AN OBJECTIVE TECHNIQUE FOR ASSESSING VIDEO IMPAIRMENTS

Stephen Voran and Stephen Wolf

Institute for Telecommunication Sciences, U.S. Department of Commerce
NTIA/ITS.N3, 325 Broadway, Boulder, Colorado 80303

ABSTRACT

The Institute for Telecommunication Sciences is deriving new techniques for assessing impairments induced by video transport and storage systems. These techniques are based on digital image processing operations performed on digitized original and impaired video sequences. Measurements that quantify perceptual video attributes in both the spatial and temporal domains are extracted from the digitized video. These measurements are then used to compute a single score that quantifies the perceptual impact of the impairments present in the video sequence. This objective impairment score is well-correlated (r=.92) with impairment assessments made by human viewers. Thus, it can be used to augment, or possibly to replace, the expensive and time-consuming subjective viewing tests that are typically used to evaluate video coding and transmission techniques.

1. INTRODUCTION

When faced with the task of assessing the video impairments in a video storage or transport system in a meaningful and repeatable way, the video engineer has relatively few practical options. First of all, carefully conducted subjective tests can provide very useful data, but these tests must be conducted in a controlled viewing environment, and they require the participation of tens of viewers. Secondly, while many digital image processing techniques for assessing the displayed quality of a *single* image have been proposed, these techniques do not easily extend to cover the *continuing sequence* of images that are generated by a video signal. Thirdly, the objective video test signals typically used in the analog television industry (e.g., multiburst, color bars) are either static or are deterministic functions of time. While these signals and the associated measurements are useful for the characterization of the electrical performance of time-invariant analog video systems, the measurements often do not correlate well with the impairment level perceived by the users of the video system. Furthermore, video signals are now commonly transmitted and stored in compressed digital form and impairments may arise from the compression process or the presence of errors in the transmitted bit stream. Effective compression algorithms are dynamic, with the input video signal dictating the overall behavior of the algorithm through sub-algorithms that perform motion prediction, adaptive transforms, and adaptive quantization, to name only a few. The resulting systems are clearly time-varying and signal dependent. Static, deterministic test signals cannot provide an accurate characterization of their performance on real video sequences.

These observations motivated us to design a video impairment assessment technique that uses actual video signals. This approach provides a realistic measurement environment and allows measurements to be made while a video system is in use. The design process is described in Section 2. The process uses data from an extensive series of subjective video impairment tests, which is then correlated against a family of proposed objective video impairment measurements. Two effective measurements are presented in Sections 3.1 and 3.2. A measure of spatial impairment is based on normalized energy differences of Sobel-filtered video frames and a complementary temporal impairment measurement is based on the first-order frame difference sequence. A linear combination of these measurements generates a single objective impairment score that is well correlated with the subjective test data. In Section 3.3, we discuss the performance of this video impairment assessment technique and identify areas for continuing research.

2. SUBJECTIVE AND OBJECTIVE VIDEO IMPAIRMENT TESTS

We are seeking an objective assessment technique that quantifies video impairments in a way that correlates with subjective evaluations made by those who actually view the impaired video. It is clear that we must incorporate knowledge of human perception in the design of the assessment technique. Toward this end, we have conducted an extensive set of subjective video impairment tests. The tests were conducted following a methodology and viewing laboratory environment specified in the CCIR Recommendation 500-3[1]. Forty-eight viewers rated a collection of 132 impaired video sequences. The viewers, grouped 3 at a time, were shown a 9-second

original video sequence followed by an impaired version of that sequence and were then asked to rate the impact of the impairment. The five possible responses were: 5=imperceptible, 4=perceptible but not annoying, 3=slightly annoying, 2=annoying, and 1=very annoying. The responses of the 48 viewers were then averaged to obtain a mean value on a 5-point scale for each of the 132 impaired sequences.

The impaired video sequences were selected from a collection of 36 video sequences that were passed through a group of 28 video systems, both analog and digital. Digital systems range from slightly compressed (45 Mbps) to highly compressed (56 Kbps). Controlled bit errors were introduced into some of the digital video systems. Analog impairments included NTSC encode/decode cycles, VHS record/play cycles, and a noisy RF channel. In order to derive the most general system possible, the video sequences contained widely varying amounts of spatial and temporal information. Examples include sports events, newscaster, still shots, and graphics. The spatial and temporal information content of test scenes are important considerations because, in digital systems, information content determines sample rates and plays a crucial role in determining the level of impairment that is suffered when the video is transmitted over a fixed-rate digital channel. Similarly, in analog video systems, the spatial information content governs the relationship between system passband and video impairment level.

In addition to the subjective impairment tests, a set of candidate measurements based on digital image processing techniques were applied to digitized versions (24 bits/pixel, 756 pixels/line, 486 lines/frame, approximately 30 frames/second) of the original and impaired video sequences. In an exact parallel to the subjective tests, the objective measurements were all differential. That is, they involved both the original and the impaired versions of each video sequence. 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, multicolored, multidimensional signal. Useful impairment measures must take note of this human interpretation and mimic it to the extent possible. Thus, the candidate set of objective measures included those designed to measure temporal, luminance, chrominance, and spatial impairments to the video image sequence.

The design process hinges on a joint statistical analysis of the *objective* and *subjective* video impairment data sets. This analysis reveals which of the objective video impairment measurements are meaningful, and how they might be combined to create a video impairment assessment technique that correlates well with subjective test results. An effective pair of spatial and temporal video impairment measurements is discussed in the following section. Because color impairments in the data set were very small relative to other impairments, the measurements discussed below are applied only to the luminance portion of the video signal. More complete descriptions of the subjective and objective video impairment experiments can be found in [2], [3], and [4].

3. RESULTS AND DISCUSSION

The scanned nature of video signals makes the true separation of their spatial and temporal dimensions a rather complicated interpolation issue. While a single frame is not truly "a set of spatial samples at a fixed time instant," we make that approximation. Accordingly, we acknowledge that for the "spatial" and "temporal" video measurements presented here, the separation is not complete and the classification is only approximate.

3.1. Spatial Video Impairment Measurement

The selected measurement of spatial impairment is based on normalized energy differences of Sobel-filtered video frames. It is given by

$$m_s = \frac{\left| mean^2_{time}(x_t) - mean^2_{time}(y_t) \right|}{mean^2_{time}(x_t)}, \qquad (1)$$

$$x_t = std_{space}(Sobel(X_t)), \quad y_t = std_{space}(Sobel(Y_t)),$$

$$X_t: \text{Original Video Frame } t, \; Y_t: \text{Impaired Video Frame } t.$$

In Equation 1, std_{space} denotes a standard deviation calculation conducted over the visible portion of the pixel array. The temporal means are calculated over the entire 9-second sequence that was subjectively rated. Since the Sobel filter operation emphasizes edges or high-frequency content of a video frame[5], the measurement tracks changes in edge content or spatial resolution, quantities that are known to be perceptually important. Due to the absolute value function, either an increase (e.g., noise, false edges) or a decrease (e.g., reduced resolution) in these quantities causes a positive deflection of m_s. As the impairment level is reduced to zero, y_t must approach x_t, and m_s goes to zero. If one defines x_t to be the spatial information content of frame X_t, then m_s is a normalized measure of the change in spatial information content caused by the video system under test. Computations of m_s over a large set of video sequences have indicated that the Sobel filtering operation can be replaced with a "pseudo-Sobel" operation without significant impact on m_s. This substitution can provide a significant savings in computations since, for every pixel, the true Sobel filter combines the outputs of the vertical and horizontal convolutions via $[h^2+v^2]^{1/2}$, while the "pseudo-Sobel" approximates this with $|h|+|v|$.

3.2. Temporal Video Impairment Measurement

As mentioned earlier, we have only approximately separated the temporal and spatial dimensions of the video signal.

Because of this overlap, the temporal impairment measurement, m_t, that best complements the spatial impairment measurement m_s, may not be optimal for quantifying pure temporal impairments. Rather, the measurement was selected to correlate with the overall subjective impairment variations that are not tracked by m_s. This allows a linear combination of m_s and m_t to track the subjective data set as closely as possible in the mean-squared error sense.

The temporal impairment measurement is based on the first-order frame difference sequence. First, define

$$\Delta F_t = \text{mean}_{\text{space}}(|F_t - F_{t-1}|), \qquad (2)$$

where F_t is a video frame at time t, and the absolute value operation is applied to each pixel of the frame difference. When two successive frames are identical, ΔF will be zero. When motion is present in the video sequence, successive frames are different and this is reflected by ΔF. Thus, ΔF_t tends to track the amount of motion present in the video sequence at time t. Figure 1 demonstrates how the ΔF sequences provide a wealth of information regarding frame rate, continuity of motion, and noise levels, all of which are highly relevant to human perception. The broken line in Figure 1 shows values of ΔF for 180 frames ($\approx$6 seconds) of an original video sequence. The solid lines indicate values of ΔF for two impaired versions of that same sequence. The lower line displays runs of up to 10 zeros, indicating that this digital video encoder achieves bit-rate reduction by reducing its frame rate from approximately 30 frames/second to as low as 3 frames/second. Notice that the frame rate is adaptive. The contents of the video sequence determine how many bits are required by a frame update, and hence how often updates may be encoded. This temporal impairment is perceived as a variable level of "jerkiness" in the video sequence. The upper solid line in Figure 1 shows the output of a noisy video system that always generates approximately 30 frames/second. Here the random noise adds a somewhat constant amount of "false motion" to the video sequence.

When we compute the sequence defined in Equation 2 for a pair of original and impaired video sequences, we refer to them as ΔX_t and ΔY_t, respectively. To measure the dissimilarity between the two sequences, we compute temporal statistics of their log ratio and form the temporal impairment measurement m_t

$$m_t = [\max_{\text{time}}(s_t) - \min_{\text{time}}(s_t)] + \tfrac{3}{4}\,\text{mean}_{\text{time}}(s_t),$$

$$\text{where} \quad s_t = \log_{10}\left(\frac{\Delta Y_t}{\Delta X_t}\right). \qquad (3)$$

Here the spread in the log ratio is augmented by a fraction of the mean of the log ratio to form the overall measurement of temporal impairment. Notice that as impairments become small, ΔX_t will match ΔY_t, the log ratio will become zero, as will m_t. If one defines ΔF_t to be the temporal information content of frame F_t, then m_t is a measure of the "spreading" in temporal information content caused by the video system under test.

Since the "spreading" or "change in smoothness" of temporal information is a good indicator of temporal impairment, the flatness ratio should also be considered:

$$\frac{\text{flatness}(\Delta Y_t)}{\text{flatness}(\Delta X_t)} = \frac{\left[\prod_{i=0}^{N-1} \Delta Y_{t-i}\right]^{\frac{1}{N}} \left[\frac{1}{N}\sum_{i=0}^{N-1} \Delta Y_{t-i}\right]^{-1}}{\left[\prod_{i=0}^{N-1} \Delta X_{t-i}\right]^{\frac{1}{N}} \left[\frac{1}{N}\sum_{i=0}^{N-1} \Delta X_{t-i}\right]^{-1}} \cdot (4)$$

Here flatness(ΔF_t) is simply a discrete version of the spectral flatness measure[6]. It is also the ratio between the geometric mean and the arithmetic mean of ΔF_t. The flatness measure is a positive quantity that attains its maximum value of 1 when the sequence of ΔF_t maintains a constant value over the N sample measurement window. We have found that a window size of N=5 allows the temporal statistics of the flatness ratio to provide similar information to m_t but at somewhat greater computational expense.

We have also investigated a more complex algorithm that performs weighted integrals of ΔX_t between the spikes in ΔY_t. This attempt to more directly characterize the amount of motion that is "lost" or "displaced" due to frame rates below 30 frames/second shows good correlation to overall subjective impairment levels, but does not serve as an effective companion to m_s.

3.3. Discussion

Recall that the subjective assessment results were averaged across viewers to form a single score on the 5-point scale. The linear combination of the spatial and temporal impairment measurements (m_s and m_t) that minimizes the mean-squared error between the objective scores and the subjective scores is

$$\text{objective score} = 4.95 - 3.41 \cdot m_s - .46 \cdot m_t \qquad (5)$$

Since m_s and m_t are positive quantities that tend toward zero for unimpaired video sequences, the maximum objective score produced by Equation 5 is 4.95. This is consistent with our observation that when presented with identical video sequences, a small fraction of viewers will respond that the second of the two is visibly impaired relative to the first, resulting in a mean subjective impairment value that is slightly less than 5. The negative coefficients for m_s and m_t cause the objective score to decrease as spatial or temporal impairments increase. Estimates of the variance of the 3 regression coefficients in Equation 5 are .01 for the constant, .09 for the coefficient of m_s, and .01 for the coefficient of m_t. These estimates include a variance inflation factor of 1.7 to account

for the correlation between m_s and m_t. Thus, one would expect only small changes in the regression coefficients when the tests are repeated on a statistically similar data set. On the other hand, different coefficients may result when tests are conducted on data sets that have more restricted ranges of impairment levels.

For each of the 132 video sequences, the objective score is plotted against the subjective score in Figure 2. The RMS error between the two data sets is .54 impairment units, and the coefficient of correlation is r=.92. Examination of the errors reveals that the error distribution is nearly Gaussian, and has a constant variance. The objective assessment technique does not systematically err on a particular video system. This random nature of the errors means that the averaging of N results will tend to decrease the error by a factor of $N^{1/2}$. This is demonstrated in Figure 3, where groups of 3 to 7 objective and subjective scores have been averaged to generate a single objective score and a single subjective score for each of the 28 video systems in the data set. The mean value of the averaging factor, N, is 132/28 = 4.7, so we would expect the RMS error to decrease to $.54/4.7^{1/2} = .25$. In fact, Figure 3 reveals an RMS error of .25 impairment units, and a correlation of r=.98. These results are very encouraging, especially in light of the wide range of video sequences, video systems, and resulting impairment levels included in the data set.

The refinement, improvement, and implementation of this assessment technique continues. While the video digitizing and processing presented here were conducted in batch mode on a fairly powerful computer, we are currently constructing a PC-based implementation of the assessment technique. The PC is outfitted with a pair of commercially available image acquisition and processing cards that will enable it to digitize original and impaired video signals and to evaluate the perceptual impact of video impairments in real time. Additional details are available in [4].

Several additional facets of this work deserve further discussion. The first is the issue of time alignment. When testing video systems, there is always some delay between the system input and output. For the objective assessment technique to perform at its best, this delay must be accounted for when computing m_s and m_t. In compressed digital systems, the delay is often a function of the input video sequence. As a further complication, Figure 1 demonstrates that input frames might not generate a corresponding output frame. We have found that shifting smoothed versions of the ΔX_t and ΔY_t sequences to attain a minimum in the standard deviation of the difference sequence $(\Delta X_t - \Delta Y_{t-\tau})$ provides a fairly reliable and robust time alignment technique.

The measurements presented here are calculated on entire video frames. We are currently investigating their application over specific regions of interest. As an example, the video frame could be divided into a grid of 12 rectangular regions. The measurements m_s and m_t could then be applied in the region with the most motion and the region with the least motion, as determined by regional evaluations of ΔX_t. This approach is motivated by the fact that human resolution judgements and noise assessments are more critical in still areas than in moving areas, but human attention is normally drawn to moving objects. Thus, when properly combined, possibly through some dynamic weighting functions, the resulting four measurements might offer improved assessments over the existing technique. Computing a multiplicity of localized measurements has the additional advantage that individual errors may cancel as measurements are averaged or otherwise combined. Also, localized processing of very specifically shaped regions (e.g., narrow horizontal or vertical stripes) could allow the impairment assessment technique to detect and account for the different types of structure in video impairments. As an example, noise that is periodic in space or time tends to have a perceptual impact that differs from the perceptual impact of noise that is random.

When a video system has a constant non-unity gain or non-zero bias, these values must be normalized out of y_t and ΔY_t before m_s and m_t are calculated. Removing constant system gain and bias from the objective assessment technique does not destroy relevant impairment information since this is exactly what a viewer does when adjusting the contrast and brightness of a video display.

The 9-second video sequence length used in this work was selected for ease of subjective testing. In order to match the subjective tests, objective assessments were also performed over this 9-second window. Separate experiments should be conducted to determine what assessment window is most appropriate for continuous impairment assessment of extended video sequences. One possible experiment would ask viewers to note impairment changes as they perceive them throughout a 15-minute video selection with controlled impairment levels. A series of variable time-lag correlations between the subjective responses and the controlled impairment levels might then reveal what "temporal assessment windows" the viewers used. It would then be natural to use these same windows in the objective assessment technique.

4. CONCLUSION

We have presented an objective video impairment assessment technique that appears to quantify the perceptual impact of video impairments in an accurate way. The development uses 132 impaired video sequences that cover a remarkably wide range of motion, detail, impairment type, and impairment level. The objective impairment measurements show a correlation of r=.92 with carefully conducted subjective assessments. We have demonstrated that this correlation can be improved by repeating assessments and averaging the results. These results are very encouraging. Refinement and

improvement of the assessment technique continues, and several directions for continuing research have been noted. In addition, we are constructing a PC-based implementation of the technique. This instrument will digitize original and impaired video signals and evaluate the perceptual impact of the impairments in real time. The video impairment assessment technique is being considered for inclusion in the video teleconferencing performance standard that is being drafted by the ANSI Accredited Standards Committee T1, Working Group T1A1.5.

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 and National Communications System. In addition to the authors, research participants include Arthur Webster, Coleen Jones, Margaret Pinson, and Paul King, who are members of the System Performance Standards Group.

REFERENCES

[1] CCIR, Recommendation 500-3, *Method for the Subjective Assessment of the Quality of Television Pictures*, 1986.

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

[3] S. Voran and S. Wolf, "The Development and Evaluation of an Objective Video Quality Assessment System that Emulates Human Viewing Panels", *International Broadcasting Convention Technical Papers*, Amsterdam, The Netherlands, July 1992.

[4] A. Webster, et. al., "An Objective Video Quality Assessment System Based on Human Perception", *Proceedings of IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science & Technology*, San Jose, CA, US, February 1993.

[5] A.K. Jain, *Fundamentals of Digital Image Processing*, Prentice-Hall Inc., 1989.

[6] N.S. Jayant and P. Noll, *Digital Coding of Waveforms*, Prentice-Hall Inc., 1984.

Figure 1: ΔF_t for Original and Impaired Video Sequences

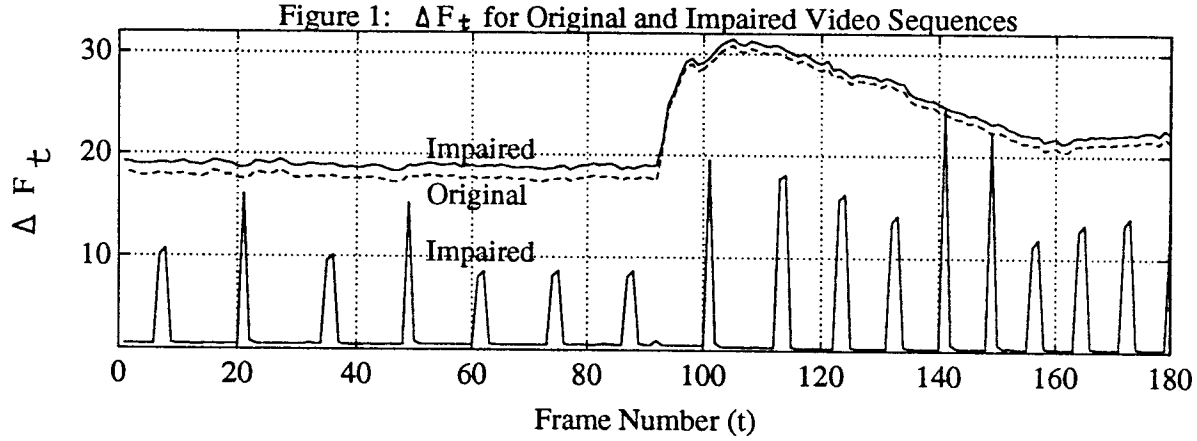

Figure 2: Objective vs Subjective
One Scene per Video System

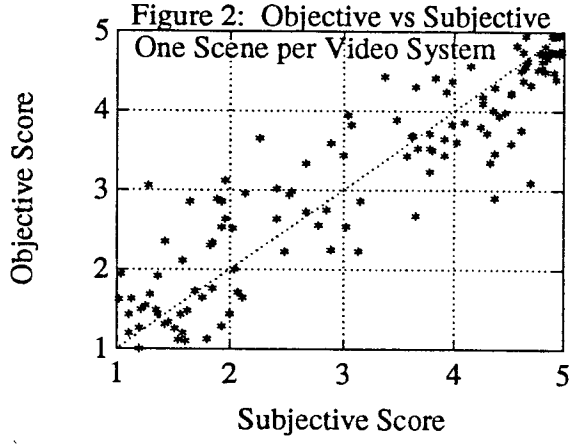

Figure 3: Objective vs Subjective
4.7 Scenes per Video System