An objective video quality assessment system based on human perception

Arthur A. Webster, Coleen T. Jones, Margaret H. Pinson,
Stephen D. Voran, Stephen Wolf

Institute for Telecommunication Sciences
National Telecommunications and Information Administration
325 Broadway, Boulder, CO 80303

ABSTRACT

The Institute for Telecommunication Sciences (ITS) has developed an objective video quality assessment system that emulates human perception. The system returns 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 the capability for making meaningful video quality evaluations without convening viewer panels. The issue is timely because compressed digital video systems present new quality measurement questions that are largely unanswered.

The perception-based system was developed and tested for a broad range of scenes and video technologies. The 36 test scenes contained widely varying amounts of spatial and temporal information. The 27 impairments included digital video compression systems operating at line rates from 56 kbits/sec to 45 Mbits/sec with controlled error rates, NTSC encode/decode cycles, VHS and S-VHS record/play cycles, and VHF transmission. Subjective viewer ratings of the video quality were gathered in the ITS subjective viewing laboratory that conforms to CCIR Recommendation 500-3. Objective measures of video quality were extracted from the digitally sampled video. These objective measurements are designed to quantify the spatial and temporal distortions perceived by the viewer.

This paper presents the following: a detailed description of several of the best ITS objective measurements, a perception-based model that predicts subjective ratings from these objective measurements, and a demonstration of the correlation between the model's predictions and viewer panel ratings. A personal computer-based system is being developed that will implement these objective video quality measurements in real time. These video quality measures are being considered for inclusion in the Digital Video Teleconferencing Performance Standard by the American National Standards Institute (ANSI) Accredited Standards Committee T1, Working Group T1A1.5.

1. INTRODUCTION

The need to measure video quality arises in the development of video equipment and in the delivery and storage of video and image information. Although the work described in this paper is concerned specifically with NTSC video (the distribution television standard in the United States), the principles presented can be applied to other types of motion video and even still images. The methods of video quality assessment can be divided into two main categories: subjective assessment (which uses human viewers) and objective assessment (which is accomplished by use of electrical measurements). While we believe that assessment of video quality is best accomplished by the human visual system, it is useful to have objective methods available which are repeatable, can be standardized, and can be performed quickly and easily with portable equipment. These objective methods should give results that correlate closely with results obtained through human perception.

Objective measurement of video quality was accomplished in the past through the use of static video test scenes such as resolution charts, color bars, multi-burst patterns, etc., and by measuring the signal to noise ratio of the video signal.[1] These objective methods address the spatial and color aspects of the video imagery as well as overall signal distortions present in traditional analog systems. With the development of digital compression technology, a large number of new video services have become available. The savings in transmission and/or storage bandwidth made possible with digital compression technology depends upon the amount of information present in the original (uncompressed) video signal, as well as how much quality the user is willing to sacrifice. Impairments may result when the information present in the video signal is larger than the transmission channel capacity. However, users may be willing to sacrifice quality to achieve a substantial reduction in transmission and

storage costs. But, how much quality is sacrificed for how much cost savings? We propose a set of measurements that offers a way to begin to answer this question. New impairments can be present in digitally compressed video and these impairments include both spatial and temporal artifacts.[2] The old objective measurement techniques are not adequate to assess the impact on quality of these new artifacts.[3]

After some investigation of compressed video, it becomes clear that the perceived quality of the video after passing through a given digital compression system is often a function of the input scene. This is particularly true for low bit-rate systems. A scene with little motion and limited spatial detail (such as a head and shoulders shot of a newscaster) may be compressed to 384 kbits/sec and decompressed with relatively little distortion. Another scene (such as a football game) which contains a large amount of motion as well as spatial detail will appear quite distorted at the same bit rate. Therefore, we directed our efforts toward developing perception-based objective measurements which are extracted from the actual sampled video. These objective measurements quantify the perceived spatial and temporal distortions in a way that correlates as closely as possible with the response of a human visual system. Each scene was digitized (at 4 times sub-carrier frequency) to produce a time sequence of images sampled at 30 frames per second (in time) and 756 x 486 pixels (in space).

2. DEVELOPMENT METHODOLOGY

Figure 1 presents a graphical depiction of the development process for the ITS quality assessment algorithm. A set of video scene pairs (each consisting of the original and a degraded version) was used in a subjective test. These scene pairs were also processed on a computer that extracted a large number of features. Statistical analysis was used to select an optimal set of quality parameters (obtained from features) that correlated well with the viewing panel results. This optimal set of parameters was then used to develop a quality assessment algorithm that gives results that agree closely with viewing panel results.

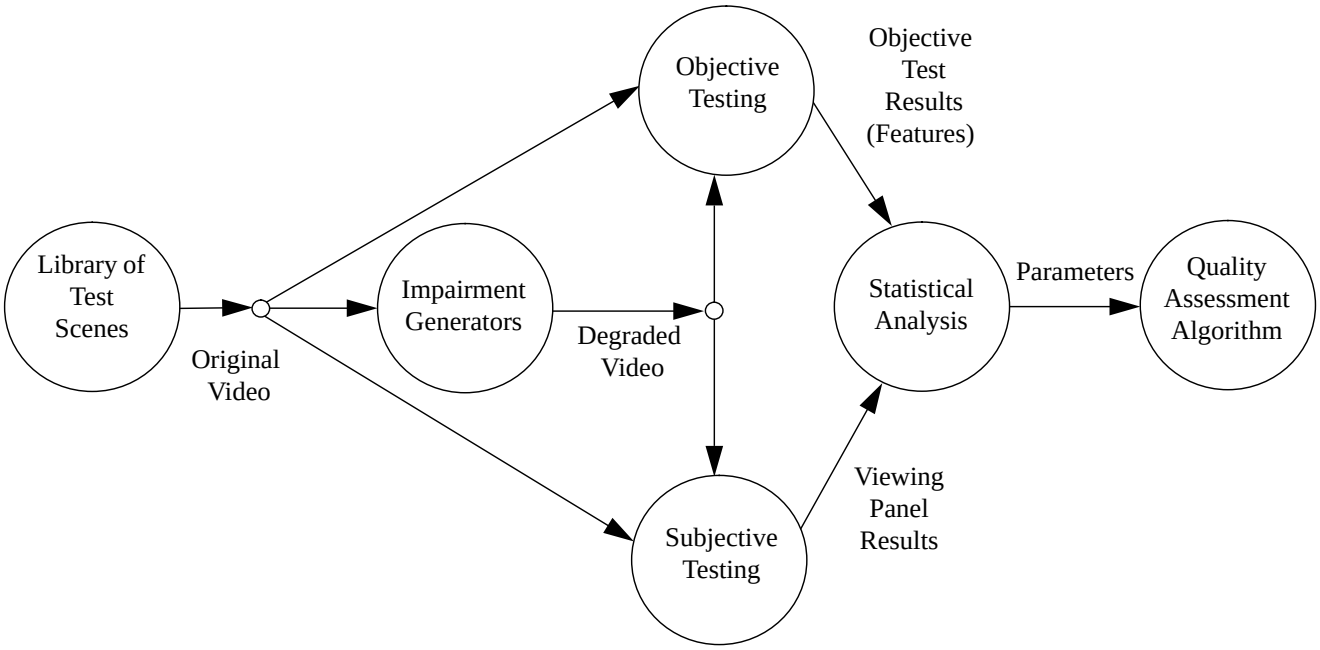

Figure 1. Development Process for Video Quality Assessment Algorithm

2.1 Library of test scenes

Several scenes, exhibiting various amounts of spatial and temporal information content, are needed to characterize the performance of a video system. Even more scenes are needed to guard against viewer boredom during the subjective testing. A set of 36 test scenes was chosen for the experiment. The test scenes spanned a wide range of user applications including still scenes, limited motion graphics, and full motion entertainment video.

2.2 Impairment generators

Twenty-seven video systems (plus the 'no impairment' system) were used to produce the degraded video that was used in the tests. The original video for this test was component analog video. The digital video systems included 11 video codecs (coder-decoders) from 7 manufacturers operating at bit rates from 56 kbits/sec to 45 Mbits/sec including bit error rates of 10^{-6} and 10^{-5}. Also included were analog video systems such as VHS and S-VHS recording and playback, and noisy RF transmission. All video systems except the 'no impairment' system included NTSC encoding and decoding.

2.3 Objective testing

Both the original video and the degraded video were digitized and processed to extract a large number of features. The processing included Sobel filtering, Laplace filtering, fast Fourier transforms, first-order differencing, color distortion measurements[4], and moment calculations. Typically, features were calculated from each original and degraded frame of the video sequence to produce time histories. Some features required the entire original and degraded video image (e.g., the variance of the error image calculated from the difference between the original and the degraded images). Other features required only the statistics of the original and degraded video images (e.g., the change in image energy obtained from the differences between the original and the degraded image variances). The time histories of the features were collapsed by various methods, e.g., maximum (MAX), root mean square (RMS), standard deviation (STD), etc., to produce a single scalar value (or parameter) for each test scene. These parameters defined the objective measurements and were used in the statistical analysis step shown in Figure 1.

2.4 Subjective testing

The subjective test was conducted in accordance with CCIR Recommendation 500-3.[5] A panel of 48 viewers were selected from the U.S. Department of Commerce Laboratories phone book in Boulder, Colorado. Each viewer completed four viewing sessions during a single week, attending one session per day. Each session lasted approximately 25 minutes and required viewing of 38 or 40, 30-second test clips. A clip is defined as a test scene pair consisting of the original video and the degraded video. The viewer was first shown the original video for 9 seconds followed by 3 seconds of grey and then 9 seconds of the degraded video. 9 seconds was allowed to rate the impairment on a 5 point scale before the next clip was presented. The viewer was asked to rate the difference between the original video and the degraded video as either (5) Imperceptible, (4) Perceptible but Not Annoying, (3) Slightly Annoying, (2) Annoying, or (1) Very Annoying. This scale covers a wide range of impairment levels and is specified as one of the standard scales in the CCIR Recommendation 500-3. Impairment testing was used since we were interested in measuring the change in video quality due to a video system. A mean opinion score was generated by averaging the viewer ratings.

The selection of 158 clips used in the test (out of 972 clips available) was made both deterministically and randomly. Random selections were made from a distribution table that paired video teleconferencing systems with more video teleconferencing scenes than entertainment scenes, and entertainment systems with more entertainment scenes than video teleconferencing scenes. The viewers rated 132 unique clips from the 158 actually viewed because some were used for training and consistency checks.

2.5 Statistical analysis and quality assessment system

This stage of the development process utilized joint statistical analysis of the subjective and objective data sets. This step identifies a subset of the candidate objective measurements that provides useful and unique video quality information. The best measurement was selected by exhaustive search. Additional measurements were selected to reduce the remaining objective-subjective error by the largest amount. Selected measurements complement each other. For instance, a temporal distortion measure was selected to reduce the objective-subjective error remaining from a previous selection of a spatial distortion measure. When combined in a simple linear model, this subset of measurements provides predicted scores that correlate well with the true scores obtained in the subjective tests. In constructing the linear model we looked for p measurements $\{m_i\}$ and $p+1$ constants $\{c_i\}$, that allowed us to estimate the subjective mean opinion score. The estimated subjective mean opinion score is

given by

$$s \approx \hat{s} = c_0 + \sum_{i=1}^{p} c_i m_i, \tag{1}$$

where s is the true subjective mean opinion score and $\hat{s}$ is the estimated score.

3. RESULTS

For the results presented here, three complementary video quality measurements (p=3) were selected. These three complementary measures (m_1, m_2, and m_3) have been used to explain most of the variance in subjective video quality that resulted from the impairments used in this experiment. The investigations and research that produced the m_1, m_2, and m_3 video quality metrics also provided insight into how the human perceives the spatial and temporal information of a video scene.

3.1 Spatial and temporal information features

The difficulty in compressing a given video sequence depends upon the perceived spatial and temporal information present in that video sequence. Perceived spatial information is the amount of spatial detail in the video scene that is perceived by the viewer. Likewise, perceived temporal information is the amount of perceived motion in the video scene. Thus, it would be useful to have approximate measures of perceived spatial and temporal information. These information measures could be used to select test scenes that appropriately stress the video compression system being designed or tested. Two different test scenes with the same spatial and temporal information should produce similar perceived quality at the output of the transmission channel. Measures of distortion could also be obtained by comparing the perceived information content of the video before and after passing through a video system. Although it is recognized that spatial and temporal aspects of vision perception cannot be completely separated from each other, we have found spatial and temporal features that correlate with human quality perception of spatial detail and motion. Both of these features require pixel differencing operations, which seem to be basic attributes of the human visual system. The spatial information (SI) feature differences pixels across space while the temporal information (TI) feature differences pixels across time. Here, both the SI and TI features have been applied to the luminance portion of the video.

3.1.1 Spatial information (SI)

The spatial information feature is based on the Sobel filter.[6] At time n, the video frame F_n is filtered with the Sobel operators. The standard deviation over the pixels in each Sobel-filtered frame is then computed. This operation is repeated for each frame in the video sequence and results in a time series of spatial information values. Thus, the spatial information feature, $SI\,[F_n]$, is given by

$$SI\,[F_n] = STD_{space}\{Sobel\,[F_n]\}, \tag{2}$$

where STD_{space} is the standard deviation operator over the horizontal and vertical spatial dimensions in a frame, and F_n is the n^{th} frame in the video sequence. Figure 2 shows a time sequence of 3 contiguous video frames for an original scene (top row) and degraded version of that scene (second row). These images were sampled at the NTSC frame rate of approximately 30 frames per second. The degraded version of the scene was obtained from a 56 kbits/sec codec. The third row of Figure 2 shows the Sobel filtered version of the original scene and the fourth row shows the Sobel filtered version of the degraded scene. The highly localized, clearly focussed edges in the third row produce a large STD_{space} since the standard deviation is a measure of the spread in pixel values. On the other hand, the non-localized, blurred edges shown in the fourth row produce a smaller STD_{space}, demonstrating that spatial detail has been lost. This is particularly evident for the images in the third column.

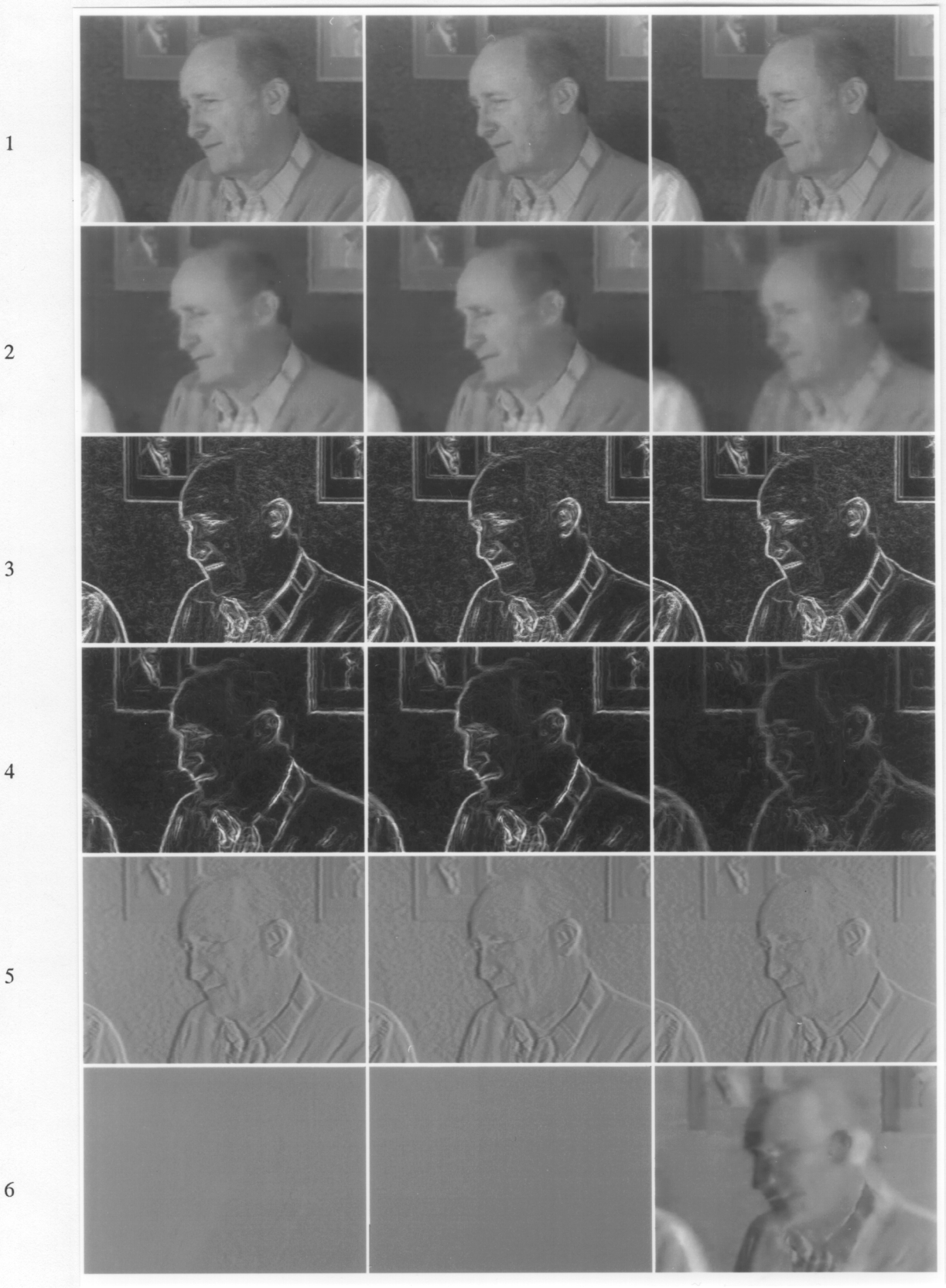

Figure 2. Video Processed to Demonstrate Perceived Spatial and Temporal Information

DOCKET ALARM

Explore Litigation Insights

Docket Alarm provides insights to develop a more informed litigation strategy and the peace of mind of knowing you're on top of things.

Real-Time Litigation Alerts

Keep your litigation team up-to-date with **real-time alerts** and advanced team management tools built for the enterprise, all while greatly reducing PACER spend.

Our comprehensive service means we can handle Federal, State, and Administrative courts across the country.

Advanced Docket Research

With over 230 million records, Docket Alarm's cloud-native docket research platform finds what other services can't. Coverage includes Federal, State, plus PTAB, TTAB, ITC and NLRB decisions, all in one place.

Identify arguments that have been successful in the past with full text, pinpoint searching. Link to case law cited within any court document via Fastcase.

Analytics At Your Fingertips

Learn what happened the last time a particular judge, opposing counsel or company faced cases similar to yours.

Advanced out-of-the-box PTAB and TTAB analytics are always at your fingertips.

API

Docket Alarm offers a powerful API (application programming interface) to developers that want to integrate case filings into their apps.

LAW FIRMS
Build custom dashboards for your attorneys and clients with live data direct from the court.

Automate many repetitive legal tasks like conflict checks, document management, and marketing.

FINANCIAL INSTITUTIONS
Litigation and bankruptcy checks for companies and debtors.

E-DISCOVERY AND LEGAL VENDORS
Sync your system to PACER to automate legal marketing.