the image.

2. $\{i_L, \ldots, i_n, \ldots, i_U\}$, the range of contiguous input fields (lower, …, current, …, upper) to examine for a match with output field o_m.

3. *ROI* = {*top*, *left*, *bottom*, *right*}, the input field sub-region (region of interest) over which to perform the comparison, *left* and *right* are in pixels, *top* and *bottom* are in lines. Note: *ROI* may be a manually determined input to the calibration algorithm or an appropriate *ROI* could be automatically calculated (see STEP 1 - Select the Region of Interest).

4. $\{h_L, \ldots, h_s, \ldots, h_U\}$, the range of possible horizontal shifts (lower, …, current, …, upper) of the output field in pixels, where a positive shift indicates that the output is shifted to the right with respect to the input.

5. $\{v_L, \ldots, v_s, \ldots, v_U\}$, the range of possible vertical shifts (lower, …, current, …, upper) of the output field in lines, where a positive shift indicates that the output is shifted downward with respect to the input.

6. *g*, an initial guess for the transmission channel gain as defined in ANSI T1.801.03 (nominally set to 1.0).

3.4.2.2 Comparison Function

Given the above definitions, a variance comparison function for comparing output field o_m to input field i_n is defined as:

$$\mathbf{var}(o_m, i_n, h_s, v_s, g) = \left\{ \frac{1}{P} \sum_{i=top}^{bottom-1} \sum_{j=left}^{right-1} \left[\frac{1}{g} o_m(i+v_s, j+h_s) - i_n(i,j) \right]^2 \right\} - \left[\mathbf{mean}(o_m, i_n, h_s, v_s, g) \right]^2$$

where

$$\mathbf{mean}(o_m, i_n, h_s, v_s, g) = \left\{ \frac{1}{P} \sum_{i=top}^{bottom-1} \sum_{j=left}^{right-1} \left[\frac{1}{g} o_m(i+v_s, j+h_s) - i_n(i,j) \right] \right\},$$

$$P = (bottom - top)(right - left),$$

and h_s, v_s, and g are some hypothesized horizontal shift, vertical shift, and gain of the output field. The point (i_n, h_s, v_s, and g) where the comparison function is minimized is defined as the global calibration point for output field o_m. Using the variance instead of the mean square error for the comparison function has several advantages. One advantage is the reduction of time alignment errors resulting from changes in scene brightness levels. The variance comparison function is more likely to use true scene motion for time alignment of the input and output images rather than changes in scene lighting conditions or transmission channel level offset. The variance comparison function also eliminates the transmission channel level offset from the search, and allows this calibration quantity to be directly computed after the other calibration quantities are determined.

3.4.2.3 Algorithm Description

Figure 9 presents a flow diagram of the search algorithm that is used to find the desired global calibration point for output field o_m. The algorithm uses the following steps which are applied as shown in the figure.

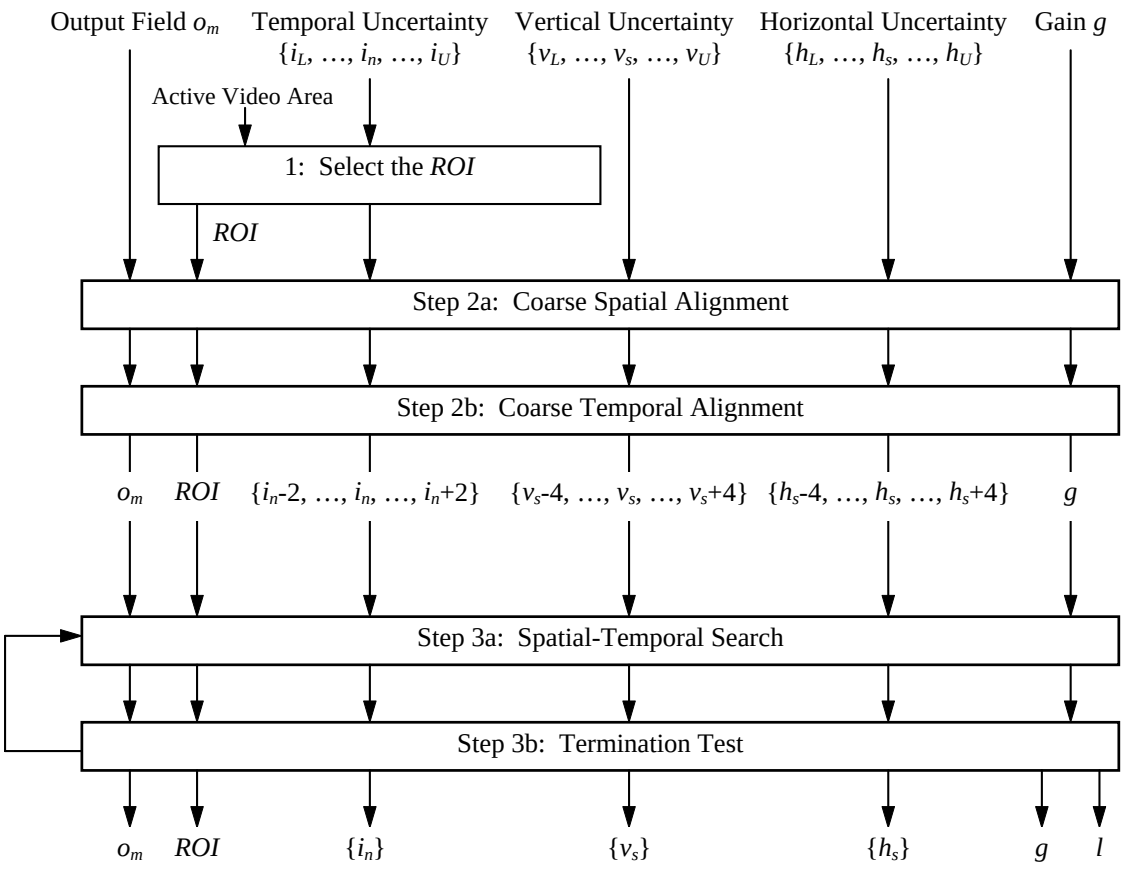


Figure 9 Calibration Algorithm Flow Diagram

STEP 1 - Select the Region of Interest (*ROI*)

The first step is to select a region of interest (*ROI*) upon which to base the comparison function calculations. This is an important step to assure that the comparison function is minimized at the true global calibration point. The *ROI* can be manually or automatically selected depending upon the following important considerations:

1. The *ROI* should be chosen such that it is contained within the active video area. [4]

2. The *ROI* should include both horizontal and vertical edges to assure proper spatial registration of the input and output fields. The spatial information (*SI*) features in section 6.1.1.1 of ANSI T1.801.03 can be applied to the input sequence to determine if horizontal and vertical edges are present.

[4] The active video area is defined in section 5.3 of ANSI T1.801.03-1996 as that rectangular portion of the input active video that is not blanked by the transmission service channel. Technically, the active video area cannot be calculated before the active video shift is known. However, one can choose a conservative *ROI* well within the estimated active video area.

3. The *ROI* should include both still and motion areas to assure proper temporal registration of the input and output fields. The temporal information (*TI*) features in section 6.1.1.2 of ANSI T1.801.03 can be applied to the input sequence to determine if motion and still areas are present.

4. The size of the *ROI* should be carefully considered. Input to output field comparisons will be faster if a smaller *ROI* is selected. Too small an *ROI* might miss important alignment information while too large an *ROI* might create difficulties in temporal registration for scenes that contain small amounts of motion.

5. The *ROI* should contain only the valid scene area or that portion of the input scene that contains picture. For example, the *ROI* should be reduced for scenes that are in the letterbox format.

6. The *ROI* must be no larger than the intersection of the active video area (point 1 above) and the valid scene area (point 5 above), and must account for the horizontal and vertical shift uncertainties (i.e., $\{h_L$ to $h_U\}$, $\{v_L$ to $v_U\}$).

STEP 2 - Coarse Spatial and Temporal Alignment

Since images are often oversampled from Nyquist both spatially and temporally, a coarse spatial and temporal alignment search (i.e., a search that does not include every pixel and field) can be used to effectively reduce the initial spatial and temporal uncertainties (i.e., $\{h_L, \ldots, h_s, \ldots, h_U\}$, $\{v_L, \ldots, v_s, \ldots, v_U\}$, and $\{i_L, \ldots, i_n, \ldots, i_U\}$). The coarse search parameters are selected to be fine enough so that the search algorithm will not miss the global calibration point (i.e., the point at which the comparison function is a global minimum). Coarse registration to within (and subsequent fine registration over) ±4 pixels, ±4 lines, and ±2 fields is sufficient to insure that the desired global calibration point is achieved.[5]

For efficiency, the coarse spatial and temporal search is itself performed as a two step process as follows:

a) Coarse Spatial Alignment

Coarse spatial alignment of output field o_m is performed using the current best guess for the matching input field. The comparison function is computed for: output field o_m, input

[5] The spatial search limits of ±4 pixels and lines are based on scenes with a moderate amount of motion. To assure that the fine registration algorithms converge to the proper input field, these spatial search limits should be chosen to include the maximum amount of motion between two sequential fields (i.e., field 1 and the next field 2). A temporal uncertainty of ±2 fields allows for the possibility of being off by one field of the same type as the current field (for example, consider the case where o_m is an NTSC "field 1", the current i_n is an NTSC "field 1", but the correct input time alignment is an NTSC "field 1" at time location i_n-2).

field i_n (current best guess) [6], horizontal shifts $\{h_L, \ldots, h_s\text{-}4, h_s, h_s\text{+}4, \ldots, h_U\}$, vertical shifts $\{v_L, \ldots, v_s\text{-}4, v_s, v_s\text{+}4, \ldots, v_U\}$, and g equal to the current guess for the transmission channel gain. The horizontal and vertical shifts (h_s and v_s) are updated to that point which minimizes the comparison function. An updated estimate for the transmission gain g is then computed using the calibration equations in section 5.1.2 of ANSI T1.801.03 and the updated spatial alignment.

b) Coarse Temporal Alignment

Coarse temporal alignment of output field o_m is performed using the spatial alignment and gain found in step 2a. The comparison function is computed for: output field o_m, input fields $\{i_L, \ldots, i_n\text{-}2, i_n, i_n\text{+}2, \ldots, i_U\}$, the updated horizontal shift h_s from step 2a, the updated vertical shift v_s from step 2a, and the updated gain g from step 2a. The best matching input field i_n is updated to that field which minimizes the comparison function. An updated estimate for the transmission gain g is then computed using the calibration equations in section 5.1.2 of ANSI T1.801.03 and the updated input field.

STEP 3 - Fine Spatial and Temporal Alignment

Fine spatial and temporal alignment of output field o_m is performed using the coarse calibration estimates and reduced uncertainties (±4 pixels, ±4 lines, ±2 fields) from step 2. The fine search algorithm uses the comparison function to examine all possible spatial and temporal shifts within the reduced uncertainties. The fine search algorithm is applied repeatedly until convergence is reached (i.e., i_n, h_s, and v_s remain the same from one iteration to the next).

a) Spatial-Temporal Search

The comparison function is computed for: output field o_m, input fields $\{i_n\text{-}2, i_n\text{-}1, i_n, i_n\text{+}1, i_n\text{+}2\}$, horizontal shifts $\{h_s\text{-}4, \ldots, h_s\text{-}1, h_s, h_s\text{+}1, \ldots, h_s\text{+}4\}$, vertical shifts $\{v_s\text{-}4, \ldots, v_s\text{-}1, v_s, v_s\text{+}1, \ldots, v_s\text{+}4\}$, and transmission channel gain g. The horizontal and vertical shifts (h_s and v_s) are updated to that point which minimizes the comparison function over the above range of inputs. An updated estimate for the transmission gain g is then computed using the calibration equations in section 5.1.2 of ANSI T1.801.03 and the updated spatial-temporal alignment.

b) Termination Test

The values of i_n, h_s, and v_s at the end of step 3a are compared to their previous values at the beginning of step 3a. If there is any difference, then step 3a is repeated with the new calibration values. Otherwise, stop because the search algorithm has finished. The level

[6] Caution should be observed near a scene cut to assure that input field i_n is the same scene as the output field o_m. One could examine the input sequence for scene cuts using the techniques presented in [17, 18]. These techniques locate large changes, or spikes, in the temporal information (TI) sequences which are indicative of scene cuts.

offset l is then calculated using the current values of i_n, h_s, v_s, g, and the equations in section 5.1.2 of ANSI T1.801.03-1996.

3.4.3 Multiple Application of the Basic Calibration Algorithm

The basic calibration algorithm shown in Figure 9 can be applied to more than one output field.[7] The two primary reasons for doing this are to:

1. Compute more robust estimates of the calibration quantities for static (i.e., not time varying) transmission systems.

2. Continuously update the calibration quantities for transmission systems that change their behavior over time (e.g., the calibration changes from one scene to the next).

When the calibration quantities are static, the calibration algorithm can be applied to multiple output fields o_m (m=1, 2, 3, …, M) and the results can be filtered to produce robust estimates for the gain g, level offset l, horizontal shift h_s, and vertical shift v_s. A median filter is recommended for gain g and level offset l since the median is generally more robust than the mean and not as sensitive to outliers. A mean filter can be used for the horizontal shift (h_s) and the vertical shift (v_s) if one desires to estimate sub-pixel or sub-line shifts in the output image. If nearest pixel or nearest line registration is desired, a median filter should be used.

A digital video system may vary its contrast and color saturation levels over time. This might result from system drift or from scene dependent behavior of the digital coding system. Time varying changes in the calibration quantities can be tracked by repeated application of the calibration algorithm. If filtering of the calibration results is used to produce smoothly varying time estimates for gain g, level offset l, horizontal shift h_s, and vertical shift v_s, this filtering operation should not cross scene cut boundaries.

3.4.4 Calibration Test Results

The calibration algorithm described above was applied to field 1 and 2 of every 30th output video frame (i.e., once per second per field type) from each of the HRCs on the MPEG 1+ and MPEG 2 test tapes. The following observations were noted:

1. There were no significant differences between the calibration quantities for field 1 and field 2.
2. Gain and level offset were not in general constant for an HRC but instead varied dynamically from scene to scene and even within a scene. Scene to scene gain variations on the order of 30% were measured for some HRCs. Smaller within scene gain variations on the order of 10% were measured. The gain and level offset did not vary significantly for the cable simulation HRCs (i.e., SNRs of 34, 37, and 40 dB).

---

[7] For the current MPEG studies, multiple application of the calibration algorithm was used for both of the reasons cited here - see Calibration Test Results section.

However, the VHS record and playback cycle HRC did exhibit dynamic changes from scene to scene. The exact reason for this behavior is not known. It may be due to some form of contrast enhancement being performing by the VCR.

3. Some HRCs had active video shifts that varied from scene to scene (only the horizontal shift contained this variability). However, the active video shift remained fixed throughout a given scene. The reason for this variability is unknown but it may be partly due to the tape editing process that was used to generate the viewing clips.[8]

4. Temporal warping (i.e., variable video delay) of up to 3 video frames was observed for two of the HRCs (MPEG 1 systems operating at 1.5 Mb/sec and 2.2 Mb/sec). These two systems were also the only ones that dropped video frames.

5. Spatial warping (a stretching of the video from right to left by about 28 horizontal pixels) was found on one HRC (an MPEG 2 system operating at 3.0 Mb/sec) for every scene. It is unclear as to the cause of this impairment but a likely source might be a faulty A/D or D/A clock on the codec. For this HRC, the calibration algorithm produced a horizontal shift estimate that wandered randomly around 14 horizontal pixels (i.e., half of the horizontal stretch).

Table 1 gives a summary of the median filtered calibration quantities for 9 of the 10 MPEG systems that were included in the tests (the HRC that horizontally stretched the video is not included in the table). The median filtering was performed over all test scenes for each HRC. The analysis has 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. In particular, note that active video shifts up to 8 horizontal pixels and 9 vertical field lines (i.e., 18 vertical frame lines) were measured.

Table 1 Measured Calibration Quantities for MPEG Systems

MPEG System	Gain, g	Level Offset, l	H shift, h_s (pixels)	V shift, v_s (field lines)
MPEG 1+ 3.9 Mb/s MPEG 1+ Test	.95	-0.2	0	-8
MPEG 1+ 5.3 Mb/s	.96	-0.9	-7	-8

[8] The reason tape editing is suspected for the time varying portion of the horizontal shift is because all HRCs on the MPEG 2 tape (including the VHS and cable simulations) had scene to scene changes. None of the HRCs on the MPEG 1+ tape had dynamic changes to their horizontal shifts.

MPEG 1+ 8.3 Mb/s	.95	-1.4	3	-9
MPEG 1 1.5 Mb/s	1.17	8.3	-7	1
MPEG 1 2.2 Mb/s	1.17	7.7	-8	1
MPEG 1+ 3.9 Mb/s MPEG 2 Test	.90	-3.8	4	-8
MPEG 2 3.9 Mb/s	.98	2.6	-7	1
MPEG 2 5.3 Mb/s	.99	2.0	-7	1
MPEG 2 8.3 Mb/s	.99	2.2	-7	1

In light of the above observations, it was decided to compute a separate gain g, level offset l, horizontal shift h_s, and vertical shift v_s for each clip (i.e., each HRC-scene combination) 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.

3.5 Calculation of Processing Sub-region

For a given scene, the objective measurements were computed over the same video area for each HRC. This area was determined as follows. First, the valid scene area was determined (some scenes were letterbox format) as that portion of the input scene that contained valid picture. Next, the active video area of each HRC was determined (keeping in mind that this active video area is referenced to the input according to ANSI T1.801.03-1996, so that these calculations must remove the active video shift). The

processing sub-region was then determined by the intersection of all the HRC active video areas with the valid scene area. This method provided the largest image sub-region that could be safely used for all the HRCs.

3.6 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 (these methods were first introduced in [15]). 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.[9] 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. Figure 10 presents a pictorial representation of this concept for a 10 frames per second (fps) transmission system. The solid lines give the input and output frame pairs for computation of objective parameters for the constant alignment case while the dashed lines give these pairings for the variable alignment case.

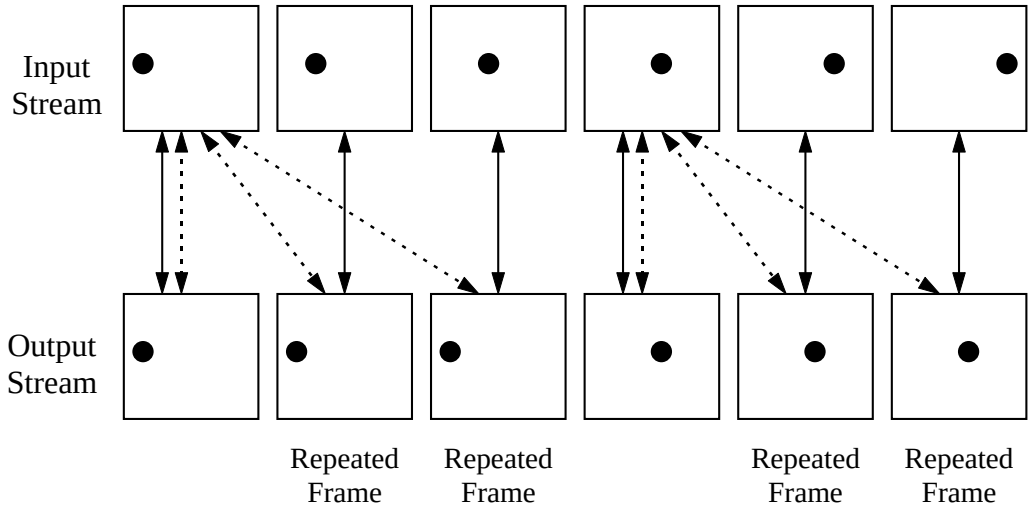


Figure 10 Constant alignment vs variable alignment

[9] One variable alignment method is given by [19], where output frames are categorized as active (i.e., unique or different) or repeated (i.e., same as previous) and the video delays of only the active output frames are estimated.

3.6.1 Constant Alignment (Constant Video Delay)

Section 6.4.1 of ANSI T1.801.03-1996 provides one method for performing constant alignment. This method can temporally align the input and output video streams to a resolution of 1/60 second or one NTSC field. Spatial registration of the input and output NTSC frames (an NTSC frame is composed of two interlaced fields) is used to determine how the output video frame is shifted horizontally and vertically with respect to the input video frame. If a one field time shift is present in the output video (i.e., the vertical spatial shift is an odd number of lines - see note in section 6.2 of ANSI T1.801.03), the output NTSC video framing is shifted by one field. Next, the temporal information (*TI*) features are calculated for the input and output video streams. These two *TI* feature streams, computed at a rate of 30 samples per second, quantify the amount of motion in the input and output video streams. Cross correlation of the *TI* streams is then used to produce an estimate of the constant alignment.

Figure 11 presents a method for directly computing input and output *TI* feature streams at a rate of 60 times per second (this method was first introduced in [6]). An advantage of using this method is that spatial registration is not required in order to achieve an accuracy of 1/60 of a second or one NTSC field. In Figure 11, *TI* is computed separately for each NTSC field type (field 1, field 2) and the results are interleaved to produce a 60 Hz sampling. The standard alignment algorithms given in section 6.4.1 of ANSI T1.801.03 are then used to temporally align the input and output *TI* streams.

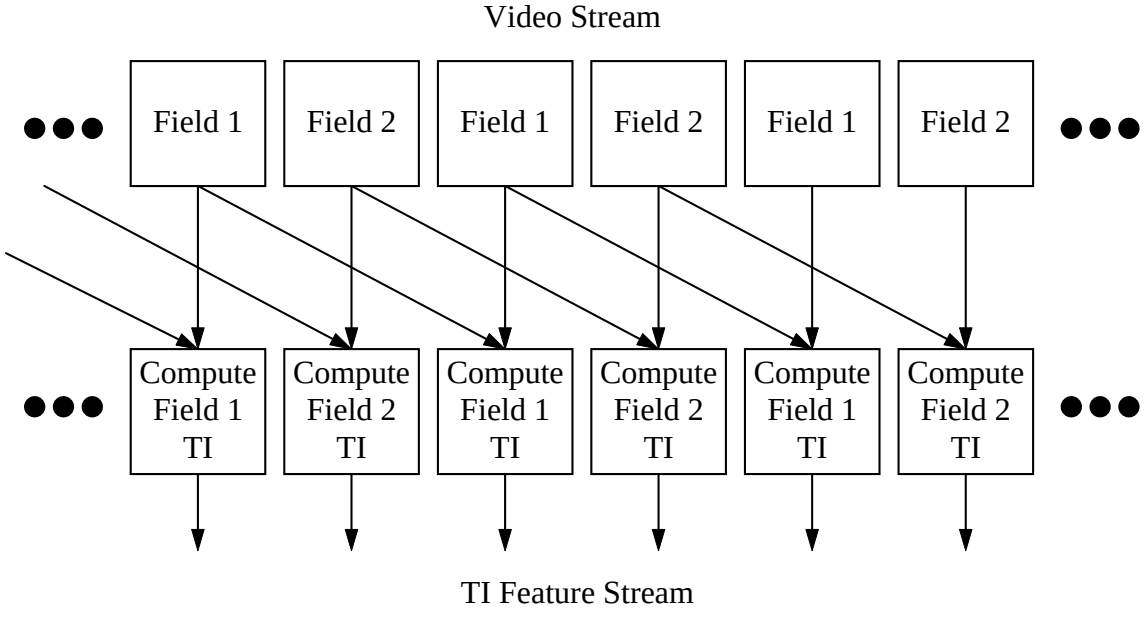


Figure 11 Interleaved fields method for calculating TI

3.6.2 Variable Alignment (Variable Video Delay)

The ITS video quality software is capable of performing variable alignment on each and every output video field. This is accomplished by the use of a minimum MSE matching algorithm to find the best matching input field for every output field. Variable alignment comparisons are based upon NTSC fields rather than frames because an output frame can be composed of two non-sequential input fields. This is illustrated by output frame 3 in Figure 12. The variable alignment results for each field are computed only once and stored for later reference.

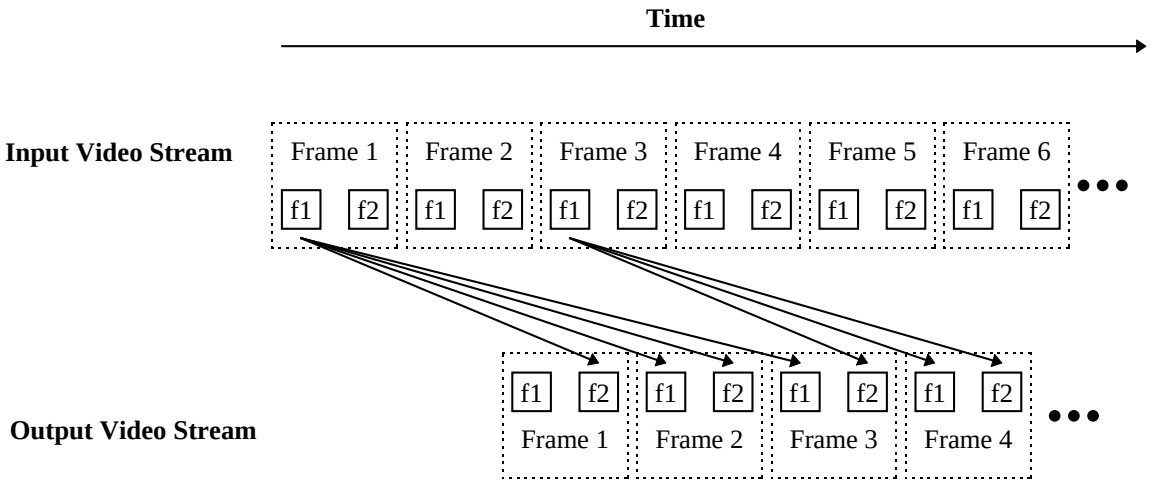


Figure 12 An example of why field comparisons are used for variable alignment

3.6.3 Temporal Alignment Test Results

For high quality NTSC transmission systems like MPEG, the constant alignment method presented in Figure 11 has proven to be an excellent and simple technique for measuring video delay. [10] It has the added advantage of being an "in-service" method of measurement for video delay. For transmission systems that repeat frames, drop frames, or perform temporal warping, this alignment method produces a temporal alignment that reflects the average alignment of the ensemble of output video frames being examined. For the current studies, this alignment technique was chosen as the one to use for computation of the objective parameters.

It was observed that PSNR computed with constant alignment tended to over-penalize the two HRCs with temporal warping and dropped frames. Thus, the use of variable alignment was examined for computation of the matrix objective parameters (i.e., PSNR, Negsob, Possob), since it was thought that precise temporal alignment of input and

[10] 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.

output fields might improve their correlations to subjective score. However, for all three matrix metrics, variable alignment produced objective parameter values with a poorer correlation to subjective score then constant alignment. One possible reason for this behavior seemed to be that variable alignment removed all penalties for temporal warping and dropped frames.

The variable alignment techniques were not able to compute reliable output to input frame matching for the HRC which horizontally stretched the video (an MPEG 2 system operating at 3.0 Mb/sec). However, the constant alignment techniques presented here and in ANSI T1.801.03 were able to determine the correct video delay. The *TI* motion computations used for constant alignment are robust with respect to changes in spatial scaling while the output to input frame matching computations based on mean square error (MSE) are not.

3.7 Summary of Objective Parameters 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	Method of Measurement
711	Section 7.1.1 of ANSI T1.801.03 (maximum added motion energy)
712	Section 7.1.2 of ANSI T1.801.03 (maximum lost motion energy)
713	Section 7.1.3 of ANSI T1.801.03 (average motion energy difference)
714	Section 7.1.4 of ANSI T1.801.03 (average lost motion energy with noise removed)
715	Section 7.1.5 of ANSI T1.801.03 (percent repeated frames)
716	Section 7.1.6 of ANSI T1.801.03 (maximum added edge energy)

717	Section 7.1.7 of ANSI T1.801.03 (maximum lost edge energy)
718	Section 7.1.8 of ANSI T1.801.03 (average edge energy difference)
719	Section 7.1.9 of ANSI T1.801.03 (maximum HV to non-HV edge energy difference)
719_60	Section 7.1.9 using an r_{min} of 60 instead of 20 (maximum HV to non-HV edge energy difference, threshold=60)
719a	Section 7.1.9 using feature comparison function in section 6.5.1.5 (minimum HV to non-HV edge energy difference)
719a_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 (minimum HV to non-HV edge energy difference, threshold=60)
7110	Section 7.1.10 of ANSI T1.801.03 (added edge energy frequencies)
7110a	Section 7.1.10 using modified feature comparison function to sum the missing frequencies (i.e., sum positive part instead of negative part) (missing edge energy frequencies)
721	Section 7.2.1 of ANSI T1.801.03 (maximum added spatial frequencies)
722	Section 7.2.2 of ANSI T1.801.03 (maximum lost spatial frequencies)

732 Section 7.3.2 of ANSI T1.801.03
(minimum peak signal to noise ratio)

733 Section 7.3.3 of ANSI T1.801.03
(average peak signal to noise ratio)

Negsob Mean of the negative part of the input minus output pixel by pixel differences of SI_r values (see section 6.1.1.1 of ANSI T1.801.03), mean $[\text{Sobel(input)-Sobel(output)}]_{np}$ ($[X]_{np}$ defined in section 6.5.1.9) (negative Sobel difference)

Possob Mean of the positive part of the input minus output pixel by pixel differences of SI_r values (see section 6.1.1.1 of ANSI T1.801.03), mean $[\text{Sobel(input)-Sobel(output)}]_{pp}$ ($[X]_{pp}$ defined in section 6.5.1.7) (positive Sobel difference)

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.
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.

4. Subjective Data

4.1 Methods Used to Collect Subjective Data

The method used to collect subjective data was a variant of the method used in the 1994 T1A1.5 multi-lab study [20, 21]: 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 method used in the 1994 T1A1.5 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 "standard" method. 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 specs for the video disc machine are marginally poorer 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 labs employees. They were chosen to be cable TV customers, familiar with the signal quality of cable TV, and also familiar with paying for TV. 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 people participated in both studies, but the studies were separated by nearly a year, more than enough time for people to forget fine details of visual stimuli.

4.2 Summary of Subjective Data

The basic subjective data 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 2 shows the mean ratings for the MPEG 1+ study and Table 3 shows the mean ratings for the MPEG 2 study. The standard errors of the values in Table 2 are on the order of 0.7, and in Table 3 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 2 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 3 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

5. Statistical Analyses

5.1 Methods

5.1.1 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.

5.1.2 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.

5.1.3 Reliability

The reliability issue is important because it limits the statistical fit of even a perfect objective measure (see [22, 23]). That is, if the subjective judgments have noise in them (as we know they certainly will), then even perfect objective measures will not be able to predict the subjective judgments perfectly. The 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 [23]).

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

We used the method of analyzing repeated measurements to compute estimates of the statistical reliability[11] 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 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 (compare Figure 1 with Figure 2), as well as the digitizing and computing (Figure 8).

5.1.4 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.)

---

[11] 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.

5.2 Results

5.2.1 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).

5.2.2 Reliability of objective and subjective variables

Table 4 shows the results of the reliability analyses. The R^2 values each represent the covariance of a variable with respect to itself across the two studies, divided by the variance of the variable (i.e., mean square) pooled across the two studies (see [23]). Each reliability was computed from 80 data points (eight scenes by five HRCs in each of the two data sets). In the case of the subjective ratings, each of the 80 data points is the mean of the ratings of 30 consumers.

Table 4 Reliability of objective and subjective measures of video quality across two studies, proportion of variance accounted for

Measure	Reliability
711	0.995

7110	0.769
7110a	0.921
712	0.952
713	0.995
714	0.793
715	No variation
716	0.934
717	0.910
718	0.922
719	0.994
719a	0.982
719_60	0.989
719a_60	0.990
721	0.979
722	0.945
732	0.942
733	0.956
Possob	0.981
Negsob	0.982
Subjective ratings	0.890

Note that the reliability of the subjective ratings here is apparently somewhat higher than that reported for the three-lab T1A1.5 study [20]. We say "apparently" because the designs of the two studies were quite different. In the T1A1.5 study there were very few repeated trials, and these trials were not distributed in a way that promoted averaging across subjects. Therefore, the T1A1.5 reliability of 0.84 for subjective judgments may have been artificially low because it was based on data for individual subjects.

Comparing Figure 1 with Figure 2, one can see that differences in the MPEG 1+ and MPEG 2 objective measurement paths included differences in manufacturer's equipment, differences in transcoding, and differences in tape generation. With each tape generation and transcoding, SNR and frequency response decrease slightly (frequency responses are concatenated, and hence multiplied to find the total response). Some of the objective parameters (i.e., 714, 717, and 7110) seem to be more sensitive to these analog processing differences than others.

5.2.3 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.

5.3 Interpretation of results

5.3.1 Which measures work best

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

After Negsob, the ability to predict increases with the addition of another two or three variables. Exactly which ones are 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.

5.3.2 Comparison to results of T1A1.5 multi-lab study

The T1A1.5 multi-lab study [20] used objective measures of video quality to predict subjective judgments. How do the results of the two studies compare? The answer is: The results of the two studies generally agree with each other, but the studies differed enough that it is difficult to say precisely how well the results agreed. We consider three areas of possible agreement between the two studies, (a) how well objective parameters predicted subjective judgments, (b) which objective parameters predicted best, and (c) the array of HRCs and scenes used in the two studies.

(A) How well. The degree of statistical fit was very similar in the two sets of studies, as described in Section 5.2.3. The statistical fit in the current study was somewhat better.

(B) Which parameters. In the T1A1.5 study, the best-fitting group of objective parameters were identified as P6, P9, and P13. P6 was a measure of lost motion, somewhat comparable to the measure called 714 in the current study. P9 was a measure of change in edge energy; several measures in the current study were related to changes in edge energy, including Negsob and the 719 family. P13 was a measure of change in spatial frequency, which is related to the 721 measure in the current studies. The measures 714, 719, and 721 were good predictors, and models involving sets of multiple parameters usually included something like them.

However, the T1A1.5 study had no measures comparable to the Negsob and Possob measures in the current study. (The reason is that computing these measures requires massive data storage that became available to us only recently). Negsob in particular is highly correlated with both the 719 family of measures and with 721. And, Negsob is more highly correlated with the subjective judgments than any other objective measure we tried. In the regression models Negsob assumes primary importance. Therefore, two of the measures that correspond most nearly with those from the previous study are essentially subsumed by Negsob and are pre-empted by Negsob from appearing in their own right. Qualitatively, though, the results of the T1A1.5 study and the current studies are quite similar.

(C) HRCs and scenes. The current studies were designed to be complementary to the T1A1.5 multi-lab study. The HRCs used here fit into the gap in the T1A1.5 study between 1.5 and 45 Mb/s, and the scenes were chosen to be appropriate for the HRCs. However, the studies differ in two very important ways.

- One is that the current study did not include the great variety and severity of impairments present in the T1A1.5 study. In the current study even the least

powerful HRCs were nearly able to handle the video material. By contrast, in the T1A1.5 study there were many glaring mismatches between low bit rate HRCs and the video material they were asked to handle. The various objective measures are designed to be sensitive to particular artifacts and impairments. Since the present study and the previous study differ in terms of the number and severity of impairments, one would expect differences between the two studies in the sorts of objective measures that performed well.

- The second major difference between the studies is in the range of variables, which affects the fit of regression models generally. Since the T1A1.5 study covered a great range of HRCs, bit rates, and video material coding difficulty, it also covered a great range in both objective and subjective measures of video quality. The present study probably covered a smaller range of variables (although it is difficult to be certain because the measures were not identical). Larger ranges of variables lead to better statistical fits of regression models, other things being equal. Therefore, the fit of objective to subjective data should have been better in the T1A1.5 study (other things being equal). The fact that the objective measures predicted the subjective measures better in the current study is impressive, given the smaller range of variables.

5.3.3 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, recall that in the large T1A1.5 study ([20], 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 ([23], 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.)

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 0.845, the observed 0.763 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.

5.3.4 What "should" our measures measure?

The MPEG data sets included one HRC with a unique impairment that has stimulated discussion and influenced the thinking on how video performance parameters should be developed and related to human perception of impairment. This particular impairment was a smooth horizontal stretching or scaling of the output video by about 30 pixels. This impairment may have resulted from an A/D sampling clock or a D/A output clock that was operating at the wrong frequency. However the impairment was introduced, the important observation was that the naïve viewers (and subsequent expert viewers looking at the output video) did not notice the impairment at all! It was only detected in the laboratory after the calibration programs failed to produce reliable spatial and temporal registrations of the sampled output and input video images. As previously described, these calibration programs utilized mean square error between the output and input images to determine the best spatial and temporal alignment.

The preceding observation raised the question as to whether objective video quality parameters should detect such an impairment. One line of reasoning dictated that since the viewer is unable to detect the impairment, the objective parameters should respond similarly and not be sensitive to the impairment. Advocates of this school of thinking believe that the ultimate goal of objective parameters is to replace subjective assessments. A different line of reasoning argued that the objective parameters *should* detect the impairment even though the viewers did not because some users would certainly want to know if their output video was undergoing such a distortion. In particular, it was felt that a technician responsible for maintaining a large broadcast network would gasp at the thought of such an impairment and would desire to isolate and replace the faulty piece of equipment.

Clearly, there appears to be a need to develop a hierarchy of objective measurements with differing degrees of sensitivity to subjective judgments of distortion. This hierarchy should encompass the concerns of end-users as well as service providers and provide the flexibility to specify performance parameters for a wide range of systems and applications. Figure 13 gives a first attempt to describe this hierarchy. The horizontal axis at the bottom of the figure is an attempt to represent, on a one dimensional scale, all of the perceptual dimensions of video quality. For illustration purposes, examples of independent perceptual dimensions of video quality might be spatial resolution, temporal resolution (e.g., transmitted frame rate), and color fidelity. Experiments should be run to determine these perceptual dimensions of video quality and objective parameters should be developed to independently quantify each of these dimensions.

This approach is preferable to developing one global parameter that is sensitive to all dimensions of video quality since the end-user and service provider are given maximum flexibility for specification of system attributes that match the targeted application. For example, medical imagery might require very high spatial resolution while desktop video teleconferencing might only require moderate spatial resolution to be subjectively rated "excellent" quality. The vertical axis in the figure specifies the degree of sensitivity of the objective measurements to subjective judgments of distortion. At the lowest level (detectable), an objective parameter is detecting distortions that are not visually perceptible. This level of sensitivity would be useful for maintenance and monitoring of

system performance and would address the concerns mentioned earlier regarding the technician and the impairment that horizontally stretched the video.

The next level (Perceptible) would take the human visual system into account in that impairments which are detectable but not perceptible would be filtered out by human visual system filters. For example, several dropped video frames that are detected by a temporal resolution metric might be filtered out if they occur immediately after a scene cut or a point where the human visual system is known to exhibit reduced sensitivity. Note that this approach of filtering each perceptual dimension of video quality (see Figure 13) is somewhat different than that proposed in [24], where the approach is to develop one global parameter that is sensitive to all dimensions of video quality.

The third level (Meaningful) would allow only those perceptible impairments that affect the end-users' perception of quality. This level is the hardest to quantify since higher order judgmental processes in the brain are involved. The human judgmental filters used to transform perceptible measures of distortion into meaningful measures of distortion must take into account the targeted application and viewer population. For example, the expectations of broadcasters using a "contribution" quality system (i.e., a system used to transport very high quality video from one studio to another) are much more rigorous than the expectations of someone using a "freebie" Internet videophone.

Most subjective experiments are designed to produce results that are meaningful and thus these experiments include higher order judgmental processes that are not very well understood or controlled. This could explain why the "random" or unexplained variance between two identical sets of video clips rated in separate subjective experiments appears to be significantly larger than what might be expected from the confidence intervals calculated on the basis of only one experiment. It is not clear how much of this higher order information the objective quality parameters should endeavor to explain.

In Figure 13, the dimensions of video quality are assumed to be independently quantifiable by objective parameters. However, in reality there will always be some degree of dependence between the objective parameters. Additional processes could be introduced to account for these interactions or to transform dependent objective parameters into a set of orthogonal or independent objective parameters.

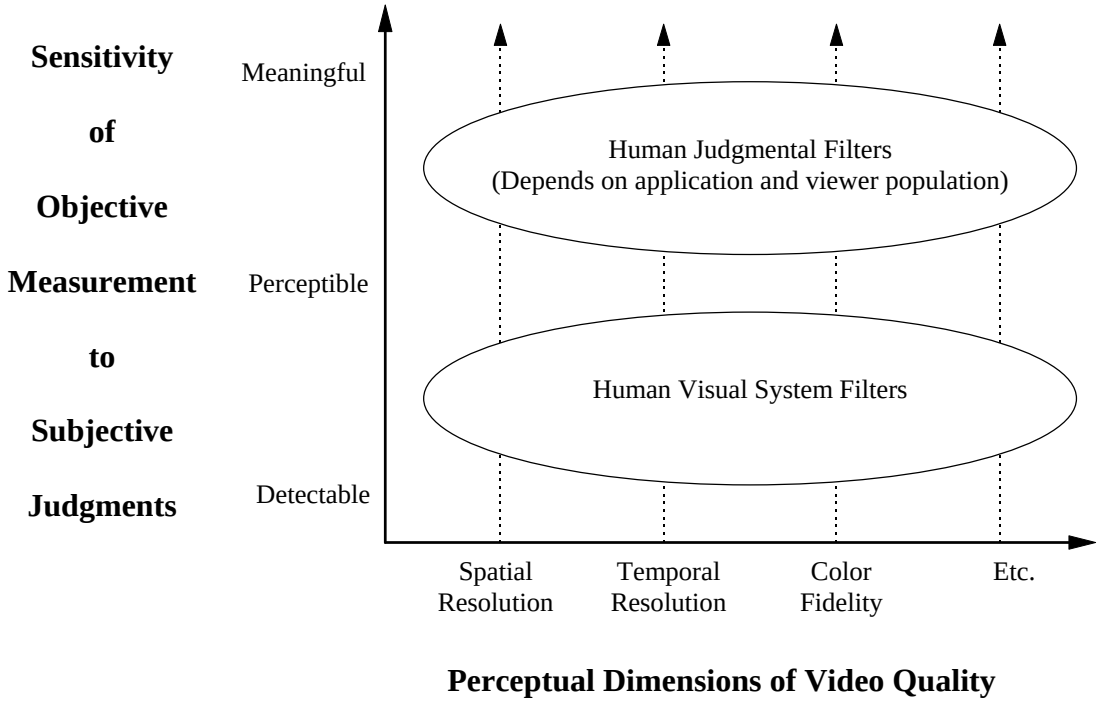


Figure 13 Hierarchy for Developing Objective Video Quality Parameters

6. Future Directions

6.1 Solving the Time Collapsing Problem

The time collapsing problem was mentioned briefly in section 3.2.3. This is the process that is used to time collapse frame-by-frame objective parameter values to produce a composite value that reflects perceived quality changes. Subjective tests conducted in accordance with CCIR Recommendation 500 produce one subjective mean opinion score (MOS) for each HRC-scene combination (normally a 10 to 15 second period). However, consumers in the home viewing environment are continuously noticing quality changes in their received picture. Recent advances in subjective testing methodology [25] are being developed that would allow continuous (in time) subjective sampling. These techniques give the viewer freedom to rate the perceived quality "on-the-fly" by adjusting the position of a slider. The position of the slider is sampled several times per second to yield a continuous stream of subjective scores. When these subjective techniques are standardized, they may be used to collect continuously sampled subjective data, which, together with the continuously sampled objective data, can be used to develop better time collapsing functions.

6.2 Measuring Distortions in the Chrominance (Color) Signal

The objective measures are able to account for a surprisingly large amount of the subjective variance using only the luminance portion of the video signal. It is reasonable that some portion of the unexplained subjective variance is due to color distortions. Investigations into psychometrically uniform color spaces such as those developed by the International Commission on Illumination (CIE) [26] might produce useful objective measures for distortions in the chrominance or color signal. These perceptually uniform color transformations could be used in conjunction with the measurement methodologies presented in ANSI T1.801.03 to investigate color distortion metrics.

Studies of subjective response to color distortion will be especially important: Recent results indicate that consumers may not only tolerate color distortion in video, but actually prefer some kinds of color distortion! This kind of result is most clearly seen in studies that include VHS as an HRC [1, 2, 20]. Every example of VHS we have tested with consumers has had quite distorted color and yet surprisingly good ratings. (Readers familiar with the multi-lab T1A1.5 study [20, 21] should recall how the brightly-colored blouses in the scenes "vtc1nw," "vtc2mp," and "vowels" appeared in the VHS, Null, and 45 Mb/s HRCs.)

6.3 Building Objective Video Quality Models

For the purposes of this paper, a video quality model is defined as a mapping of the objective parameter values to a single figure of merit, such as an estimated subjective mean opinion score (MOS). Building objective video quality models involves conducting simultaneous subjective and objective tests and determining how objective parameter values can be used to predict the subjective viewer responses. This model building process is necessary for determining the overall accuracy of the objective parameters and for identifying the portion of the subjective responses explained by the objective parameters. However, developing useful video quality models for operational video transmission systems is a much more complicated process.

Two simple examples illustrate the complexity involved. For the first example, consider two different applications: transmission of high-resolution graphics imagery with pointer capability, and transmission of sign language. Here, these two fundamentally different applications require different performance characteristics for the various dimensions of video quality (e.g., spatial resolution, temporal resolution, or color reproduction accuracy). The graphics application requires very high spatial resolution with low frame rates while the sign language application requires high frame rates at a lower spatial resolution. An objective model that produces overall quality estimates from a set of fundamental objective parameters would have to account for these application-specific effects.

For the second example, consider two different viewer populations: the naïve or non-expert viewer, and the critical expert viewer. In this case, the expert viewer may tend to downgrade the quality more than the naïve viewer for the same amount of video impairment. For an actual example that illustrates this viewer population effect, see [6]. Precisely the opposite effect of expertise has also been observed: In the GTE Labs portion of the multi-lab T1A1.5 study, half the viewers were quite experienced video

teleconferencing users. These experienced users were more forgiving of the teleconferencing HRCs in the study [20, pg. 22].

Another influence on modeling accuracy is the changing expectations of people over time. This is particularly true for digital video systems where the technology is improving rapidly and the cost is decreasing rapidly.

For these reasons, objective video quality modeling is valid only if the application and viewer population are well defined. Given sufficient time and effort, objective parameters can be used to develop effective video quality models for a large number of video applications and viewer populations. References [6, 16, 27] present several such preliminary models for various applications. The subjective and objective data in this paper can also be used to produce a model of video quality for broadcast applications.

6.4 Setting Guaranteed Levels of Service

Many users want assurances that they will receive some guaranteed level of service. This will involve the determination of appropriate thresholds for individual parameter values or video quality model outputs. Different thresholds could be used to define levels or grades of service (e.g., low, medium, high). If the thresholds are violated, the user will have a mechanism to resolve the problem. It is expected that the establishment of standard threshold levels will require a multiyear effort and cooperation by manufacturers, carriers, and users.

6.5 Determining subjective dimensions of video quality

The technical community recognizes a number of quality impairments for compressed video [see 12, 27]. These impairments correspond to what Section 5.3.4 and Figure 13 refer to as the Perceptible layer of a hierarchy of objective measurements. Work that needs to be done at this level of the hierarchy is of three types:

- Systematize the perceptual impairments. The impairments are certainly related to each other by means of common causes in a video system. The perceptual impairments are probably related to each other in subjective terms as well. The subjective terms, or "dimensions," that relate the various impairments do not necessarily map one-to-one with the objective causes of the impairments. Finding the subjective dimensions that unite and relate the many specific perceptual impairments will help in understanding the relationship between objective video quality and subjective video quality.

- Catalog novel impairments. Having a system for cataloging impairments should facilitate recognizing a novel impairment when it appears, just as having a system for cataloging stars or a system for cataloging butterflies helps in detecting when a new one has been discovered.

- Relate the subjective dimensions to objective measures. Objective measures currently are used to predict a single overall subjective quality judgment. As discussed in Section 5.3.4, a finer level of prediction would be useful in order to take account of specific contexts in which video is used. The subjective

dimensions of video quality would probably be at about the right level of analysis to handle context and usage effects.

6.6 Comparing subjective video quality to overall subjective value

Section 5.3.4 discusses even more high-level video attributes or impairments under the title "Human Judgment Filters." Although Committee T1 may not claim jurisdiction over some of the higher level attributes of judgment, we note that many of the member organizations will ultimately have to be interested in these higher-order variables. Examples are

- Overall subjective value of a video system to a consumer
- Willingness to pay for a video system
- How important picture quality is for the consumer compared to other system attributes in determining overall system value -- attributes such as usability of the user interface, number of channels, quality of content available, price, sound quality, and even cabinet design.

Although some work has been done relating video quality to other high-level variables such as audio quality [28], much more work remains to be done.

7. Conclusions

- The current generation of objective video quality measures has achieved good prediction 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 must 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.
- Areas in which further work is recommended:
  - Time collapsing: Integrating a time series of quality measurements to form a single overall measurement,

- Computing resources; the best current objective measures require a lot of computer resources,
- The effect of color fidelity on overall judged video quality,
- Assessing objective measures in specific domains, such as teleconferencing vs. entertainment,
- Discovering unifying dimensions of subjective quality/impairment, as opposed to specific individual impairments,
- Relating picture quality to overall value of a video system.

8. References

[1] Cermak, G. W., Teare, S. K., Tweedy, E. P., and Stoddard, J. C. (1996a), "Consumer acceptance of MPEG2 video at 3.0 to 8.3 Mb/s." Paper for presentation at SPIE conference, Nov. 1996 and for publication in conference proceedings.

[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. [12]

[3] CCIR Recommendation 654, "Subjective quality of television pictures in relation to the main impairments of the analogue composite television signal," Recommendations and Reports of the CCIR, 1986.

[4] CCIR Report 405-5, "Subjective assessment of the quality of television pictures," Recommendations and Reports of the CCIR, 1986.

[5] CCIR Report 959-1, "Experimental results relating picture quality to objective magnitude of impairment," Recommendations and Reports of the CCIR, 1986.

[6] S. Wolf and A. Webster, "Objective and subjective video performance testing of DS3 rate transmission channels," ANSI T1A1 contribution number T1A1.5/93-060, Apr 1993.

[7] 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.

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

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

[12] 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.

[10] 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).

[11] 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.

[12] 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.

[13] 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.

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

[15] Stephen Wolf, “Features for Automatic Quality Assessment of Digitally Transmitted Video,” NTIA Report 90-264, US Department of Commerce, June, 1990.

[16] Stephen Voran, “The Development of Objective Video Quality Measures that Emulate Human Perception,” 1991 IEEE Global Telecommunications Conference, Phoenix, Arizona, Dec 2-5, 1991.

[17] Stephen Wolf, Margaret Pinson, Coleen Jones, Arthur Webster, “Objective Video Performance Testing,” ANSI T1A1 contribution number T1A1.5/93-152, Nov 1993.

[18] Stephen Wolf, Margaret Pinson, “Corrections and Extensions to T1A1.5/93-152,” ANSI T1A1 contribution number T1A1.5/94-110, Jan, 1994.

[19] Alfred Morton, “Draft ANSI T1 Standard on Multimedia Communications Delay, Synchronization, and Frame Rate Measurement,” ANSI T1A1 contribution number T1A1.5/96-101, Oct, 1996.

[20] 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.

[21] Alfred Morton, “Subjective test plan (ninth draft),” ANSI T1A1 contribution number T1A1.5/94-118, Oct, 1993.

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

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

[24] Jeffrey Lubin, “A Visual Discrimination Model For Imaging System Design and Evaluation,” Report published by David Sarnoff Research Center, 1995.

[25] N.K. Lodge and D. Wood, “New Tools for Evaluating the Quality of Digital Television - Results of the MOSAIC Project,” 1996 IEEE Broadcast Symposium, Washington DC, Sept 26-27, 1996.

[26] International Commission on Illumination (CIE), “Recommendations on Uniform Color Spaces, Color Difference Equations, and Psychometric Color Terms,” Supplement No. 2 to CIE Publication No. 15 (E-1.3.1), TC-1.3 Colorimetry, Bureau Central de la CIE, Paris, France, 1978.

[27] B. Cotton, “An objective model for video quality performance,” ANSI T1A1 contribution number T1A1.5/96-105, Mar 1996.

[28] B. Cotton, "Combined A/V Model With Multiple Audio and Video Impairments,” Contribution ITU-T, COM 12-54-E, Question 22/12, June, 1995.