NTIA Technical Memorandum TM-04-412

The Impact of Monitor Resolution and Type on Subjective Video Quality Testing

Margaret H. Pinson
Stephen Wolf

technical memorandum

U.S. DEPARTMENT OF COMMERCE · National Telecommunications and Information Administration

NTIA Technical Memorandum TM-04-412

The Impact of Monitor Resolution and Type on Subjective Video Quality Testing

Margaret H. Pinson
Stephen Wolf

U.S. DEPARTMENT OF COMMERCE
Donald L. Evans, Secretary

Michael D. Gallagher, Acting Assistant Secretary
for Communications and Information

March 2004

DISCLAIMER

Certain commercial equipment and materials are identified in this report to specify adequately the technical aspects of the reported results. In no case does such identification imply recommendations or endorsement by the National Telecommunications and Information Administration, nor does it imply that the material or equipment identified is the best available for this purpose.

CONTENTS

Page

THE IMPACT OF MONITOR RESOLUTION AND TYPE ON SUBJECTIVE VIDEO QUALITY TESTING

Margaret H. Pinson and Stephen Wolf[1]

This memorandum compares subjective video quality test results from a professional cathode ray tube (CRT) television monitor with that of a consumer liquid crystal display (LCD) video phone monitor. The CRT monitor supported the full ITU-R Recommendation BT.601 resolution (720 x 486) while the LCD monitor only supported Common Intermediate Format (CIF) resolution (352 x 288). The subjective results from the two tests are very similar, with the only significant difference being that the CIF monitor masks impairments that appear in only one of the two interlaced fields.

Key words: CIF; image quality; ITU-R Recommendation BT.601; monitor resolution; subjective testing; video quality

1. INTRODUCTION

Concern has been expressed regarding the use of a professional cathode-ray tube (CRT) monitor when conducting a subjective video quality test of systems intended for multimedia (MM) applications. Typical MM applications (e.g., internet video) utilize personal digital assistants (PDAs) and cellular telephones that display video on small liquid crystal display (LCD) panels. The influence of monitor resolution and type (CRT vs. LCD) on the subjective video quality ratings has not been investigated. This document describes and compares results from a subjective video quality experiment that was conducted using two monitors: a professional CRT monitor with ITU-R Recommendation BT.601 resolution (720 x 486) and a consumer LCD monitor with common intermediate format (CIF) resolution (352 x 288).

2. SUBJECTIVE EXPERIMENT DESIGN

The subjective experiment described in this document will be named data set thirteen.[2] Data set thirteen employs the Single-Stimulus Continuous Quality Evaluation (SSCQE) method [1]. Data set thirteen uses hidden reference removal, a second stage in post-processing of the SSCQE scores that is being proposed by VQEG for the upcoming Reduced Reference No Reference Television (RRNR-TV) test.[3]

---

[1] The authors are with the Institute for Telecommunication Sciences, National Telecommunications and Information Administration, U.S. Department of Commerce, Boulder, CO 80305.

[2] This data set number was chosen to be consistent with the numbering of other video quality data sets in previous NTIA/ITS publications.

[3] With hidden reference removal, the original video sequences are presented, but the viewers are not aware that they are evaluating the original video. The quality score is computed by taking 100 plus the viewer's opinion of the processed video sequence minus the viewer's opinion of the original video sequence, analogous to the Double Stimulus Comparison Scale (DSCS) method. The quality scores range from 0 (worst quality) to 100 (best quality), with some chance of excursions above 100 (e.g., when the processed video is perceived as being of higher quality than the original video).

This subjective test uses twenty original video sequences: one 45-sec video sequence, three 30-sec video sequences, and sixteen 15-sec video sequences. The 45-sec video sequence is a movie trailer containing rapid scene cuts. The three 30-sec video sequences each contain a concatenation of three 10-sec video sequences with similar video content (e.g., outdoor scenes containing ducks and water).

Data set thirteen uses sixteen hypothetical reference circuits (HRCs). Four HRCs used an internet video-phone codec operating at rates ranging from 64 kbits/s to 384 kbits/s; three of these video-phone HRCs included added traffic impairments in the digital transmission channel. One HRC used a video-phone that was designed to operate over an analog phone line (28 kbits/s). One HRC used an older proprietary codec operating at 384 kbits/s with Gaussian distributed digital transmission errors at a bit error rate of $8 \cdot 10^{-5}$. Eight impairments were produced using one of three different software MPEG-2 encoders operating at bit-rates ranging from 2 Mbits/s to 8 Mbits/s with varying encoding options. The software MPEG-2 file was then written to a DVD and played using a consumer grade DVD player. Two of these eight HRCs contain errors that were produced by fingerprints on the DVD. The remaining two impairments used a software MPEG-2 encoder operated in low-resolution mode at 1 Mbits/s and 3 Mbits/s.

Data set thirteen did not utilize a complete matrix of the above scenes and HRCs. Thirty-four video clips that contained interesting error conditions were included in the subjective test. One of the MPEG-2 encoders (denoted as "s25") contained an implementation error that will be described later in this document. Five scenes that exhibited that implementation error were included in the test. The remaining video sequences were primarily chosen such that the range of quality within the subjective test was evenly distributed for low to high quality. Where possible, secondary considerations in the test design were to match each scene with either three or four HRCs. However, one scene was matched with only two HRCs and three scenes were associated with more than four HRCs. As a result of these criteria, HRCs were matched with between three and ten scenes, for a total of 110 processed video sequences.

To reduce the amount of data involved in the analyses and minimize autocorrelation of the samples, the SSCQE data with hidden reference removal was sub-sampled in time. Subjective ratings were retained every five seconds, beginning ten seconds into the video sequence. Thus, two subjective ratings were retained for the 15-sec video sequences, six subjective ratings for the 30-sec video sequences, and eight for the 45-sec video sequences. Experiments [2] have shown that these sub-sampled SSCQE subjective ratings are closely correlated to ratings produced by a Double Stimulus Continuous Quality Scale (DSCQS) test [1]. These experiments were performed using the ten seconds of video prior to each of the retained SSCQE subjective ratings (i.e., divide the SSCQE video sequence into 10-sec video sequences that overlap by 5-sec, and then perform a DSCQS test).

Four randomized viewing orderings were created, ensuring that no scene or HRC appeared twice in a row. Viewers were evenly distributed among the four randomized orderings. The original video sequences are included in the viewing sessions (for the hidden reference removal). The viewing sessions were 31 minutes in duration. A short training session was developed using video sequences created for but not used in this subjective test. Viewers were chosen at random from the U.S. Department of Commerce site phonebook.

3. MONITOR COMPARISON

Two sets of viewers were used for data set thirteen: one set of viewers used a 20 inch professional CRT monitor, and one set of viewers used a 5 ½ inch consumer LCD monitor (a CIF resolution internet video-

phone with an NTSC video input[4]). Data were collected from twenty viewers for each monitor, using the same viewing room with identical lighting conditions, viewing distance (approximately 4 times the monitor height), etc. Subjective results from the professional CRT monitor will be referred to as the "CRT High Resolution" and subjective results from the IP video-phone will referred to as the "LCD Low Resolution."

Figure 1 depicts a scatter plot of the two sets of subjective ratings, averaged over all viewers in each set.

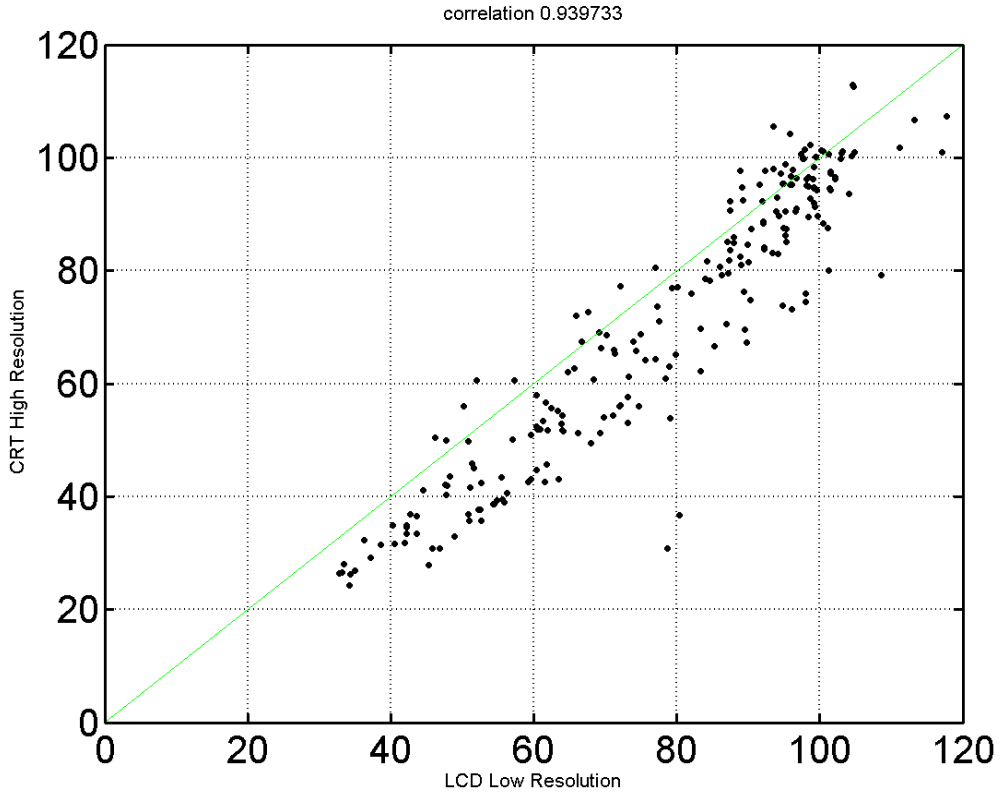

Figure 1. Scatter plot comparing subjective ratings from CRT high resolution and LCD low resolution monitors.

From Figure 1, one can observe an obvious DC shift and gain between the subjective results from the two monitors, where the CRT high resolution (CRT_high) and LCD low resolution (LCD_low) monitor results are related as follows:

$$CRT\_high = -12.5584 + LCD\_low * 1.0657 \quad (1)$$

Such systematic differences in scores commonly occur when multiple subjective experiments are conducted [3].

[4] National Television Systems Committee (NTSC) is the 525-line analog color video composite system adopted by the US and most other countries (excluding Europe).

Figure 2 depicts a scatter plot of the CRT high resolution subjective ratings against the LCD low resolution ratings, after scaling the LCD low resolution ratings by Equation 1. Data points plotted in blue are low resolution systems, most of which utilize a CIF or QCIF encoding scheme. Data points plotted in red are HRCs that preserved both interlaced fields (field 1 and field 2). A visual examination of the plot indicates that there is no obvious bias for this subset of field-preserving HRCs.

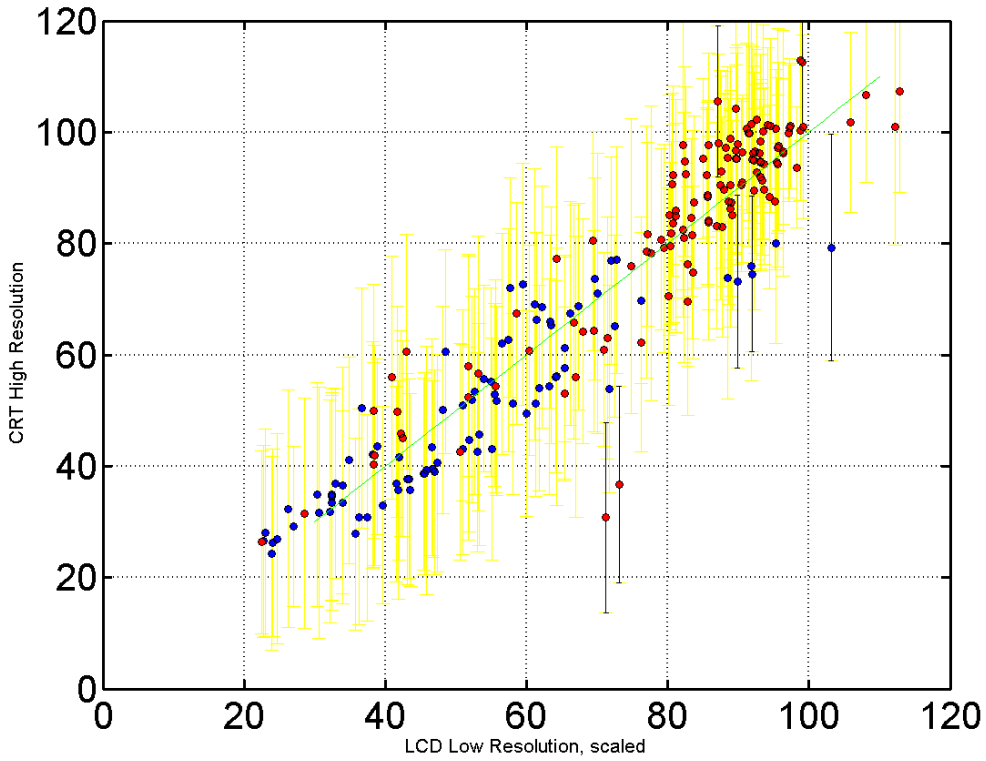

Figure 2. CRT high resolution and LCD low resolution MOS, plotted with summed error bars.

Each clip is plotted with an error bar that indicates the sum of the CRT and LCD monitors' 97.5% confidence intervals (CI), using the Student T-test and therefore without a presumption of normally distributed data.[5] Using those individual CIs, we have joint 95% confidence that 212 of the 219 data points have clip means that are equivalent; these clips are plotted with yellow error bars. The seven remaining data points are plotted with black error bars.

The worst two outliers in Figure 2 are both from the same scene and codec. The scene is a 15-sec segment of the ANSI scene "football" [4] and the impairment was created using the faulty MPEG-2 encoder mentioned earlier (HRC s25). These two clips display an impairment judged to be "poor" by the CRT high resolution viewers (i.e., approximately 30 on the SSCQE scale), whereas the LCD low resolution viewers judge the impairment to be "good" (i.e., around 70 on the SSCQE scale). When the video is viewed simultaneously on both monitors, the above ratings are found to reflect real perceived differences. The s25 impairment for the football scene seems to be related to a temporal field ordering problem (i.e., late field output *before* early field) that appears in wide horizontal bands across the screen.

[5] Error bars are summed to simplify the visual representation.

The impairment is particularly visible in high motion portions of the video sequence. The impairment is not perceived on the LCD low resolution monitor. An investigation determined that the LCD low resolution monitor discards NTSC field two, thus rendering the impairment invisible.

The five remaining outliers come from 15-second scenes, where the SSCQE score was sampled at the 10 and 15 second points. Three of these five outliers came from scenes where only one of the two SSCQE scores was an outlier (i.e., the other SSCQE score was not an outlier). The other two outliers both came from the same 15-second scene (i.e., SSCQE samples at both the 10 and 15 second points were outliers). This 15-second scene was a black and white weather satellite imagery scene passed through a low resolution MPEG-2 encoder at 3 Mbits/s. The satellite video was updated 10 times each second.

The correlation between the CRT and LCD viewers' subjective ratings is 0.940. The correlation without the two outliers mentioned above is 0.951. For comparison purposes, there are three subjective experiments that have been performed by standards committees for which lab-to-lab correlations are available: a T1A1 subjective test [5], a VQEG Full Reference Television Phase I test [6], and a VQEG Full Reference Television Phase II test [7]. The correlation between the CRT and LCD monitor subjective ratings lies within the range of correlations established by these lab-to-lab correlations, shown in Table 1.

Table 1. Lab-to-Lab Correlations for Video Quality Experiments Performed by Standards Committees

Test	**Labs**	**Total Viewers**	**Lab-to-lab Correlation**
T1A1	3	30	0.926, 0.952, and 0.958
VQEG FRTV Phase I, 50Hz/low quality	4	73	0.942, 0.946, 0.950, 0.956, 0.945, and 0.948
VQEG FRTV Phase I, 50Hz/high quality	4	71	0.882, 0.892, 0.909, 0.882, 0.851, and 0.876
VQEG FRTV Phase I, 60Hz/low quality	4	80	0.747, 0.913, 0.933, 0.807, 0.727, and 0.935
VQEG FRTV Phase I, 60Hz/high quality	4	73	0.790, 0.854, 0.831, 0.818, 0.837, and 0.880
VQEG FRTV Phase II, 60Hz	2	66	0.97

We can also make a histogram plot of the difference between the CRT and LCD monitors' MOS. This histogram is shown in Figure 3. If the two football / s25 outliers are eliminated, this distribution passes the Bera-Jarque hypothesis test of composite normality [8].

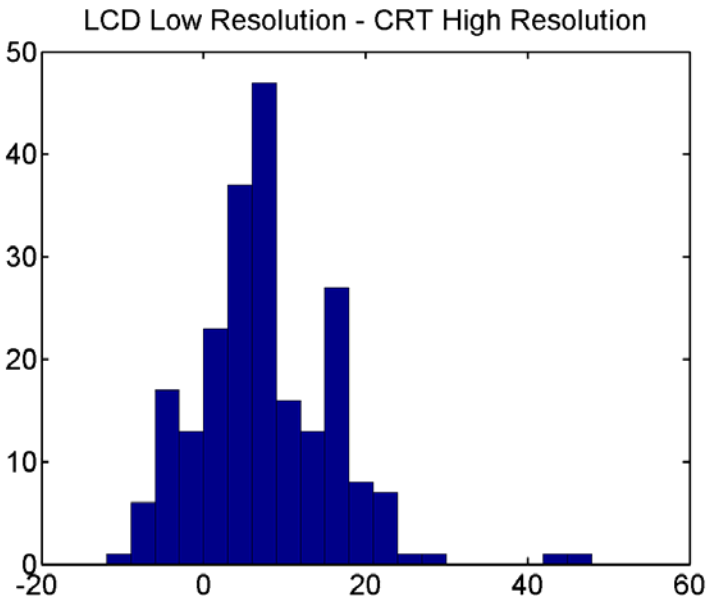

Figure 3. Histogram of LCD minus CRT monitor averaged viewer ratings.

4. CONCLUSION

Subjective results from the professional CRT and consumer LCD monitors were to a large extent statistically equivalent. The correlation between CRT and LCD scores was comparable to lab-to-lab correlations; 97% of the video sequences had equivalent clip means after compensating for the systematic differences in scores given by Equation 1; and, once two football outliers were explained and eliminated, the difference between CRT and LCD scores had a normal distribution.

The results from our monitor comparison study indicate that a CRT high resolution monitor probably can be used to emulate the subjective experience of viewers utilizing an LCD low resolution monitor, provided some caution is exercised. The differences between the CRT and LCD monitors (e.g., response time, artifact visibility, resolution, color calibration) did not significantly impact the subjective ratings of most sequences. The choice of a double stimulus subjective test (i.e., measuring the difference between the original and processed video sequences) rather than a single stimulus subjective test (i.e., rating only the processed video sequence) probably reduced the impact of monitors on the final subjective scores.

A significant deviation between the subjective responses for the two monitors occurred for an impairment that was only visible on monitors that are capable of displaying both NTSC field one and field two. The impairment in our experiment with this attribute was created by a faulty codec that confused the proper NTSC interlaced field ordering. While the impairment was readily visible on the CRT monitor, the LCD monitor negated the impairment by discarding one of the NTSC fields. We expect impairments of this type to be rare in practice. One other area of caution might be in testing the perceptibility of very brief transient impairments, since LCDs typically have slower response times than CRTs. However, our subjective experiment did not address this area of concern.

5. REFERENCES

[1] ITU-R Recommendation BT.500, "Methodology for subjective assessment of the quality of television pictures," Recommendations of the ITU, Radiocommunication Sector.

[2] M. Pinson and S. Wolf, "Comparing subjective video quality testing methodologies," SPIE Video Communications and Image Processing Conference, Lugano, Switzerland, Jul. 8-11 2003.

[3] M. Pinson and S. Wolf, "An objective method for combining multiple subjective data sets," SPIE Video Communications and Image Processing Conference, Lugano, Switzerland, Jul. 8-11 2003.

[4] ANSI T1.801.01-1995, "American National Standard for Telecommunications – Digital Transport of Video Teleconferencing/Video Telephony Signals – Video Test Scenes for Subjective and Objective Performance Assessment," American National Standards Institute.

[5] C. Jones, N. Crow, S. Wolf, and A. Webster, "Analysis of T1A1.5 Subjective and Objective Test Data," ANSI T1A1 contribution number T1A1.5/94-152, Oct 1994.

[6] Video Quality Experts Group (VQEG), "Final report from the Video Quality Experts Group on validation of objective models of video quality assessment," 2000 VQEG. Available: www.vqeg.org

[7] Video Quality Experts Group (VQEG), "Final report from the Video Quality Experts Group on validation of objective models of video quality assessment, phase II," 2003 VQEG. Available: www.vqeg.org

[8] G. G. Judge, R. C. Hill, W. E. Griffiths, H. Lutkepohl, and T.-C. Lee, *Introduction to the Theory and Practice of Econometrics*, New York: Wiley, 1988.

FORM **NTIA-29** U.S. DEPARTMENT OF COMMERCE
(4-80) NAT'L. TELECOMMUNICATIONS AND INFORMATION ADMINISTRATION

BIBLIOGRAPHIC DATA SHEET

1. PUBLICATION NO. TM-04-412	2. Government Accession No.	3. Recipient's Accession No.
4. TITLE AND SUBTITLE The Impact of Monitor Resolution and Type on Subjective Video Quality Testing		5. Publication Date March 2004 6. Performing Organization Code
7. AUTHOR(S) Margaret H Pinson and Stephen Wolf		9. Project/Task/Work Unit No. 3141000-300
8. PERFORMING ORGANIZATION NAME AND ADDRESS Institute for Telecommunication Sciences National Telecommunications & Information Administration U.S. Department of Commerce 325 Broadway Boulder, CO 80305		10. Contract/Grant No.
11. Sponsoring Organization Name and Address National Telecommunications & Information Administration Herbert C. Hoover Building 14th & Constitution Ave., NW Washington, DC 20230		12. Type of Report and Period Covered

14. SUPPLEMENTARY NOTES

15. ABSTRACT (A 200-word or less factual summary of most significant information. If document includes a significant bibliography or literature survey, mention it here.)

This memorandum compares subjective video quality test results from a professional cathode ray tube (CRT) television monitor with that of a consumer liquid crystal display (LCD) video phone monitor. The CRT monitor supported the full ITU-R Recommendation BT.601 resolution (720 x 486) while the LCD monitor only supported Common Intermediate Format (CIF) resolution (352 x 288). The subjective results from the two tests are very similar, with the only significant difference being that the CIF monitor masks impairments that appear in only one of the two interlaced fields.

16. Key Words (Alphabetical order, separated by semicolons)

CIF; image quality; ITU-R Recommendation BT.601; monitor resolution; subjective testing; video quality

17. AVAILABILITY STATEMENT	18. Security Class. (This report)	20. Number of pages
☐ UNLIMITED.	Unclassified	7
	19. Security Class. (This page) Unclassified	21. Price:

NTIA FORMAL PUBLICATION SERIES

NTIA MONOGRAPH (MG)

A scholarly, professionally oriented publication dealing with state-of-the-art research or an authoritative treatment of a broad area. Expected to have long-lasting value.

NTIA SPECIAL PUBLICATION (SP)

Conference proceedings, bibliographies, selected speeches, course and instructional materials, directories, and major studies mandated by Congress.

NTIA REPORT (TR)

Important contributions to existing knowledge of less breadth than a monograph, such as results of completed projects and major activities. Subsets of this series include:

NTIA RESTRICTED REPORT (RR)

Contributions that are limited in distribution because of national security classification or Departmental constraints.

NTIA CONTRACTOR REPORT (CR)

Information generated under an NTIA contract or grant, written by the contractor, and considered an important contribution to existing knowledge.

JOINT NTIA/OTHER-AGENCY REPORT (JR)

This report receives both local NTIA and other agency review. Both agencies' logos and report series numbering appear on the cover.

NTIA SOFTWARE & DATA PRODUCTS (SD)

Software such as programs, test data, and sound/video files. This series can be used to transfer technology to U.S. industry.

NTIA HANDBOOK (HB)

Information pertaining to technical procedures, reference and data guides, and formal user's manuals that are expected to be pertinent for a long time.

NTIA TECHNICAL MEMORANDUM (TM)

Technical information typically of less breadth than an NTIA Report. The series includes data, preliminary project results, and information for a specific, limited audience.

For information about NTIA publications, contact the NTIA/ITS Technical Publications Office at 325 Broadway, Boulder, CO, 80305 Tel. (303) 497-3572 or e-mail info@its.bldrdoc.gov.