PROCEEDINGS OF SPIE
SPIE—The International Society for Optical Engineering

Human Vision and Electronic Imaging V

Bernice E. Rogowitz
Thrasyvoulos N. Pappas
Chairs/Editors

24–27 January 2000
San Jose, USA

Sponsored by
IS&T—The Society for Imaging Science and Technology
SPIE—The International Society for Optical Engineering

Published by
SPIE—The International Society for Optical Engineering

SPIE
P
PROCEEDINGS SERIES

Volume 3959

Toward Developing a Unit of Measure and Scale of Digital Video Quality: IEEE Broadcast Technology Society Subcommittee on Video Compression Measurements

John M. Libert[a], Leon Stanger[b], Andrew B. Watson[c] and Ann Marie Rohaly[d]

[a]National Institute of Standards and Technology, Gaithersburg, MD 20899-8114
[b]DirecTV, El Segundo, CA 90245
[c]NASA Ames Research Center, Moffett Field, CA 94035-1000
[d]Tektronix, Inc., Beaverton, OR 97077

ABSTRACT

Development of video quality metrics has taken support from experimental vision data mainly at two levels of abstraction. On the one hand are the carefully controlled tests of human visual response to well-defined, controlled visual stimuli, such as the ModelFest [15] study. On the other hand are experiments in which viewers rate the global quality of "natural" video sequences exhibiting impairments of loosely-controlled composition and amplitude, as in the Video Quality Experts Group study [8]. The IEEE Broadcast Technology Society Subcommittee on Video Compression Measurements has initiated an intermediate level approach to video quality assessment aimed toward developing a scale of video impairment and unit of measure by which to describe video distortion in both perceptual and engineering terms. The proposed IEEE study will attempt to define a scale of video impairment in terms of multiple measurements of the just-noticeable-difference (JND) of compression-induced video impairments. A paired comparison psychophysical method will be used to define a psychometric function of the visual sensitivity to compression-induced video impairments of various amplitudes. In this effort, quality assessment is related directly to visual perception of video impairments rather than to the more "atomic" visual stimuli as used in many human vision experiments. Yet the experimenters' control over the stimuli is greater than that used in much of contemporary video quality testing.

Keywords: video compression measurement, IEEE standards, video quality, video fidelity, numerical category scaling, JND

1. INTRODUCTION

Within the organization of the Institute of Electrical and Electronics Engineers, Inc. (IEEE), the Video Compression Measurements Subcommittee, G-2.1.6, was commissioned by the Audio-Video Techniques Committee, G-2.1, of the IEEE Broadcast Technology Society. The subcommittee's scope document [1] calls for it to "... *investigate and recommend methods of directly quantifying image sequence impairments resulting from compression and decompression cycles that are well correlated to results of subjective comparison.*" Indeed, the Broadcast Technology Society's interest in video quality measurement may have been captured quite succinctly by one member whose observation might be paraphrased as "*...the TV program sponsor may or may not be impressed by the technical challenges of digital high definition television (HDTV) -- but if his commercial doesn't look good, he doesn't pay.*" Accordingly, the broadcaster is interested in tools by which to ensure picture quality and by which to monitor, adjust and document the performance of the digital broadcast system. A standard is sought by which to specify and test new equipment used in television production and broadcasting and to direct manufacture of test equipment to be used in verifying system and component performance both before and after installation.

In executing its charter over the past several years, the Video Compression Measurements Subcommittee (hereafter referred to in this paper as "the subcommittee") has held quarterly meetings, generally in coordination with the Video Distribution and Processing Subcommittee (G-2.1.4) and the T1A1[1] subcommittee of the American National Standards Institute (ANSI). It has solicited presentations and technical papers from developers of video quality measurement models and has maintained

[1] Committee T1 provides standards needed for the planning, design and operations of global end-to-end telecommunications and related information services. Subcommittee T1A1, Performance and Signal Processing, develops and recommends standards and technical reports related to the description of performance and the processing of speech, audio, data, image and video signals, and their multimedia integration, within U.S. telecommunications networks.

In *Human Vision and Electronic Imaging V*, Bernice E. Rogowitz, Thrasyvoulos N. Pappas, Editors, Proceedings of SPIE Vol. 3959 (2000) • 0277-786X/00/$15.00

communication with other standards bodies such as the International Telecommunications Union (ITU) and its "offspring" working groups, including the Video Quality Experts Group (VQEG). The goal of the subcommittee is to define and recommend to the parent body a standard for digital video quality measurement. The recommendation can be either wholly defined by the subcommittee or simply reference a standard defined by another standards organization and found by the subcommittee to satisfy the needs of the IEEE. Another paper [2] published in this proceedings volume summarizes the activities and relationships among the various standards groups involved with video quality measurement.

The subcommittee has approached defining a standard from two directions. It has examined the computational models being proposed to supplant subjective testing. It also has attempted to identify a set of standard test materials to be used in evaluating the relative performance of quality metrics and to serve as validation and calibration benchmarks. Over the course of numerous discussions, a combination of these two directions evolved into the subcommittee's current approach to a video quality measurement standard – an attempt to derive a scale of video quality through controlled psychometric testing and to produce and distribute a suite of perceptually-referenced video materials for testing and calibration of computational metrics. The rationale for this approach and implementation options are described in the remainder of this paper.

2. BACKGROUND

The subcommittee agreed that useful objective measurement methods must correlate well with subjective assessments of quality. It was observed, however, that the objective measures it had examined appeared to correlate only weakly to the subjective ratings. Subsequent discussions led to the conclusion that the state of computational metrics may not have evolved sufficiently to declare a standard. Also, the subcommittee found that the nature of subjective testing typically used in the video industry [3], itself, may be part of the problem. That is, in order to be considered successful, an objective measure would have to correlate with subjective assessments at least as well as subjective tests correlate with each other. Some results examined by the committee showed correlation of only 0.91 between two subjective test subject groups, a value that was discouraging to some members looking for methods to support expectations of 99% reliability for commercial broadcast systems. Even if such correlation values were reasonable for a subjective test, concern was raised over repeatability of such tests.

The turning point of the subcommittee's quest was initiated by a presentation by Leon Stanger [4] in which he described the need for a unit of measure and means to calibrate picture quality degradation. Stanger observed that while currently deficient, quality measurement algorithms would continue to evolve. Rather than attempting to standardize a method prematurely, Stanger encouraged the subcommittee to focus attention instead on defining a unit of measure of video quality. In the discussions which followed, it was agreed that the subcommittee might contribute more to video quality measurement were it to develop a standard set of video materials, each having associated subjective measures of quality, if possible, locatable on an interval scale of quality. Acknowledging that defining an actual "visual volt" might be ambitious, Stanger suggested that a scale marked by increments of a "just-noticeable-difference" (JND) threshold might be a sufficiently stable and repeatable means by which to evaluate video quality. Subcommittee discussions over subsequent meetings resulted in proposal of an experiment which will be described below. Some details have yet to be worked out and in some key areas, the subcommittee has been offered alternative approaches to consider.

3. PROPOSED STUDY

3.1 Objectives

The subcommittee aims to develop a suite of video sequences exhibiting artifacts typical of those resulting from discrete cosine transform (DCT) based compression systems to be used for system calibration and testing; to develop an industry standard scale of quality, such as multiple just noticeable differences (JNDs), by which to quantify levels of video degradation; and to disseminate the test sequences and subjective fidelity measurements to interested parties to serve as benchmarks by which to calibrate automated video analysis tools and to test digital video components.

In discussing the utility of such materials the distinction between "video quality" and "video fidelity" was acknowledged. The distinction is discussed at length in [5]. It was decided that the critical eyes of video engineers would be more sensitive to image defects irrespective of their context and more likely to yield a "fidelity" assessment. Hence, even as the term "quality" is used in subcommittee discussions and appears in many of its documents, "fidelity" is the more appropriate term for the interests of this standards body.

The nominal yield of the activity will be a suite of video materials exhibiting video impairments discretely sampled over a range of interest and the impairment level of each located at incremental positions on a continuous scale of perceptibility. The scale would be anchored at the visibility threshold, i.e., 1.0 JND, and subsequent sequences would be impaired further such that each is just noticeably different from the preceding sequence. A possible outcome may be an impairment visibility function or set of functions, each characteristic of a particular impairment type. The subcommittee is well aware of the dependence of the impairment behavior on the peculiarities of both the encoder and the source video. Accordingly, a completely general function may not be achieved. However, for purposes of a calibration standard, even a fidelity function limited to a standard set of test sequences would be sufficient.

3.2 Video Test Materials

The selection of source video material for testing has brought out some differences of opinion among those in the video industry. The opposing notions are that of "realism" and that of "control." For many in the video industry, a useful test sequence should be representative of the type of material that will actually be processed by the system and shown to the viewer. It should be interesting to look at, and present as many encoding challenges as possible. The opposing view is that "interesting" video complicates the analysis of quality measurements. Since it is difficult to characterize such stimuli, it is also difficult to interpret the results. It is possible that both positions are correct to some extent.

For example, "realistic" stimuli are likely to present impairments in a "natural" context of visual masking elements and also to influence attention so as to either enhance or diminish the importance of picture defects. Of course, these factors greatly increase the number of unquantified or uncontrolled variables in an experiment.

Synthetic images are easier to control, but may not generate realistic impairments. In this regard, it has been observed by Fenimore, Libert, and Roitman [6] that without appropriate filtering prior to encoding, computer-generated graphics can generate some very unrealistic responses on the part of encoders. Moreover, the experimental context of the impairments may not be representative of "real television," if for example one's goal is to determine the limit of "acceptable" degradation for the home viewing. One should not lose sight of the fact that to a great extent, the video quality measurement technology is driven by the need to provide a picture that is "slightly better than just good enough" so as to free the remaining bandwidth for other purposes.

Weighing these considerations, the subcommittee settled on using actual video clips, as opposed to simple, computer-generated images. However, it was also decided to attempt to limit the composition of video impairments as much as possible to a single type, e.g., blocking, blurring, "mosquito noise," etc.

Measurement of visual thresholds requires video test sequences exhibiting increasing degradation of at least 3.0 JNDs. Moreover, if a paired comparison method is used for threshold estimation, it is necessary to control the degree of impairment finely and precisely, particularly in the neighborhood of the threshold. Experiments by Libert, Fenimore, and Roitman [7] detail two methods by which this can be accomplished. A 2^{nd} or 3^{rd} order polynomial can be fit via regression analysis to relate some objective measure of distortion, e.g. peak signal-to-noise ratio (PSNR), to mixing coefficients such that an impaired sequence can be linearly combined with its source to yield visually realistic impairments at any targeted level. This can be done in "real-time" using moderate computing power if needed to support adaptive threshold measurement schemes.

In order to simplify the interpretation of the threshold measurements, it also was decided to try to limit impairments of any test sequence to a single type of artifact. Methods for simulating impairments [8] were examined but were rejected over concerns that synthetic impairments thus generated might not be sufficiently realistic, either subjectively or objectively, to serve as a calibration benchmark. Accordingly, it was decided to limit impairments through the judicious selection of video source material and through the setting of encoding parameters. One approach might be to capture video content specifically expected to induce particular types of impairments. Suggested candidates include:

blocking -- a fixed camera position on a rapidly flowing stream of water. To a compression system, the moving water will present a moving picture with little correlation. To a human observer, there will be a highly predictable pattern. Squares or straight lines resulting from compression will be easily seen as compression artifacts.

mosquito noise -- a still image with text characters superimposed or keyed over a background image. The text will consist of white characters over a uniform dark gray background on some portions of the scene. The human observer will look for artificial vertical or horizontal edges surrounding the text characters.

contouring -- a still image with a smooth gradation in the luminance level. A scene containing a background wall with non-uniform lighting is suitable for this test. The human observer will look for contour lines on the background wall compared to the original scene which has a smooth change in the gray scale. If noise is present, the contour line will appear ragged.

Another possibility under consideration by the subcommittee is to apply its new measurements to the video material used in the recently completed VQEG study [8,9,11]. The VQEG video clips may not exhibit the desired homogeneity of impairment types, although the impairment composition may be found sufficiently limited in any given clip to satisfy this requirement. Clearly, some of the hypothetical reference circuits (HRCs) used in processing the source video might not be of interest to the IEEE subcommittee but many are of interest. The VQEG material offers the additional advantage that it already has associated with it an extensive database of subjective quality ratings. Even though these quality ratings are derived from a different method of category scaling, using the same material for threshold measurements may yield, for example, interesting data concerning those elements of video quality that are not solely perceptual in nature. Thus, the new threshold measurements may both contribute value to and extract value from the VQEG database.

Additional requirements are that source video be available to the public on a royalty-free basis, that it be distributed in Rec. 601 [11] digital video format and that the sequences be archived on a medium suitable for distribution with minimal generational loss such as digital video tape or high-capacity computer tape.

3.4 Subjective Testing

The details of the subjective testing protocol have yet to be fully specified, although guiding principals have been set. The primary objective of the subcommittee, is that the repeatability of the measurements be maximized and the uncertainty of the testing and subsequent analysis be minimized.

Some decisions have been made irrespective of the testing protocol. Inasmuch as the results would have to support critical viewing, the decision was made to use trained viewers, though not necessarily video experts. Subjects would be provided advanced direction as to the types of defects they might observe in each sequence. In order to "set" internal scales, subjects would be shown video covering the range of impairments at the beginning of test sessions. Other conditions aimed at supporting the video professional include using a large, professional quality monitor, 19" to 32" diagonal, and at least one viewing distance of three times the screen height. Other conditions of the subjective testing are included in a document by Stanger [13].

Several options are under consideration relative to the testing protocol as well as to the ultimate nature of the quality/fidelity scale that will result. Some examination of the options follows.

Early discussions of the desired type of subjective measurement tended to view the JND scale as a sequence of perceptual steps. Starting with an unprocessed source video sequence, a gradual and monotonic increase in stimulus intensity (impairment in the present case) would be introduced until the difference between the present and original state of the stimulus became just noticeable, i.e., 1.0 JND. Further increase in the impairment level would be undetectable until finally a level was reached at which the new state could be differentiated from the previous stimulus intensity, or 2.0 JNDs. Additional JND steps would be found by continuing this process, stepping through a sequence of JNDs, each JND serving as the "base level" for the next set of comparisons. Perhaps the stimulus intensity adjustment for each trial might be made discontinuously and at random distances from the base level.

If the goal of the experiment is to generate for each source video sequence a set of processed sequences at 3.0 JNDs, then this method may suffice provided that the rate of change in perceptibility is such that 3 or more JNDs of impairment remains within a useful interval for the application of the resultant standard. A potential problem with this discrete approach may be that only several calibration points are provided within the useful impairment range. For validation of an objective model or for its calibration, one would like to have more than just several subjective ratings over the range of impairment appropriate to the application. Depending upon the error variance of the subjective measure, one might like to have a large number of measurements or even an incremental threshold function describing for any degree of impairment the change of impairment just perceptible, e.g., $f(I) = \Delta I / I$, *where* I = *impairment level.*

Such a relationship might be determined by making a number of threshold measurements relative to each of a number of base level stimuli selected at random from the useful range of impairments. The relationship in JNDs between any two of the base level stimuli is not important as the degree to which at each level the impairment must be increased or decreased so as to

be differentiated from the base level. Once the constraint of discrete JND steps is lifted, it becomes obvious that any stimulus may serve as the comparison base level and that either positive or negative changes from this base level carry useful information. Taking this notion further, one realizes that the designation of a "base level" or comparison standard is completely arbitrary and that any stimulus may be compared to any other in deriving the $\Delta I / I$ relationship.

The disadvantage to threshold measurements of the sort discussed above is that only the comparisons of stimuli close to the visual threshold actually provide useful information if one considers only that A > B or vice versa. That is, with only the inequality as a guide, the adjustment of the stimulus value can only be selected according to some algorithm using the direction of the inequality. Without an *a priori* expectation of the location of a threshold, one can expect to "waste" a relatively large number of trials while converging on the threshold. The direction of change is indicated, but no guidance is provided as to the magnitude of the required adjustment. (Although, it should be noted that the inequality provides more information than simply $A \neq B$.)

In the present case, the need for efficiency is amplified by the intent to measure multiple thresholds. Therefore, a large number of trials may be necessary. Unlike measurements using instruments having unlimited endurance, those involving human observers must consider fatigue as a significant factor to be controlled.

The information provided by each comparison can be increased by calling for the subject to estimate the distance between the two stimuli along a numerical scale. A scale from +20 to –20 or +10 to –10 can provide information about the relative degree of perceived difference between the comparison stimuli. The sign indicates the direction of change from stimulus A to stimulus B. In an experiment working with one subject per session, the numerical estimate can guide the adjustment of the stimulus, improving the efficiency of each trial.

Further efficiency might be realized by limiting the degree of stimulus adjustment needed to arrive at each JND. Whereas a fixed interval staircase approach or even an adaptive procedure such as described in [14] might involve a number of trials to converge on a particular JND, a numerical category scaling method such as that described in [16,17] with appropriate statistical treatment of the data might provide an efficient solution.

4. SUMMARY AND CONCLUSIONS

The IEEE Compression Measurements Subcommittee was commissioned to evaluate appropriate methodologies and video test materials in order to recommend a standard for measurement of digital video quality. It has taken a course that it hopes will lead eventually to definition of a quantitative scale of video quality and unit of measure based on the threshold, or just-noticeable-difference between video sequences exhibiting various degrees of compression-induced impairment. Eventually, the subcommittee may identify a measurement algorithm as the basis for an IEEE standard. In the near-term, its goal is to support quality metric development by supplying perceptually-referenced video materials with which to validate and calibrate objective quality or fidelity metrics. Several experimental approaches are under consideration by the subcommittee.

5. ACKNOWLEDGEMENTS

The authors wish to acknowledge Alan Godber, Interim Chairman of G-2.1.6 and to extend special thanks to recording secretary, Doug Lung for maintaining the subcommittee Web page, including his excellent meeting records, referred to often in preparing this paper. Moreover, recognition is extended to the members of IEEE G-2.1.6 for their continuing contributions to measurement standards development.

6. REFERENCES

1. Leon Stanger. "Detailed Scope of Activities: Institute of Electrical and Electronics Engineers Broadcast Technology Society, Audio-Video Techniques Committee G-2.1, Compression and Processing Subcommittee G-2.1.6," Revision 1.1, 5 January 1998.
2. David Fibush. " Overview of standardization activities related to digital television," *Human Vision and Electronic Imaging V*, SPIE Vol. 3959, January 2000. [Preprint]
3. Recommendation ITU-R BT.500-9, "Methodology for the subjective assessment of the quality of television pictures," ITU-R 1974-1998.

4. Leon Stanger. "The Need for A Unit of Measure and Calibration for Picture Quality Degradation," Broadcast Technology Society, Audio-Video Techniques Committee G-2.1, Compression and Processing Subcommittee G-2.1.6, Doc. G-2.1.6/74, 25 January 1998.
5. D. A. Silverstein and J. E. Farrell. "The relationship between image fidelity and image quality," *Proceedings of the IEEE International Conference on Image Processing*, Lausanne Switzerland, 881-884 1996
6. Charles P. Fenimore, John M. Libert, and Peter Roitman. "Mosquito noise in MPEG compressed video: test patterns and metrics," *Human Vision and Electronic Imaging V* , SPIE Vol. 3959, January 2000. [Preprint]
7. John M. Libert, Charles Fenimore, and Peter Roitman. "Simulation of graded video impairment by weighted summation: validation of the methodology," *Multimedia Systems and Applications II,* SPIE Vol. 3845,
8. Recommendation ITU-T P.930, *Principals of a reference impairment system for video,* ITU-T 1996.
9. P. Corriveau and A. Webster, "VQEG evaluation of objective methods of video quality assessment," SMPTE Journal, 108, pp. 645-648, 1999.
10. Philip J. Corriveau , Nikolaus Walch , and Alexander Schertz. *VQEG Subjective Test Plan.* Draft 15, 8/18/99. [Download from VQEG web-site, http://www.crc.ca/vqeg/]
11. Philip J. Corriveau, Arthur A. Webster, Ann Marie Rohaly, and John M. Libert. "Video quality experts group: the quest for valid objective methods," *Human Vision and Electronic Imaging V* , SPIE Vol. 3959, January 2000, San Jose, CA.
12. Recommendation BT.601-5, *Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios.*
13. Leon Stanger. "Requirements for a JND-based video quality measurement method," Isreqc.doc. 21 June 1999, IEEE G-2.1.6 Web-site,
14. Andrew B. Watson and Denis G. Peli. "QUEST: A Bayesian adaptive psychometric method*," Perception & Psychophysics 1983*, 33 (2), 113-120.
15. Thom Carney, Stanley A. Klein, Christopher W. Tyler, Amnon D. Silverstein, Brent Beutter, Dennis Levi, Andrew B. Watson, Adam J. Reeves, Anthony M. Norcia, Chien-Chung Chen, Walter Makous, and Miguel P. Eckstein. "The development of an image/threshold database for designing and testing human vision models," *Human Vision and Electronic Imaging IV* , SPIE Vol. 3644, 25-29 January 1999, San Jose, California, 542-551.
16. D. Amnon Silverstein and Joyce E. Farrell. "Quantifying perceptual image quality," The Society for Imaging Science and Technology, Image processing quality and capture, May 1998.
17. A. M. van Dijk, J. B. Martens, and A. B. Watson. "Quality assessment of coded images using numerical category scaling," SPIE Vol. 2451, January 1999.
18. J. M. Neter, M. H. Kunter, C. J. Nachtsheim, and W. Wasserman. 1996. *Applied Linear Statistical Models,* 4th Edition, WCB/McGraw-Hill, Boston, MA.