A Performance Metric for Evaluating Conformance of Medical Image Displays with the Proposed ACR/NEMA Display Function Standard

Bradley M. Hemminger*, Keith Muller#

*Department of Radiology, University of North Carolina,
#Department of BioStatistics, University of North Carolina,
Chapel Hill NC 27599, bmh@rad.unc.edu

ABSTRACT

ACR/NEMA (now DICOM) has released for public comment the Display Function Standard document. This standard, produced by ACR/NEMA Working Group XI, has been developed in order to solve the problem of standardizing the response of grey scale display systems. This paper presents a methodology proposed in the Display Function Standard for quantitatively calculating the conformance of a display device to the Standard Display Function based on statistical measures, which are referred to as the *linearization uniformity measure* s (LUM). There are two LUM measures, R^2, or global uniformity, and Root Mean Square Error (RMSE), or local uniformity. The derivation of both measures is described. Two additional measures that provide a better description of the achievable dynamic range of a display device than simply its luminance range are also described, the theoretical number of Just Noticeable Differences (JNDS) of the display, and the realized number of JNDS of the display. Currently available medical image display systems are analyzed using each of these measures, to examine their shortcomings, and suggest what changes might be desirable in the design of medical image display systems in the future.

Keywords Image Display, Display Function Standard, Visual Models, Conformance, Display Standards, ACR/NEMA, DICOM

2. INTRODUCTION

Standardization of display systems has assumed an increasingly important role in medical imaging, especially since the acceptance of the DICOM medical imaging communications standards in clinical radiology practice. More and more, physicians are being required to view images on multiple display systems, and not just from one laser hardcopy device dedicated to a specific acquisition device. Radiologists now view films from multiple networked laser printers, high quality diagnostic CRT video displays, lower quality video displays on less expensive personal computers or teleradiology systems, and reflective hardcopy (paper). With this proliferation of viewing environments and mediums has come a cost. Each different display device may produce significantly different luminance distributions given the same digital inputs. As a result, the same image viewed on different display devices may not appear similar, and may cause different judgments because of differences in the reproduction of the images. In order to address this problem, ACR/NEMA (now DICOM) working group XI has written the Display Function Standard[1], which will standardize the response of all display systems. It defines a Standard Display Function to be used by all display devices, to insure similar reproduction of images across different display devices. The Standard Display Function is defined mathematically as a curve (and also in tabular form). It corresponds to the human contrast sensitivity response as predicted by the Barten visual model.[2,6,7,8,9] In order for the standard to work effectively, we must be able to calculate quantitative measures of how well a given display system conforms to the standard. Hemminger[2,3,4] has previously proposed a statistical methodology that allows quantitative comparisons of display systems characteristic curves with the Standard Display Function, as well as defining the theoretical and realized Perceived Dynamic Range (PDR). This paper presents the complete derivation of the improved version of these previously proposed statistical tests, which is the version currently proposed in the Display Function

Standard as a quantitative method for evaluating conformance of display system's to the Standard Display Function. The statistical tests are then used to evaluate current film and video display systems, and suggest improvements in design for future display systems.

3. METHODS

Statistical Measures

In order to compare the Standard Display Function and the display system's approximation to it, a common measure is required. The Standard Display Function is defined in terms of Just Noticeable Differences (JNDs).[10,11,12] A JND is the minimum amount of luminance difference (ΔL) is required to perceive an object on a uniform background luminance of L. If we apply this to the display system, the natural measure would be to define a ΔL with respect to a luminance L, where ΔL is the difference between luminance (L_i) produced by the display system for Digital Driving Level DDL_i and the luminance (L_{i+1}) for DDL_{i+1}. This is depicted in figure 1. Thus, each interval defined by the luminances produced by adjacent DDLs of the display system, produces a JND value, from these values we can plot the Luminance Intervals versus JNDs for the display system (figure 2).

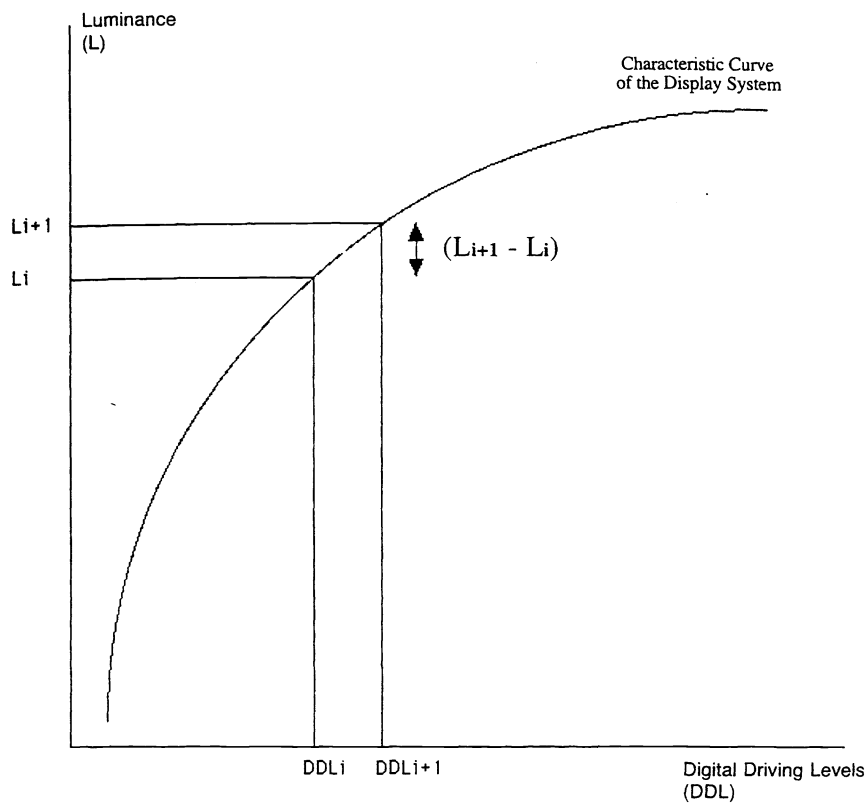

Figure 1. Definition of JND for display systems based on the characteristic curve for the display system. The horizontal axis is the digital driving levels of the display system. The vertical axis shows the luminance produced by the display system at each digital driving level.

Previously Hemminger described a similar methodology using contrast thresholds.[3] He proposed the improvement to this methodology at the end of that paper[3], which we now describe in more complete detail. The revised metric, which plots the number of JNDs (vertical axis) versus $DDL_{interval}$ (horizontal axis) more directly captures the issue of the size of each step of a luminance distribution, without resorting to the remappings of ratios of contrast thresholds used in the previous version.

The Luminance Intervals versus JNDs plot of a standardized system would be a horizontal line, with a horizontal intercept equal to the JND value that is constant across the luminance intervals of the test luminance distribution of the display system. Figure 2 shows the Luminance Intervals versus JNDs plot of a high brightness video display system with 8 bit Digital to Analog Converter (DAC). The 8 bit DAC supports 2^8, or 256 discrete levels of contrast. Two key measures are required to determine conformance: first, does the display system's characteristic curve match the Standard Display Function in a global sense, i.e. is it of similar shape in the correct location; and second, is it a close match at small scale, i.e. does it maintain equal steps sizes in perceived brightness according to the Standard Display Function across the luminance range. The statistical test of multiple linear regression provides an appropriate measure of how similar the Luminance Intervals versus JNDs curve of the display system is to the standard display function in both the large scale and small scale sense. After normal multiple linear regression assumptions are verified for the data[5], two related tests in the regression analysis summarize the results.

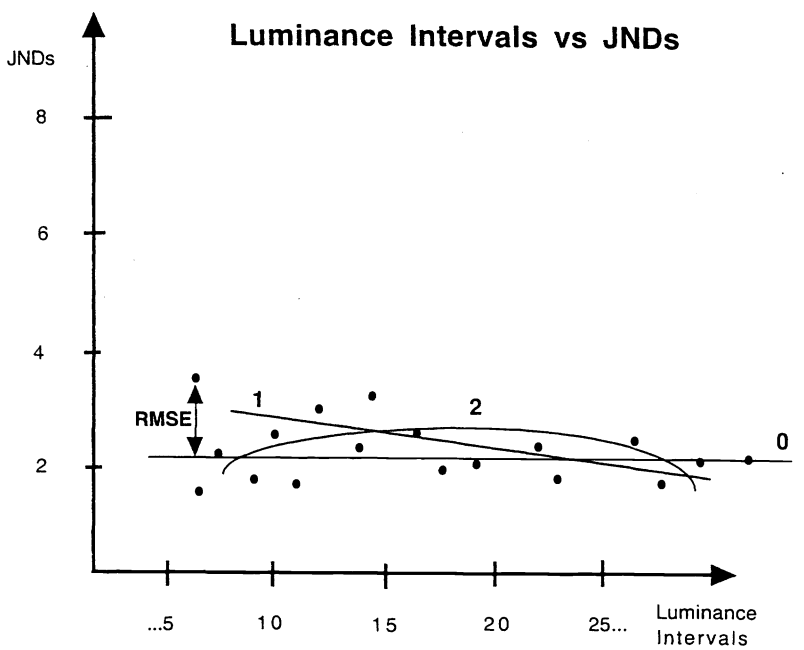

Figure 2. Example JNDs versus Luminance intervals plot for CRT monitor display system. The curves labeled 0, 1, and 2, correspond with the R^2 0 order, 1st order, and 2nd order curves matched to the dataset, respectively. The RMSE arrow shows the distance (error) measured at a single luminance interval. Note that the LUM RMSE is a single value calculated from all of the intervals.

The first test, named the LUM R^2 test, matches the Luminance Intervals versus JNDs curve of the test luminance distribution against different order polynomial fits. In statistical terms this means performing a standard multiple regression statistical analysis to compute the R^2 values of B_0 (intercept only, i.e. a horizontal line with no slope), B_1 (linear), B_2 (quadratic), and B_3 (cubic) fits. The R^2 value of the B_0 fit produces a number from 0.0 (best) to 1.0 (worst) which defines the quality of the standardization of any luminance distribution. The regression analysis should test comparisons through third order curves (cubics). Additionally, the intercept, B_0, indicates the mean number of JND steps per interval, which is the best measure of contrast resolution. If the Luminance Intervals versus JNDs curve of a display system's characteristic curve best matches an intercept only horizontal line with constant JND values, then it is considered to globally match the Display Function Standard. If the curve corresponding to the display system's characteristic curve instead matches any higher order curve rather than the intercept only horizontal line then the distribution does not match the Standard Display Function. The R^2 values for B_1, B_2, B_3 show how much of an improvement is made by fitting the higher order curves, i.e. 1st order, 2nd order, 3rd order, respectively. Thus, if any of the R^2 values for B_1, B_2, B_3 are significantly different from zero then the B_0 fit was not the best, and the display system would not be considered standardized. In the example distribution depicted in figure 2, the distribution is fairly tight, and a 0 order curve (curve labeled 0) might be the best fit. Distributions suggestive of 1st and second order fits are shown as well (labeled 1 and 2, respectively). An example of how the RMSE error is calculated is shown by the distance between the mean of the 0 order curve and the first sample point. Calculating the RMSE across all the points would yield the LUM RMSE value.

The second test, the LUM *RMSE*, analyzes whether the size of luminance steps are uniform in perceptual size (i.e. JNDs) across the luminance range. This is measured by the Root Mean Square Error (RMSE) of the curve fit of an intercept only model to the Luminance Intervals versus JNDs plot of the test luminance distribution. The RMSE in this case is simply the standard deviation of the Luminance Intervals versus JNDs plot of the test luminance distribution from a plot of its mean JND. Smaller LUM values would indicate better fits. At this time, there is no value scale attached to the LUM values. Until clinical experience can correlate a value scale with the LUM scale, we should be cautious in applying it, except to compare mathematically how close a display systems characteristic curve is to the display function standard curve.

Both the LUM R^2 and LUM RMSE measures can be conveniently calculated with standard statistical packages. To summarize, a luminance distribution is considered standardized when its Luminance Intervals versus JNDs plot is best matched by the zero-order intercept only curve. For a display system meeting the standardization requirement, the quality of the match of the display system to the Standard Display Function is given by the LUM RMSE value.

As described above, if the test luminance distribution passes the LUM R^2 test, then the measure of quality of the distribution is determined by the single quantitative measurement (LUM RMSE) of the standard deviation of the JNDs from their mean. An important related factor is the number of output digitization levels of the display system. For instance, the LUM RMSE could be improved by using fewer output levels, at the cost of decreasing contrast resolution. While the LUM RMSE is influenced by the choice of the number of input digitization levels with respect to the number of output digitization levels in the display system, the appropriate number of input digitization levels is determined by the clinical application, including possible grey scale image processing that may occur independently of the display function standardization. Thus, this standard does not proscribe a certain number of grey levels of input or output digitization levels. However, in general, the larger the number of distinguishable grey levels available, the higher the possible image quality because the contrast resolution is increased. It is recommended that the number of necessary input digitization levels for the standardized distribution be determined prior to standardization of the display system (based on clinical applications of display system), so that this information can be used when calculating the standardized distribution to avoid using distributions with fewer output digitization levels than needed. In general, it is likely that contrast resolution is a more

significant factor in the overall quality of the image presentation, and should not be traded off for small improvements in the LUM RMSE, unless there is clear benefit.

Measures of Quality of Display System

There are four measures of quality of the display system defined in this section. The first two are the result of the LUM R^2 test, which specifies what curve fit is the best matched to the test luminance distribution (intercept only, linear, quadratic, or cubic), and the LUM RMSE test, which signifies the amount of error in matching the correctly standardized intercept only model. In addition to these two tests which describe the accuracy of the standardization of the display system, two further values are produced during the standardization that provide useful measures for comparing display systems. The first is the number of *theoretically achievable JNDs* of the display system. This number can be determined directly from Table 1 in Section 6 of the Display Function Standard, or indirectly via the Barten formula in section 6. The second is the number of *realized JNDs* of the display system. The number of theoretically achievable JNDs is simply the number of JNDs predicted by the visual model given the luminance range of the display device used. The more useful number of realized JNDs, describes how many JNDs are actually realized given the specifics of the display system (i.e. the number of grey levels of contrast resolution and the distribution of luminance values). This number is calculated beginning at the minimum luminance of the display system, and then stepping one JND in luminance from the current luminance value, and choosing the smallest DDL that achieves a step at least that large. Repeating this through all the available DDLs will produce a sequence of steps, all at least 1 JND apart, and the length of this sequence of steps is then the number of realized JNDs of the display system. Shown below is a C program segment for calculating the number of realized JNDs for a display system with characteristic curve DPY, and N digital driving levels.

```
i = 0;
L[i] = DPY[0];
/* DPY[i] is the luminance at DDL[i] of the Display System */
while ((NextLum = L[i] + BartenCT ( L[i] )) < L[n] )
 {
    i = i + 1;
    L[i] = find_minimum_DDL_step(NextLum);
    /* returns value DPY[k], where DPY [k-1] < NextLum <= DPY[k] */
 }
NumberAchievableJNDs = i;
```

Procedure for Measuring Conformance of Display System to Standard

Measuring the luminance values of a standardized display system provides a quantitative measure of the quality of the standardization of that display system, and provides the guarantee of a specific level of conformance to a standard. Sections D.1.1 (emissive displays), D.2.1 (transmissive displays), and D.3.1 (reflective hardcopy) in the Display Function Standard discuss how to makes measurements of the characteristic curve of the display system, and give illustrative examples. Although the measurement techniques differ slightly between display methods, the same conformance technique can be used by all display systems, including emissive (video), transmissive film on lightbox, and reflective hardcopy.

1) Measure the characteristic curve (DDLs versus Luminance) of the display system in its standardized state at all (or as many as is feasible) DDLs of the display system. If only a subset of luminance levels are measured, then normalize the luminance intervals to single steps size increments, and the JND values accordingly.

2) Use these JND values for each luminance interval as a sample set, and calculate the LUM measures of R^2 and RMSE with the statistical package on the sample set.

4. RESULTS

The results of analyzing both a single 8 bit DAC video display system in-depth, as well as a survey analysis of multiple video and film display systems are presented below. Table 1 shows the calculated theoretical JNDs and realized JNDs for the 8 bit DAC video display system. Table 2 shows the LUM R^2 and LUM RMSE calculations for the 8 bit DAC video display system under four different configurations. These results were generated with the SAS (SAS Inc, Cary NC) statistical software package.

DDL	Luminance (cd/m^2)	JNDs
0	0.667	56
255	354.7	656
Theoretically Achieveable JNDs		600
Realized JNDs		197

Table 1. The theoretical and realized JNDs for a 8 bit DAC video display system.

```
DEFAULT # Input/Output digitization level=256/256
        _IN_       _RSQ_      _RMSE_     INTERCEP

          .          .         1.6376       3.258
          1       0.3948       1.2765       5.445
          2       0.8389       0.6599       3.475
          3       0.9096       0.4953       3.117


STANDARDIZED  # Input/Output digitization level=256/256
        _IN_       _RSQ_      _RMSE_     INTERCEP

          .          .         1.8924       2.926
          1       0.0000       1.8961       2.907
          2       0.0001       1.8998       2.875
          3       0.0001       1.9035       2.863


STANDARDIZED  # Input/Output digitization level=128/256
        _IN_       _RSQ_      _RMSE_     INTERCEP

          .          .         1.4576       5.876
          1       0.0001       1.4633       5.851
          2       0.0001       1.4692       5.824
          3       0.0001       1.4751       5.821


STANDARDIZED  # Input/Output digitization level=256/1024

        _IN_       _RSQ_      _RMSE_     INTERCEP

          .          .         0.6244       3.278
          1       0.0000       0.6257       3.281
          2       0.0001       0.6269       3.296
          3       0.0003       0.6281       3.310
```

Table 2. The LUM R^2 and LUM RMSE results for four different configurations of the same display system. The first configuration (DEFAULT) is for the uncorrected characteristic curve of the display system. The second is with the number of input and output digitizations levels both equal to 256. The third is with the input digitization levels equal to 128, and the output levels equal to 256. The fourth and last example output is for 256 input digitization levels and 1024 output digitization levels.

From table 2 it is clear that the DEFAULT (non-standardized) state of the display system does not conform to the display function standard. This is because the R^2 value is 0.39, which is non-zero and indicates an improved match, for the 1st order curve fit compared to the 0 order curve fit. Similarly, the 2nd order curve fit (.83) is even better than the 1st order curve fit . And the 3rd order curve fit (0.9) is a slight improvement over the second order fit . On the other hand, each of the other three standardized luminance distributions are conform well. The R^2 values are essentially zero, indicating no improvement by higher order curve fits. When we compare the RMSE values for these we find the best result (0.6) when 256 input digitization levels and 1024 output digitization levels (256/1024) are used. The worst RMSE value (1.8) is when we use 256 input digitization levels and 256 output digitization levels (256/256).

Assuming LUM is a good metric for measuring the quality of the standardization, two important conclusions can be drawn from table 2. First, that while standardization with the same number of input digitization levels as output digitization levels meets the global criteria (LUM R^2), it does not significantly improve the LUM RMSE of the system compared to the default values. For instance, although the DEFAULT luminance distribution does not pass the LUM R^2 test, its RMSE value (1.6) is lower than that for the standardized 256/256 case (1.8). This result is expected due to the lack of flexibility in choosing DDLs when the output range is the same as the input range[2]. Second, that when the number of input digitization levels is smaller than the number of output digitization levels to chose from, the standardization improves. Using the 8/8 DAC but limiting our number of output levels to 128 shows an improvement over the 256/256 standardization for each display system. Similarly, the 256/1024 shows a substantial improvement over both the 256/256 and the 128/256 cases. Previously, Hemminger suggested that as a rule of thumb, the larger the number of potential DDL choices versus the number utilized, the better the possible standardization. With the LUM metric defined, we investigated what the optimal choice of input digitization level would be for a given output digitization level. To determine this we calculated the LUM RMSE values for all possible input digitization levels for a given display system (with a fixed output digitization level). We did this for CRTs and film printers with different output digitization levels (256 corresponding to 8 bit DACs, 1024 corresponding to 10 bit DACs, and 4096 corresponding to 12 bit DACs). The result for a single 8 bit CRT is shown in figure 3. The horizontal axis is the number of input digitization levels used, from 1 to all of the levels in the display system. The vertical axis is the LUM RMSE error at that particular number of input digitization levels. What is striking is that the LUM RMSE values have minimum values when the input digitization levels are 1/2, 1/4, 1/8, etc. of the number of output digitization levels. This result also maps to all the other displays tested. Figure 4 shows all the display systems measured, but on normalized axes of output digitization levels and RMSE values, making salient the similarity of response of all the display systems with respect to LUM RMSE scores versus the ratio of input to output digitization levels. All the display systems exhibit the similar behavior with the smallest (best) LUM RMSE scores occurring when the input digitization resolution is 1/2, 1/4, or 1/8 of the output digitization resolution. It now appears that while having the number of output digitization levels be larger than the number of input digitization levels is important, the main criteria is that the number of output digitization levels be a multiple of the number of input digitization levels.

The trade off of better LUM RMSE values is increasing the cost of the display system (more DAC bits to achieve more output digitization levels) or choosing instead to decrease the number of input digitization levels, which would result in a loss of contrast resolution. Thus, as discussed earlier, it is important to know the contrast resolution requirements of clinical applications. For instance, we could achieve a LUM RMSE of 1.8 for a 256/256 standardization, while we could achieve a LUM RMSE of 1.4 with a 128/256 mapping. Many clinical applications, however, would see a loss of contrast information if only 128 distinct input digitization levels were utilized. The third important point, is that with an 8/10 DAC, and the 256/1024 standardization, the LUM values were significantly improved in all cases, while maintaining the

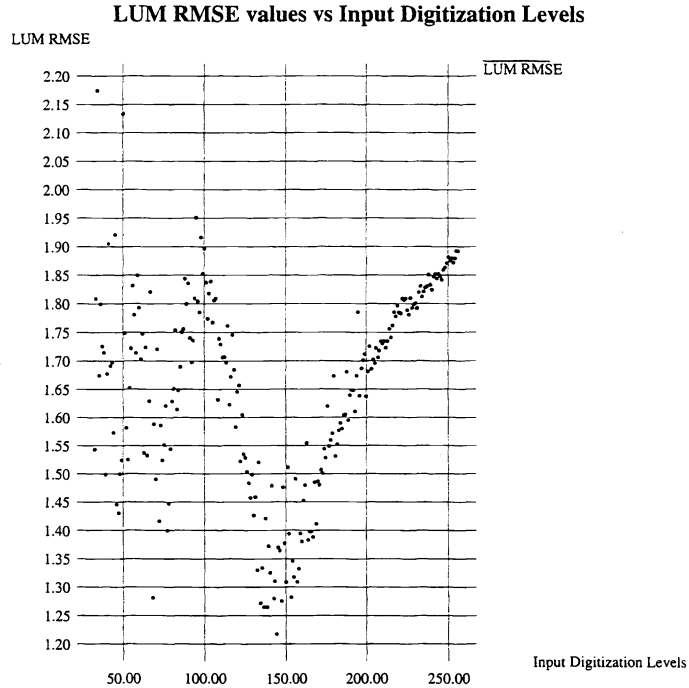

Figure 3. Plot of LUM RMSE values for 8 bit DAC video display system, for all possible input digitization levels.

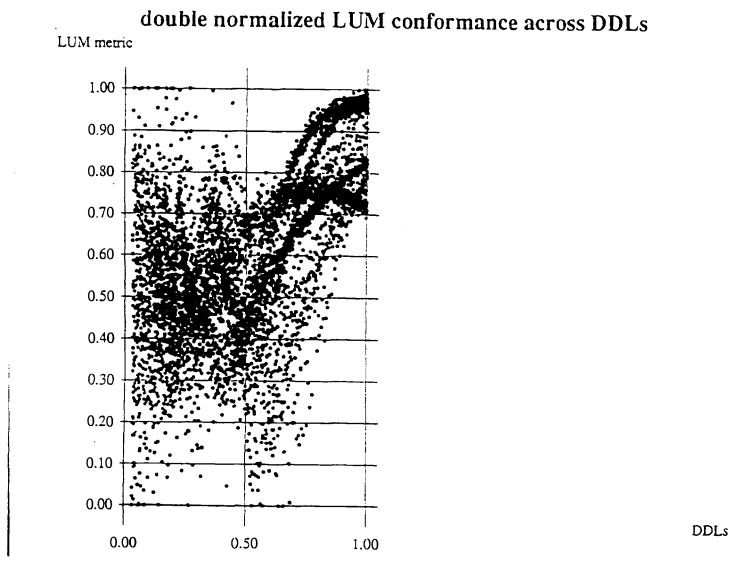

Figure 4. Plot of LUM RMSE values for multiple film and video display systems, for all possible input digitization levels of each system. Because of the differences in output digitization resolutions (from 256 levels to 4096 levels) due to different DAC resolutions both axes have been normalized (0 - 1) so that trends are more salient.

full 256 output levels. Thus 8/10 or 10/10 DACs are superior to the 8/8 DACs in achieving a standardized display system while maintaining the full contrast resolution (256 levels) of the display system. Note, however, that the contrast thresholds achieved by the 8/10 DAC are still coarser than the visual capabilities of the human observer. This supports previous arguments[2,6,7,13] that the appropriate DAC resolution is in the range of 10-12 bits, and suggests the next step would be for manufacturers to produce a 10/12 bit DAC, which may turn out to be sufficient for medical imaging display purposes. In the future, we plan to conduct formal observer studies to evaluate the effect of the LUM metric with respect to observer performance and observer preference.

5. DISCUSSION

A metric, the LUM, is presented for mathematically calculating conformance with the ACR/NEMA working group XI Display Function Standard. These methods have been applied to high brightness CRT display systems, standard workstation CRT displays, laser printed film on lightbox displays, and reflective hardcopy displays. Preliminary analysis suggests that display systems with output digitization resolutions larger than input digitization resolutions by a factor of 2 or 4 may be the most ideal choices for standardized displays.

6. ACKNOWLEDGMENTS

Hartwig Blume provided important comments and feedback regarding this work, in addition to proposing some measures of his own. Gene Johnston and Robert Hemminger provided financial assistance that allowed me to attend the Standards committee meetings. This work was supported in part by NIH grants P01-CA47982, R01-CA60193, and R01-CA44060.

7. FOLLOW-UP WORK

Correlate clinical outcomes with the LUM measures, especially RMSE, to facilitate comparison of display systems, and display system configurations using the LUM metric.

8. REFERENCES

1. ACR/NEMA Display Function Standard, Public Comment Draft version 1.2, Working Group XI, Hartwig Blume chairman.

2. Hemminger BM, Johnston RE, Rolland JP, Muller K, "Perceptual Linearization of video display monitors for medical image presentation", Proceedings of Medical Imaging 1994: Image Capture, Formatting, and Display, eds Kim Y, SPIE vol 2264, 1994.

3. Hemminger BM, Blume HR, "Are Medical Image Display Systems Perceptually Optimal: Measurements before and after Perceptual Linearization", SPIE Medical Imaging Vol 2707-58, Feb 1996.

4. Hemminger BM, "Minimum Perceptual Error Calculation for Perceptual Linearization", UNC Chapel Hill Dept of Computer Science Technical Report, TR94-032, 1994.

5. Kleinbaum DG, Kupper LL, Muller KE, Applied Regression Analysis and Other Multivariable Methods, Duxbury Press, 2nd Edition, pp45-49, 1987.

6. Blume H, Roehrig H, Browne M, Ji TL, "Comparison of the Physical Performance of High Resolution CRT Displays and Films Recorded by Laser Image Printers and Display on Light-Boxes and the Need for a Display Standard", Proceedings of Medical Imaging: Image Capture, Formatting and Display, SPIE volume 1232, pp97-114, 1990.

7. Blume H, Daly S, Muka E, "Presentation of Medical Images on CRT displays A Renewed Proposal for a Display standard", Proceedings of Medical Imaging: Image Capture, Formatting and Display, SPIE volume 1897, pp215-231, 1993.

8. Barten PGJ, "Physical Model for the Contrast Sensitivity of the Human Eye", Proceedings of Human Vision, Visual Processing, and Digital Display III, SPIE vol 1666, 1992.

9. Barten P, "Spatio-Temporal Model for the Contrast Sensitivity of the Human Eye and its Temporal Aspects", Proceedings of Human Vision, Visual Processing, and Digital Display IV, SPIE Vol 1913, pp2-14, 1993.

10. Pizer, SM, Chan, FH, "Evaluation of the number of discernible levels produced by a display", *Information Processing in Medical Imaging*, R DiPaola and E. Kahn, Eds., Editions INSERM, Paris, pp. 561-580, 1980.

11. Pizer, SM, "Intensity mapping to linearize display devices", *Computer Graphics and Image Processing*, 17, pp 262-268, 1981.

12. Wyszecki, G, Stiles, WS, Color Science: Concepts and Methods, Quantitative Data and Formulae, pp. 567-570, 2nd Edition, John Wiley and Sons, New York, 1982.

13. Sezan MI, Yip KL, Daly SJ, "Uniform Perceptual Quantization: Applications to Digital Radiolography", IEEE Transactions on Man, Machine, and Cybernetics, Vol SMC-17, No 4, pp622-634, 1987.