THE DEVELOPMENT AND CORRELATION OF OBJECTIVE AND SUBJECTIVE VIDEO QUALITY MEASURES

Stephen D. Voran and Stephen Wolf

U.S. Department of Commerce
NTIA/ITS.N3
325 Broadway
Boulder, CO 80303

Abstract

A primary focus of video quality assessment work is to obtain objective measures of performance that correlate well with subjective test scores. The Institute for Telecommunication Sciences, National Telecommunications and Information Administration, maintains and operates a subjective viewing laboratory that complements the objective video quality measurement program. The subjective viewing laboratory conforms with CCIR Recommendation 500-3, the international specification for conducting subjective viewing tests. The selection of source scenes for the subjective and objective tests is an important issue. Of particular importance are the spatial and temporal information content of each scene. These parameters play a critical role in determining the amount of video compression that is possible and, consequently, the level of video quality that is attainable when the video scene is transmitted over a fixed-rate digital channel. The test scenes are passed through a variety of distortion inducing devices and then rated by a panel of viewers in the viewing laboratory. The objective video quality measurement system is applied to the same test scenes. The result is a set of objective measures coupled with a set of subjective ratings of impairment. These data sets provide the basis for the derivation of a mapping from objective measures to subjective ratings. Such a mapping or algorithm provides an automated means for predicting subjective video quality without the use of viewer panels.

Introduction

Traditional objective measurements of the performance of video transport and storage systems often do not correlate well with video quality perceived by video system users. As an example, a system with a weighted signal-to-noise ratio of 20 dB can produce video of poor or acceptable quality, depending on the power spectrum of the noise and the characteristics of the video signal. This paper describes a method for deriving objective measurements of video quality that do agree with the human perception of video quality. Because these measurements provide meaningful information without the expense and effort of a test of a large group of human subjects in a controlled viewing environment, they are valuable to designers of video components and systems and to persons involved in standards work.

The development process presented here is based on three components, diagramed in Figure 1. First, a wide-ranging library of video scenes is digitized and processed. The processing extracts and quantifies the attributes of the video scenes that are important to the human visual and perceptual systems. The resulting data forms the objective data set. The details of this first component are described in a companion paper[1]. The second component involves the collection of viewer quality ratings for this same video library. This results in a set of subjective data. The final component of the process is a simultaneous statistical analysis of the subjective and objective data sets. The results of this analysis will be used to build a prediction algorithm that generates accurate predictions of subjective quality based on objective measurements. This final component is currently being developed.

This paper focuses on the second component of the process. First, the facility used to gather viewer quality ratings, The Institute for Telecommunication Sciences (ITS) Video Viewing Laboratory, is described. Next, the test methodology is presented, along with a discussion of the selection of video scenes for the tests. A final section describes plans for the integration of subjective and objective data sets.

Video Viewing Laboratory

The ITS Video Viewing Laboratory provides ITS with a location for conducting subjective measurements of video quality in accordance with CCIR Recommendation 500-3. Recommendation 500-3 specifies a standard visual environment for conducting picture quality assessment. The recently completed ITS facility conforms to these recommendations in all respects except for the color temperature of room illumination. The 14 x 11 x 8 foot room is finished with full length white drapes on three sides and grey drapes on the fourth (rear) wall. Grey carpet completes the subdued visual environment and provides noise reduction. An offset stud wall reduces noise transfer from an adjacent office area and a specially designed air handling system operates with a minimum of noise.

The six light fixtures in the laboratory are independently steerable and dimmable. This flexible lighting feature allows one to obtain the specified screen-to-background luminance ratio while simultaneously providing comfortable lighting conditions for writing. Recommendation 500-3 indicates that a light source with a color temperature of 6500 degrees Kelvin is preferred, presumably to match NTSC white. Several problems (power requirements, heat dissipation, bulb life, fixed light output) seem to be intrinsically linked to such sources. As a practical matter, ITS chose to use color-corrected incandescent bulbs. The tinted bulbs presently in use have a maximum color temperature of 4000 degrees Kelvin. This means that the "white" room illumination does not exactly match NTSC white. Given the adaptive nature of human color perception, effects of this fixed background color error are expected to be minimal. If necessary, color correction filters may be added in order to more closely approximate 6500 degree sources.

Video scenes are displayed on a broadcast quality monitor. The monitor can accept component video signals and display them with a maximum of 900 lines of horizontal resolution on a 19-inch screen. The monitor contains an option that allows repeatable monitor setup which is referenced to digitally stored values. This feature eliminates monitor setup as a potential source of variation in ITS video quality assessment tests. Comfortable seating for three viewers is provided at a distance of six picture heights from the monitor screen. By restricting the number of viewing locations to three, the viewers in the end locations are only 20 degrees off-center and potential test variations due to viewer location are minimized. The laboratory is equipped with hidden studio quality speakers to allow for recorded voice instructions as well as future subjective video-with-audio quality assessment tests.

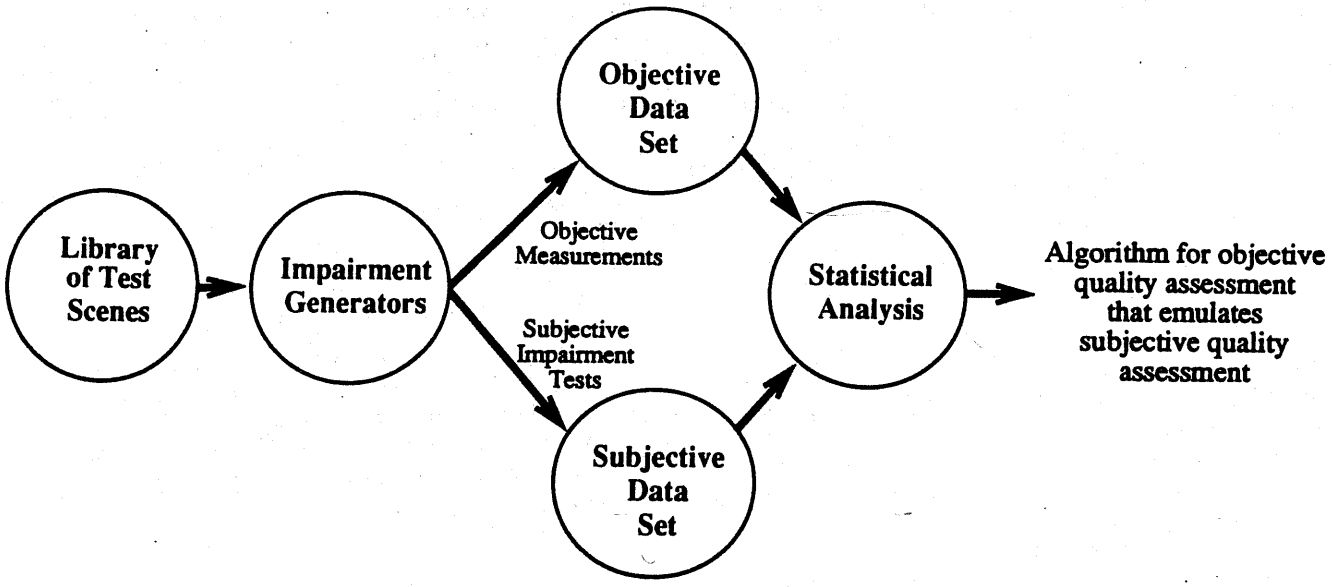

Figure 1. Algorithm Development Process

Test Methodology

ITS will conduct viewing sessions in the new video viewing laboratory to gather subjective video quality measurements. These measurements will build subjective data sets which complement the objective data sets. The methodology described here is based on standard subjective video testing procedures augmented by the results of initial viewing sessions conducted at ITS. Those tests used a viewer questionnaire to determine the testing rates and scales most likely to provide accurate responses.

Subjects for the viewing sessions will be selected at random from the site telephone directory. This gives a pool of 1,750 prospective viewers. The pool of viewers includes maintenance workers, office workers, administrators, scientists, and engineers. The use of a large, random sample of viewers from this diverse pool should substantially reduce any occupational bias that might be present. By recording the occupation of each viewer as a potential correlate, subjective quality measures within occupation can be calculated and results can be weighted to approximate those of other populations. In addition, each subject's exposure to several types of video, including broadcast and cable television, video teleconferencing and computer video displays will be recorded as potential correlates. In accordance with CCIR Recommendation 500-3, each viewer will be given a visual acuity test and a color vision test. The results of these tests along with the age and sex of the viewer will become part of the viewer's data record.

Each viewer will participate in four twenty-minute sessions. These sessions will be conducted on four consecutive days. The first third of the initial session is a training period. The training period exposes the subjects to a wide range of video impairments, many of which may be new to the subjects, as well as a wide range of scene types which can also affect the perceived severity of the impairments. The final portion of the training period allows the subjects to practice marking the response form between scenes. The remainder of the initial session and the subsequent three sessions contain 30-second impairment tests. Here the subject is presented with nine seconds of a scene, three seconds of grey screen, nine seconds of the same scene as distorted by some transmission channel or storage medium, and finally a nine second period in which to mark the response form. The subjects are instructed to decide on and mark the level of impairment in the second scene, using the first scene as a reference. The five possible responses offered are: imperceptible, perceptible but not annoying, slightly annoying, annoying, and very annoying. This scale intentionally covers a very wide range of impairment levels in a nonlinear fashion. By including reference scenes, impairment tests take advantage of the fact that the human eye excels at making comparisons. To reduce unwanted comparison effects, the order of scene presentation is randomized. Impairment tests also tend to reduce inter-laboratory testing variances.

After allowance for training periods, rest intervals, and some redundancy (to provide consistency checks), the cumulative 80 minutes of testing allow for the viewing and rating of 127 test scenes. In order to hold the subjects' interest, this body of 127 scenes is composed of roughly 30 distinct scenes. Thus, an average of 4 impairments per scene can be investigated. These distorted or degraded scenes are drawn from a larger pool of 14 different distortions caused by digital encoding, compression, and transmission, or analog encoding, followed by a bandlimited or noisy transmission channel or storage medium. Examples from the digital class include video codecs operating over real or simulated digital networks with line rates that range from 45 Mbps down to 64 Kbps. Examples from the analog category include NTSC encode/decode cycles, VHS record/play cycles, and laboratory-controlled low signal strength RF transmission links.

Selection of Test Scenes

The selection of the scenes used in the impairment tests is an important issue. In particular, the spatial and temporal information content of a scene are critical parameters. These parameters play 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.

Video compression schemes attain various degrees of compression by the removal of the spatial and temporal statistical redundancies of the video signal. For a highly detailed scene, where the scale of the details is comparable to the scale of the spatial sampling grid and the details are hard to predict, the video signal has very little spatial statistical redundancy. In this case, little compression can be gained by removing the small amount of spatial statistical redundancy. On the other hand, a scene with few details or with highly predictable details can often be greatly compressed. A parallel situation exists in the time domain. Scenes

with large amounts of unpredictable motion have little redundancy in their temporal statistics and little temporal compression can be expected. Scenes with no motion or limited motion are highly redundant in their temporal statistics and can often be greatly compressed.

In light of the direct links between spatial and temporal information content, potential for compression, and potential for transmission at a given rate, fair and relevant video quality measurements must use video scenes with spatial and temporal information content that is consistent with the video services that the device or system under test was intended to provide. As an example, video scenes of a soccer game contain too much spatial and temporal information to be useful in testing the performance of a codec designed to provide video teleconferencing services over 64 Kbps lines. In order to make measurements across the wide range of impairment types mentioned above, the impairment tests use scenes with widely varying amounts of spatial and temporal information. Figure 2 shows the relative amounts of spatial and temporal information for some possible test scenes. Inappropriate combinations of scene and device-under-test will be eliminated before a random scene selection process is implemented, resulting in a pseudorandom scene selection process.

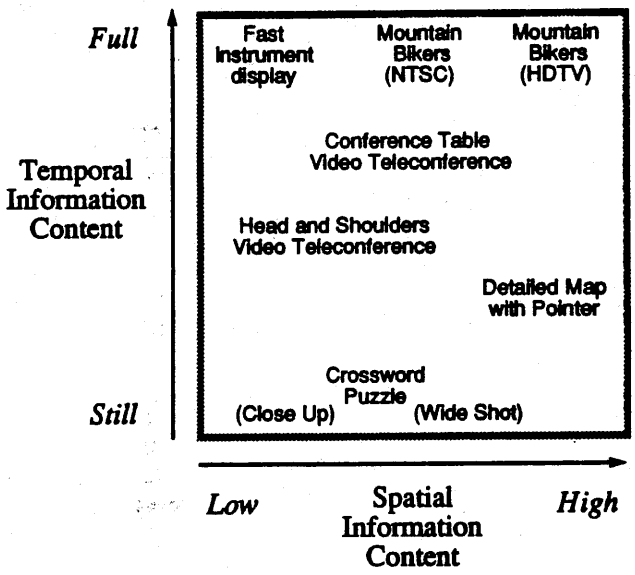

Figure 2. Library of Test Scenes

Integration of Subjective and Objective Data Sets

After all subjects have participated in all viewing sessions, the distribution of impairment ratings for a given scene is used to calculate a mean impairment rating and a confidence interval for that mean value. In initial test sessions, a panel of 41 viewers used a 5-point quality scale to rank 25 scenes. The variance of the scores awarded by this panel was such that the 95 percent confidence intervals for the mean values typically had a length of one-half quality unit. This allows eight confidence intervals to fit into the entire quality range, resulting in up to eight statistically distinguishable mean values. ITS considers this an appropriate number.

Once mean subjective impairment scores with acceptable confidence intervals are acquired, the mean scores will be compared with objective measurements, or parameters, that have been extracted from the same video scenes. Some of these parameters are described in reference 2. Based on this comparison, a mapping from the multi-dimensional objective measurement space to the one-dimensional subjective score space will be derived. Any objective measurements that prove to be redundant or not significant in the mapping will be discarded. By selecting a set of video scenes with widely varying amounts of spatial and temporal information and subjecting them to a wide range of impairments, ITS intends to generate a mapping that is accurate over a wide spread of quality levels for many different scene types. This mapping, along with the automated objective video measurement hardware and software, will provide a fully automated, objective estimate of perceived video quality.

Conclusion

The ITS video quality measurement system extracts objective measurements from carefully selected digitized test scenes. The same test scenes are evaluated by a panel of viewers under controlled viewing conditions. The resulting subjective and objective data sets provide the foundation for the design of an automated objective video quality assessment algorithm that provides accurate predictions of perceived video quality without the effort and expense of polling a large group of human subjects in a controlled environment. Such an algorithm will be valuable to persons involved in standards work as well as those involved in video equipment and system design.

References

[1] S. Wolf, M. Pinson, S. Voran, and A. Webster, "Objective Quality Assessment of Digitally Transmitted Video," in Proceedings of IEEE Pacific Rim Conference on Communications, Computers, and Signal Processing, 1991.

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