Quantitative Evaluation of Overall Electronic Display Quality

Nicholas J. Hangiandreou, Kenneth A. Fetterly, Scott N. Bernatz,
Laurie J. Cesar, Debra S. Groth, and Joel P. Felmlee

THE USE OF ELECTRONIC display devices for primary diagnostic interpretation of digital images in Radiology is becoming commonplace. For all equipment involved in image acquisition or display, it is typical practice in radiology that quantitative measurements of physical parameters affecting image quality be made for the purposes of new equipment selection, acceptance testing, and quality control (QC). Electronic displays are problematic in that routine measurement of display parameters other than luminance are difficult and expensive to obtain, especially in the field. To address this difficulty, we are investigating a method for quantitatively evaluating the overall quality of electronic displays. This technique is very similar to traditional contrast-detail (CD) methods[1-3] in that many factors affecting image quality (including contrast sensitivity, noise and modulation transfer) are included in an overall quality measure. A technique acceptable for the purposes of new equipment selection, acceptance testing, and QC would allow the quantitative evaluation of a device in 30 minutes or less with a precision of about 10%. Such a technique would also allow quantitative evaluation of the effects of display set-up and environment, such as maximum display luminance and ambient room lighting.

The purpose of the current work is to make an initial pilot investigation of the feasibility of using contrast-detail techniques to provide overall quantitative quality evaluations of electronic displays. Our data may also allow us to make a preliminary evaluation of the importance of maximum display luminance.

METHODS

Contrast-detail methods[1-3] involve the presentation of simple targets on a uniform background to a group of observers. The targets have varying contrast and size. For each target size, the observer indicates the contrast threshold at which the target is just visible, and in most cases, interpolation between actual image contrast is encouraged. We incorporated the idea of requiring the observer to indicate the area of the test image in which the target is seen as proof that the target is actually visualized.[4]

Six sets of 8-bit per pixel test images were created. Each set consisted of eight images, each image corresponding to one of eight possible square target sizes. The target sizes were 1, 2, 3, 4, 7, 11, 17, and 27 pixels. In each test image, eight rows of four test areas were present. Each row corresponded to one of eight possible target contrasts. The target contrasts were 1, 2, 3, 4, 7, 11, 17, and 27, denoted as the pixel value difference between the target and the background. For object size of 1 pixel, larger contrasts were used. All targets were darker than the background to enhance sensitivity to cathode ray tube (CRT) glare. Each test area presents the target in one of four randomly selected quadrants. The test areas are defined by low contrast borders. A test image set is schematically illustrated in Fig 1.

The test image sets were created using a common PC paint application (PaintShop Pro, Version 4, JASC Incorporated, PO Box 44997, Eden Prairie, MN, 55344). Three test image sets were created with 15% video backgrounds (pixel value = 38), and three test image sets had 85% video backgrounds (pixel value = 217). The pixel matrix of each test image was 1518 × 1758, and was selected to match 1:1 with the video memory matrix size of the workstation used to present the images to the observers (SCID 2A diagnostic workstation and PACS, General Electric Company, Mount Prospect, IL, 60056).

A group of five observers consisting of two medical physicists and three radiology quality control technologists was recruited. For each viewing session, three independent test image sets with 15% video backgrounds, and three independent test image sets with 85% video backgrounds were presented. During the viewing session, the observer was asked to inspect each test area and indicate either the quadrant in which the target was observed, or that no target was seen. A score of 1 was assigned for each correct target location indicated, 0 was assigned for each incorrect response, and 0.25 was assigned for each "no target seen" response (the average of that expected from guessing). To minimize data collection time, observers were allowed to start at any row in which all four targets were correctly located, and to proceed from there to lower-contrast rows. The viewing time required for each set of test images was recorded.

Three display conditions were tested. In all cases, the same workstation and CRT monitor were used (Megascan Model UHR4212P, Raytheon E-Systems, 11 Executive Park Drive, Billerica, MA, 01862). The room lighting was held constant at 4.7 ± 0.6 lux, corresponding to the approximate ambient light level found in one soft-copy reading room on our campus.[5] The three display conditions corresponded to maximum monitor luminance settings of 30 ± 0.5, 50 ± 0.5 and 70 ± 0.5 foot-Lamberts (ft-L). The minimum luminance was 0.042 ± 0.002 ft-L in each condition. The monitor display function was measured in each condition. Monitor luminance and room illumination measurements were made with a standard lumi-

From the Department of Diagnostic Radiology, Mayo Clinic and Foundation, Rochester, MN.

Address reprint requests to Nicholas J. Hangiandreou, Ph.D., Diagnostic Radiology, East 2, Mayo Clinic, 200 First Street SW, Rochester, MN 55905.

0897-1889/98/1103-1056$8.00/0

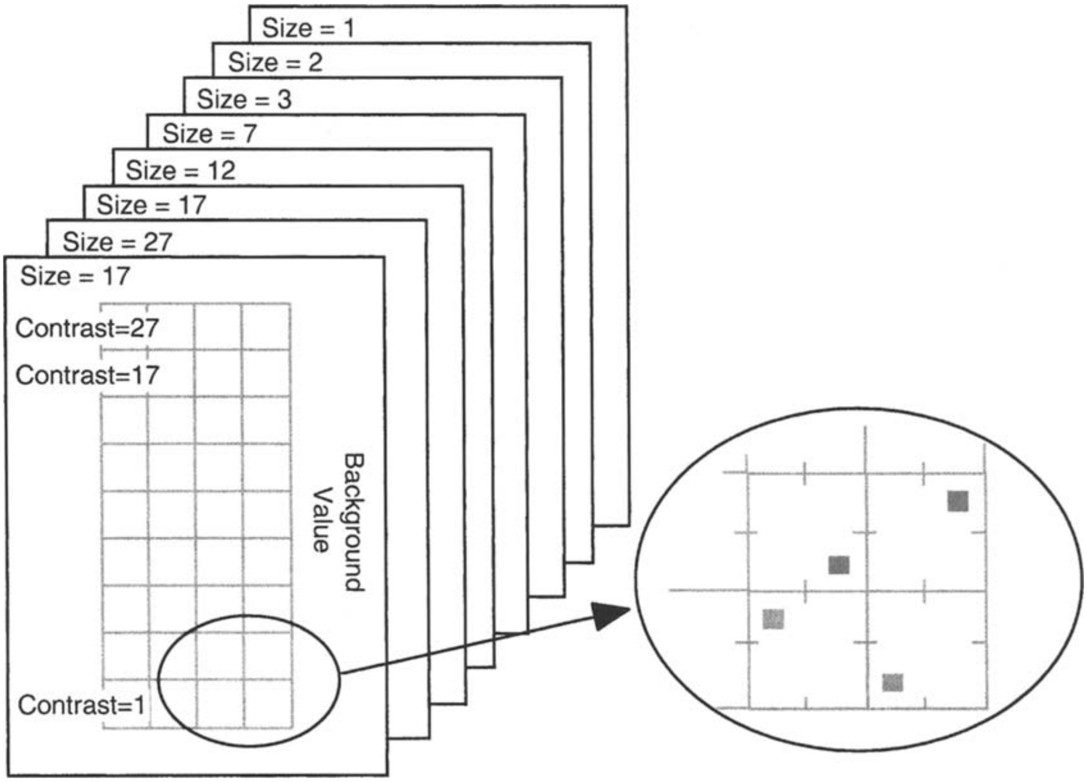

Fig 1. Schematic diagram of the eight images comprising one test image set.

nance head (model 265), cosine filter (model 211), and photometer (model 371, United Detector Technology, Graseby Optronics, 12151 Research Parkway, Orlando, FL, 32828).

Analysis of the data collected using each set of eight test images resulted in a determination of threshold contrast values for each image and target size. This process began with the tabulation of a score for each row in the image by summing the scores for the four test areas. Then, starting at the lowest-contrast row, the row-scores were examined for the first instance where a critical value of 2.5 was exceeded. The threshold contrast was obtained via logarithmic interpolation between the contrasts bracketing the critical value. The critical value 2.5 was selected as the midpoint between perfect observation of the test objects with a row-score of 4, and the average row-score of 1 expected if no targets were actually seen and responses were due to guesses. Our results were found not to vary as a sensitive function of the actual critical value used.

Each of the eight images in a test set will potentially yield a (target size, threshold contrast) ordered pair or data point, so each test set will yield a maximum of eight data points. No data point will result if the row-scores in the image all exceed the critical value. The data points from each test set are graphed on a log-log (base-10) plot. These data points are also fitted to a line with a fixed slope of -1 (as predicted by the Rose Model[6]), resulting in a best-fit y-intercept. The y-intercept is raised to the power of 10. Three test image sets were observed during each viewing session, resulting in three measurements of $10^{\text{y-intercept}}$. The mean value (and standard deviation) of these three measurements is computed, and is taken as a figure-of-merit for the overall display quality of the monitor. This figure-of-merit has units of contrast and may be interpreted as the maximum threshold contrast (MTC) of the display corresponding to the smallest (1-pixel) target that may be presented. Lower MTC values indicate better display performance.

Data were collected from all five observers for a particular monitor condition over a period of about a week. The monitor condition was then changed and data collection was repeated for the five observers. The order of data collection was 50 ft-L, 30 ft-L, 70 ft-L, and 50 ft-L. The notation "50(1)" and "50(2)" will be used to distinguish data and results from the first and second 50 ft-L sessions. Four additional 50 ft-L data sets were collected over a subsequent period of about a month for observers A and B. During each viewing session, data were collected using test images with 15% background pixel values and 85% background pixel values. The notation "dark background" and "bright background" will be used to refer to the 15% and 85% cases, respectively.

RESULTS

Figure 2 shows sample graphs of data collected for observer A, with a maximum monitor luminance of 50 ft-L. The fixed-slope lines of best fit are also shown for the data collected from each test image set. The data sets were commonly found to deviate from the ideal linear, fixed-slope model. In general, data collected using the bright background

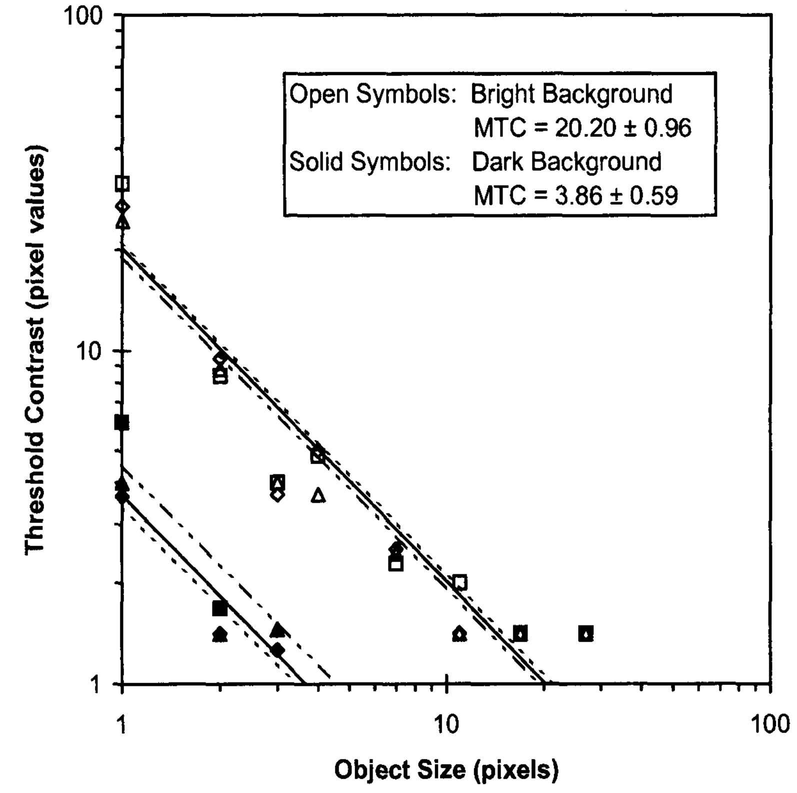

Fig 2. Sample threshold contrast data collected for observer A, during one observation session, with 50 ft-L maximum luminance. Three test image sets for both dark and bright backgrounds were used, each resulting in one data set.

test images indicated poorer display performance (larger threshold contrast values) than for the dark background. The bright background data sets also contained more data points (average n = 23) than the dark background data sets (average n = 8). The observation time necessary to review the test sets was measured to be 6.4 ± 1.8 minutes and 11.1 ± 3.0 minutes for dark and bright background images, respectively.

Figure 3 presents a summary of the *single observer* MTC measurements obtained from each of the five observers under the four monitor conditions for both dark and bright background values. The averages over all observers and monitor conditions of the MTC, standard deviation and percent standard deviation values were 4.02 ± 0.45 (11.2%) and 22.05 ± 1.49 (6.8%), for the dark and bright background cases, respectively. The percent standard deviation is interpreted as an indication of the *precision* of the MTC measurement. The *sensitivity* of the MTC measurement is obtained by computing the minimum relative change in the average MTC measurements that is expected to be demonstrated as statistically significant (*t* test, 95%). MTC sensitivity values of 25.4% and 15.2% were computed for the dark and bright background cases, respectively.

The six MTC measurements made for observers A and B over a two month period were analyzed for reproducibility by computing mean and standard deviation values (Table 1). Percent standard deviation values are seen to be about 10% and 5%, for the dark and bright background cases, respectively.

Average MTC measurements over the group of five observers for the various monitor conditions were computed, and are also shown in Fig 3. The sensitivities of these average MTC values are 35.9% and 35.4% for the dark and bright background cases, respectively. The poorer sensitivities of these group averages, as compared to the single observer sensitivities noted above, are due to the larger standard deviations of the group measurements as seen in Fig 3. These result directly from the variation in absolute performance between the five observers, also evident in Fig 3.

An alternative method of analyzing group data is through paired difference analysis. This involves computing differences between MTC measurements obtained by individual observers under different monitor conditions, and then computing the group average of these difference values. Results of the *group paired difference* MTC analysis for comparisons between the four monitor conditions are shown in Fig 4. A negative MTC difference

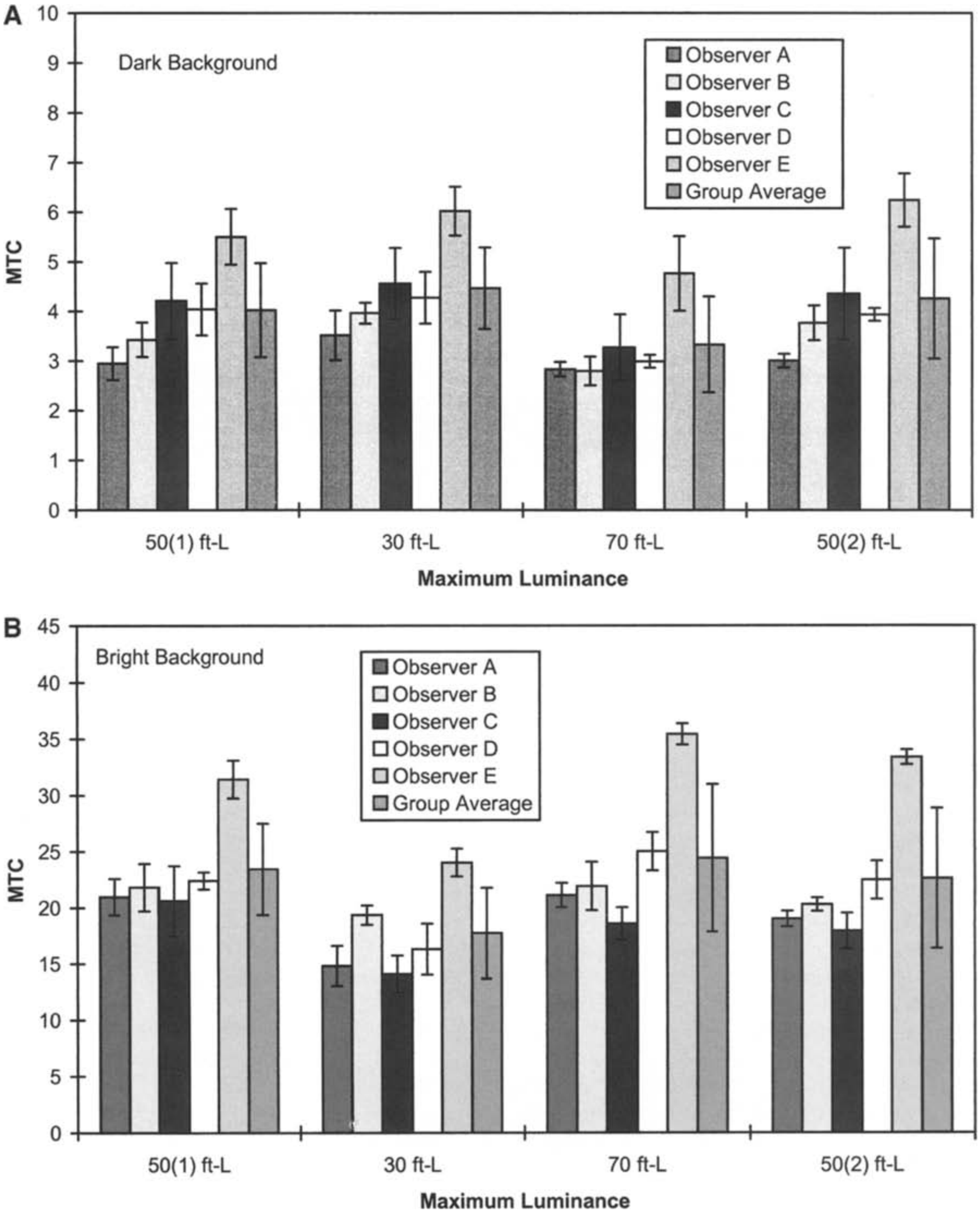

Fig 3. Maximum threshold contrast (MTC) measurements for the five observers and four monitor conditions, for (A) dark and (B) bright background values. Average measurements for each group are also shown. Error bars represent one standard deviation. Note the different Y-axis scales for (A) and (B).

Table 1. MTC Mean and Standard Deviation Values Computed from the 50 ft-L Data Collected for Observers A and B Over a Two Month Period

		Time Span of Measurements		
	Observer	3 Days (n = 3)	2 Months (n = 3)	2 Months (n = 6)
Dark Background	A	3.70 ± 0.19	3.13 ± 0.27	3.41 ± 0.37
	B	3.97 ± 0.36	3.61 ± 0.17	3.79 ± 0.32
Bright Background	A	20.85 ± 0.58	20.21 ± 1.06	20.53 ± 0.84
	B	23.77 ± 0.58	21.77 ± 1.47	22.77 ± 1.48

indicates superior display performance for the first monitor condition in a pairing. For example, the negative differences observed when comparing the 30 and 50(1) ft-L conditions for bright backgrounds indicate superior display performance of the 30 ft-L condition. The precision values of the group paired difference MTC measurements are 7.5% and 9.0% for the dark and bright background cases, respectively. The sensitivity values are 9.3% and 11.1% for the dark and bright background cases, respectively. From the 95% confidence intervals shown in Fig 4 for each of the group measurements, it is seen that statistically significant differences are shown for all comparisons made, except for that between the 50(1) and 50(2) data, for which no difference is expected. These group results are also corroborated by the individual observer comparisons.

DISCUSSION

As seen in Fig 2, the plots of threshold contrast versus target size were visibly non-linear, especially for the bright background data. This behavior is likely due to the effects of blur, glare, and structured noise from the monitor phosphor and scan lines. In spite of the non-linear behavior, the intercept of the best-fit fixed-slope line (maximum threshold contrast, or MTC) provides a reasonable measure of overall display quality. More detailed comparisons between devices can be made by examining the data obtained for specific object sizes.

It is evident from Fig 3 that display performance measured using the dark background test images was superior to that measured using the bright background images. This is likely due to several factors. The effects of monitor glare will be increased when displaying bright background images. Also, inspection of the monitor display functions and correlation of this data with the Barten model of human visual system contrast sensitivity[7] predicted that observers would be more sensitive to contrasts displayed against the dark backgrounds as compared to the bright backgrounds for all three monitor maximum luminance conditions examined. Since it is desirable to sensitize our measurements to the effects of glare, the bright background measurements are better indicators of overall display performance. It was also noted that the dark background MTC data are less precise, less sensitive, and less stable over time than the bright background data. In general, the bright background MTC measurements are more useful than the dark background data for overall display quality assessment. Future applications of this technique for overall display quality assessment will involve test image sets with a single background value. This background level will be chosen to approximate the average pixel value of a set of representative clinical images, and to produce a reasonable number of experimental data points. A background corresponding to a video level of 50-70% should generate 15-20 data points per experiment, and should result in viewing sessions of 30 minutes or less.

The MTC measurements for individual observers presented in Fig 3 also slows a range of absolute performance between observers. In spite of this range, when group measurements were obtained by pairing measurements from each observer for two conditions being compared, the group statistic (with n = 5 observers) is predicted to be sensitive to changes in MTC of 11.1%. Including more observers should increase the sensitivity of the technique even more. As shown in Figure 4, the validity of the paired group MTC measurements is supported by the fact that all individual measurements agree with the group preference for one condition over the other, with the exception of the cases where the group MTC difference was not statistically significant.

To compare display performance under the 30, 50, and 70 ft-L maximum luminance conditions, our experiments were designed to mimic conditions in a soft-copy diagnostic reading room in routine use on our campus.[5] The ambient room lighting and monitor vendor and model were identical. The reading room CRT displays are set up for 70 ft-L maximum luminance, which matches one of our experimental monitor conditions. Experience over the past two years has provided a high degree of confidence in the acceptability of these soft-copy reading conditions. The bright background data is

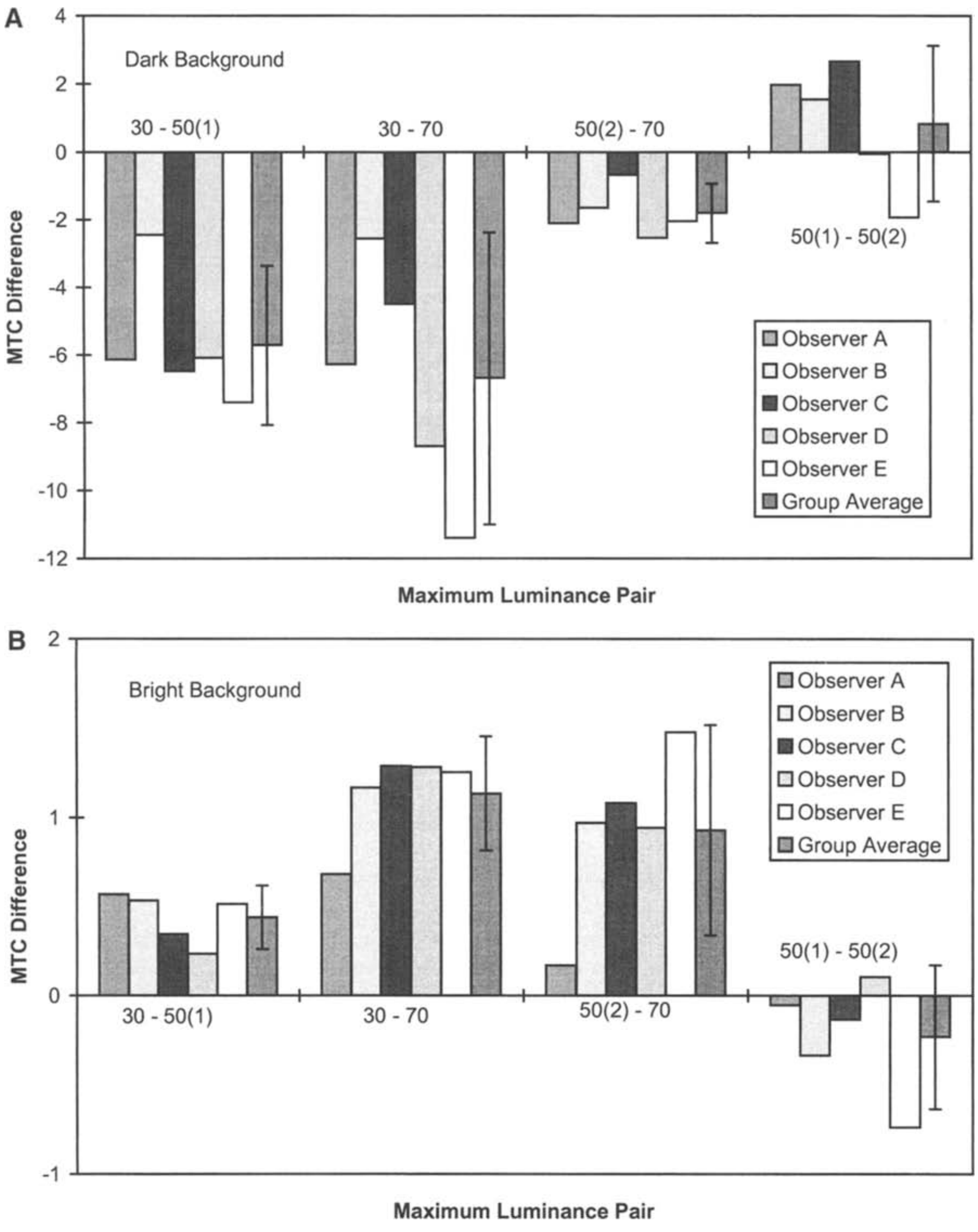

Fig 4. MTC difference measurements for the five observers and combinations of the four monitor conditions, for (A) dark and (B) bright background values. Average difference measurements for each group are also shown. Error bars represent 95% confidence intervals. Note the different Y-axis scales for (A) and (B). Statistically significant differences are indicated by the lack of overlap between the 95% confidence interval and the X-axis.

expected to be most useful for assessing overall display performance for reasons discussed above. Fig 4B indicates that lower luminance levels produce better display quality, as measured by each individual, as well as the group. The difference between the 50 and 70 ft-L measurements is small compared to that between 30 and 50 ft-L. It appears that 50 ft-L would be a reasonable operating point for diagnostic displays in areas with dim, well controlled room lighting. Although 30 ft-L may provide better performance (as indicated by the MTC measurements), sensitivity to even small changes in ambient room lighting and viewer fatigue are concerns at this low absolute luminance. These measurements provide some assurance that diagnostic tasks which must be performed under very low maximum luminance conditions (eg, ultrasound workstations with color CRT displays) do not necessarily suffer from degradation due solely to the low luminance levels, as long as room lighting is carefully controlled. Additional MTC measurements at various room lighting levels are necessary to draw further conclusions.

CONCLUSIONS

This study indicates that contrast-detail data should be very helpful in providing quantitative measurements of overall electronic display quality. The method would be suitable for new equipment selection, acceptance testing, and quality control. The recommended protocol would only involve observer data obtained using test images with mid-range background pixel values. Improvements to the current linear curve fit may also provide increased levels of measurement precision and sensitivity. To put the measurements in proper context, MTC measurements of a group of displays currently in use and deemed acceptable for the display task in question (e.g. primary diagnosis or clinical display) should be obtained by a group of observers, if possible.

When making quantitative recommendations regarding equipment selection, or display configuration (eg, maximum display luminance or ambient room lighting levels), a group of observers should be used, since the decisions made will presumably affect a large number of radiologists, technologists or clinical physicians using the display workstations. With a group of five observers, and using the group paired difference analysis technique, measurement precision will be 9.0%, and sensitivity to MTC changes will be 11.1%. Each set of raw data for a measurement of MTC can be collected and analyzed for each observer in approximately 30 minutes, so data sufficient for a comparison of two devices could be collected and analyzed within an hour.

When making measurements for equipment acceptance testing or routine QC measurements (eg, on a quarterly or twice-yearly basis), measurements from a single observer should suffice since the goal is an assessment of the relative performance of an individual device. Precision of the single observer MTC measurements will be 6.8%, and sensitivity will be 15.2%. Measurements made over a period of time should have a reproducibility of about 5%.

REFERENCES

1. Cohen G, DiBianca FA: The use of contrast-detail-dose evaluation of image quality in a computed tomographic scanner. J Comput Assist Tomogr 3:189-195, 1979

2. Constable RT, Henkelman RM: Contrast, resolution and detectability in MR imaging. J Comput Assist Tomogr 15:297-303, 1991

3. Dobbins JT, Rice JJ, Beam CA, Ravin CE: Threshold perception performance with computed and screen-film radiography: implications for chest radiography. Radiology 183:179-187, 1992

4. Krupinski EA, Roehrig H, Yu T: Observer performance comparison of digital radiographic systems for stereotactic breast needle biopsy. Acta Radiol 2:116-122

5. Hangiandreou NJ, King BF, Swensen AR, Webbles WW, Jorgenson LL: Picture archive and communication system implementation in a community medicine practice. J Dig Imaging 10 (suppl 1):36-37, 1997

6. Rose A: The sensitivity performance of the human eye on an absolute scale. J Soc Opt Am 38:196-208, 1948

7. Blume H, Members of ACR/NEMA Working Group XI: The ACR/NEMA proposal for a grey-scale display function standard. Proc SPIE 2702:344-360, 1996