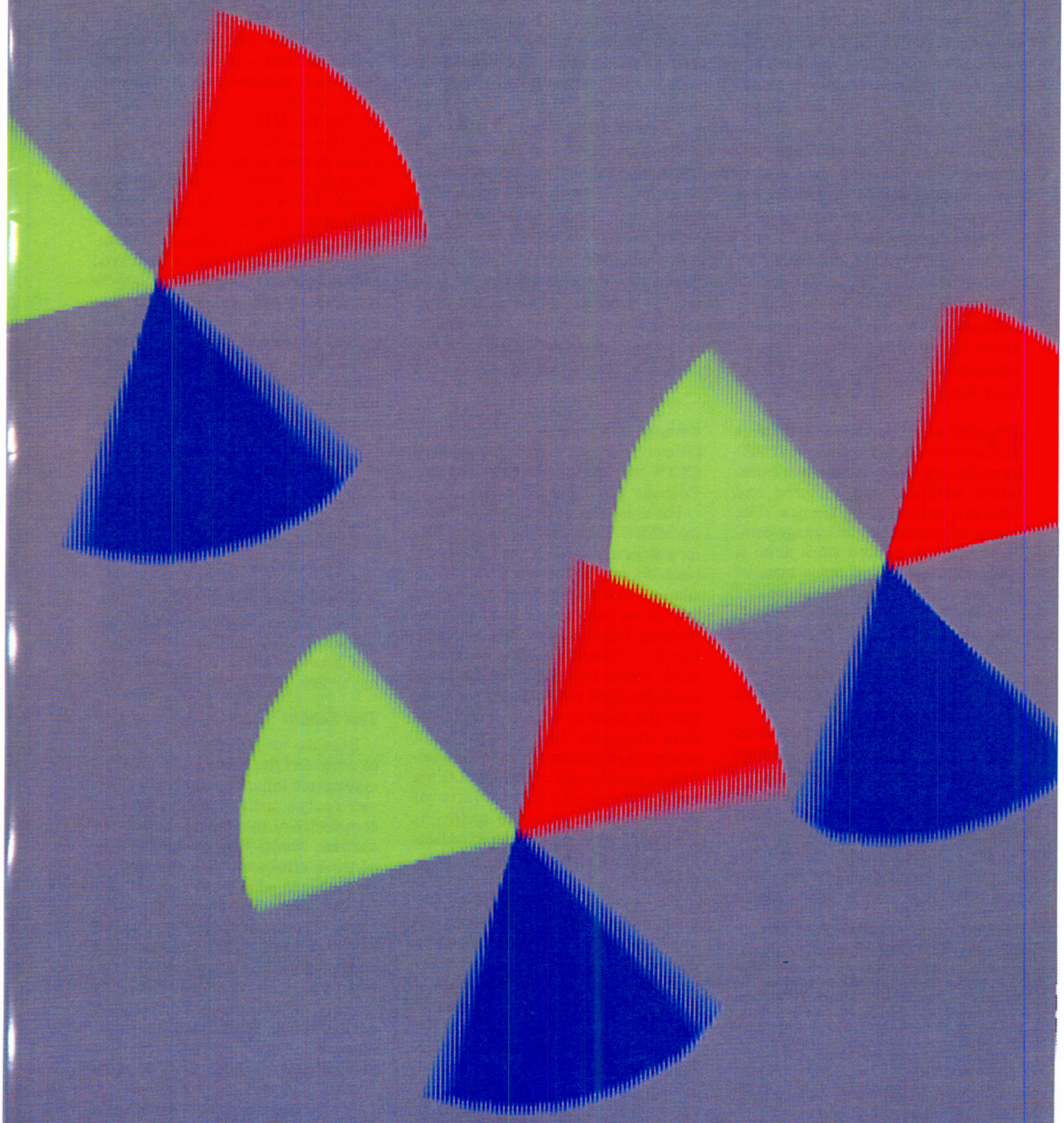
SMPTE Journal

MARCH 2000



Perceptual Effects of Noise in Digital Video Compression

By Charles Fenimore, John Libert, and Stephen Wolf

This paper presents the results of subjective viewer assessment of the quality of MPEG-2 compressed video containing wideband Gaussian noise. The video test sequences consisted of seven clips (both classical and new materials) to which noise with a peak signal-to-noise ratio (PSNR) of 28 to 47 dB was added. Software encoding and decoding was performed at five bit rates ranging from 1.8 to 13.9 Mbits/sec. A panel of 32 viewers rated the difference between the noisy input and the compression-processed output. For low noise levels, the subjective data suggests that compression at higher bit rates can actually improve the quality of the output, effectively acting as a low-pass filter. Defining an objective and a subjective measure of scene criticality allows finding the two meaures that correlate for the data. For difficult-to-encode material (high criticality), the data suggests that the effects of compression may be less noticeable at mid-level noise. In contrast, for easy-to-encode video (low criticality), the addition of a moderate amount of noise to the input led to lower scores. This suggests that either the compression process may have reduced noise impairments or a form of masking may occur in scenes that have high levels of spatial detail.

Digital video compression systems achieve bit rate reduction (BRR) by exploiting image information correlation within a single frame and between neighboring frames. The degree of correlation (and image compressibility) is reduced when noise is introduced. Sources of noisy material include archival material collected with low signal-to-noise ratio (SNR) tube cameras; modern digital, low-noise cameras operating in a low-light environment; and other degraded signal sources such as aging original film or videotape.[1]

In this study, the effects of noise on an MPEG-2 compression system were investigated. The experimental setup for the measurements was based on ITU Rec. 500.[2] The input test scenes were chosen for variety, although they do not necessarily represent the full range of video interest. Of the seven test clips used, one is in the public domain and available from NIST *(Wheels)* and two others are standard CCIR test materials (*Mobile and Calendar* and *Ballet Dancer*).

For noisy test scenes, the output of the MPEG-2 decoder can produce better subjective quality than its input, since discrete cosine transform (DCT) filtering and higher order coefficient truncation can behave as a low-pass filtering function. For this reason, a bipolar subjective quality scale was used, where the quality of the input could be rated either higher or lower than the decoder output. Indeed, the data suggests that compression enhancement occurs, although the statistical significance of the effect is not especially high. The effects of increased noise on video quality is ambiguous; two possible mechanisms are identified as sources of the ambiguity.

For some of the test materials, the compression is nearly transparent, in a statistical sense. The criticality (difficulty of compression) of the video sequences has some predictive power for the bit rate in which transparent coding occurs.[3,4] The Appendix to this paper details the basis for the definition of criticality.

Presented at the 140th SMPTE Technical Conference (paper no. 140-59), in Pasadena, CA, October 28-31, 1998. Charles Fenimore and John Libert are with the National Institute of Standards and Technology, Gaithersburg, MD, and Stephen Wolf is with the Institute for Telecommunications Sciences. An unedited version of this paper appears in *DTV Reality Check—The First Year,* SMPTE, 1998.

Overview of the Test Plan

The primary purpose of the subjective experiment was to collect subjective viewer response data used to construct an objective model of video quality for MPEG-2 video systems. For this experiment, the MPEG-2 video system consisted of one pass through an MPEG-2 coder-decoder chain. The video input and output (I/O) of the system conformed to ITU-R Rec. BT.601.[5] In addition to examining the effect of bit rate on perceived quality, another design factor largely ignored in past experiments was included, i.e., the effect on subjective quality of adding increasing amounts of noise to the input material. The range of added noise power was selected to produce a just-perceptible to slightly-perceptible change in video transmission quality. Viewers were given the task of rating the difference in quality between the I/O video. Figure 1 presents a conceptual block diagram of how each video clip pair (I/O) was generated for the subjective viewing experiment.

Experimental Variables

Four experimental variables contributed to the variability in the subjective data: test scene, noise level, coding bit rate, and viewer.

Test Scenes

Because the subjective perception of noise and the behavior of MPEG-2 systems are influenced by scene attributes such as spatial detail, amount and complexity of motion, brightness, and contrast, scenes that spanned a range of these attributes were selected for the study. In addition to "natural" test scenes, one computer-generated test scene specifically designed to accent MPEG-2 systems was included. This computer-generated scene was selected so that it was viewable (that is, the range of motion and spatial detail was not excessive). Readily available input material of the highest quality was selected. The input material included

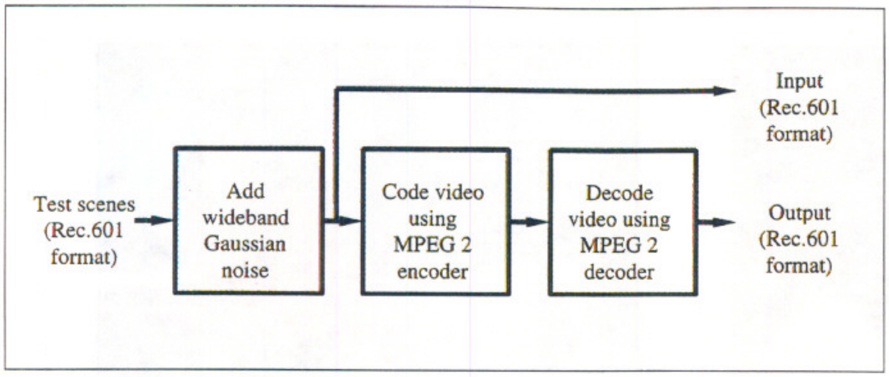


Figure 1. Block diagram for generating subjective test material.

some test scenes from the original Rec. 601 tests,[6] scenes produced with professional cameras and recorded onto 1/2-in. professional tape using a component format and a computer-generated test pattern.

Table 1 gives a description of the seven scenes used for the experiment. Figures 2 and 3 display one frame from each of the clips used, except *Ballet,* for which frames from both cuts are included. The length of each scene was 10 sec, but the viewers only observed the middle 9-sec interval. The first and last 15 frames of each scene were eliminated to avoid possible coder transients at the beginning and ending of each clip.

Noise Levels

Different levels of PSNR for the test scenes were achieved by adding wideband Gaussian noise to the Y (luminance) component channel. To assess the increased MPEG-2 coding difficulty on the high-data-rate luminance component, noise was not added to the C_B and C_R chrominance channels. The primary interest was to investigate the levels at which noise begins to produce perceptible, but slight, changes in MPEG-2 video system quality.

The direct method of Abramowitz and Stegun[7] generated zero-mean Gaussian noise (that is, N [0, σ2]). The noise samples were added to the Y channel of the Rec. 601 video stream after conversion to floating point. The Rec. 601 format has some headroom (that is, Y ranges from 16 to 235 for the 8-bit range [0, 255]) so small amounts of noise can be introduced without significant clipping effects.

Two independent Gaussian noise samples, n_1 and n_2, were generated from uniformly distributed [0, 1] noise samples u_1 and u_2 by:

$$n_1 = \sqrt{-2\sigma^2 \ln(u_1)}\,\cos(2\pi\, u_2)$$
$$n_2 = \sqrt{-2\sigma^2 \ln(u_1)}\,\sin(2\pi\, u_2) \qquad (1)$$

The floating point Y-channel video samples with Gaussian noise added were rounded to the nearest integer and clipped at levels 1 and 254 (0 and 255 are reserved for synchronizing data in Rec. 601).

PSNR is often used to specify the SNR of a video signal. This method has the advantage of removing scene-to-scene variation of the signal power, varying from scene to scene, from the SNR calculation for a given SNR, indicative of some fixed amount of noise power. Calculate PSNR according to the following formula:

$$PSNR = 20 \log_{10}\left[\frac{V_{peak}}{\sigma}\right] \qquad (2)$$

Table 1—Description of Test Scenes Used in the Experiment

Scene Name (Abbreviation)	Description	Source
Mobile and calendar (Mobile)	Independent motion of many objects (for example, red ball, toy train, calendar) against a highly detailed colorful background with a camera pan	Rec. 601 test material
Ballet dancer (Ballet)	Two ballet dancers against blue or brown backgrounds with camera pans and scene cuts	Rec. 601 test material from film
Grand Prix start (Start)	Start of a Grand Prix race — colorful cars in foreground with detailed crowds in background and random camera motion	1/2-inch professional tape
Water bubbling (Water)	Ground level close-up of a bubbling stream in a forest with random camera motion	1/2-inch professional tape
One duck (Duck)	Close-up of a duck swimming and preening with scene cuts	1/2-inch professional tape
Taos boy with zoom (Boy)	Boy in Taos, NM in winter — close-up shot with zoom-out to snow and blue sky	1/2-inch professional tape
Spinning color wheels (Wheels)	Three paddles in red, green, and blue form wheels that spin and move against a background with time-varying gray intensity levels	Computer-generated

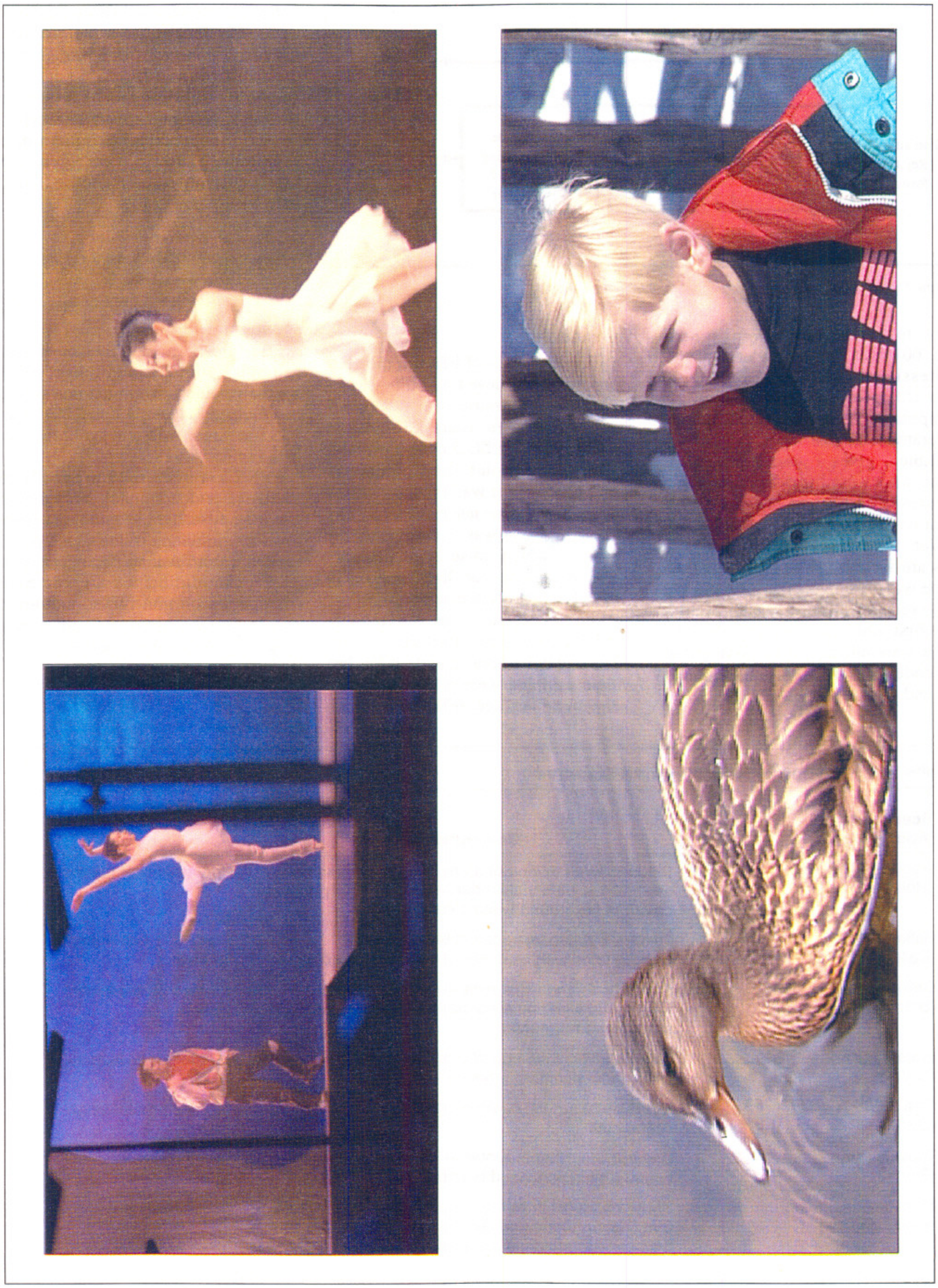


Figure 2. Two frames from the Ballet clip and single frames from Boy and Duck

Figure 3. Single frames from test clips Start, Mobile, Water, and Wheels.

Table 2—Noises σ's of the Test Scores

Noise Condition	Noise σ (Rec. 601 units)	Unweighted PSNR (dB)
1 (original source)	1 (estimated)	47.4
2	3.0	37.9
3	9.0	28.3

Table 3—Subjective Rating Scale

Score	-3	-2	-1	0	1	2	3
Subjective classification	Much worse	Worse	Slightly worse	The same	Slightly better	Better	Much better

in which σ is the standard deviation of the added Gaussian noise and V_{peak} = 235 is "peak white," following the convention of ANSI T1.801.03-1996.[8] Alternative formulas for calculating SNR use true signal power, maximum peak-to-peak signal amplitude (which for the present case would be 235-16 = 219), and frequency-weighted noise. For SNRs based on weighted noise, the frequency-weighting function is normally based on the NTC7 filter.[9] While weighted noise is sometimes used because the human visual system is less sensitive to high-frequency noise than low-frequency noise, the PSNR figures in this paper are presented as unweighted numbers for simplicity.

A total of three noise levels were included in the subjective experiment. The maximum PSNR was limited by the 8-bit sampling of Rec. 601 and the inherent noise level of the input scenes before digital sampling. Table 2 summarizes the three noise levels used (σ's in the above equations) used.

Coding Bit Rates

To generate the MPEG-2 impairments, Test Model 5 (TM5) software encoder (main level, high profile, interlaced mode of operation) and the corresponding decoder provided by the MPEG Software Simulation Group was used. The MPEG-2 video target bit rate was varied to generate five different MPEG-2 conditions: (1) 1.8 Mbits/sec; (2) 3.0 Mbits/sec; (3) 5.0 Mbits/sec; (4) 8.3 Mbits/sec; and (5) 13.9 Mbits/sec. These bit rates were selected to concentrate more systems at the lower bit rates (bit rates above 8 Mbits/sec were expected to produce nearly-imperceptible impairments).

Viewers

A total of 32 viewers were randomly drawn from a pool of 2,000 employees working at the U.S. Department of Commerce's Boulder Laboratories site. Randomly selected viewers were pre-tested to verify having normal visual acuity and color vision.

Subjective Testing

A full factorial design was used for the subjective experiment (that is, all possible combinations of test scene, noise level, and coding bit rate were rated by all the viewers). This yielded 7 x 3 x 5 = 105 conditions that were rated by each viewer. In addition, three test conditions were repeated to obtain a measure of session and viewer variability, for a total of 108 conditions.

Subjective Rating Scale

The goal of the subjective experiment was to measure the change in perceived quality between the input and the output as shown in Fig. 1. This is equivalent to measuring video transmission quality, rather than the absolute video quality of the output. For noisy input test scenes, it was thought that the output of the MPEG-2 decoder might actually have better subjective quality than the input. This was because preprocessing and/or DCT filtering (for example, higher order coefficient truncation) in the MPEG-2 encoder could behave like a low-pass filter function and act to remove visible noise in the input. In view of this consideration, the quality comparison scale given in Table 5 of CCIR Rec. 500-5,[2] and reproduced in Table 3 was selected for the subjective experiment. With this scale, the viewers are shown two versions of each clip (first A, then B) and asked to rate the quality of the second version (B) using the first version (A) as a reference. A subjective rating that falls on the zero point or center of the scale represents the condition where the first and second presentations are perceived as being of identical video quality. To assure that the viewers made full use of both sides of the scale, the order of presentation of the I/O was randomized so that the input appeared first

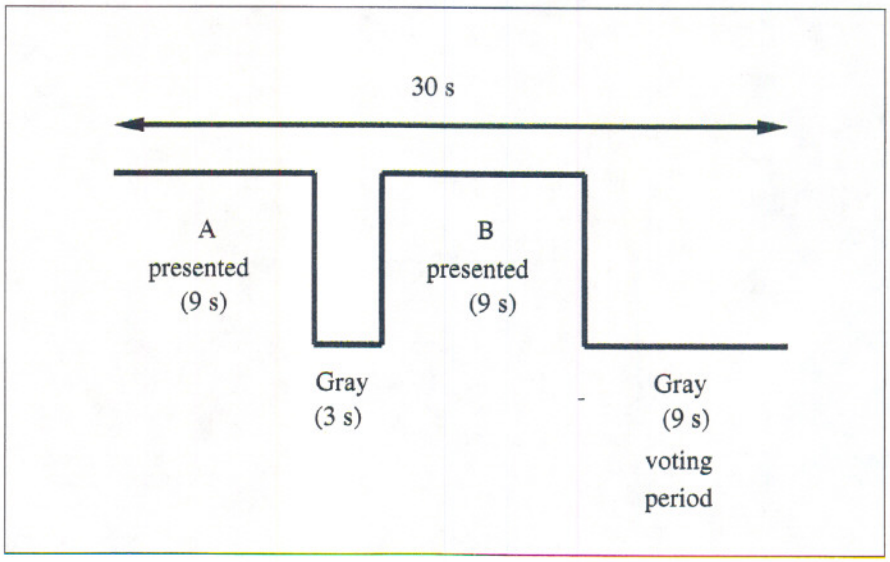


Figure 4. Layout of each clip pair on the videotape.

half of the time and the output appeared first the other half of the time.

Presentation Ordering and Scene Length

Figure 4 details how the A-B clip pairs were shown to the viewers. To reduce clip ordering effects that might result from having all the viewers see the clips in the same order, two random orderings were used (a "red" R, randomization, and a "green" G randomization). To reduce fatigue, each viewing was further split into two half-hour sessions separated by a break. For this purpose, the R and G randomizations were each spread over two viewing tapes having 54 clips each, with three repeated test conditions appearing in both sessions. The four tapes, R1, R2, G1, and G2, provided four possible clip orderings that were shown to a particular viewer (R1R2, R2R1, G1G2, and G2G1). Each viewer was randomly assigned a particular clip ordering. For balance, eight of the 32 viewers saw each of the four possible clip orderings.

Training

The viewers were given a brief training session (less than five minutes) at the beginning of the test, exposing them to the range of impairments in the test and allowing them to gain familiarity with the scoring procedure. After the training session, the experimenter checked that the test subjects understood the scoring procedure before beginning the actual test.

Test Facilities

Testing was performed using quiet audiovisual testing rooms, meeting Noise Criteria 30,[10] and associated audiovisual test equipment. The rooms were finished in light gray and measured approximately 2.7 m by 3.0 m. The viewers sat in a chair centered in front of a video monitor and placed at a distance of four times the picture height of the monitor. Viewers were tested one at a time to avoid unwanted distractions. The illumination of the back wall was adjusted to be approximately 0.15 times the peak luminance of the picture. A 20-in. broadcast-quality monitor containing SMPTE phosphors was used. The setting of the color temperature setting was D65 and the monitor was calibrated with a color analyzer probe and SMPTE color bar.

Table 4—Subjective and Objective Measures of Scene Criticality for Seven MPEG-2 Scenes

Scene	Mobile	Start	Water	Boy	Wheels	Duck	Ballet
Subjective criticality, *s*	-1.34	-1.21	-1.15	-0.84	-0.78	-0.76	-0.62
Objective measure, *o*	3.86	3.71	3.33	3.13	2.77	3.32	2.67

Data Analysis

Analysis of the subjective data proceeded by determining the behavior of the data for each of the experimental variables: viewer variability; compressibility of the various scenes using a measure of scene criticality; changes in quality as the compression bit rate increases; and effect of increasing noise level on the quality of the video. The analysis used the mean opinion score (MOS) averaged over the viewer responses and the half-width 95% confidence interval (two standard deviations of the MOS), C_{95}, for each of the 108 test clips. The randomization of the order of the I/O clips dictated a reordering of the data.

Table 5—Criticality Ranking also Ranks by Transparent Coding Rate Threshold for Seven MPEG-2 Scenes

Scene	Mobile	Start	Water	Boy	Wheels	Duck	Ballet
Subjective measure, *s*	-1.34	-1.21	-1.15	-0.84	-0.78	-0.76	-0.62
Transparent coding bit rate (Mbit/sec)	>13.9	>13.9	5.0 to 8.3	5.0 to 8.3	5.0 to 8.3	5.0 to 8.3	3.0 to 5.0

Consistency of Viewer Ratings

With few exceptions, C_{95} varied from a low of 0.11 quality units to a high of 0.37; the average was 0.24.

In the first case, errors in writing one of the test tapes led to repeating the first field in place of the second field on four clips. Cross-comparisons with viewer scores of the same clips on other test tapes indicated an average negative offset of about 1 quality unit. Therefore, an adjustment of +1 quality units was made to scores on those four clips for the eight viewers who rated the output video worse than the input video. This adjustment affected less than 1% of the data. None of the conclusions depends on the adjustment.

In the second case, it appeared a single viewer suffered momentary confusion and reversed the ordering of the pair of clips. Evidence of this was a single score deviating by 5.65 quality units from the MOS for the clip. Other deviations did not exceed 3.25 quality units. The viewer was retested for this scene, and the new score was not an outlier and it was used in the data analysis. No other corrections were applied. The narrowness of the confidence bounds demonstrates a high degree of consistency across viewers in this subjective experiment.

Scene Criticality and Compressibility

Criticality is a measure of the difficulty of encoding a scene. Two measures of criticality were employed. One was a subjective measure that was derived from the subjective data while the other was an objective measure that was derived from computer-based processing of the sampled video images. The objective measure of criticality *(o)* is detailed in the Appendix[3] and given by

$$o = \log_{10}\left\{mean_{time}\left[\mathrm{SI}\,(t_n)^*\,TI(t_n)\right]\right\} \quad (3)$$

where *SI* measures spatial detail, *TI* measures frame-to-frame image changes, and t_n indexes the frames of the video clip. The objective measure of criticality *(o)*, which was developed using a set of ANSI-standardized test scenes (see Appendix) was evaluated using the set of MPEG-2 test scenes described in this paper. The subjective measure of criticality *(s)* was calculated by taking the absolute value of the averaged MOS for each test scene with a noise level $\sigma = 1$ (that is, MOSs were averaged over bit rates for each scene), namely

$$s = \overline{MOS} \tag{4}$$

Table 4 presents the results of comparing *s* and *o*. Here, higher *s* numbers indicate more impairment and, therefore, scenes that are more difficult to code. Figure 5 presents a scatter plot of the results. The coefficient of correlation was -0.89, indicating a fairly strong correlation between *s* and *o*. Most of the remaining unexplained variance is due to a single outlier *(Duck)*. The elimination of the scene *Duck* lowers the coefficient of correlation to -0.96. In this scene, the duck's feathers contain high spatial information that changes rapidly. However, the rapid motion that produces this change prevents the eye from tracking the spatial detail.

Transparent Coding Bit Rates

Using the subjective MOS and the half-width 95% confidence interval, the transparent coding bit rates (TBCR) were determined. TBCR is the range of bit rates in which MOS first goes to zero (where MOS is not statistically significantly different from 0 at the 5% level). The subjective measure of criticality properly orders the sequences with respect to this bit rate, as seen in Table 5. This measure is admittedly crude and might be shown to be less effective with more refined increments in the encoding bit rate.

Quality May Be "Improved" by Compression

For low criticality scenes, the data suggests possible improvements to the video by compression. Table 6 shows the MOS for each test clip in the low

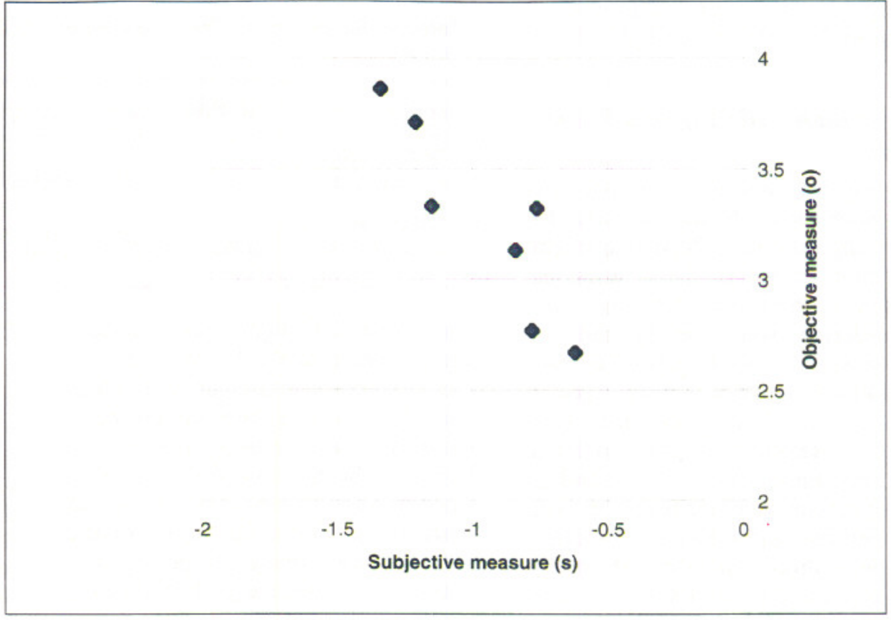


Figure 5. Plot of subjective and objective measures of scene criticality for seven MPEG-2 scenes.

Table 6—Mean Opinion Scores (and Half-width 95% Confidence Intervals) for Each Low Noise Test Scene at the Five Bit Rates

Bit rate (Mbit/sec)	1.8	3.0	5.0	8.3	13.9
Mobile	-2.78 (0.17)	-1.94 (0.29)	-1.16 (0.31)	-0.59 (0.24)	-0.25 (0.18)
Start	-2.69 (0.16)	-1.38 (0.30)	-0.84 (0.27)	-0.69 (0.24)	-0.44 (0.26)
Water	-2.84 (0.16)	-1.81 (0.27)	-0.69 (0.19)	-0.16 (0.21)	-0.25 (0.23)
Boy	-2.28 (0.35)	-1.56 (0.32)	-0.25 (0.20)	-0.09 (0.14)	0.00 (0.20)
Wheels	-2.75 (0.15)	-0.84 (0.37)	-0.19 (0.16)	-0.19 (0.26)	0.06 (0.26)
Duck	-2.34 (0.23)	-1.00 (0.22)	-0.22 (0.15)	-0.16 (0.18)	-0.06 (0.17)
Ballet	-2.66 (0.17)	-0.59 (0.26)	0.16 (0.22)	0.13 (0.29)	0.03 (0.14)

Table 7—Subjective Measures of Scene Quality (MOS) (and Half-width 95% Confidence Intervals) Averaged over Five Compression Bit Rates for Seven MPEG-2 Scenes, Shown at Three Noise Levels.

Scene	MOS and C_{95} (noise $\sigma = 1$)	MOS and C_{95} (noise $\sigma = 3$)	MOS and C_{95} (noise $\sigma = 9$)
Mobile	-1.34 (0.7)	-1.23 (0.9)	-1.29 (0.9)
Start	-1.21 (.10)	-1.03 (0.9)	-1.33 (0.9)
Water	-1.15 (0.8)	-1.18 (0.8)	-1.12 (0.8)
Boy	-0.84 (0.7)	-0.80 (0.8)	-0.87 (.10)
Wheels	-0.78 (.10)	-0.70 (.07)	-0.64 (0.8)
Duck	-0.76 (.07)	-0.81 (.06)	-0.79 (.07)
Ballet	-0.59 (.08)	-0.76 (.10)	-1.07 (.11)

noise case, $\sigma = 1$. In the lower right hand corner, at higher bit rates, there are several positive MOSs, although the data does not support statistical significance at the 95% level. For scenes with high criticality, the MOS does not go to zero at these bit rates. The data supports the use of the bipolar quality scale in these subjective quality measurements. Without positive going scores the MOS scores have a negative bias.

Effect of Noise Level

For some of the scenes, a combination of high spatial detail and motion leads to relatively high criticality, particularly for scenes *Mobile and Calendar* and *Start*. In these scenes, a suggestion of improvement in the bit rate averaged MOS as the level of the input noise is increased from $\sigma = 1$ to $\sigma = 3$ (Table 7). For scenes with lower criticality the only effect of increasing noise is to decrease quality, particularly for scene *Ballet*. For this low criticality scene, the compression impairments generated by the addition of noise are very noticeable. This suggests that in high criticality scenes either noise is being reduced in the compression process or compression impairments are being masked. At the highest noise level the effects of compression were generally no less noticeable to the panel than at low noise.

Conclusion

The results suggest that the effect of noise on the perceived quality of compressed digital video is not described by a simple monotonic function. In some cases, the detail in an image masks the impairments introduced by the compression process. For the low-criticality scenes ($s > -1.0$) studied, the *MOS* becomes positive for low noise at some of the higher bit rates, although no single combination of scene, noise level, and bit rate is statistically significant at the 95% confidence level. The data suggests that for a larger range of test materials and bit rates, one may find that the quality measurement process will rate compression-"impaired" video as superior to the input material. In a practical sense, the subjective measurement process can detect this effect by employing a bipolar measurement scale such as that used in the experiment described here. If the effect is deemed significant, a more fundamental problem arises concerning objective measurement technology. It is common for such techniques to rate any image change an impairment while viewer preference may rate such change an improvement. This conflict will have to be addressed by new objective measurement techniques.

Acknowledgments

This research was supported by intramural funding from the Digital Video on Information Networks Program of the Advanced Technology Program at the National Institute of Standards and Technology (NIST).

References

1. J. Hamalainen, et al., "Facts and Fiction: Some Aspects Regarding the Design of Digital Television Cameras Using CCD Image Sensors," *Intl. J. Imaging Sys. Tech., 5*: 314-322, 1994.
2. CCIR Rec. 500-5, "Method for the Subjective Assessment of the Quality of Television Pictures," 1992.
3. S. Wolf, "Subjective and Objective Measures of Scene Criticality," ITU Experts Meeting, SG 12 Document no. JRG010, Turin, 1997.
4. I. Yuyama, et al., "Objective Measurement of Compressed Digital Television Picture Quality," *SMPTE J., 107*:348-352, June 1998.
5. ITU-R Rec. BT.601, "Encoding Parameters of Digital Television For Studios," Recommendations of the ITU, Radio Communication Sector.
6. ITU-R Rec. BT.802-1, Test Pictures and Sequences for Subjective Assessments of Digital Codecs Conveying Signals Produced According to ITU-R Rec. BT.601-4.
7. M. Abramowitz and I. Stegun, *Handbook of Mathematical Publications*, Dover 1972.
8. ANSI T1.801.03-1996, "Digital Transport of One-Way Video Telephony Signals—Parameters for Objective Performance Assessment," Alliance for Telecommunications Industry Solutions, Washington, DC.
9. E. B. Crutchfield, ed., *National Association of Broadcasters Engineering Handbook, 7th ed.*, Washington, DC, (Natl. Assoc. of Broadcasters, 1995) 4.1-38-4.1-40.
10. L. Beranek, *Noise and Vibration Control*, McGraw-Hill, 1971.

Appendix

Measures of Scene Criticality

The difficulty of coding a video scene to achieve a constant perceived quality level increases with the amount of spatial detail and motion. This Appendix describes preliminary results of an investigation to derive a combined spatial-temporal metric for estimating scene criticality, or coding difficulty. This objective metric of scene criticality has several potential uses, including use as a tool for systematically selecting an appropriate range of test material without unnecessary duplication, and as a method for performing dynamic bit rate allocation in a "constant-quality, variable bit rate," statistically-multiplexed transmission channel.

The basis for the investigation was to determine if scene criticality could at least be roughly estimated from the set of low bandwidth spatial information (SI) and temporal information (TI) features.[A1, A2] The advantage of using these particular features is that they are simple to compute in realtime and can be readily transmitted or stored as digital side information because of their extremely low bandwidth and data storage requirements. Thus, they may be used for automatically controlling and monitoring the behavior of digital video transmission systems. The SI feature examined here is given by

$$SI\,(t_n) = \mathrm{rms}_{space}\left[\mathrm{Sobel}\,(F(t_n))\right] \quad \text{(A1)}$$

while the TI feature examined is given by

$$TI\,(t_n) = \mathrm{rms}_{space}\left[F(t_n) - F(t_{n-1})\right] \quad \text{(A2)}$$

where $F(t_n)$ is the luminance-only video frame at time t_n, Sobel is the Sobel filter,[A3] and rms_{space} is the root mean square function over the entire valid image subregion. Preliminary results indicate that a coarse model of scene criticality can be derived using these simple image features. Obvious refinements made to improve this model include the use of more localized estimates of SI and TI, scene-cut masking, object segmentation, and object motion tracking (including the randomness of the direction of motion) that emulates human perception.

Subjective Measure of Scene Criticality

In 1995, ANSI-accredited committee T1A1 undertook an extensive experiment that involved the subjective evaluation of 25 test scenes[A4] injected into 24 different digital video systems for a total of 600 scene-system combinations. Most of the digital video systems were video teleconferencing systems that included a range of bit rates from 64 kbits/sec to 1.5 Mbits/sec. Video home system (VHS) recorded scenes and 45 Mbits/sec encoded scenes were also used as two reference conditions. To obtain a subjective estimate of the scene criticality, an average of the subjective scores for each scene across all viewers and digital video systems was used in the test. This computed average is referred to as the scene main effect by statisticians and provides a measure of the portion of the MOS that is due solely to the test scene. Since a wide range of digital video systems was used in this test, the scene main effect should also provide an estimate of the scene criticality. Scenes that are the most difficult to code will have a lower scene main effect, or average MOS, while scenes that are easy to code will have a higher scene main effect.

Table A1 presents a summary of the subjective measure of scene criticality *(s)* for the 25 test scenes. Since the subjective scores were derived using an impairment scale that ranged from 1 to 5 (where, 5 = "imperceptible"; 4 = "perceptible but not annoying"; 3 = "slightly annoying"; 2 = "annoying"; and 1 = "very annoying"), the table shows that the scene main effect varied from "annoying" to somewhere between "slightly annoying" and "perceptible but not annoying." As expected, the football scene *(ftball)* was the most difficult to code while a head and shoulders scene *(disguy)* was the easiest to code. The 25 points in Fig. A1 were used to develop the objective model of scene criticality that is presented in this paper.

Objective Measure of Scene Criticality

Of several objective measures of scene criticality that were considered, the simplest that was developed, *o*, is given by the model

$$o = \log_{10}\left\{mean_{time}\left[\mathrm{SI}(t_n)^* TI(t_n)\right]\right\} \tag{A3}$$

Values for this model were computed using a time window that was the same as the length of the video clips used in the subjective testing (nine seconds). The model measures the average value (over time) of the instantaneous frame-by-frame product of SI and TI. When a large amount of spatial-temporal gradient energy is present, the scene is difficult to code. The criticality number for this simple model is given in column *o* (objective measure) of Table A1, while a plot of the performance of the model is given in Figure A1. The

Table A1—Subjective and Objective Measures of Scene Criticality for 25 ANSI Scenes

Scene Abbreviation	Scene Description	s (subjective measure)	o (objective measure)
Ftball	Football game	2.05	3.4
Cirkit	Circuit diagram, camera pan	2.16	3.75
2wbord	Two people at white board, scene cuts	2.33	2.69
Rodmap	Road map with hand and pen motion, camera pan	2.56	3.18
Smity2	Salesman at desk with magazine	2.56	3.43
Smity1	Salesman at deck with box	2.58	3.36
Flogar	Flower garden with windmill, camera pan	2.62	3.74
Washdc	Washington, DC, map with hand and pointer	2.63	2.82
Ysmite	Yosemite map & hand motion (intensity fluctuations)	2.73	2.77
Fredas	Fred Astaire tap dancing (black and white)	2.73	2.84
Split6	Split screen, six people	2.77	2.83
Intros	Introductions of people sitting at table, camera pans	2.8	2.69
Boblec	Bob's lecture at chalkboard	2.86	2.59
3inrow	Men at table, camera pan	3.02	2.70
Vowels	Woman at whiteboard teaching vowels	3.1	2.85
Vtc2zm	Woman standing next to map with pointer, zoom and	3.14	2.88
Inspec	Woman at document camera	3.14	2.34
3twos	Two pairs of people, scene cuts	3.17	2.51
Susie	Susie on telephone	3.28	2.56
5row1	Five people in a row sitting at a table	3.37	2.44
Filter	Filter diagram on yellow pad with hand motion	3.51	2.43
Disgal	Female announcer	3.65	2.19
Vtc1nw	Woman sitting reading news story	3.66	2.13
Vtc2mp	Woman standing next to map	3.67	2.43
Disguy	Male announcer	3.68	2.16

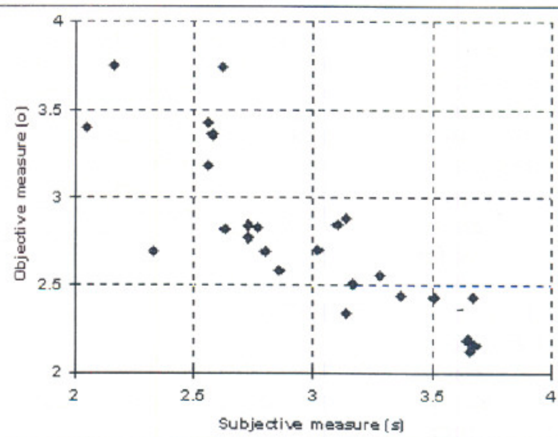


Figure A1. Plot of subjective vs. objective measures of scene criticality for 25 ANSI scenes.

coefficient of correlation between the objective and subjective measures is -0.82 (here, the objective model is negatively correlated to the subjective score since higher subjective scores indicate easier to code test scenes). Most of the remaining unexplained variance results from several outliers.

Elimination of just one of these outliers (scene *2wbord*), which contains several scene cuts, lowers the coefficient of correlation to -0.87. The magnitude of the correlation achieved in the training phase is comparable to the correlation found in the test materials discussed in the body of this paper (0.89). The effect of scene cuts on coding difficulty cannot be explained by the simple objective model presented here.

Appendix References

A1. ITU-T Rec. P.910, "Subjective Video Quality Assessment Methods for Multimedia Applications," Recommendations of the ITU (Telecommunication Standardization Sector).

A2. ANSI T1.801.03-1996, "Digital Transport of One-Way Video Telephony Signals—Parameters for Objective Performance Assessment," Alliance for Telecommunications Industry Solutions, Washington, DC.

A3. R.C. Gonzalez and P. Wintz, *Digital Image Processing*, 2nd ed., Addison-Wesley: Reading, 1987.

A4. ANSI T1.801.01-1995, "Digital Transport of Video Teleconferencing/Video Telephony Signals —Video Test Scenes for Subjective and Objective Performance Assessment," Alliance for Telecommunications Industry Solutions, Washington DC.

THE AUTHORS

Charles Fenimore graduated from Union College with a BS degree and from Berkeley with a Ph.D. in mathematics. He has worked on computational models for fluid and electrically driven flows at the Lawrence Berkeley Lab and at NIST.

For several years, Fenimore has worked on measurements and models for quality assessment of moving digital imagery. This includes the development of measurements as well as the production of synthetic imagery and the acquisition of natural imagery as "probes" of compression systems.

Stephen Wolf received a BSEE from Montana State University in 1979 and an MS in electrical and computer engineering from the University of California at Santa Barbara in 1983. Since 1988, he has been project leader for the Video Quality Standards Project at the Institute for Telecommunication Sciences, an agency of the National Telecommunications and Information Administration in Boulder, CO. During this time, he developed innovative methods for performing in-service digital video quality measurements for which he was awarded two U.S. patents.

Wolf, an active participant and contributor to the standardization activities of both ANSI and the ITU, has served as technical editor for two video performance standards, ANSI T1.801.01-1995 (Video Test Scenes) and ANSI T1.801.03-1996 (Objective Video Performance Parameters).

John M. Libert received a BS degree in experimental psychology and an MS in quantitative geology from the University of Maryland in 1970 and 1981, respectively. Joining Vitro Corp. in 1986 as principal investigator of its Signal and Image Analysis Laboratory, he led research to develop computational vision methods for image motion perception and stereopsis. He consulted with Sandia National Laboratories in developing a moving target detector for an infrared broad area surveillance system.

In 1997, Libert joined the Electronics and Electrical Engineering Laboratory of the National Institute of Standards and Technology where he currently conducts research in video quality measurements a part of NIST's Video Technology Project.

SMPTE Journal®

VOL. 109, NO. 3 MARCH 2000

SMPTE Journal (ISSN: 0036-1682) published monthly by the Society of Motion Picture and Television Engineers, 595 West Hartsdale Ave., White Plains, NY 10607. Periodicals postage paid at White Plains, N.Y., and at additional mailing office. Statements and opinions expressed in articles or editorials in the *SMPTE Journal* are expressions of contributors and do not necessarily represent the policies or opinions of the Society. Advertisements appearing in this publication are the sole responsibility of the advertiser. Subscription rates: SMPTE members, $20.00 included in membership dues; nonmembers, U.S., $125.00; elsewhere, $140.00; airmail charge in addition to nonmember rates is $65.00. Single copy, $10.00 plus shipping. MEMBERSHIP in SMPTE is available according to qualifications at the following annual dues: Active Member, $95.00; Student Member, $25.00. All members receive the *SMPTE Journal* as part of membership.

POSTMASTER: Send change of address to *SMPTE Journal*, 595 West Hartsdale Ave., White Plains, NY 10607-1824.