HOW TO TEST COMPRESSED DIGITAL TELEVISION

Guy W. Beakley, Ph.D.
StellaCom, Inc.
Arlington, VA 22209

Abstract

Conventional analog video test measurements are not adequate for compressed video signals. This is because digital video distortion and artifacts often are nonlinear, discontinuous and depend on picture content. Most analog measurements assume that errors are of a continuous, linear nature. The only alternative to date has been subjective testing. Formal subjective tests (e.g., CCIR 500) can provide reliable, relative measures of video quality. However, such testing is time-consuming and expensive. Objective testing methods are needed to provide efficient, repeatable measures of video quality. Presently, no standardized objective measures exist for digital video. We have implemented measures from a number of sources and others of our own design on a low-cost workstation. These measures utilize complex digital image processing techniques to analyze differences between source and processed video sequences. This paper presents some of these measurements and describes our implementation of an automated system to capture and test digital video quality.

I. Why Compress Television?

Television is rendered in digital form for the inherent quality and lack of degradation in replication and transmission. It is compressed so that it can easily be transmitted and stored. For example, the Armed Forces Radio and Television Service (AFRTS) intends to procure and operate a compression system for the delivery of six video channels with associated stereo audio, three to six additional stereo audio channels, 1.5 Mbps T-1 data and two FM talk radio channels via 36 Mhz satellite transponders to AFRTS downlink locations worldwide. Before compression they transmitted only one video and four audio channels in a transponder.

II. Why Test Video Quality of Digital Television Systems?

The limitations of standard video testing methods for the evaluation of codecs have been demonstrated so frequently that an alternative is often employed. The codec evaluator simply looks at the video using a favorite test tape and makes a judgment based on experience and intuition. For this to be effective the test tape has to be extensive and contain a number of scenes that may be difficult for the codec to process. In addition, the codec evaluator must have the correct viewing conditions. The only alternative has been to assemble a panel of viewers and run subjective tests under controlled conditions.

Objective measures, which are repeatable and do not depend on viewing conditions or the mood of the viewer, need to be used in addition to subjective testing to compare compression systems. Codecs not only remove redundant information (lossless compression), but also modify the picture (lossy compression). Unfortunately, the picture distortion may result in perceptible impairments. Objective measures aid the codec evaluator in identifying and understanding codec artifacts. The complex processing of today's codecs requires complex testing that can be done only with a computer[1]. Also, the computer allows flexibility for objective test evolution as standards evolve for testing codecs.

Computer tests[1] can be done for spatial and temporal distortion, horizontal, vertical, and diagonal resolution, image build-up, bit-error-rate and various detail and motion impairments. These tests are based upon work done by NASA, NTIA, ACATS, CRC and others to quantify the quality of digital video codecs. The computer is used to predict how the average viewer rates a sequence on the CCIR-500 Impairment Quality Rating scale. The measures have been correlated with subjective testing by the NTIA. In addition, an absolute error signal is calculated, so that the computer user can see what has actually been lost or gained by codec processing. The computer allows the user to view the original video, codec processed video and error signal side-by-side with the error measures. These measures are also useful for analyzing bit-error and concatenation effects. In fact, the impairment quality rating measure has been correlated with subjective tests that include impairments due to bit errors and concatenation.

While conventional analog video measurements are needed, they prove little about a codec's performance with complicated motion. Dynamic tests are essential because most high-performance codecs use motion compensation as part of the compression algorithm. The need to measure resolution in several dimensions is new to the testing of digital codecs. The standards for such tests are not yet established. The use of moving zone plates to measure resolution is successful in revealing information not obtained from any other test. Other measures not widely known, such as spatial distortion and temporal distortion, are also necessary for describing a codec's performance.

The only way to completely test a codec is to play a wide range of real video materials through it. Test materials must be selected to represent the full range of video content for which a particular codec (or class of codecs) is intended. Test signals and related measurements can still serve a purpose, but their utility is more limited than with analog systems for the reasons outlined above. The tester needs to compile an extensive test tape consisting of many short video sequences, play the video through the compression system, and record the video output. Then, the output and input video can be compared, both visually and by computer analysis.

A combination of subjective and objective test methods is the most effective way to test codecs. Subjective methods are most useful for rapid, qualitative assessment of video quality. Formal subjective testing[2] is useful for quantitative quality assessment but is often impractical due to manpower, cost and time considerations. Objective testing with computer-based measurements can provide repeatable, quantitative results for comparisons of video quality. A suite of comparative distortion measurements is needed to fully characterize video quality. Statistics can be compiled automatically for large amounts of video and the computer can be programmed to identify poor-quality sequences for further analysis (subjective and objective) and for documentation. Moreover, the objective methods described below provide additional insights into codec operation and impairments that cannot be easily discerned using subjective methods.

An important benefit of testing with real video is that common source materials can be used for both subjective and objective testing. Many codec developers and end users already have their own *killer* test tapes for codec evaluation. The Ad-hoc Group on MPEG-2 Verification Testing[3] has assembled a set of six test sequences of 525-line/60-fields/sec video plus a few 24-frames/sec film-based sequences for subjective quality testing. The National Telecommunications and Information Administration, Institute for Telecommunication Sciences (NTIA/ITS) uses a set of 36 sequences[4] for their work with the Interexchange Carriers Committee on video

quality (T1A1.5). Standard libraries exist at the CCIR as well. All such video materials can be used directly for objective quality measurements as described below.

There are many requirements for testing the quality of compressed digital video. The codec manufacturer's objective is to produce the best picture quality at a minimum encoded data rate. End users need to compare the video quality of competing codecs for purchase decisions. In addition there are a number of reasons for continual testing. One reason is identification and documentation of picture artifacts. Video quality varies with scene content. Users will want to document troublesome scenes and feed these back to the codec manufacturer. Video quality also varies with the use of a statistical multiplexer. A low-priority channel may not have the bit-rate it needs for good picture quality if the high-priority channels are carrying sports programs. What happens if the network degrades because of attenuation of carrier power? What happens when the packets are grossly out of order? Also, compression can be used at the signal origination or at the destination in addition to the communications or storage system. Then concatenation effects become important and should be tested. During the early stages of compression the vendors will make encoder improvements and there will constantly be new codec vendors with "better" products. These claims will need to be verified.

It is commonly believed that MPEG-2[5] will solve most of the users' video quality issues. However, MPEG-2 is a flexible syntax or tool kit allowing over thirty variables that affect picture quality. The use of MPEG does not guarantee picture quality. Examples of MPEG variables affecting picture quality include data-rate, use of I-, P- and B-frames, allocation of bits to I-, B- and P-frames, field/frame adaptive prediction, method and range of motion estimation/compensation, slice size, type and scale of quantizer, downloadable quantizer matrices and buffer size. Also, many different error detection, correction and concealment strategies can be employed in an MPEG decoder. Furthermore, filtering and resampling can occur before and after the encoding/decoding process. Variability also exists in the quality of the NTSC encoding/decoding process. MPEG-2 provides a flexible framework for allowing varying picture quality depending on communications and storage requirements and the intended application.

2.2. Why analog tests are not sufficient

Codecs employ complex algorithms that incorporate discrete spatial and temporal processing. They exhibit non-linear responses to changing scene content and operating conditions. Typically, codecs introduce distortion into the picture. Also, artifacts can be introduced in any of the encoding, transmission and decoding stages of a digital system. Analog measurements are not intended to characterize these types of systems and phenomena. Most analog tests employ predefined test signals and are intended to measure continuous, linear response characteristics of analog devices. For digital systems, the response to a particular test signal only describes the performance of the system for the test signal and often has little or no relation to how the system will perform on real video. For example, the commonly used method for determining signal-to-noise-ratio (SNR) employs a static, flat gray input signal. Most DCT codecs can produce very high SNR of 60 dB or higher on static scenes -- but SNR on real scenes depends on internal quantization levels for the DCT coefficients as well as motion prediction and other coding techniques. Many codecs tested at StellaCom have exhibited significant variation in SNR depending on the pedestal level used to make the SNR measurement. A measurement that varies so much with gray level is not useful for measuring the quality of codecs. In addition, typical noise and distortion for real scenes are higher than

SNR measurements imply and can vary widely by scene and by codec. Few conventional analog video measurements are useful for testing codecs because they prove little about a codec's performance with complicated motion. Dynamic tests are essential because most high-performance codecs use motion compensation as part of the compression algorithm.

Our testing system incorporates a suite of distortion measures and viewing tools on a graphics workstation. The Impairment Quality Rating (*IQR*) measurement developed by Wolf and others at NTIA/ITS,[6,7,8], is used to predict how an average viewer rates a sequence on the CCIR-500 impairment quality scale. The *IQR* measurement has been correlated with subjective testing using statistical methods. In addition, a digital error signal is computed. It can be viewed as video and used for statistical measurements such as RMS Error and Signal-to-Error Ratio (*SER*). The testing system allows the user to view video (source, codec-processed video and error signal) side-by-side with plots of measurement results on the computer display. These and other measures are described below. An important consideration is that the measures be as general as possible. The approach is to treat the encoder, transmission system, storage system and decoder as a single black box system that can introduce both analog and digital distortions. Analog distortions are typically introduced in the processes of analog-to-digital and digital-to-analog conversion, prefiltering, and NTSC encoding and decoding. The encoder can introduce analog and digital distortions as well as encoding artifacts. Transmission errors or *bit errors* for digital systems can also cause artifacts. Error correction and concealment processing in the decoder will ultimately determine the output video quality in the presence of such bit errors. Concatenation effects can be tested by analyzing the recorded output of multiple passes through the system.

III. Computer Measurements

A. Video Content Measures

Video content measures form a foundation for testing with real video. Their primary function is for classification of source video materials. Then, the overall performance of a compression system can be related to different classifications of source materials. This is important because codecs' video quality can vary dramatically depending on spatial detail, motion, colorimetry, and other content factors

Many measures can be useful for characterizing scenes. We use measures of spatial and temporal information content for luminance and chrominance. Measures designed to detect the presence of specific features may also be useful for testing purposes. Features such as scene cuts, frame repeats, 3/2 pulldown, and others can be used to organize video materials and testing results. Scene cuts are particularly important because codecs with interframe processing and motion prediction often exhibit increased distortion after a cut. Glenn[9] shows that the presence of scene cuts can mask visual sensitivity to spatial detail and motion immediately before and after the cut. Ideally, any distortion measure designed to predict perceptible image quality would need to test for the presence of scene cuts in the source video. Wolf et.al.[6] have recently proposed a distortion measurement that incorporates scene cut masking effects using a threshold on the frame-to-frame increase of a temporal content measure to indicate the presence of a cut. Other researchers have developed scene cut detection schemes based on colorimetry statistics[10,11].

B. Test-Signal Measurements

The main purpose for test signals in this testing methodology is to determine the magnitude of those effects that are invariant to scene content and operating conditions. Invariant effects such as gain and level

distortion, spatial filtering, spatial misregistration, and time delay may be introduced by either analog or digital processes. Gain and level distortions can be made in the processes of analog-to-digital and digital-to-analog conversions. Frequency response attenuation (i.e., horizontal resolution) can be affected by pre- and post-spatial filtering (analog or digital). Spatial cropping and misregistration are common in multimedia and teleconferencing codecs.

Test signals can also serve a more traditional role to determine the upper and lower limits of performance. Examples of these kinds of tests include resolution measurements[12], zone-plate loading[13], noise loading[14] and SNR measurements. We have found these types of measures to be useful for analysis of codec performance. But, in general, they do not provide quantitative predictions of video quality for real video.

1. Colorimetry

Gain and level distortions are defined as linear scaling and translation of color values in the YUV color space. These distortions are generally introduced by analog processes and so they are linear time-invariant effects. Gain and level are measured using a statistical comparison of source and degraded video test sequences in a manner similar to Wolf[6]. Gain and level are computed independently on the luminance and chrominance channels. Luminance gain and bias are analogous to contrast and brightness. Chrominance gain is analogous to color intensity. Chrominance level is usually fixed, but could potentially be modified. Gain and level are reported as separate distortion measures. The degraded video is preprocessed to compensate for the measured gain and level shifts before other distortion measures are computed. Thus, the other distortion and quality metrics do not penalize a codec for linear impairments that can be corrected by conventional analog adjustments (e.g., brightness, contrast, color, etc.).

2. Spatial Cropping And Registration

Spatial cropping and spatial misregistration are common artifacts of digital video compression systems. DS-3 codecs (45 Mbps) typically encode and pass the entire active picture. However, MPEG-2 defines a 704x480-pixel cropped subregion of the full 720x483 CCIR-601 active picture for encoding. Other low-data-rate encoders crop to a greater extent. Many codecs introduce horizontal and vertical misregistration in the picture. This problem has been seen even on 45 Mbps codecs. Measures of cropping and spatial registration are useful characterizations in their own right, and also they are necessary inputs for preprocessing prior to error signal computation.

ITS has recommended the practice of cropping to an assumed *viewable* region of 672x448 pixels to approximate typical consumer horizontal and vertical overscan. Any cropping outside the viewable region is ignored, while cropping within the region is treated as distortion. Manual inspection of the output signal (either with a digital oscilloscope or image viewing tools on a computer) is used to determine the actual active picture.

C. Comparative distortion measures

Comparative distortion measures involve direct comparisons between source video and degraded output video imagery. These measures include the discrete error signal, RMS error and signal-to-error ratio, and a CCIR-500 impairment quality rating (*IQR*) with three associated distortion measures.

Distortion measures can be characterized as either local or cumulative. A local measure is a discrete spatial-temporal sample of video distortion. The most fundamental local measure is the error signal, defined as the difference of the source and degraded digital imagery, and sampled at discrete pixel locations. A cumulative measure is defined over some

spatial-temporal block of video. Common examples of blocks include an 8x8 coding block, multiple blocks, field, frame or short video sequence. Root-mean-squared error (RMS error) is an example of a cumulative distortion measure that can be computed over any arbitrary block of video. Typically, local measures are used to form the kernels of cumulative measures. Combinations of digital filtering, weighting, integration, and statistical techniques are used to define cumulative measures.

1. Error Signal and *SER*

The error signal is a basic diagnostic tool for testing digital video systems. The error signal is computed as a pixel-by-pixel difference of the digitized CCIR-601 source and degraded video. The error is computed separately on each of the three YUV channels, where Y is luminance, U is the blue color difference (also called C_b), and V is the red color difference (also called C_r). Figures 1 and 2 show a degraded frame and the absolute value of its luminance-error signal (enhanced for printing). Examination of the error signal is often useful for detailed analysis of compression artifacts.

The simple definition of the error signal is complicated by two real-world factors. First, the degraded imagery must be accurately corrected for temporal alignment, spatial cropping and registration before the differences can be taken. Spatial cropping and registration are performed as described above. For video systems that have a fixed delay, temporal alignment is performed by visual inspection with compensation for the delay in the video capture system. However, some systems can introduce variable frame dropping/repeating (especially low-data-rate codecs). For these, a technique developed by ITS is used to determine a fixed delay to best align each pair of source and degraded sequences[15]. Second, the pictures are aligned spatially, and subpixel registration is performed. Third, the pictures are cropped.

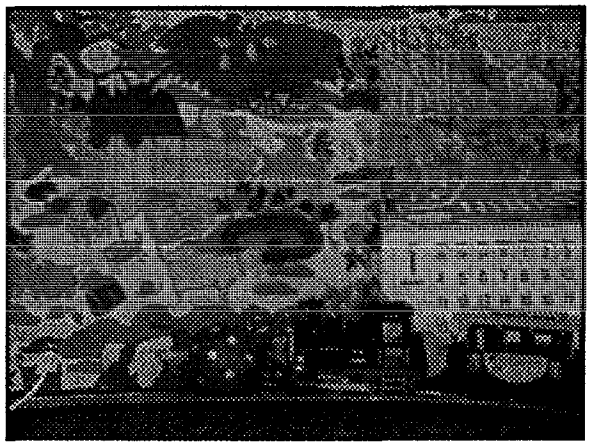

Fig. 1. Degraded version of *Mobile & Calendar* sequence,frame4.

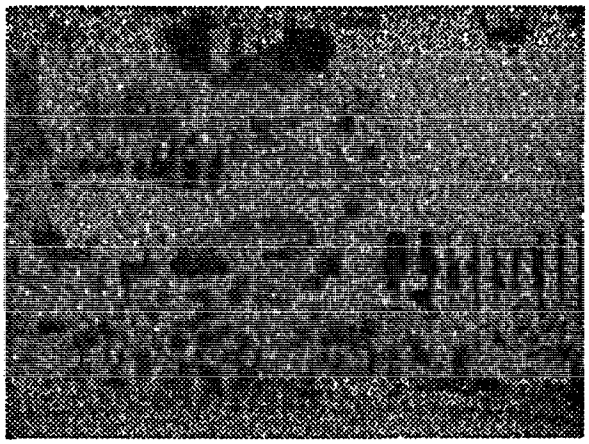

Fig. 2. Absolute Y-error (values x25) of degraded *Mobile & Calendar* sequence,frame4.

Fourth, the degraded video is corrected for any invariant gain and level distortions introduced by the video compression system. The degraded degraded imagery is corrected according to the method previously described, and the gain and level are reported separately. Finally, the error signal reflects only those distortions that are not correctable by constant analog adjustments to the signal.

Signal-to-error ratio (*SER*) is similar to the traditional signal-to-noise ratio (*SNR*) but generalized to apply to real video. It is defined using the log of the ratio of peak-to-peak signal versus the RMS error and is expressed in dB.

The *SER* is computed by field and overall for each video sequence as shown below.

$$SER_{field}(t) = 20\log_{10}\left\{\frac{V_{pp}}{rms_{space}(S(x,y,t)-D(x,y,t))}\right\} \quad (1)$$

$$SER_{sequence} = 20\log_{10}\left\{\frac{V_{pp}}{rms_{time}(rms_{space}(S(x,y,t)-D(x,y,t)))}\right\} \quad (2)$$

Signal-to-error ratio is defined in a similar way for each of the luminance and color-difference chrominance channels in the YUV color-space. Note that *SER* can be measured for any video signal and for any type of impairment -- for a gray input it approximates the traditional unweighted *SNR* measurement.

SER is an absolute measure of image reproduction accuracy. It is most significant for assessment of contribution-quality compression where it is desirable to minimize distortion of any kind. Also, SER is very useful for identifying occurrences of discrete impairments and artifacts within a codec's output sequence. Major spatial and temporal impairments cause recognizable signatures in field-by-field SER plots that are unique to each codec and test sequence combination. SER is less useful as an overall measure of perceptible impairment quality because different compression-induced distortions are perceived differently by the human visual system.

1. Impairment Quality Rating

The following is an approach developed by researchers at NTIA/ITS[6,7,8] to predict subjective impairment quality ratings (*IQR*) for pairs of degraded versus source real-video sequences. The method produces a rating on the CCIR-500 five-grade impairment scale for double-stimulus tests[2] (5 = *Imperceptible*, 4 = *Perceptible But Not Annoying*, 3 = *Slightly Annoying*, 2 = *Annoying*, 1 = *Very Annoying*). The *IQR* measurement formulation is based on a functional combination of distortion measures, optimized for correlation with available subjective test results. ITS uses a linear combination of distortion measures of the form:

$$IQR \approx c_0 - c_1 m_1 - c_2 m_2 - c_3 m_3 \quad (3)$$

The c_i are weighting coefficients determined to give the best fit to subjective test results using a least-squares-error criterion. The three distortion measures were selected from an exhaustive search of combinations of over 100 different distortion measures. These three measures, with coefficients of $c_0 = 4.7485$, $c_1 = 0.9553$, $c_2 = 0.3331$, and $c_3 = 0.3341$ produced the best correlation with the average ratings of a large subjective test involving 48 viewers, 36 nine-second test sequences, and 27 different types of analog and digital impairments. Analog impairments included NTSC encode/decode, VHS record/play, and a noisy RF channel. Digital impairments included video codecs operating in the range of 56 Kbps to 45 Mbps with simulated digital networks and controlled error rates. ITS showed correlation with CCIR-500 subjective test results with an expected accuracy for individual ratings of about 0.5 rating points.

Conceptually, m_1 is a measure of spatial distortion, m_2 is a measure of reduced motion, and m_3 is a measure of added motion. An important aspect of the NTIA/ITS approach is that these three measures do not involve direct comparison of the source and degraded video imagery. Rather, the source and degraded sequences are each characterized by small feature sets from which the measures are computed. In the case of m_1 , the feature set is the set of spatial variance values of the Sobel-filtered images of each frame in the video sequence -- essentially a measure of frame-by-frame edge content.

The method can be used to predict subjective ratings for arbitrary codecs, scenes and impairments. Predicted ratings have roughly a two-thirds probability of being accurate within 0.5 CCIR-500 rating points

based on ITS's statistical analysis. *IQR* ratings are interpreted in the context of the range of impairments and viewing conditions present in the correlated subjective test. In addition, the magnitudes of the c_1m_1, c_2m_2, and c_3m_3 distortions (we denote the c_im_i by M_i) can provide insights into the nature of impairments that are generally not available from subjective tests. The M_i are indicators of the relative contributions of spatial and temporal distortions to the overall *IQR*. The overall *IQR* results and M_i distortions are plotted by sequence for comparisons between two or more codecs. Fig. 3 shows *IQR* ratings for the same codecs and sequences shown earlier.

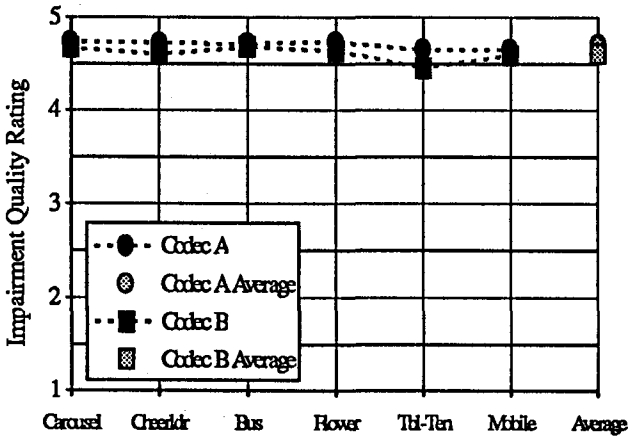

Fig. 3. Overall *IQR* for two codecs and six sequences.

The approach is well suited to overall assessment of one codec versus another on a common set of test sequences. Voran[16] has shown that average ratings, taken over several sequences give significantly better correlation with subjective tests than do individual ratings. Also, the approach is amenable to remote, real-time computation since only a minimal data set must be transmitted from the source site to the receiver site for comparison and computation of the underlying measures.

a. SER Results

SER tests have been made on a number of commercial codecs. Luminance and chrominance SER measures are calculated using the six video sequences described previously. Overall Y, U, and V SER are plotted in Figures 4, 5, and 6 for two 45-Mbps codecs.

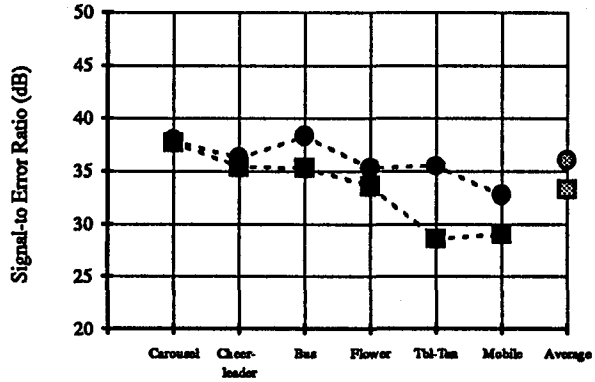

Figure 4. Luminance (Y) Signal-to-Error Ratio, Overall by Sequence.

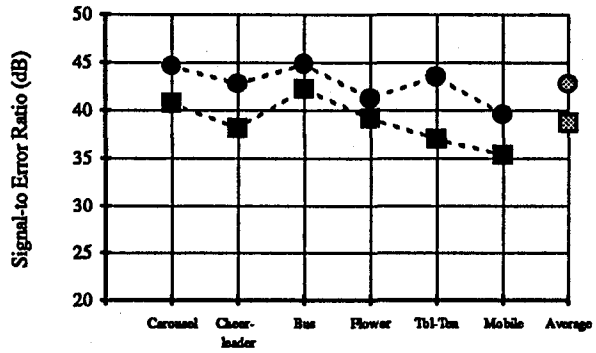

Figure 5. Blue Color Difference (U) SER, Overall by Sequence.

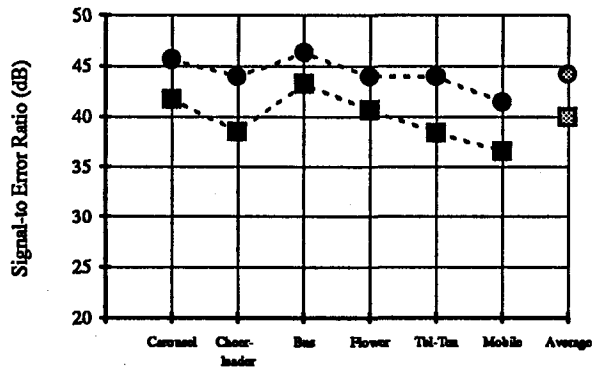

Figure 6. Red Color Difference (V) SER, Overall by Sequence.

Codec A has a distinct, consistent advantage in SER for all channels. Averaged over all sequences, codec A showed advantages of 2.7 dB, 4.0 dB, and 4.3 dB across the Y, U, and V channels, respectively. Codec A produced a better SER than codec B on all sequences. Codec A had the greatest advantage on the scenes with high spatial-temporal content, especially *Table Tennis* where it had a 7 dB advantage in luminance SER. Both codecs produced higher signal-to-error in the chrominance channels than in luminance.

SER values in the range of 30 dB to 40 dB are typical of good DS-3 codecs on these sequences. Distortions characterized by 30-40 dB overall SER are often, but not always, imperceptible. Distortions characterized by less than 30 dB are more likely to be perceptible. Distortions in the low-20's dB are likely to be perceptible and at least somewhat annoying.

The testing system also produces frame-by-frame images of codec-induced distortions that can be viewed interactively on the computer. These "error sequences" in conjunction with the field-by-field SER plots provide a valuable tool for identifying and analyzing codec artifacts. In the present case, they did indeed help to locate a particular type of artifact induced by these codecs.

IV. Instrument Measures

A. Measurement of Multi-Dimensional Codec Resolution

In conventional NTSC tests, horizontal resolution is measured as frequency roll-off using the multiburst or continuously swept test waveforms. However, vertical resolution and diagonal resolution are seldom measured. Vertical resolution in NTSC is set by the number of visible scan lines in the system, and by the Kell factor which depends upon the camera and the monitor. Transmission and processing seldom change the vertical resolution in conventional TV. However, the introduction of compression in video signal transmission changes this situation. Vertical resolution can be seriously affected by the compression and decompression processing, as can horizontal and diagonal resolution. Therefore, all three resolutions need to be measured. Systems that use predictive coding are even more complicated to test, as the various resolutions can depend upon the speed at which the scene is changing. A method of measuring these three resolutions using moving zone plates[17] can be used.

1. Signal Source and Test Equipment

The signal source is a Tektronix TSG-1001 multi-format zone-plate generator. For testing resolution, one-dimensional moving zone plates are used, with a specific set-up for each of the three directions. A single zone-plate parameter is used for measuring resolution in the horizontal or vertical direction, and a pair of parameters is used for the measurement in the diagonal direction. The equipment is configured as in Figure 7. Modulation-transfer-function data is captured by a Tektronix 2232 digital storage oscilloscope with samples externally clocked by a pulses, one per frame, from the zone-plate generator. The data is then plotted on a laser printer. Although picture monitors are present, they are not involved in the measurement, making this an accurate test of the codec resolutions.

2. Results Obtainable From Moving-zone-plate Tests

Plots taken with the setup in Figure 7 show the modulation-transfer-function (MTF) as a function of spatial frequency in each of the three principal directions. Resolution (cycles per picture height) in each direction is defined as the spatial frequency at the 3-dB point on each MTF plot. Resolution in TV-lines is double the resolution number in cycles/picture-height. The results from making these plots with many codecs are as follows: (1) Horizontal MTF for

luminance and chrominance is identical to that obtained with the stationary swept sine wave test signal or the multiburst test signal applied separately to the luminance and chrominance channels. (2) Vertical MTF is generally flat to the Nyquist frequency and beyond for both luminance and chrominance, but as there is generally no vertical prefiltering provided prior to the 2:1 decimation of the chrominance channels, chrominance signals are not protected against vertical aliasing with most codecs. (3) Diagonal MTF is likely to have ripple, caused by the quantizing matrix in DCT-based codecs.

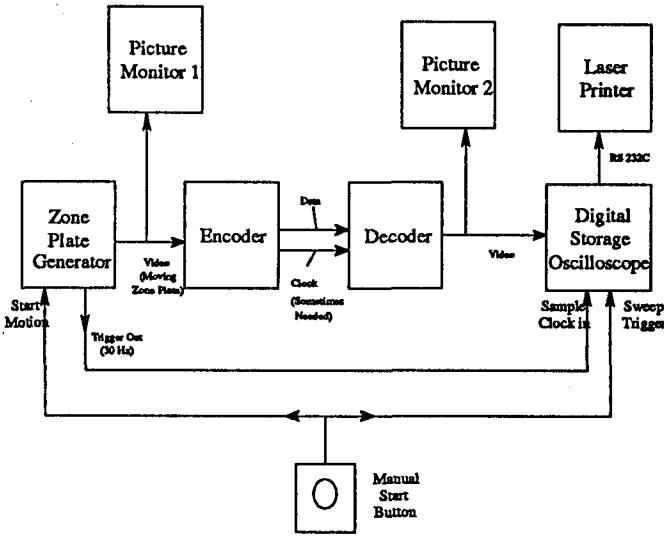

Figure 7. Equipment Configuration for Resolution Testing.

1. EIA-250-C Testing

Since no standard exists for measuring the picture quality of compressed digital video signals, the question is frequently asked, "Does the codec meet the EIA-250-C short, medium or long-haul specification?," even though it is known that the EIA measurement set does not adequately characterize a complex digital video compression codec. The Tektronix TSG-1001 multi-format video test signal generator and the VM-700A video measurement can be used to determine if there are any problems in the NTSC coding,, but will not find problems in the compression coding.

B. Subjective Testing with Zone-Plate Loading

Circular zone plates with spatial frequencies up to full Nyquist detail at the top and bottom of the picture and with various rates of motion can be applied to these codecs. The outputs are observed subjectively on a monitor. The artifacts of NTSC are always very prominent in this subjective test, making it difficult to recognize small artifacts of codec processing for zone plates with high-density features. However, some codecs have artifacts that are easily found by this subjective test, making it a worthwhile part of the testing plan.

V. Implementation of a Digital Video tester

We have developed a system for testing digital video compression systems based on a Silicon Graphics workstation. The system computes the error signal, *SER*, *IQR* and associated distortion metrics by field, by frame and by sequence. In addition, it computes temporal alignment, spatial registration, and gain and level. The tester incorporates viewing tools that allow interactive viewing of source, degraded, and error video imagery on the computer monitor. Graphs of *SER*, *IQR* and M_1, M_2, and M_3 can be displayed along side the video imagery. We use D1, D2 or Betacam tape to record source video and play it through a codec. Currently, source and degraded video are captured on an Accom WSD disk recorder. We are in the process of fully integrating video capture into our testing system using the standard video input capabilities of an SGI Indy workstation. A diagram of this system is shown in Fig. 8. The Indy system is a low cost approach and does not have many of the Accom features including capture of 30 seconds or more of 720 x 486 video.

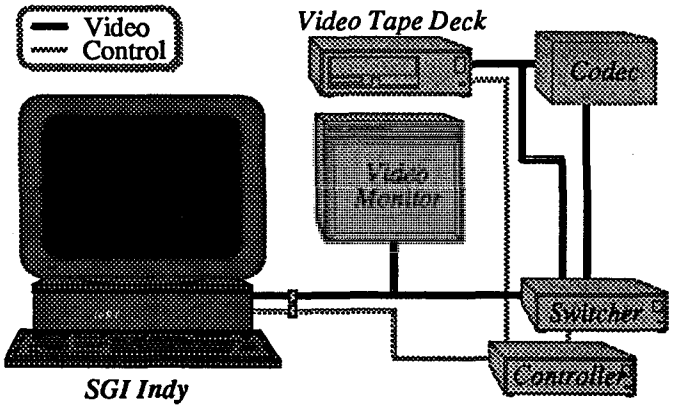

Fig. 8. Diagram of prototype testing system.

The Indy incorporates a number of video technologies that enable this system. The Indy has a standard video input subsystem (VINO) that allows burst capture at 30 frames-per-second video directly to system RAM. Video is captured at 640x480 pixel resolution in YUV 4:2:2 format. This format is similar to CCIR-601 but uses square pixels and therefore a slightly lower sampling rate. Video input is via the standard NTSC or S-video input ports. The input video is digitized one field at a time in the VINO subsystem, then transferred real-time across the Indy's internal GIO-64 bus (33 Mhz, 64-bit) to a RAM buffer. The burst capture is limited by the size of the RAM buffer -- the Indy's maximum RAM configuration of 256 MB allows capture of six-second sequences. The source and degraded imagery are downloaded to files on an internal SCSI-2 disk for later access during analysis and viewing.

Our system performs image processing computations for the various measurements completely in software. The Indy incorporates a MIPS R4000 100 Mhz RISC CPU that is capable of image processing at a usable level of performance for our applications. We use SGI's object-oriented ImageVision library for the development of our image processing software. Computation of all measures for a five-second sequence currently takes about 40 minutes. With automation, an overnight test can compute a full set of measures for about 20 five-second test sequences. We believe that with careful selection, 20 sequences of varying content is sufficient to provide adequate characterization of a codec's performance.

We expect to improve the performance in a number of ways: 1) Tuning our software should give about a 2x increase in performance, 2) We can test using one-second sequences. This allows a greater variety of test material (i.e., 100 sequences versus 20) because computation time is proportional to video length independent of the number of sequences. The main reason to test with five-second sequences is to allow subjective assessment. However, one-second sequences can be selected for objective testing from a much larger test tape used for subjective assessment. Alternatively, short objective test sequences can be played in a continuous loop or swing mode for subjective assessment. 3) Computer price-performance is advancing rapidly. Our ImageVision software will take advantage of faster CPUs and parallel processing when they become available on low-cost workstations. 4) Dedicated image processing hardware (e.g., DSP boards) can accelerate performance by 100x or more. DSP boards are available on other platforms, though not yet on the Indy. CPU performance is the main bottleneck in the existing systems performance. The GIO-64 bus provides extremely fast transfers of image data between RAM and CPU. SCSI transfers of image data from disk to RAM can limit image processing and video image display performance in some circumstances. However, the ImageVision library incorporates a innovative look-ahead, multi-tasking caching scheme that eliminates most of these bottlenecks.

Figures 10 and 11 show typical displays from StellaCom's tester. Fig. 10 shows the source video, the error signal, and the degraded video imagery being displayed simultaneously in separate synchronized windows. The three pictures are presented either still or in motion at up to 20 frames per second. A graphical user interface control panel provides VCR-like

controls for playback with random access. Graphs of *SER* and *IQR* and associated spatial and temporal distortion measures can be displayed in separate windows. Fig. 11 shows the degraded picture along with plots of the *IQR* and M_1, M_2, and M_3 frame-by-frame kernels. The overall *IQR* for the sequence appears as a dashed horizontal line. A cursor tracks the current frame on each plot during playback. The range of frames for playback and graphing may be reset to any subset of the full test sequence. Also note that standard desktop tools on the SGI allow for interactive magnification and enhancement of the video imagery for detailed assessment of distortion and artifacts.

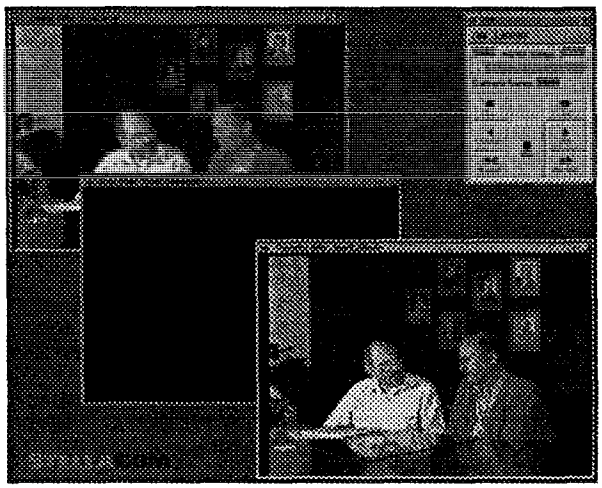

Fig. 10. Display of source video, error signal, and output video on StellaCom's tester.

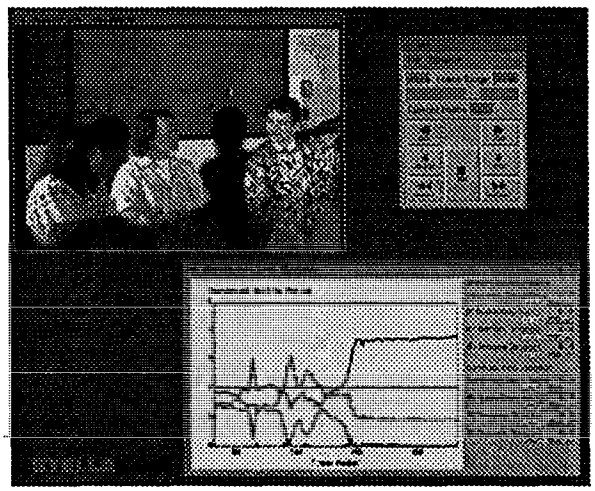

Figure 11. Display of degraded video and *IQR* measurements

VI. Summary

Compressed digital video systems require new testing methods. We have demonstrated that a graphics workstation which has been programmed with appropriate measures can suitably test such systems. We have developed a testing system that incorporates a suite of distortion measures and viewing tools for the workstation. The Impairment Quality Rating (*IQR*) measurement developed by Wolf and others at NTIA/ITS can be used to predict how an average viewer rates a video sequence on the CCIR-500 impairment quality scale. The *IQR* measurement has been correlated with subjective testing using statistical methods. Associated with the *IQR* are a spatial distortion measure and two temporal distortion measures. In addition, we compute an error signal that can be viewed as video and used for statistical measurements such as RMS Error and Signal-to-error ratio (*SER*). Our system allows the user to view video (source, codec-processed video and error signal) side-by-side with plots of measurement results on the computer display. It permits in-depth analysis of the quality of digital video compression, transmission and storage systems.

The quality of compressed video depends on a number of factors including data rate. Codecs employ complex algorithms that incorporate discrete spatial and temporal processing. They exhibit non-linear responses to changing scene content and operating conditions. Codecs working at low data rates and processing video with significant detail and motion can introduce distortion into the picture. Consider the impairment quality rating (*IQR*) for degraded versus source video sequences. Analog VHS material normally yields a rating between 3.0 and 4.5 (between slightly annoying and barely perceptible) depending on whether the source material readily shows the imperfections of VHS tape. The average rating for VHS is usually between 3.5 and 4.0. The rating for a 6 Mbps digital codec is normally

above 4 depending on the source material. Hence, a good 6 Mbps digital codec can be superior to an analog VHS tape by this measure. Another measure of picture quality is resolution. VHS video has a horizontal resolution of approximately 220 TV lines per picture height (TVL/PH). Static horizontal resolution on most 6 Mbps codecs exceeds 450 TVL/PH. The average resolution of a 6 Mbps digital codec on moving objects is less than the static resolution, but can exceed that of an analog VHS tape.

Compression of video signals is cost-effective now and will be pervasive in the future. Engineers and engineering standards bodies had better learn how to measure the quality of compressed video systems. Otherwise, we will not know the ultimate quality of the material we are producing, transmitting and displaying.

Acknowledgments

The author would like to thank Mr. Ted Dienert, president of StellaCom, for his support of this project. Appreciation is expressed to Mr. Steve Wolf of NTIA/ITS for providing video test materials, subjective and objective test results, and numerous discussions. Thanks are expressed to Chris Cressy, Dr. Elliott Kohn, Jeff Van Pelt and Javier Vega of StellaCom for significant contributions to this effort.

References

1. C. Cressy and G. Beakley," Computer-based testing of digital video quality," IS&T/SPIE Symposium, San Jose, CA, February 6-10, 1994.
2. CCIR Recommendation 500," Method for the Subjective Assessment of the Quality of Television Pictures", 1990.
3. T. Hidaka, "MPEG-2 Verification Test," ISO/IEC JTC1/SC29/WG11/MPEG93/732, August, 1993.
4. S. D. Voran and S. Wolf, "The Development and Evaluation of An Objective Video Quality Assessment System That Emulates Human Viewing Panels," *International Broadcasting Convention*, Amsterdam, July, 1992.
5. ISO/IEC CD 13818, "Coding of Moving Pictures and Associated Audio," November, 1993.
6. S. Wolf, et al, "A Summary of Methods of Measurement for Objective Video Quality Parameters Based on the Sobel Filtered Image and the Motion Difference Image, T1A1.5/93-152, November, 1993.
7. A. W. Webster et al, "An Objective Quality Assessment System Based on Human Perception," *SPIE 1993 International Symposium on Electronic Imaging: Science and Technology*, June, 1993.
8. S. Voran and S. Wolf, "An Objective Technique for Assessing Video Impairments," *Proceedings of the IEEE Pacific Rim Conference*, 1993.
9. W. E. Glenn, K. G. Glenn, and C. J. Bastian, "Imaging System Design Based on Psychophysical Data," Proceedings of *the SID*, Vol. 26/1:71-78, 1985.
10. K. Otsuji and Y. Tonomura, "Projection Detecting Filter for Video Cut Detection," *Proceedings of ACM Multimedia 93*, 1993, pp. 251-257.
11. H. Ueda, T. Miyatake and S. Yoshizawa, "IMPACT: An interactive natural-motion picture dedicated multimedia authoring system," *CHI'91 Conference Proceedings*, 1991, pp. 343-350.
12. E. S. Kohn and G. W. Beakley, "Resolution Testing of An HDTV Codec with Moving Zone Plates," 134th SMPTE Technical Conference, Toronto, November 1992.
13. G. W. Beakley, C. R. Caillouet, E. S. Kohn and J. L Van Pelt, "Testing of a Digital HDTV Codec Through the NASA Communications Network," *1992 HDTV World Conference Proceedings*, National Association of Broadcasters, April 1992.
14. G. W. Beakley and E. S. Kohn, "Transmission of Compressed Digital High-Definition Video Through the NASA Communications Network," *Proceedings of SPIE Conference*, San Diego, July 1992.
15. S. Wolf and A. Webster, "Objective Performance Parameters for NTSC Video at the DS3 Rate," T1A1.5/93-60, April 28, 1993.
16. S. Voran, "The Effect of Multiple Scenes on Objective Video Quality Assessment," T1A1.5/92-136, July, 1992.
17. E.S. Kohn and G.W. Beakley, "Resolution Testing of An HDTV Codec with Moving Zone Plates," 134th SMPTE Technical Conference, Toronto, November 1992.