In-Service Performance Metrics for MPEG-2 Video Systems

By Stephen Wolf and Margaret H. Pinson

Institute for Telecommunication Sciences
National Telecommunications and Information Administration
U.S. Department of Commerce

Abstract: *With the advent of new digital video systems that utilize compression to achieve a savings in transmission or storage bandwidth, the quality of the received output video can be dependent not only upon the inherent spatial and temporal information content of the input video but also upon the dynamic variability of the digital communications channel. Therefore, out-of-service quality measurements using video test signals or scenes may not relate at all to the resultant received quality of actual program material. Furthermore, traditional in-service quality measurements made by injecting test signals into the non-visible portion of the video signal (e.g., the vertical interval in the NTSC or PAL video standard) are not applicable to modern digital video systems. Thus, a new method is required to make in-service video quality measurements on actual program material. This paper describes a new test instrument for measuring the quality of a video transmission or storage system where the input and output of the system may be spatially separated, and when there is no a priori knowledge of the input video. The test instrument makes continuous quality measurements by (1) extracting statistics from sequences of processed input and output video frames, (2) communicating these extracted statistics between the input and the output ends using an ancillary-data channel of arbitrary bandwidth, (3) computing individual video quality parameters from the communicated statistics that are indicative of the various perceptual aspects of video quality (e.g., spatial, temporal, color), and (4) calculating a composite video quality metric by combining the individual video quality parameters. The test instrument makes coarser quality measurements (coarser in the sense that the extracted statistics come from larger spatial-temporal regions) when smaller capacity ancillary-data channels are available and finer quality measurements when larger capacity ancillary-data channels are available. The design goal for the test instrument is to make the most accurate in-service video quality measurements given the available ancillary-data-channel bandwidth (mobile telephone connections, modem connections over the Public Switched Telephone Network (PSTN), Internet connections, Local Area Network (LAN) connections, satellite connections, cable connections, etc.).*

1. Introduction

This paper describes a new video quality model that uses variable-bandwidth features[1] extracted from spatial-temporal (S-T) regions of input and output video scenes. The motivation behind the development of this model was to add fine spatial impairment measurements to the techniques described in ANSI T1.801.03 [1], while still accommodating the provisions for in-service measurements. The term "in-service" is used in the sense that the input and output ends of the video transmission/storage system under test are not at the same physical location, and there is no *a priori* knowledge of the input video. The technique that makes "in-service" measurements feasible for digital video systems is shown in Figure 1 and has been described in previous papers [2, 3] and patented by the U.S. Department of Commerce [4, 5]. Since the features extracted from the input and output video streams can have considerably less bandwidth than the ITU-R Recommendation BT.601 (referred to as Rec. 601 later in this paper) [6] video streams from which they were extracted, they can be readily transmitted and/or stored as ancillary information. "In-service" methods based on this approach are anticipated to have the widest possible

[1] A feature is defined here as a quantity of information that is associated with a specific S-T region of a video scene. Examples of features are summary statistics (e.g., mean, standard deviation) calculated using all the image pixels within the S-T region.

application in today's broadcast environment where the performance of encoders, digital transmission channels, and decoders changes depending upon scene and network loading conditions. In these cases, taking the video system "out-of-service" or using predetermined test scenes can change the measured performance characteristics of the system under test.

The nonintrusive, in-service model presented here is hierarchical in the sense that the available bandwidth for the extracted features drives the accuracy of the measurement system. In some applications, such as network control and fault monitoring, this available feature bandwidth might be quite low while in other applications, such as laboratory testing of a video codec, this available feature bandwidth might be quite large and might even approach the bandwidth of the Rec. 601 video stream. Like the human visual system, the model presented here requires that small spatial impairments have longer temporal durations than large spatial impairments to be rated as "objectionable." This basic perceptual attribute allows one to examine fairly small spatial regions while maintaining feature bandwidths that are orders of magnitude lower than the bit rate required to transmit the Rec. 601 video stream.

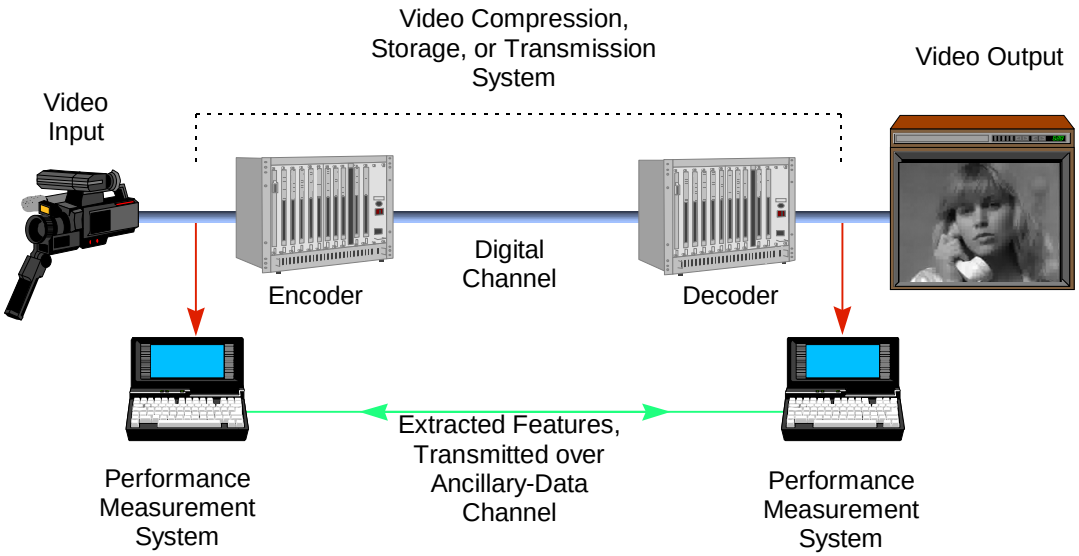

Figure 1. In-service video quality measurement approach.

2. S-T Region Sizes and Compression Factors

Figure 2 gives an illustration of two possible S-T region sizes, one of which includes 8 horizontal pixels x 8 vertical lines x 1 video frame, while the other includes 2 horizontal pixels x 2 vertical lines x 6 video frames. ANSI T1.801.03 uses an S-T region size that includes all of the valid pixels of a single video frame. This whole frame S-T region size is useful when the ancillary-data-channel bandwidth is restricted to very low bit-rates. Features, or summary statistics, extracted from such S-T regions achieve large amounts of compression while characterizing perceived distortions in spatial activity (e.g., blurring, tiling), temporal activity (e.g., dropped frames, transmission errors), and chrominance activity (e.g., cross-color). As the number of pixels encompassed by the S-T region decreases, the compression factor also decreases. This results in an increased ancillary-data-channel bandwidth. However, as section 6 will discuss, the ancillary-data-channel bandwidth does not have to approach the bandwidth of a Rec. 601 video stream (270 Mb/s) to achieve optimum correlation results. The optimal S-T region size that maximizes the correlation between the objective metric and the subjective ratings is much larger than 1 image pixel. Since the correlation worsens slowly as one moves away from this optimum point (i.e., reducing S-T granularity or increasing the S-T granularity), the objective measurement system designer has considerable flexibility in selecting appropriate S-T region sizes.

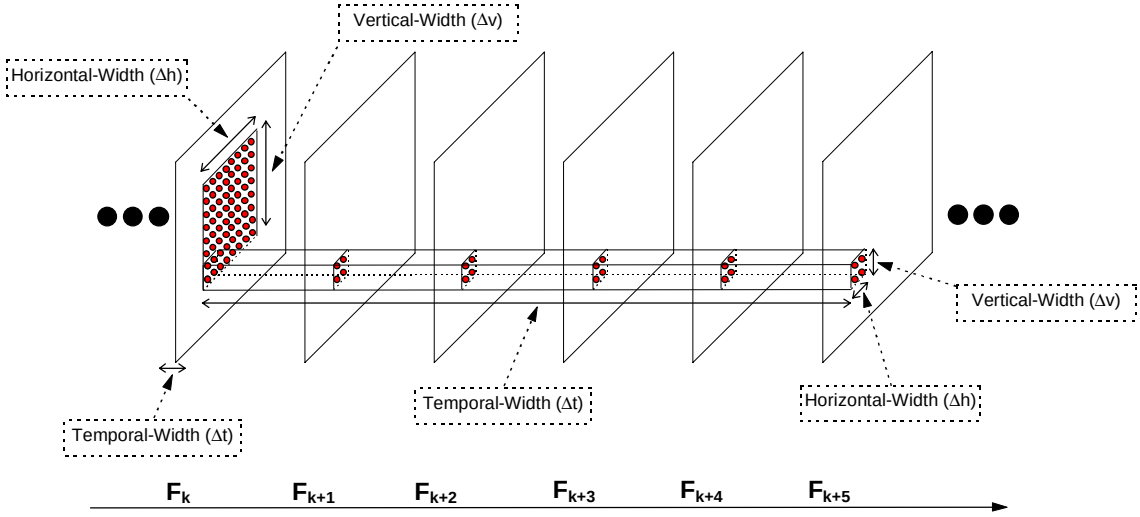

Figure 2. Illustration of spatial-temporal (S-T) regions for a video scene.

Table 1 presents compression factors for several example S-T region sizes assuming a single feature is extracted with a coding efficiency of 2 bytes per extracted feature. These S-T regions in the table have been selected to provide a degree of nesting with each other and to fit within the 525-line 720 x 486 and 625-line 720 x 576 image sampling windows of Rec. 601. A fundamental issue is the optimization of the accuracy of in-service measurements for a given ancillary-data-channel bandwidth. Is it better to increase the spatial extent or the temporal extent of the S-T regions to increase the compression factor? This important issue will be addressed in section 6 by examining the correlation between objective parameters calculated from different S-T region sizes and subjective quality ratings.

Table 1 -- Example S-T Region Sizes and Their Compression Factors

Time Duration (Frames)	H Duration (Pixels)	V Duration (Frame Lines)	Compression Factor
1	672	448 (525-line) 544 (625-line)	150528 (525-line) 182784 (625-line)
30 (525-line) 25 (625-line)	32	32	15360 (525-line) 12800 (625-line)
6 (525-line) 5 (625-line)	32	32	3072 (525-line) 2560 (625-line)
30 (525-line) 25 (625-line)	8	8	960 (525-line) 800 (625-line)
1	32	32	512
6 (525-line) 5 (625-line)	8	8	192 (525-line) 160 (625-line)
30 (525-line) 25 (625-line)	2	2	60 (525-line) 50 (625-line)
1	8	8	32
6 (525-line) 5 (625-line)	2	2	12 (525-line) 10 (625-line)

3. Video Quality Measurement System Overview

Figure 3 shows a more detailed overview of the nonintrusive in-service video quality monitoring system. The *input calibration processor* estimates and adds the video transmission system delay to the input video stream to synchronize the input and output video streams. Video delay is estimated by correlating low bandwidth temporal features (i.e., motion features) extracted from the input and output video streams by the *input and output calibration processors* [1, 4, 5]. Once temporal alignment has been achieved, the *output calibration processor* estimates the spatial shifts, gains, and level offsets of the video transmission system and applies these corrections to the output video stream. Spatial shifts, gains, and level offsets are estimated using low bandwidth spatial features extracted from the synchronized input and output video images by the input and output calibration processors.

Four *programmable filters* extract features (described in section 4) from the calibrated input and output video streams that quantify the amount of information, or activity, in four perceptual characteristics of video quality: spatial, temporal, spatial-temporal, and chroma. These filters are programmable in the sense that they can extract features from any S-T region size (see Table 1). Larger S-T regions are used when the ancillary-data-channel bandwidth is small while smaller S-T regions are used when the ancillary-data-channel bandwidth is large.

The *video quality processors* receive the extracted features from the *programmable filters* and produce a set of quality parameters (described in section 5) which are indicative of perceptual distortions in spatial, temporal, spatial-temporal, and chroma activities. These quality parameters provide the user with a detailed time history of the video transmission system performance. The *video quality processors* also produce a composite score that is indicative of the overall subjective video quality rating using a perceptual combination of the quality parameters and calibration information (e.g., video delay).

4. Extracted Features

Features are extracted at different levels of S-T granularity from the input and output video streams for storage and/or transmission over the ancillary-data channel as depicted in Figure 3. While many different types of features are possible, extensive investigations have resulted in the selection of three particularly promising features chosen for their mutual independence and high degree of effectiveness in predicting subjective quality ratings. These features characterize the behavior and activity of image motion, edges, and color, respectively.

The temporal feature $T(n)$ characterizes the activity of temporal differences or gradients between successive frames. A digital video system can add motion (e.g., error blocks) or reduce motion (e.g., frame repeats). The spatial feature $S(n)$ characterizes the activity of image edges or spatial gradients. A digital video system can add edges (e.g., edge noise) or reduce edges (e.g., blurring). The chrominance feature $C(n)$ characterizes the activity of color information. A digital video system can add color information (e.g., cross color - black edges on white backgrounds that produce color artifacts) or reduce color information (e.g., color sub-sampling).

The fourth aspect of video quality shown in Figure 3, measured by S-T activity parameters (see section 5.2), is a cross product of spatial features $S(n)$ and temporal features $T(n)$. The S-T activity filter is shown separately in Figure 3 since, in general, a unique level of granularity (i.e., S-T region size) may be used for this measurement. An example of added spatial-temporal activity in the output video is mosquito noise [7] that is visible in the stationary background around moving objects.

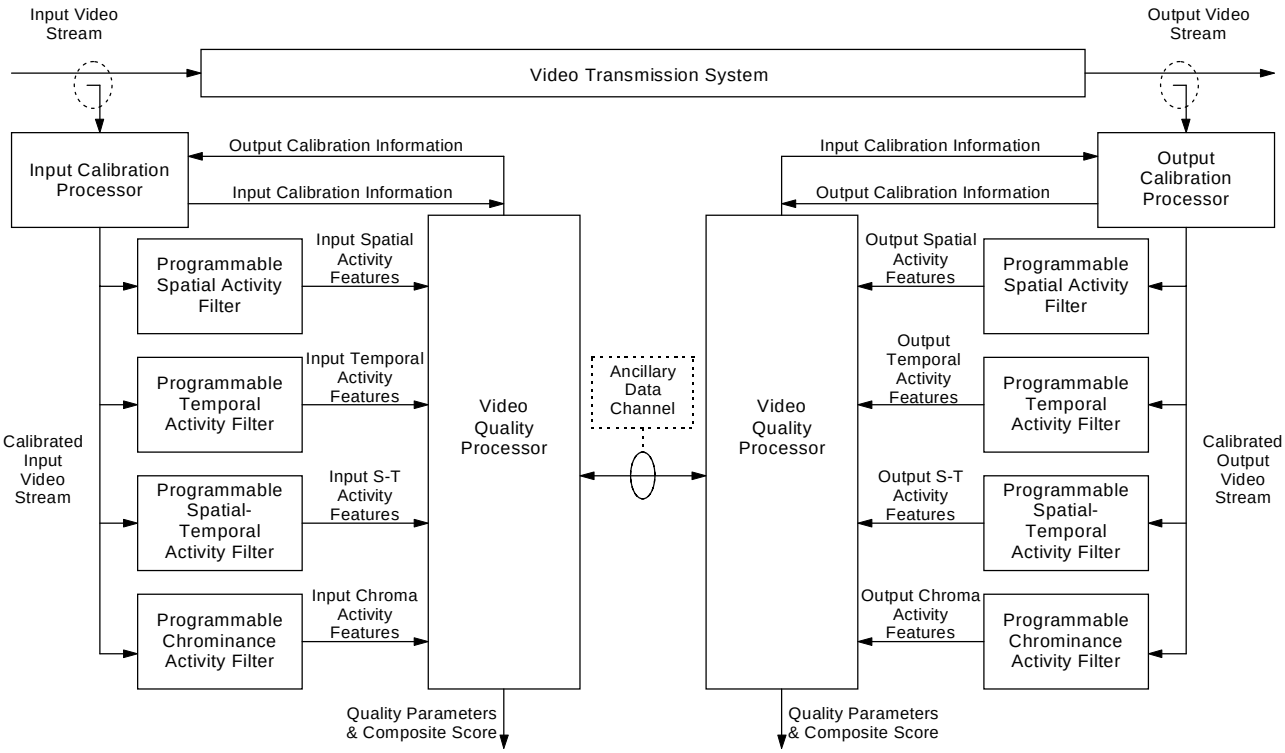

Figure 3. Overview of in-service video quality measurement system.

5. Parameters

5.1 Generalized Description of Parameter Computation

A parameter is the result that is generated by comparing two parallel streams of feature samples, one from the input and the other from the output. The two feature streams are assumed to have been pre-calibrated as described in section 3 before the feature extraction process. Gain and loss functions, represented by *gain*(*n*) and *loss*(*n*), are computed separately using all the pixels in S-T region *n*, where *n* is an index that represents the position of S-T regions within the calibrated input and output video streams. Gain and loss are examined separately for each pair of feature streams since they measure fundamentally different aspects of quality perception (e.g., loss of spatial activity due to blurring and gain of spatial activity due to noise).

The *gain*(*n*) and *loss*(*n*) functions utilize the fact that the perceptibility of spatial and temporal impairments are approximately inversely proportional to the amount of localized spatial and temporal activity in the input scene. In other words, spatial impairments become less visible as the spatial activity in the input scene is increased (i.e., spatial masking), and temporal impairments become less visible as the temporal activity in the input scene is increased (i.e., temporal masking).

Once the gain and loss of the features have been computed for each S-T region in the video clip, then gain and loss parameters are computed from the statistics of *gain*(*n*) and *loss*(*n*). Optimal statistics normally involve some form of worst case processing. This is because quality decisions tend to be based on the worst impairment that is perceivable for a given S-T region (i.e., impairments localized in space and time produce larger effects on subjective quality). Finally, a non-linear mapping is applied to the parameters to remove non-linear perceptual effects at the low and high ranges of values.

5.2 S-T Cross Product Parameters

S-T cross product parameters measure changes in the cross product of spatial features $S(n)$ and temporal features $T(n)$. These parameters allow one to account for relative impairment masking (i.e., reduced visibility of impairments) in areas of high spatial-temporal activity versus areas of low spatial-temporal activity. Secondary masking effects measured by these cross-product parameters cannot be explained by either pure spatial masking (i.e., reduced sensitivity to spatial impairments in areas of high spatial activity) or pure temporal masking (i.e., reduced sensitivity to temporal impairments in areas of high temporal activity). One useful S-T parameter imposes more severe penalties for impairments that occur in localized S-T regions of the input scene that have little motion (e.g., still background) and few edges (e.g., constant luminance) relative to those regions that have high motion and many edges.

6. Optimal S-T Region Sizes

Subjective quality ratings from three independent experiments were used to evaluate the performance of the objective video quality parameters, given in section 5, as a function of S-T region size. The three experiments included twenty-six 525-line video systems and thirty-six test scenes. The subjective test procedure for one of these experiments is documented in [8]. The twenty-six video systems included H.261 and MPEG-2 digital video systems that spanned a range of bit-rates from 768 kb/s to 36 Mb/s and analog video systems used by the broadcast community (NTSC composite, ½-inch professional component, VHS). Multiple-generations as well as simulated digital transmission errors were included. The thirty-six test scenes spanned a wide range of spatial detail, motion, contrast, and brightness.

A differential mean opinion score (DMOS) for every scene-system combination (i.e., a particular scene sent through a particular video system) was generated by averaging the responses from 32 viewers. The meaning associated with the resultant DMOSs is given in Table 2. The subjective DMOSs were cross-correlated with the objective video quality parameter values at various levels of S-T granularity. For ultra-fine grain spatial and temporal activity measurements on MPEG-2 video systems (i.e., unlimited ancillary-data-channel bandwidth), this research has shown that S-T region sizes of 8 pixels x 8 lines x 1 frame and 2 pixels x 2 lines x 6 frames are near optimal for 525-line broadcast quality systems when the input and output video streams are accurately calibrated before feature extraction. Since smaller S-T region sizes produce a decrease in objective to subjective correlation performance, there appears to be little reason to have feature compression factors that are smaller than about 10 to 30 (see Table 1). This implies that an ancillary-data-channel bandwidth of 10 Mb/s should be sufficient to make objective measurements of the highest possible accuracy for a 270 Mb/s Rec. 601 video stream.

Table 2 – Subjective DMOSs and their Associated Meanings

Subjective DMOS	Associated Meaning
1	output quality is “slightly better” than input
0	output quality is “the same” as input
-1	output quality is “slightly worse” than input
-2	output quality is “worse” than input
-3	output quality is “much worse” than input

Figure 4 demonstrates the objective to subjective correlation performance over the three subjective data sets for a single video quality parameter that measures a loss in spatial activity $S(n)$. For small S-T region sizes, a loss of spatial activity can result from blocking (i.e., tiling) as well as blurring since blocking appears as localized smearing of edge information. As can be seen in Figure 4, the optimal S-T region size that achieves the maximum correlation to subjective quality for this parameter is 8 lines x 8 pixels x 1 frame. Coincidentally, this block size is also used by the discrete cosine transform (DCT) for

MPEG-2. While MPEG-2 DCT distortions may influence the optimal choice of S-T region size, the non-MPEG-2 subjective data also produced similar results. Since there was no attempt to align the DCT blocks with the S-T regions (many of the systems included spatially shifted video and the processing regions used by the objective measurements were centered within the valid video area), an 8 pixel x 8 lines x 1 frame S-T region size may be indicative of some perceptual limit. Indeed, an 8 x 8 x 1 size may have been used for MPEG-2 to maximize perceived image quality. Also note from Figure 4 that the 4 x 4 spatial regions achieve a higher correlation for temporal extents longer than one frame. This demonstrates that the human visual system requires longer time durations to perceive smaller spatial distortions.

Extending the S-T regions to include more space and/or time to achieve a greater compression factor results in a decrease of correlation of measurement results to perceived quality. It is important to note that correlation results slowly become worse with increasing S-T region size so that accurate "in-service" measurements of video quality can be obtained even at fairly large compression factors (and hence small feature transmission bandwidths). The range of compression factors for the correlation points plotted in Figure 4 spans more than three orders of magnitude!

We have determined that optimal S-T region sizes are also influenced by the video application (broadcast television, video teleconferencing, etc.) and the accuracy of the spatial and temporal registrations of the input and output video streams. Larger S-T regions produce more optimal correlations to subjective quality ratings when the output video application has many dropped frames, variable video delays, or large spatial registration errors resulting from severe spatial impairments.

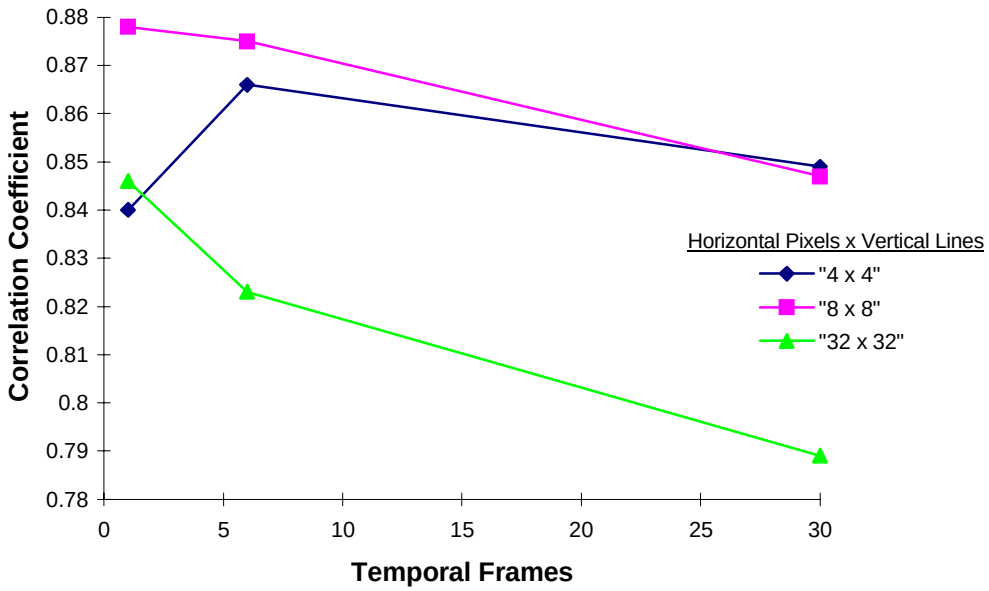

Figure 4. Correlation coefficient of spatial activity loss to DMOS for various S-T region sizes.

7. MPEG-2 Video Quality Model for Medium Bandwidth Ancillary-Data Channel

The techniques described above were used to develop an objective model of MPEG-2 video quality for submission to the International Telecommunications Union (ITU) Video Quality Experts Group (VQEG) [9]. Since the ground rules of the model submission included access to the entire Rec. 601 input and output video streams by the objective model software, we chose to develop an objective model using the optimal MPEG-2 S-T regions discussed in section 6. With these optimal S-T regions, in-service measurements are feasible using ancillary-data channels that have the approximate bandwidth of commonly available local area networks (LANs).

Five parameters were selected for the MPEG-2 video quality model. These parameters measure loss in spatial activity, gain in spatial-temporal activity, gain in chrominance activity, loss in chrominance activity, and gain in temporal activity. These five parameters explain most of the variance in the subjective data. Figure 5 shows a scatter plot of the fitted objective quality ratings and the subjective DMOSs for the three subjective experiments described in section 6. Each point in the scatter plot represents the quality of a particular scene through a particular video system (i.e., scene-system combination). The coefficient of correlation between the objective quality ratings and the subjective DMOSs is 0.95.

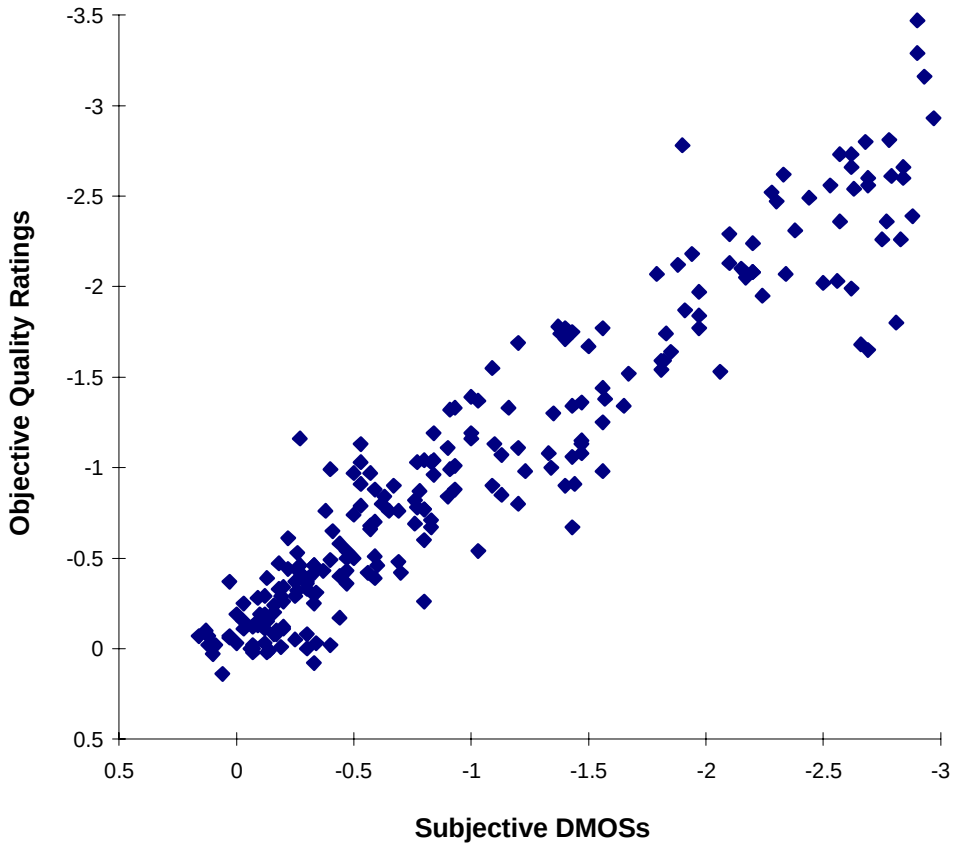

Figure 5. Objective quality vs. subjective quality (scene-system performance).

Figure 6 shows a scatter plot of the objective vs. subjective system performance for the three subjective experiments described in section 6. Here, each point in the scatter plot represents the average quality of a particular video system and was obtained by averaging the objective quality ratings and the subjective DMOSs in Figure 5 over all scenes that were injected into that particular system. The coefficient of correlation between the averaged objective quality ratings and the averaged subjective DMOSs is 0.99.

This MPEG-2 objective video quality model will be tested by the VQEG in the latter half of 1998 along with 9 other models that have been submitted by various international organizations [9]. VQEG's testing program includes conducting identical subjective experiments at several independent laboratories. Since multiple subjective laboratories will be conducting the same experiment, it will be possible to compare subjective to subjective correlations with subjective to objective correlations. These experiments and analyses will produce invaluable information to developers of objective video quality models and will determine if current objective measurement technology has sufficient accuracy to replace subjective experiments for MPEG-2 video systems.

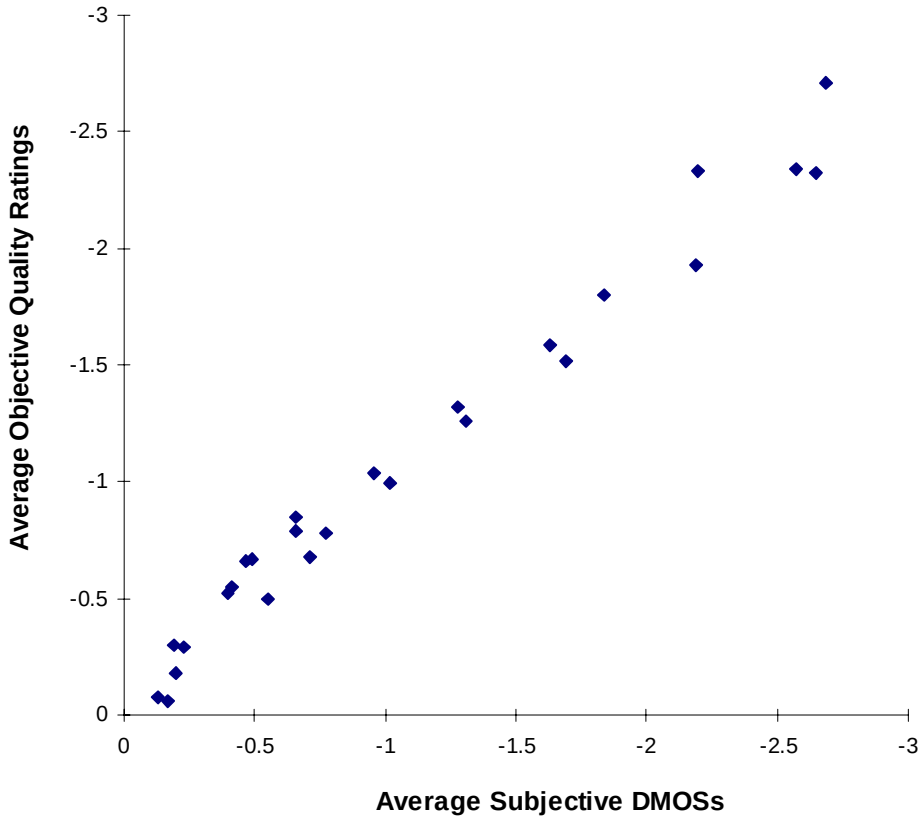

Figure 6. Objective quality vs. subjective quality (system performance).

8. Conclusion

This paper has presented a promising new approach for making in-service video quality measurements. The approach uses feature extraction from S-T regions of input and output video streams, where the extracted features are communicated between the input and output measurement points by means of an ancillary-data channel. The advantage of the proposed system is that true in-service measurements can be made without having *a priori* knowledge of the input video scene or access to the 270 Mb/s Rec. 601 input and output video streams. This approach gives all users, including broadcasters, in-service performance measurement capability using virtually any ancillary-data channel that might be available.

The accuracy of the objective measurements is commensurate with the bandwidth of ancillary-data channel, but there appears to be no measurement advantage to using bandwidths greater than about 1/20 of the bandwidth of the Rec. 601 video stream. The accuracy of the new video quality model will be independently tested by the VQEG of the ITU. Refinements to the proposed model will continue to be made as more and better subjective data sets become available.

9. References

[1]. ANSI T1.801.03-1996, "American National Standard for Telecommunications - Digital Transport of One-Way Video Telephony Signals - Parameters for Objective Performance Assessment," Alliance for Telecommunications Industry Solutions, 1200 G Street, N. W., Suite 500, Washington, DC 20005.

[2]. S. Wolf, "Measuring the End-to-End Performance of Digital Video Systems," IEEE Transactions on Broadcasting, September 1997, Volume 43, Number 3, pages 320-328.

[3]. G.W. Cermak, S. Wolf, E.P. Tweedy, M.H. Pinson, and A.A. Webster, "Validating Objective Measures of MPEG Video Quality," SMPTE Journal, April 1998, Volume 107, Number 4, pages 226-235.

[4]. United States Patent 5,446,492, "Perception-Based Video Quality Measurement System," awarded August 29, 1995.

[5]. United States Patent 5,596,364, "Perception-Based Audio-Visual Synchronization Measurement System," awarded January 21, 1997.

[6]. ITU-R Recommendation BT.601, "Encoding Parameters of Digital Television For Studios," Recommendations of the ITU, Radiocommunication Sector.

[7] ANSI T1.801.02-1996, "American National Standard for Telecommunications - Digital Transport of Video Teleconferencing / Video Telephony Signals - Performance Terms, Definitions, and Examples," Alliance for Telecommunications Industry Solutions, 1200 G Street, N. W., Suite 500, Washington, DC 20005.

[8]. C. Fenimore, J. Libert, and S. Wolf, "Perceptual Effects of Noise in Digital Video Compression," 140th SMPTE Technical Conference, Pasadena, CA, to be published October 28-31, 1998.

[9]. P. Corriveau and A. Webster, "The Video Quality Experts Group Evaluates Objective Methods of Video Image Quality Assessment," 140th SMPTE Technical Conference, Pasadena, CA, to be published October 28-31, 1998.