THE DEVELOPMENT AND EVALUATION OF AN OBJECTIVE VIDEO QUALITY ASSESSMENT SYSTEM THAT EMULATES HUMAN VIEWING PANELS

S.D. Voran, S. Wolf

Institute for Telecommunication Sciences, United States

Abstract

The Institute for Telecommunication Sciences is conducting research to develop an objective video quality assessment system that emulates human perception. The system should return results that agree closely with quality judgements made by a large panel of viewers. Such a system is valuable because it provides broadcasters, video engineers and standards organizations with a means for making meaningful video quality evaluations without convening viewer panels. The issue is timely because compressed digital video systems present new measurement questions that are largely unanswered. We describe our development procedure, present some results, and evaluate a prototype version of the video quality assessment system.

1. Introduction

The traditional performance measurements for video transport and storage systems use fixed test signals and assume that the system under test is time-invariant[1]. While these signals and the associated measurements are indispensable for the characterization of the electrical performance of time-invariant, analog video systems, the measurements often do not correlate well with video quality as perceived by the end users of the video system. For example, weighted signal-to-noise ratio does not give an accurate indication of image quality when the noise is correlated with the image, as is the case with predictive video coders[2].

In recent years, the problem has become more complicated and wide-spread. Video signals are now commonly transmitted and stored in compressed digital form with a possible resulting loss of quality. Effective compression algorithms are dynamic, with the input video signal dictating the overall behavior of the algorithm through many sub-algorithms that perform motion prediction, adaptive transforms, and adaptive quantization, to name only a few. The resulting video systems are clearly time-varying and signal dependent. Static, deterministic test signals cannot provide an accurate characterization of their performance on program material.

In a broadcast environment, digital video equipment must provide virtually flawless imagery. In other applications, larger distortion levels are often tolerated in order to allow video communications over lower bit rate digital channels. An accurate and repeatable characterization of these different distortion levels will allow users to specify and verify the appropriate video quality level for a given video application.

The forgoing observations motivate us to look for a video quality assessment system that utilizes actual video signals. This approach provides a realistic measurement environment and allows for in-service measurements of video systems. The system should work well for all possible video scenes and for a wide range of analog and digital video systems. It should mimic the human visual and perceptual system, so that measured video quality agrees with video quality as perceived by the viewer who actually receives the video signal. In short, we seek to measure the video quality of program material while it is being delivered, in a way that correlates with the subjective evaluations made by those who actually view the video.

To meet these goals, we must incorporate knowledge of human perception in the design of the assessment system. Our perception-based development process is described in Figure 1. A set of test scenes is selected and impaired by imperfect video systems. From the video, we extract a set of candidate objective measurements that seek to quantify the perceptual impact of the video scene impairments. A panel of viewers watches the same set of test scenes and their subjective judgements of the impairments are recorded. The final step in the derivation process is a joint statistical analysis of the objective and subjective data sets. This analysis reveals which objective measurements are meaningful, and how they might be combined to create an objective metric that emulates human perception.

2. Development Summary

If an assessment system is to work for a wide range of scenes, then scenes of sufficient variety must be used in its development. The spatial and temporal information content of the scenes are critical parameters. In the analog domain, the spatial information content governs the relationship between system passband and video quality. In digital systems, information content determines sample rates and plays a crucial role in determining the amount of video compression that is possible, and consequently, the level of impairment that is suffered when the scene is transmitted over a fixed-rate digital channel. In order to derive the most general possible system, our library contains 36 scenes with widely varying amounts of spatial and temporal information. Examples include sports events, news

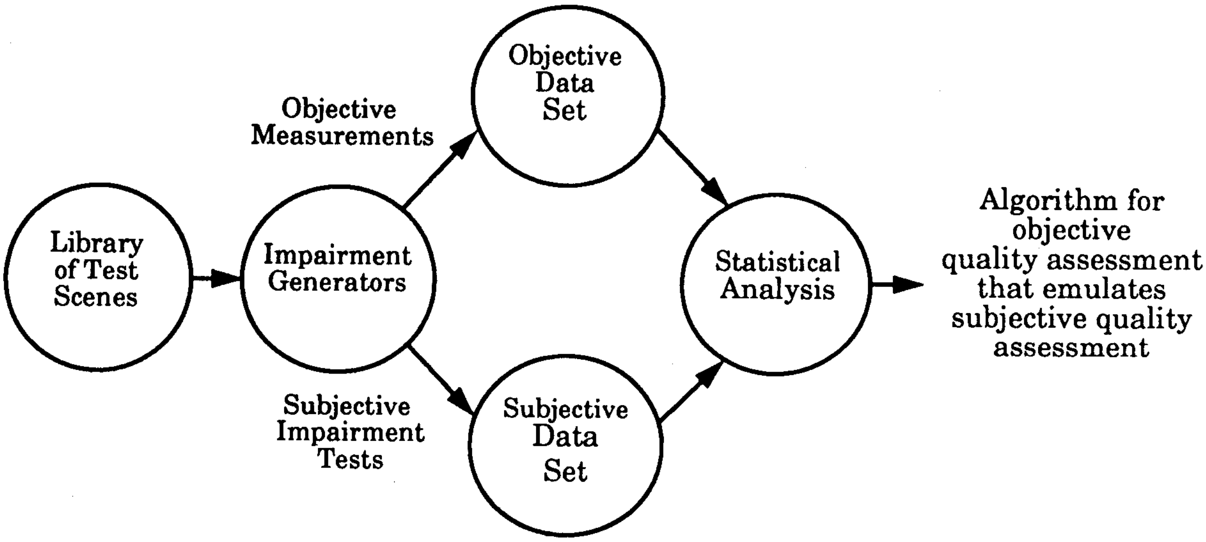

Figure 1 Perception-Based Development Process

desk, still shots, and graphics. Our library signal format is 525/60 component analog video (CAV).

Since we are striving for a quality metric that works well across a wide range of video technologies, we have included a broad family of 27 impairments in the derivation. The impairments include video codecs operating at line rates that range from 45 Mbps down to 56 Kbps over simulated digital networks with controlled error rates. Analog impairments include NTSC encode/decode cycles, VHS record/play cycles, and a noisy RF channel. All test scenes are subjected to all impairments to create a library of 972 impaired test scenes.

The subjective data set in Figure 1 is built from viewer judgements gathered in subjective impairment tests. These tests are conducted in a controlled laboratory environment. The laboratory was built using the guidelines of CCIR recommendation 500-3, which specifies the standard visual environment for conducting picture quality assessment[3]. Video scenes are displayed on a 19-inch broadcast quality monitor with CAV inputs. Seating for three viewers is provided at a distance of six picture heights from the monitor screen. The test methodology is based on standard CCIR recommendation 500-3 picture quality assessment procedures augmented by the results of our preliminary viewing sessions.

Each viewer participates in four twenty-minute sessions. These sessions are comprised of randomly ordered 30-second impairment tests. These tests use reference scenes to take advantage of the fact that the human eye excels at making comparisons. The subject is presented with nine seconds of a scene, three seconds of grey, nine seconds of the impaired version of the scene, and finally a nine second period in which to mark the response form. The five possible responses offered are: 5=imperceptible, 4=perceptible but not annoying, 3=slightly annoying, 2=annoying, and 1=very annoying. This scale covers a wide range of impairment levels in a nonlinear fashion. Due to training time, rest intervals, and some redundancy, the cumulative 80 minutes of testing allow for the viewing and rating of 128 test scenes taken from our library of scene-impairment combinations.

To date, 48 viewers have participated in these tests. Our ultimate goal is to create an assessment system that emulates the distribution of viewer responses (e.g. "80% of the viewers will report that the impairments associated with this video system are imperceptible.) For the moment however, we have elected to simplify the problem by collapsing each distribution of subjective judgements down to a single central measure according to

$$s=\sum_{i=1}^{5} i p_i \,, \tag{1}$$

where p_i is the fraction of viewers that responded in the i^{th} category. For example, s=5 indicates consensus that the impairment is "imperceptible". The problem now becomes one of modeling this central measure as accurately as possible for each of the 128 test scenes. In the next section, we refer to this central measure as the "true score".

The objective data set shown in Figure 1 is built from over 100 measurements performed on digitized (756 x 486 x 24 bit) video signals. The measurements are performed on every frame of each of the 128 test scenes. This intensive measurement approach dictates that the measurements be automated. A controlling program with a windowed user interface passes instructions to device drivers that in turn control the tape machines, frame digitizer, and video routing switcher. In an exact parallel to the subjective tests, the objective measurements are all differential. That is, they involve both the original and the impaired versions of each scene.

All video impairments can be described as distortions of the amplitude or the timing of the video waveform. When displayed on a monitor, this one-dimensional voltage waveform is interpreted as a continuously evolving, multi-colored, multi-dimensional signal. Useful measures must take note of this human interpretation and mimic it to the extent possible. Thus, our candidate set of objective measures includes those designed to measure temporal, luminance, chrominance, and spatial distortions. Chrominance distortions are characterized after transforming luminance and chrominance values to the International Commission on Illumination (CIE) LUV color space. In that 3-dimensional space, each color axis is perceptually independent and psychometrically uniform[4]. Representative measures of spatial and temporal distortion are given in the following section. As expected, it is a mixture of these measurements that provides the best overall characterization of video quality. A detailed description of the set of candidate measurements is available[5]. Further details on the development of the assessment system are also available[6].

3. The Linear Model and Selected Measurements

The final stage of the development process involves joint statistical analysis of the subjective and objective data sets. This step singles out a subset of the candidate objective measurements that provides useful and unique video quality information. When combined by an appropriate mathematical structure, this subset of measurements provides predicted scores that correlate well with the true scores obtained in the subjective tests. Potential combining structures include linear and quadratic predictors, hybrid linear-nonlinear decision algorithms (possibly adaptive), and Bayesian inferencing. At this time, we consider only the simplest of all combining functions: the linear combination. Thus, we are looking for p measurements $\{m_i\}$ and p+1 constants $\{c_i\}$, that will allow us to calculate

$$s \approx \hat{s} = c_0 + \sum_{i=1}^{p} c_i \cdot m_i \quad . \qquad \textbf{(2)}$$

For the search procedure, we adopt the least squares error criterion,

$$MIN \sum_{i=1}^{n} (s_i - \hat{s}_i)^2 \quad , \qquad \textbf{(3)}$$

where n denotes the number of video sequences involved in the test. If the p measurements are given, then the standard least-squares solution can be used. In this case, the measurements are not given, they must be selected from a larger set. Our selection algorithm uses 128 scenes and iterates between a selection step and a least-squares solution step to arrive at a nearly optimal set of measurements.

Here we present a third-order solution:

$$\hat{s} = c_0 + \sum_{i=1}^{3} c_i \cdot m_i \quad . \qquad \textbf{(4)}$$

The first measurement selected is:

$$S_{time}\left\{ 5.78 \cdot \left| \frac{std_{space}(S(O_n)) - std_{space}(S(}{std_{space}(S(O_n))} \right. \right. \qquad \textbf{(5)}$$

where O_n denotes the n^{th} frame of the original video sequence, D_n is the n^{th} frame of the degraded video sequence, S(•) indicates the Sobel filtering operation[7], and std_{space} indicates that a standard deviation of pixel values is computed. The numerical scale factor serves to normalize the variance of this measurement.

Since m_1 is computed at each of the 270 frames of the 9 second video sequence, it returns a sequence of 270 values. Each viewer returns a single measure for the entire 9 second sequence. Since our goal is the emulation of the human visual and perceptual systems, our next step is to compress the sequence down to a single value, using an algorithm that might approximate the algorithm viewers use. The selection of such "time collapsing functions" is an integral part of our measurement selection algorithm. For m_1, we find that the RMS value of the time series provides the measurement that agrees best with the subjective data.

The Sobel filtering operation enhances edges and other high frequency content in the video frame. The standard deviation returns the non-dc energy of the filtered frame. Thus, m_1 is a normalized measurement of how the high frequency spatial energy is affected by the video system under test. When D_n matches O_n exactly, the measurement value is zero. The absolute value function ensures that either a loss (eg: blurring) or a gain (eg: false edges) of high frequency image content will cause a positive swing in m_1. In light of this interpretation of m_1, we categorize it as a "spatial measurement". This helps to differentiate m_1 from m_2 and m_3, which fall into the class of "temporal measurements"; those that measure how the video system impairs the smooth flow of time and motion.

The remaining two measurements are given as:

$$m_2 = f_{time}\{\ .0934 \cdot \max\{[RMS(\Delta O) - RMS(\Delta D)]\ , 0\ \}\},$$

$$where \qquad \Delta F = F_n - F_{n-1}\ ,$$

$$f_{time}(\{x_t\}) \ = \ std_{time}(filter(\{x_t\}))\ , \qquad \textbf{(6)}$$

$$m_3 = \max_{time}\left\{\ 4.2522 \cdot Log_{10}\left(\frac{std_{space}(\Delta D)}{std_{space}(\Delta O)}\right)\ \right\}.$$

Here RMS(F) returns the RMS value of the pixels of frame F. Notice that both of these measurements are based on the input and output first-order temporal frame differences: O_n-O_{n-1} and D_n-D_{n-1}. This temporal differencing operation allows m_2 and m_3 to measure how the video system distorts time and hence motion. The time collapsing function for m_2, given by f_{time}, involves high-pass filtering the time series before calculating its standard deviation. For m_3 we simply take the maximum, or "worst case" value of the time series. It seems quite reasonable that the human perceptual system would also work on a "worst case" basis, at least for video sequences of 9 second duration.

Having selected a set of measurements, it remains only to find the four constants that define their best linear combination. In order to test the generality of our solution, we randomly select half of the 128 video scenes as training scenes and the other half become testing scenes. We calculate the linear regression of the three measurements on the true scores over the 64 scenes in the training group. The solution is:

$$c_0 = 4.7485 \qquad c_1 = -.9553 \qquad (7)$$
$$c_2 = -.3331 \qquad c_3 = -.3341 \, .$$

Since the measurements have been normalized for unit variance, we can interpret the magnitudes of c_1, c_2, and c_3 as indications of the relative importance of m_1, m_2, and m_3 in modeling the true scores.

4. Model Fit and Prediction Performance

Equations 4 through 7 together describe an objective model for subjective video impairment scores. Figure 2 is a scatter plot showing the relationship between true scores and those of the model. The spread of the 64 data points about the line y=x is a measure of the goodness-of-fit of the model. This third-order linear model explains 84.2% of the variance in the true scores. Equivalently, the coefficient of correlation between the model scores and the true scores is $(.842)^{½}$=.918. As a consequence of the least squares solution, the errors between the true scores and the modeled scores have zero mean. The standard deviation of the errors is .53 impairment units.

Now we consider our model as a video quality assessment system that provides predictions of true subjective scores and we evaluate its performance on the 64 testing scenes. Since these scenes did not enter into the regression equations, they can help us to verify the generality of our result. First we apply a clipping function to prevent predictions outside the valid range:

$$Clip(\hat{s}) = 1 \, , \; when \; \hat{s} < 1 \; , \; Clip(\hat{s}) \; = \; 5 \; , \; when \; \hat{s} > 5$$
$$Clip(\hat{s}) \; = \; \hat{s} \; , \; otherwise \; . \qquad \textbf{(8)}$$

Figure 3 shows the relationship between true and predicted scores for the 64 testing scenes. For this particular set of scenes, the predictions are slightly biased to the low side, with an average error of .27 impairment units. The error standard deviation is .43 impairment units. The coefficient of correlation is .947, which happens to be higher than the model correlation on the training data.

This commendable predictive performance indicates that the model fits more than just the training scenes. It is clear, however, that there are scenes that present a significant measurement challenge. This is not surprising in light of the wide range of scenes and impairments that these 64 data points represent. As an example, bit errors in a highly compressed digital video link can cause a large colored block to appear for 1 or 2 seconds. This impairment is dramatically different from effects of low SNR in an analog system, often described as "snow". Yet the assessment system described by equations 4 through 8 is dealing with both of these conditions along with many others.

When the video system under test is a distribution or broadcast channel, access to both input and output video must be considered. In order to implement the assessment system, one must transmit a small amount of side information from one end of the channel to the other. In particular, for each frame we can send the three scalar values; $std_{space}(Sobel(O_n))$, $RMS(O_n-O_{n-1})$, and $std_{space}(O_n-O_{n-1})$, from the input to the output. This requires an uncoded data rate of roughly 1.5 Kbps. Depending on the type of system being tested, this data might be carried in the vertical blanking interval. The measurements m_1, m_2, m_3 and the predicted score can then be computed at the output end of the channel.

5. Concluding Remarks

We have presented a method for deriving an objective video quality assessment system that emulates human perception. The prototype system shows much promise, with prediction errors on the order of .5 impairment units when tested on a set of 64 scenes. We continue with research to improve our model, strengthen and test its predictive power, and enhance its generality. Components of this assessment system are currently being considered for inclusion into the draft standard of "Analog Interface Performance Specifications for Digital Video Teleconferencing/Video Telephony Service" by the American National Standards Institute (ANSI) Accredited Standards Committee T1, Working Group T1Q1.5, Video Teleconferencing/Video Telephony Sub-working Group. The research described here is being conducted at the Institute for Telecommunication Sciences in Boulder, Colorado under sponsorship of the U.S. Department of Commerce. In addition to the authors, research participants include Arthur Webster, Margaret Pinson, Coleen Jones, and Paul King, who are members of the System Performance Standards Group.

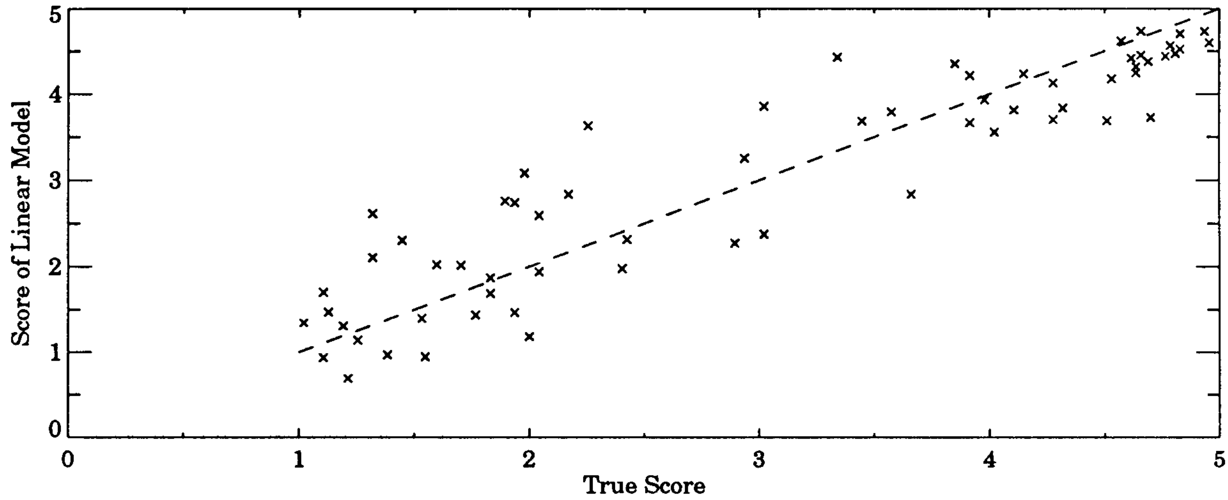

Figure 2 Model Fit, 64 Training Scenes

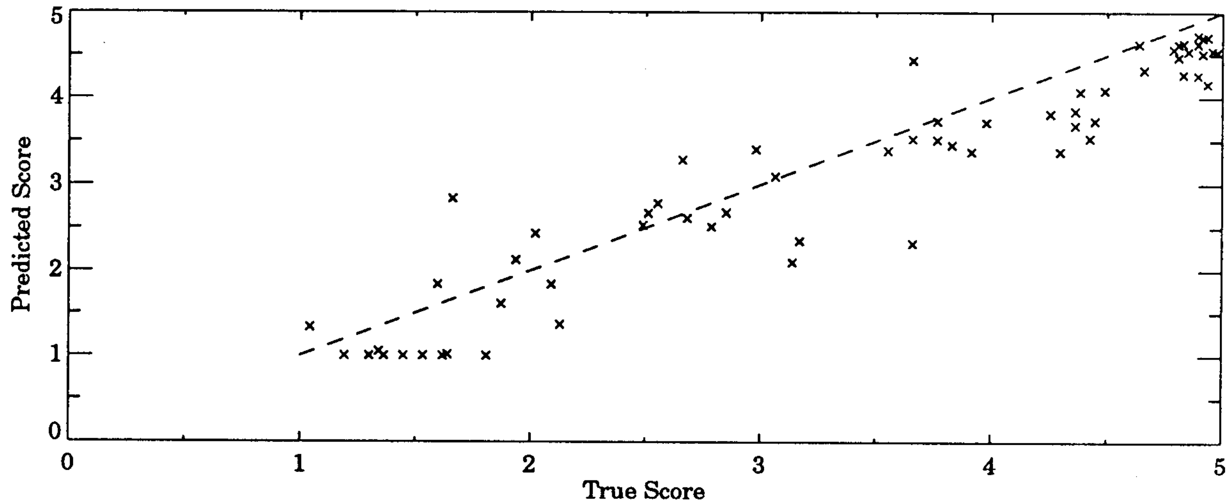

Figure 3 Prediction Performance, 64 Testing Scenes

References

1. EIA-250-B Electrical Performance Standard for Television Relay Facilities, Electronic Industries Association, Washington, D.C., US, 1976.

2. A. N. Netravali, and B. G. Haskell, Digital Pictures: Representation and Compression, Plenum Publishing Corporation, 1988.

3. CCIR Recommendation 500-3, Method for the Subjective Assessment of the Quality of Television Pictures.

4. Recommendations on Uniform Color Spaces - Color Difference Equations, Psychometric Color Terms, CIE Supplement No. 2 to CIE Publication No. 15 (E-1.3.1) 1971/(TC-1.3), 1978.

5. S. Wolf, Features for Automated Quality Assessment of Digitally Transmitted Video, U.S. Department of Commerce, National Telecommunications and Information Administration Report 90-264, June, 1990.

6. S. Wolf, et al., "Objective Quality Assessment of Digitally Transmitted Video" and "The Development and Correlation of Objective and Subjective Video Quality Measures", in Proceedings of IEEE Pacific Rim Conference on Communication, Computers and Signal Processing, Victoria, BC., Canada, May 1991.

7. A.K. Jain, Fundamentals of Digital Image Processing, Prentice-Hall Inc., Englewood Cliffs, New Jersey, US, 1989.