AD_____

Award Number: DAMD17-99-1-9347

TITLE: Quantitative Image Quality Analysis of Soft Copy Displays

PRINCIPAL INVESTIGATOR: Dev P. Chakraborty, Ph.D.

CONTRACTING ORGANIZATION: University of Pennsylvania
Philadelphia, Pennsylvania 19104-3246

REPORT DATE: September 2001

TYPE OF REPORT: Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	September 2001	Annual (1 Sep 00 - 31 Aug 01)

4. TITLE AND SUBTITLE
Quantitative Image Quality Analysis of Soft Copy Displays

5. FUNDING NUMBERS
DAMD17-99-1-9347

6. AUTHOR(S)
Dev P. Chakraborty, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

University of Pennsylvania
Philadelphia, Pennsylvania 19104-3246

E-Mail: chakrabo@rad.upenn.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES
Report contains color

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. Abstract ***(Maximum 200 Words)*** *(abstract should contain no proprietary or confidential information)*

This is a report on work performed in the second year of the project "Quantitative Image Quality Analysis of Soft Copy Displays". The purpose is to make physical image quality measurements on two video display monitors such as are used in digital mammography. There is great interest in optimizing displays for medical imaging, as it is known that displays are the weak link in image quality. In this period we successfully applied the CAMPI (Computer Analysis of Mammography Phantom Images) method to the evaluation of phosphor type of two new monitors. In preliminary results we showed that the P-45 phosphor has advantages over the newly proposed P-104 phosphor. These measurements are a first, and when they are confirmed, will represent a significant contribution to the state of our knowledge regarding soft copy displays.

14. SUBJECT TERMS
soft copy displays, measurements, signal-to-noise-ratio, CAMPI

15. NUMBER OF PAGES
53

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

Abbreviations: CCD = charge coupled device; CRT = cathode ray tube; CAMPI = Computer Analysis of Mammography Phantom Images, MTF = Modulation Transfer Function, NPS = Noise Power Spectra, CC = Cross Correlation, SNR = Signal to Noise Ratio.

INTRODUCTION

This is a report on work performed in the second year of the project "Quantitative Image Quality Analysis of Soft Copy Displays". This project has been granted a 1 year no cost extension and the new end date is 9/30/2002.

The scope of this project is soft copy optimization. The purpose is to make physical image quality measurements on two video display monitors such as are used in digital mammography. Phantom images will be displayed on the monitors, and images acquired using a state-of-the-art CCD (charge-coupled device) camera. The images will be analyzed using the PI's CAMPI (Computer Analysis of Mammography Phantom Images) methodology.

PROGRESS REPORT

The original approved tasks are summarized below.

1. Optimize the CCD method (optimal magnification, effect of non-uniformity etc.), these are detailed starting on page 11.
2. CAMPI modifications: This task is completed. Programs were written in IDL to (a) create test patterns and (b) analyze the CCD images of these test patterns. These are described in detail starting on page 6.
3. Image acquisitions and analysis: The need for varying the contrast transfer function and LUT was eliminated, as explained in the year 1 report. Preliminary experiments were conducted to test the programs. Finally a large study involving creating 414 test patterns and analyzing their CCD images was performed. These are detailed starting on page 6.
4. Bridging calculations: this will be done in the next period. See also page 10.
5. Quantification of image quality degradation due to the CCD camera: these are detailed starting on page 11.

Modifications to the original plan are described on page 31.

DETAILED REPORT ON TASKS 2 and 3

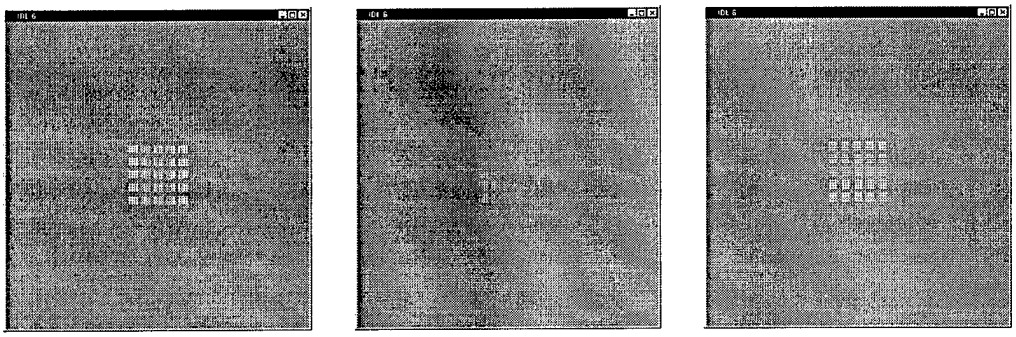
Experiment to evaluate effect of phosphor type

Aim: The purpose of this experiment was to evaluate the effect of two phosphors that are available in medical imaging monitors. These are P-45 (which is widely used) and P-104, which has been proposed as a viable alternative. The multi-component P-104 phosphor can deliver greater luminances at the cost of increased noise. There is currently great interest in image quality measurements comparing these phosphors.

Monitors Studied: Two 5-megapixel (2048 x 2560 x 8 bits) monitors manufactured by Siemens were studied. These monitors were identical in all respects except Monitor A had a P-45 phosphor, Monitor A had a P-104 phosphor. Each was calibrated according to the DICOM standard. The monitors were driven by DOME hardware.

Test Images: The following 512 x 512 images were sent to U of A: (1) 15 blank images at display driving level (DDL) values of 55, 127 and 200, where 255 represents peak white, which were used for noise measurements. (2) Three test patterns consisting of a periodic array of *phantom elements*, dots, horizontal lines or vertical lines arranged in a periodic array.

Figure 1: Shown from left to right are the three test patterns used in this work: shown are 5 x 5 arrays of vertical lines, dots, and horizontal lines. The spacing parameter (see text) is 4 pixels, and the contrast is 128 DDLs on a background of 127 DDLs (the contrast is artificially inflated to optimize the printed appearance of these images). Each sub-pattern is 13 x 13 pixels. Each image is 512 x 512, and was imaged by the CCD at 8-fold magnification after centering on the monitors.



Each test pattern was characterized by a *spacing* parameter, representing the space in pixels between neighboring elements. Corresponding to each background value six (6) different contrast levels (10, 15, 20, 25, 30, 35) of each test patterns were produced. Three spacing values were used in each case, corresponding to 2, 3 and 4 pixels. The target containing patterns are summarized in Table 1. In all 207 images (45 background and 162 containing targets) were sent to the U of A team. Each image was acquired at a magnification factor of 8 on the respective monitors and 414 (2 x 207) images, each 1316 x 1036, 14 bits per pixel, were returned on CD-ROMS to the U of Penn for analysis. Since the images were of different brightness values, to satisfactorily capture the information the CCD exposure time was varied. Separate calibration data (CCD digitizer value Vs. DDL) was acquired for each exposure time. This was used in the linearizing step, see below.

Table 1: Parameters of the target containing images. Legend: VL = Vertical lines, HL = Horizontal Lines, MC = microcalcification, CNT = contrast, BKG = background, Type = type of target, SPC = spacing. In addition 45 blank images were produced for the three background values.

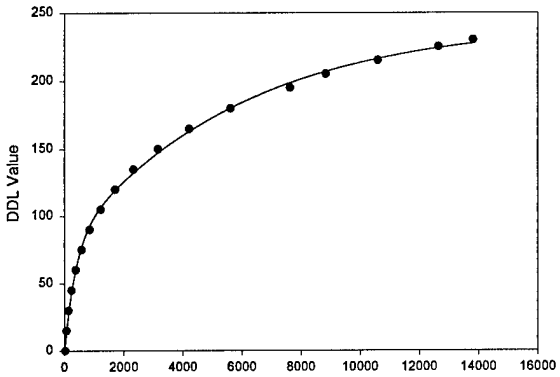
CNT	BKG	TYPE	SPC	CNT	BKG	TYPE	SPC	CNT	BKG	TYPE	SPC	CNT	BKG	TYPE	SPC
35	200	VL	4	35	55	MC	4	35	55	VL	3	35	55	HL	4
30	200	VL	4	30	55	MC	4	30	55	VL	3	30	55	HL	4
25	200	VL	4	25	55	MC	4	25	55	VL	3	25	55	HL	4
20	200	VL	4	20	55	MC	4	20	55	VL	3	20	55	HL	4
15	200	VL	4	15	55	MC	4	15	55	VL	3	15	55	HL	4
10	200	VL	4	10	55	MC	4	10	55	VL	3	10	55	HL	4
35	200	HL	4	35	200	VL	3	35	55	HL	3	35	127	HL	2
30	200	HL	4	30	200	VL	3	30	55	HL	3	30	127	HL	2
25	200	HL	4	25	200	VL	3	25	55	HL	3	25	127	HL	2
20	200	HL	4	20	200	VL	3	20	55	HL	3	20	127	HL	2
15	200	HL	4	15	200	VL	3	15	55	HL	3	15	127	HL	2
10	200	HL	4	10	200	VL	3	10	55	HL	3	10	127	HL	2
35	200	MC	4	35	200	HL	3	35	55	MC	3	35	127	MC	2
30	200	MC	4	30	200	HL	3	30	55	MC	3	30	127	MC	2
25	200	MC	4	25	200	HL	3	25	55	MC	3	25	127	MC	2
20	200	MC	4	20	200	HL	3	20	55	MC	3	20	127	MC	2
15	200	MC	4	15	200	HL	3	15	55	MC	3	15	127	MC	2
10	200	MC	4	10	200	HL	3	10	55	MC	3	10	127	MC	2
35	127	VL	4	35	200	MC	3	35	200	VL	2	35	55	VL	2
30	127	VL	4	30	200	MC	3	30	200	VL	2	30	55	VL	2
25	127	VL	4	25	200	MC	3	25	200	VL	2	25	55	VL	2
20	127	VL	4	20	200	MC	3	20	200	VL	2	20	55	VL	2
15	127	VL	4	15	200	MC	3	15	200	VL	2	15	55	VL	2
10	127	VL	4	10	200	MC	3	10	200	VL	2	10	55	VL	2
35	127	HL	4	35	127	VL	3	35	200	HL	2	35	55	HL	2
30	127	HL	4	30	127	VL	3	30	200	HL	2	30	55	HL	2
25	127	HL	4	25	127	VL	3	25	200	HL	2	25	55	HL	2
20	127	HL	4	20	127	VL	3	20	200	HL	2	20	55	HL	2
15	127	HL	4	15	127	VL	3	15	200	HL	2	15	55	HL	2
10	127	HL	4	10	127	VL	3	10	200	HL	2	10	55	HL	2
35	127	MC	4	35	127	HL	3	35	200	MC	2	35	55	MC	2
30	127	MC	4	30	127	HL	3	30	200	MC	2	30	55	MC	2
25	127	MC	4	25	127	HL	3	25	200	MC	2	25	55	MC	2
20	127	MC	4	20	127	HL	3	20	200	MC	2	20	55	MC	2
15	127	MC	4	15	127	HL	3	15	200	MC	2	15	55	MC	2
10	127	MC	4	10	127	HL	3	10	200	MC	2	10	55	MC	2
35	55	VL	4	35	127	MC	3	35	127	VL	2				
30	55	VL	4	30	127	MC	3	30	127	VL	2				
25	55	VL	4	25	127	MC	3	25	127	VL	2				
20	55	VL	4	20	127	MC	3	20	127	VL	2				
15	55	VL	4	15	127	MC	3	15	127	VL	2				
10	55	VL	4	10	127	MC	3	10	127	VL	2				

Analysis

The analysis consisted of the following steps:

1. **Linearizing Step**: the CCD image pixel values were linearized using fitting functions developed under SigmaStat (available from SPSS Software). A typical fit and the calibration data points are shown below. The fitting function in this case was f=y0+a*(1-exp(-b*x))+c*(1-exp(-d*x)), where **x** is the CCD value, and **f** is the DDL value. The remaining parameters are constants found by the non-linear fitting routine.

Figure 2: Typical calibration data. The curve is the fit to these points used to linearize the CCD data.



2. **Magnification Correction**: each test pattern (e.g., one of the 25 regions shown in Figure 1) image was magnified by a factor of 8, to compensate for the optical magnification used at the U of A. [The images returned by the U of A were much larger, and the field of view was smaller, due to this magnification difference. In a separate experiment where digital magnification was varied around 8, to maximize the correlation match, the factor of 8 was determined to be correct to within the experimental error.]

3. **Location identification**: for each CCD image, the location of the center of the top-left pattern in Figure 1 was identified manually. Due to the periodicity this determined all the other centers. Since the monitor is not moved relative to the camera during image acquisition, this need be done once only, and the analysis program was highly automated. Also the orientation angle of the acquistion matrix was practically zero degrees (the rows of the test pattern was perfectly aligned with the raster lines) so a ROI rotation in step 4 below was unnecessary.

4. **Location fine adjustment**: the location of the center of the top-left pattern on the CCD image was varied by small amounts in x and y, by applying computerized shifting using bilinear interpolation (in Year 1 we determined that this was adequate). The cross-correlation (CC) was calculated between the shifted pixel values and the true pixel values (the test pattern). The CC was maximized to fine-tune the initial manual x and y locations and determine their optimal values.

5. **Peak CC value**: The maximum cross-correlation (CC) value was computed.
6. **Repeat for 25 replicates**: The steps 4 and 5 were repeated for each of the 25 replications of the basic test pattern, see Figure 1. This yielded 25 numbers, representing the peak CC value for each pattern replica. These values were averaged to yield the *signal* quantity **S**.

Figure 4: Plot of contrast normalized SNR for the two monitors studied. The correlation seen is significant, in spite of the measurement noise. Hence the paired t-test.

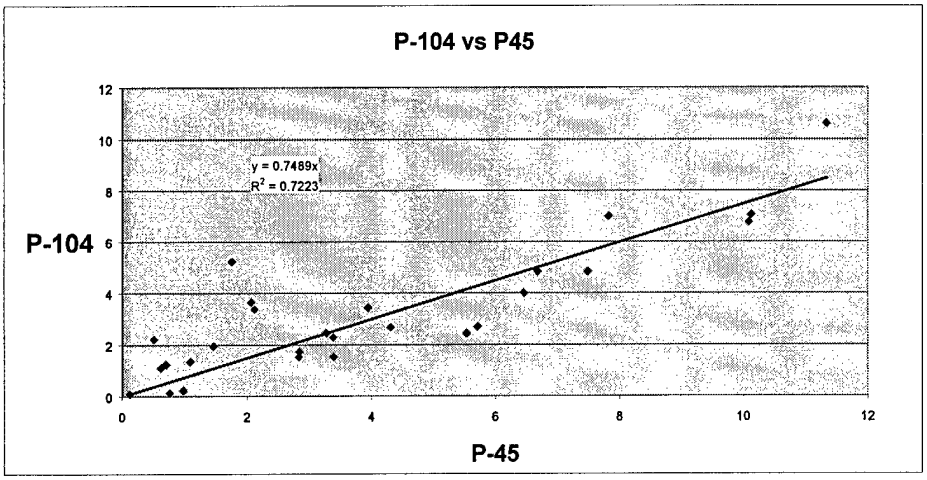


Table 2: Final results from the analysis. The quantity **m** represents the contrast normalized SNR. The other headings are as in Table 1. avg = average, std = standard deviation, cv = std/avg.

TYP	BKG	SPC	m(P-45)	m(P-104)
MC	55	2	3.94	3.44
MC	127	2	3.39	2.30
MC	200	2	0.51	2.22
MC	55	3	5.53	2.44
MC	127	3	3.39	1.54
MC	200	3	0.70	1.25
MC	55	4	4.31	2.68
MC	127	4	2.84	1.74
MC	200	4	1.09	1.35
HL	55	2	10.08	6.78
HL	127	2	7.49	4.84
HL	200	2	2.07	3.65
HL	55	3	11.33	10.61
HL	127	3	7.82	7.00
HL	200	3	1.76	5.23
HL	55	4	10.12	7.05
HL	127	4	6.67	4.85
HL	200	4	2.12	3.38
VL	55	2	0.98	0.23
VL	127	2	0.76	0.14
VL	200	2	0.12	0.10
VL	55	3	5.71	2.70
VL	127	3	2.83	1.54
VL	200	3	0.62	1.12
VL	55	4	6.46	4.00
VL	127	4	3.27	2.46
VL	200	4	1.46	1.96
		avg	3.98	3.21
		std	3.25	2.48
		cv	0.82	0.77
		p-value	0.0243	(paired t-test)

Meaning of the measurements and further avenues for research: our measurements seem to indicate that the P-45 phosphor is better than the P-104 phosphor. It yielded about 33% greater SNR than P-104 for our test patterns. This number represents an average over 3 target types, 3 background values and 6 contrast values. We have no way presently of comparing our results to others, as the latter do not exist to the best of our knowledge. [Note added: Based on an alternate analysis of the same data the differences are insignificant at the 5% level. This point will be explored further.]

There are finer details evident in the results shown in Table 2. For example, it shows that (a) the normalized SNR falls as luminance goes up, presumably due to increasing noise and decreasing subject contrast; (b) the normalized SNR decreases as spacing decreased, due to increased MTF degradation. We will also ask Drs. Robert Wagner, Robert Gagne and Robert Jennings at the FDA to review our results, as they have considerable experience in this area.

This data is relatively recent and we will spend the next few months verifying it and writing a manuscript describing this work. We are reluctant to rush this publication as it could influence the development of technology, and we need to be very certain. Also Prof. Roehrig feels that the P-104 sample monitor we obtained was sub-optimal, and is negotiating with Siemens for another monitor. When this is obtained we will repeat some of our measurements.

Among the other checks we are contemplating are (1) an observer performance test, perhaps using the DCT method of the PI (included as Appendix 1); (2) inclusion of another target type (the disk object in particular is needed to simulate the nodules); (3) correlation of the data with the physical measurements conducted at the U of A; (4) use of a smaller spacing value of 1 pixel. [Note: DCT = Degradation Comparison Threshold]

Task 4: Bridging calculations. This is the main remaining task for the next period. Some progress on this is reported in Appendix 2. We will make calculations of the *expected* SNR for the test patterns employed above. These will employ the known MTF's and NPS curves of the monitors, as measured at the Univ. of Arizona. We will calculate the expected values by numerical simulations. For example, a digital test pattern will be magnified by a factor of 8, degraded by the MTFs and NPS functions, sampled by the CCD matrix, and the cross correlation values calculated. These lead directly to the SNR (non-pre-whitening numerical observer) estimate as noted previously. These will be averaged over many noise realizations, and in this way we can systematically follow the expected variations and compare to those observed. Note that this calculation would be much more difficult with ACR phantom films as originally proposed, as the noise and MTF degradation of the x-ray system, the film-screen system, etc., would also have to be modeled. Uncertainties in these other values would make it more difficult to make precise estimates of monitor degradation and detract from the aims of this project.

DETAILED REPORT ON TASKS 1 and 5

(**NOTE**: Figs. 5 through 10 are intentionally missing. This made it easier to include the Arizona Report, as we did not have computer versions of most of their figures.)

These tasks are optimizing CCD Imaging Techniques and Quantifying degradations introduced by the CCD Camera. The approach was as follows:

1. We took images at different distances between the CCD camera and the CRT.

2. We developed and improved software for deriving CRT performance characteristics such as Modulation Transfer Functions (MTF) and Noise Power Spectra (NPS).

3. We experimented with various stimuli like "squarewave patterns", "lines" and "white noise", needed for finding MTFs.

4. We acquired instrumentation such as an x-y-z Translation Stage, an optical table and a lightbox.

5. We experimented with various methods to generate "flatfields", images necessary to correct for CCD and lens non-uniformities.

The accomplishments are summarized as follows.

1. Taking images at different distances between the CCD camera and the CRT

When taking images at different distances, this distance was quantified in terms of effective linear magnification, i.e., the ratio of linear CRT pixel dimension to linear CCD pixel dimension. A value of Mag = 8 means that the linear dimension of a CCD pixel, when projected onto the CRT, is 1/8 of the linear dimension of a CRT pixel, or conversely, when the CRT pixel is projected onto the CCD, its linear dimension is 8 times the linear dimension of a CCD pixel. Consequently a total of 8 x 8 = 64 CCD pixels will cover one CRT pixel. Magnification values investigated were Mag = 4.5, Mag = 6, Mag = 8 and Mag = 10.

Fig. 11 shows MTFs of the CCD camera for three of these four values of magnification (Mag = 4.5, Mag = 6, and Mag = 10) plus an extra one, namely Mag = 5.1. Notice that the MTFs improve noticeably from Mag = 4.5 to Mag =10 with a cut-off at about 30 lp/mm. Marked in the graph is also the Nyquist frequency of a typical high performance CRT, which is about 3.5 lp/mm. The MTF of the CCD camera for this CRT Nyquist frequency is between 85 and 90 percent. It is clear that even for a Mag = 4.5 and particularly a Mag = 8, most commonly used in our lab, the influence of the CCD camera on the MTFs of CRTs measured with this CCD camera is very small. Nevertheless corrections of the measured CRT MTFs are in order.

Fig. 12 shows MTFs of a high performance Siemens monitor with a P104 phosphor at a low luminance level (input command level only 113 ADU on an 8 bit scale) in the horizontal direction for all four magnification levels and the vertical direction for the two magnification levels Mag+8 and Mag=10. Notice that there don't seem to be major differences in the MTFs for different magnification values. We suspected this from inspection of Fig 11, the MTFs of the CCD camera for

different magnification values. While the MTFs of the CCD camera depend strongly on the magnification value, particularly at their cut-off frequencies from 18 to 35 lp/mm, they do not vary significantly at 3.5 lp/mm, which is the Nyquist frequency of most high performance CRTs.

Incidentally, the MTFs shown in this Fig. 12 were evaluated with the computer program, otherwise it would have been impossible to do this part of the study in a timely fashion.

We also made measurements of the CRT noise at the four magnification levels Mag = 4.5, Mag = 6, Mag = 8 and Mag = 10 by taking images of uniform fields of the CRT. The CCD images then were subjected to a Fourier Transform Techniques resulting in the Noise Power Spectrum. It is important to remember that the noise of a CRT is mainly determined by the granularity of the phosphor layer, and temporal noise rarely plays a role.

Fig. 13 shows the four power spectra for the case that the luminance of the uniform area was determined by a command level of 127 ADU. As expected, the different magnification ratios result in different Nyquist frequencies for the respective power spectra, which is what would be expected. However at this time, it is not clear, why the magnitudes are so different and more analysis is appropriate.

2. Development and improvement of software.

One of the big problems of CRT evaluation, particularly with respect to finding the MTF, is the fact that the CRT is a non-linear device, that means the input stimuli need to be small in amplitude in order for the non-linear CRT to act as a linear device.

An additional problem in the evaluation was the actual evaluation, i.e. the analysis of the images with the various stimuli. This was always done manually. As a result the evaluation took always a long time. As an example, finding the MTF as derived from the squarewave response would take a skilled evaluator 3 days.

We therefore developed a computer program which would do the evaluation of the images automatically. Needless to say this was a major effort, which actually is not completed and needs continued refinements.

The program is described briefly in the following.

Operation:

The input to the program is a file of type ***.img** (with pixels of depth 14 bits and a 10 byte header. The header stores the size(rows and columns) of the image and also if the pixels are arranged in the little "ENDIAN" fashion). The image should be a square wave pattern(a one dimensional profile across the image should approximately be a square wave). The path and the name of the file to be analyzed is written into the file "**PROFILE.TXT**"(in the form "D:\temp\new_siemens\ratio10\rotated\200_10.img") from which the program reads the image and writes the output into a file called "**MTF.TXT**" in the same directory.

When the program is run(the **.exe file executed), it asks for the pixel's size of the CRT. The pixel size on the CRT is** ***1/(2*Nyquist Frequency)*** i.e. the Nyquist frequency of the CRT. It also asks for the number of CCD pixels used for every CRT pixel which basically is the number of pixels with which a CRT pixel is oversampled (the above mentioned magnification value Mag).

The program by default works for images with horizontal bar patterns. To calculate the horizontal MTF from Vertical bar patterns, the image has to be rotated 90 degrees and saved and its path written into "PROFILE .TXT". The program is "Windows 95/98,NT, 2000" compatible.

ALGORITHM:

The value of pixels along five profiles taken across different cross sections of the image is averaged. The edges of the square waves in the profile are detected to ensure integral number of cycles are Fourier Transformed. The Nyquist frequency is then detected and depending on the ratio of the fundamental frequency and the Nyquist Frequency, the harmonics in the Fourier Transform data are picked up and written into MTF.TXT.

Fig. 14 (a) and (b) show MTFs of a high performance Siemens monitor (with a P45 CRT), evaluated from the squarewave response. The MTFs in Fig. 14 (a) were evaluated manually, while the MTFs in Fig. 14 (b) were evaluated with the program. Notice the close agreement of the MTFs.

3. Experiments with various stimuli like "squarewave patterns", "lines" and "white noise", needed for finding MTFs

A major problem in the performance evaluation of CRTs was always the fact that MTFs derived from different stimuli or different methods such as squarewave-response, line-response and broadband-response (resulting from transmission of white noise) usually differed.

Fig. 15 illustrates the 3 kinds of stimuli used for determination of the MTFs. Notice the small signals, identified as Δ ADU, which should allow application of Fourier Transform Techniques for the "quasi linear system"

Concentrating on the factors which caused the differences, we found for instance that even small differences in the mean value of the luminance would cause differences in the MTFs derived from the squarewave response and from the line response. Another factor are the small signals, identified as Δ ADU, which would permit to consider linearity for the CRT. Paying attention to these details we found equivalence in the MTFs of several CRTs for the three methods of stimuli illustrated in Fig. 15.

Fig. 16 (a), (b), and (c) show the result, measured (and evaluated manually) on a medium resolution 1600 x 1200 monitor. Fig. 16 (a) shows MTF derived from the squarewave response, Fig. 16 (b) shows MTFs derived from the line response, and Fig. 16 (c) shows the MTFs derived from the response to white noise, i.e., the broadband response. Note that the agreement of all MTFs for

squarewave response, line response and the MTFs of the broadband response for 113 and 200 ADU of the horizontal MTF as well as the 200 ADU of the vertical MTF are very close.

The much lower MTF for vertical direction and low luminance value (113 ADU) of the broadband response (the uppermost graph) is a mystery up to this time despite all efforts to find the cause for it and requires more investigation.

4. Acquisition of instrumentation necessary for Task 5

We acquired an x-y-z translation stage, an optical table and a lightbox. And we mounted a CCD camera. Fig. 17 is a photograph of this set-up. So far it has been used to determine the MTFs of the CCD camera, described in Fig. 11 above.

5. Using a new method to generate flat fields.

An important and very popular method to correct for CCD and lens non-uniformities is the method of "flat-fielding". Here an image of a uniform field is taken by the CCD camera for the specific imaging geometry (i.e. magnification and lens f-stop) This image is stored in computer memory. After taking the image of the desired scene (i.e. the raw image), this raw image is devided by the flatfield pixel by pixel and afterwards multiplied by a factor proportional to the mean value of the flatfield to result in the corrected image of the desired scene.

One of the big problems in CRT evaluation with a CCD camera is the fact that the CRT by itself cannot serve as the flat field being that there are the CRT raster lines, which are rarely stable. There is always some small jitter and two successive CCD images of the rasterlines can rarely be subtracted. On the other hand, there is no space for a uniform light box or integrating sphere.

A surprisingly simple solution was to acquire a highly diffusing opal glass and mount it right in the center of the CRT right on the CRT's faceplate. The opal glass is very uniform and extremely smooth (no apparent granularity). It diffuses the light from the CRT and no CRT raster lines are visible.

Fig. 18 is a photograph of the opal glass and Fig. 19 is a photograph of the CCD-CRT combination with the opal glass mounted on the CRT during the procedure to generate the flatfield. In order to perform a flat fielding procedure, the CCD camera is moved away from the CRT by the distance between the CRT phosphor and the surface of the opal glass.

Fig. 20 is the actual flatfield, stored in the computer and used by the program to do flat field correction. Notice that it looks almost identical to the raw image, but it is different as far as noise is concerned, because it is generated by averaging of 10 raw images.

Fig. 21 is an uncorrected ("raw") CCD image of the opal glass, mounted in front of the CRT and back illuminated by the light emitted by the CRT as seen in Fig. 19. Notice the non-uniformity, which mainly stem from the lens and to a lesser degree from the gain fluctuations of the CCD array. Fig. 22 is the corrected ("flatfielded") CCD image of the same portion of the opal glass. Now all the

non-uniformities are gone. The mean value of 10500 ADU is that of the raw image minus the CCD dark level, and the standard deviation is about 70. This standard deviation is due to the photon noise of that portion of light from the opal glass collected by the lens and recorded by the CCD for the particular quantum efficiency of the CCD (somewhere around 35 %).

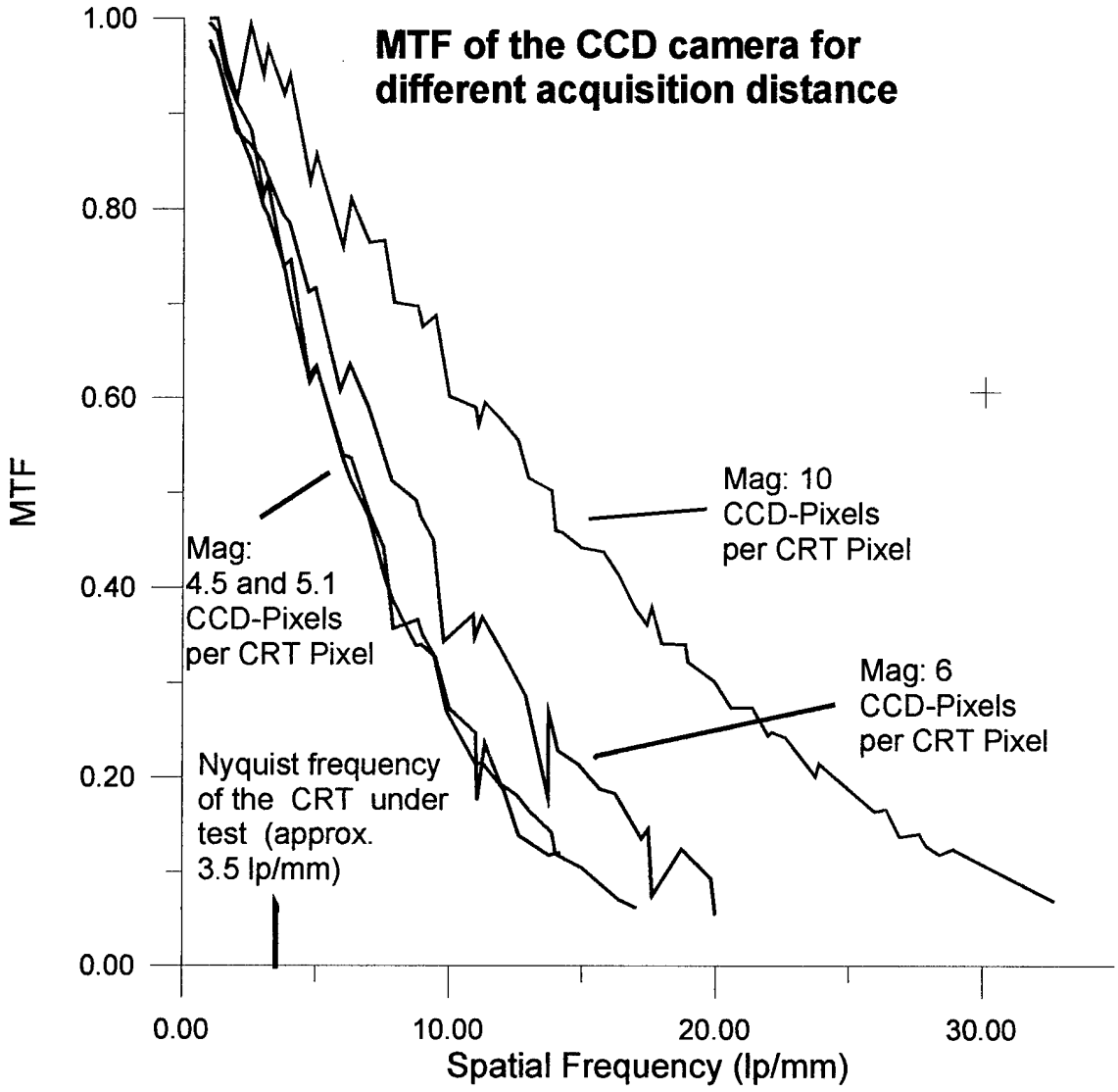


Figure 11 MTF of the CCD camera for different values of magnification

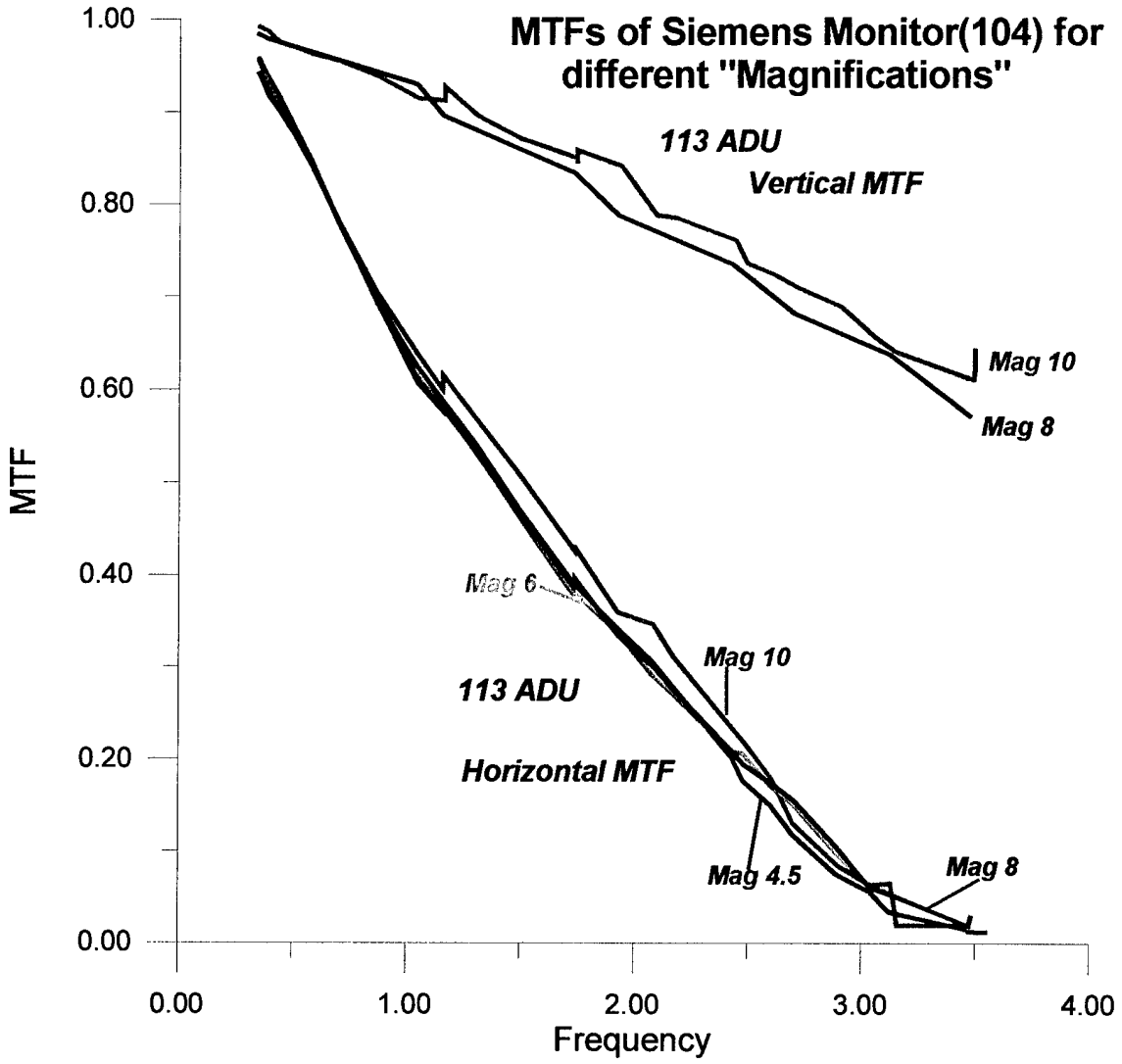


Fig 12 Horizontal and vertical MTFs at low luminance levels (command level = 113 ADU) of a Siemens monitor with a P104 phosphor for different magnification values.

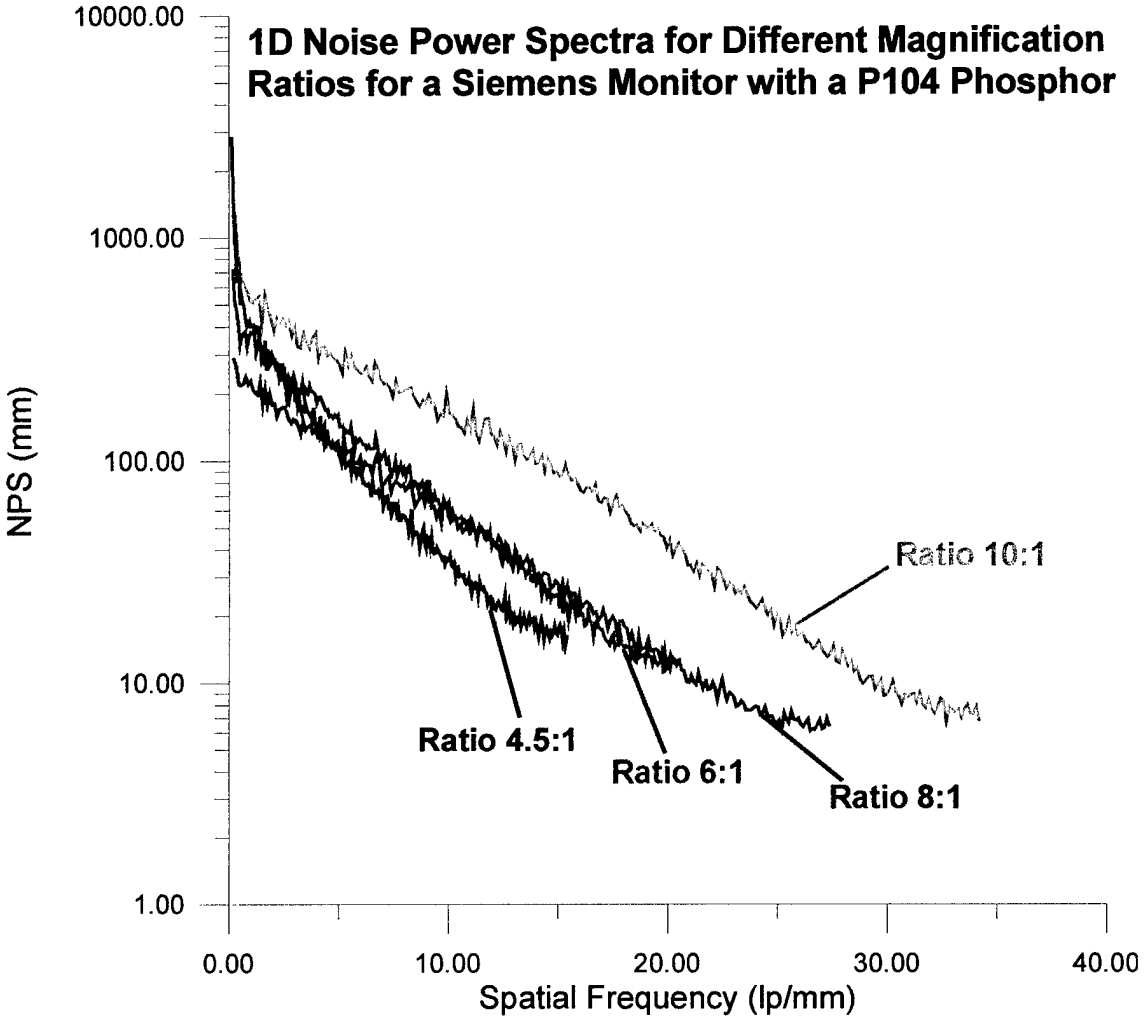


Fig 13 Noise power of a Siemens monitor with a P104 phosphor, measured at different magnification values (identified here as "Ratio")

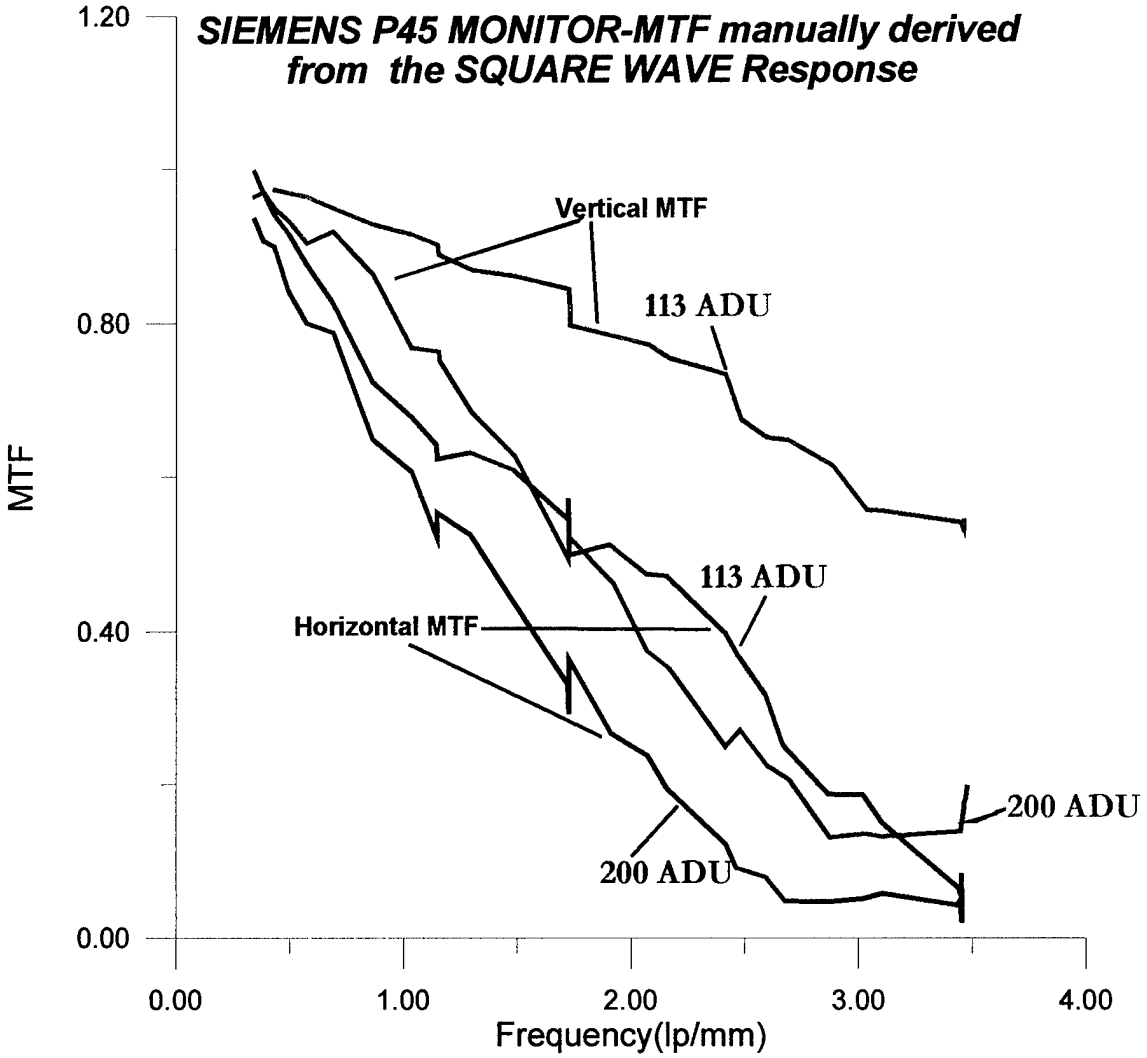


Fig 14a MTFs of a high performance Siemens monitor with a P45 phosphor, evaluated manually

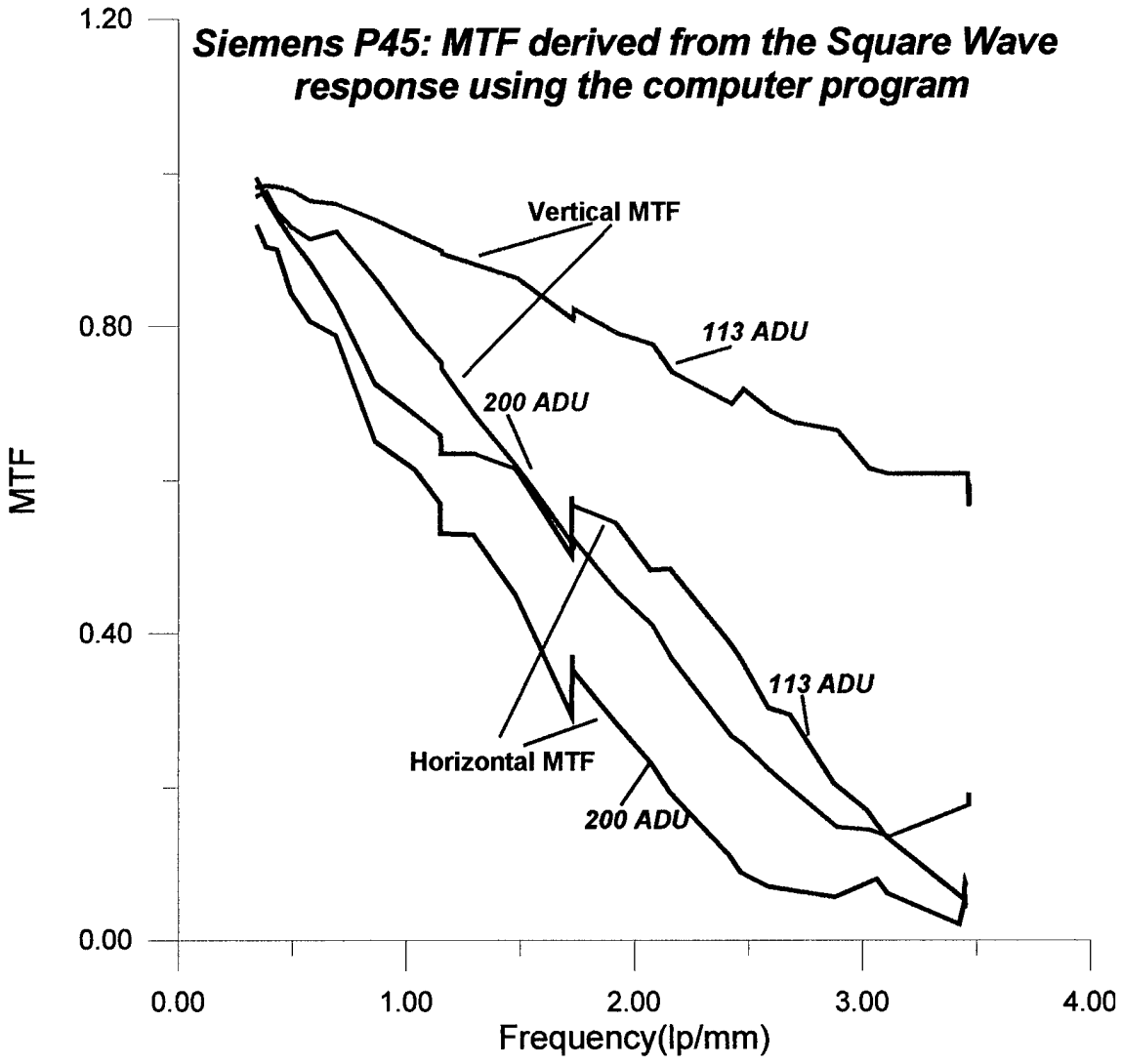


Fig 14b MTFs of a high performance Siemens monitor with a P45 phosphor, evaluated by the computer program

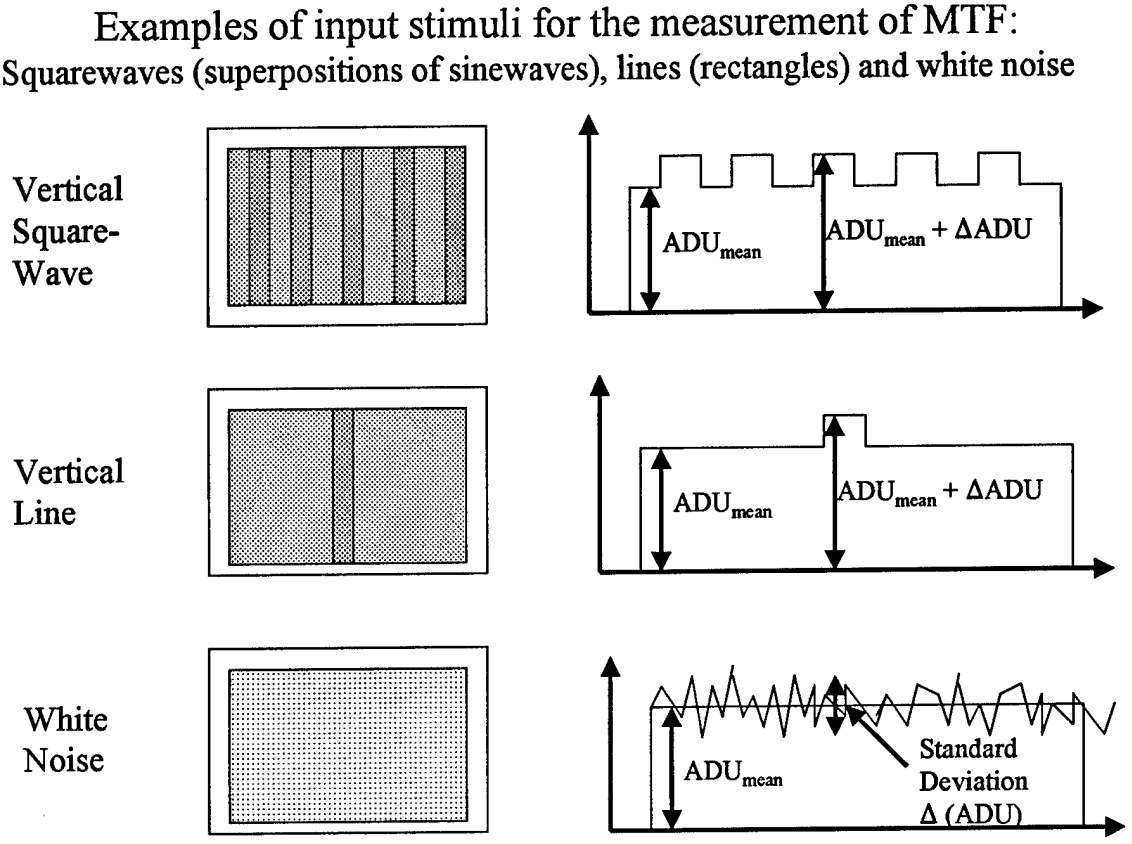


Fig 15 Schematic illustrating the various input stimuli used for the measurement of MTFs of CRTs. Notice the small signals, identified as Δ ADU, which should permit to consider the CRT as "quasi linear" and allow use of Fourier Transform Tedhniques.

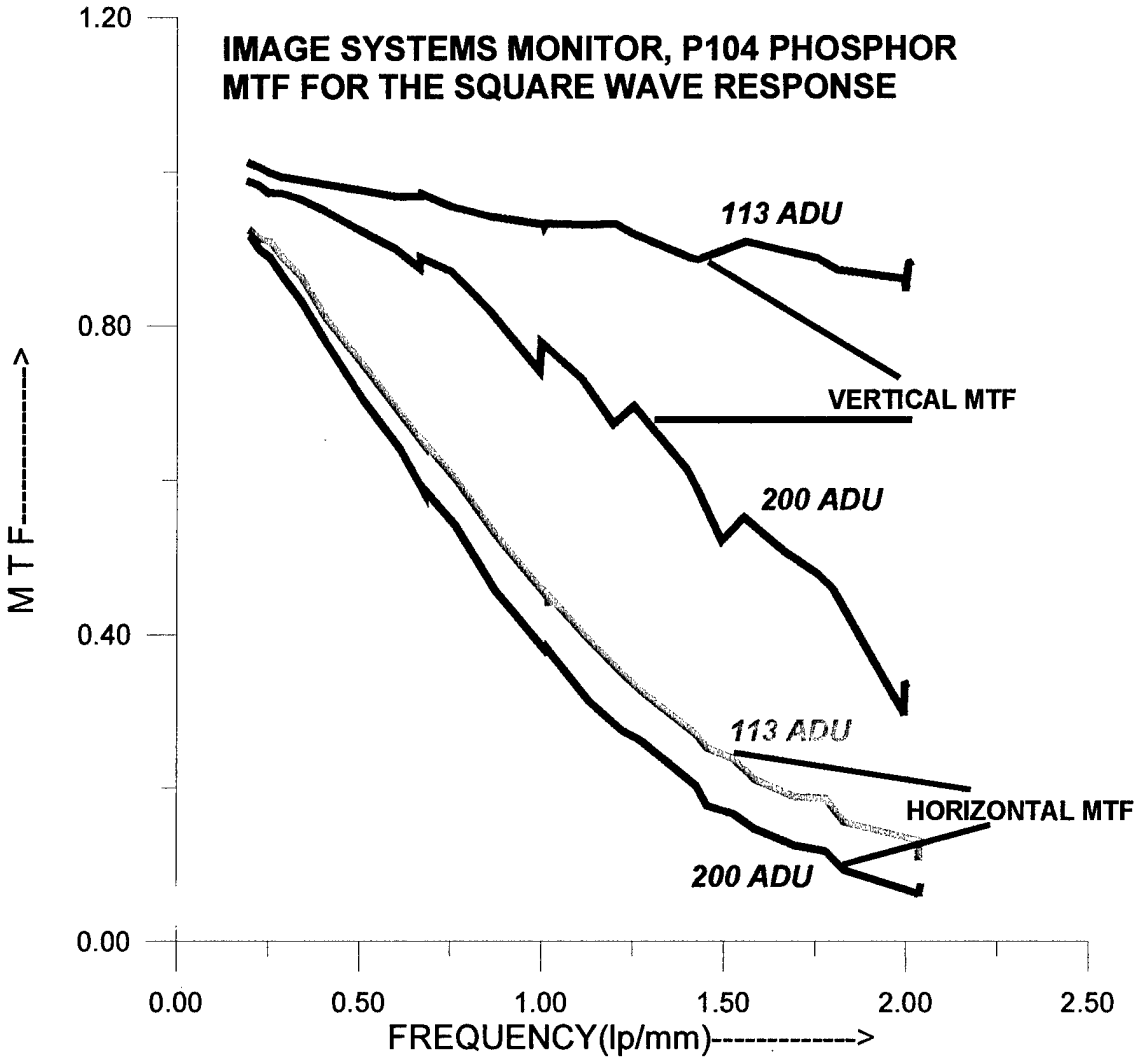


Fig 16a MTFs derived manually from the squarewave response

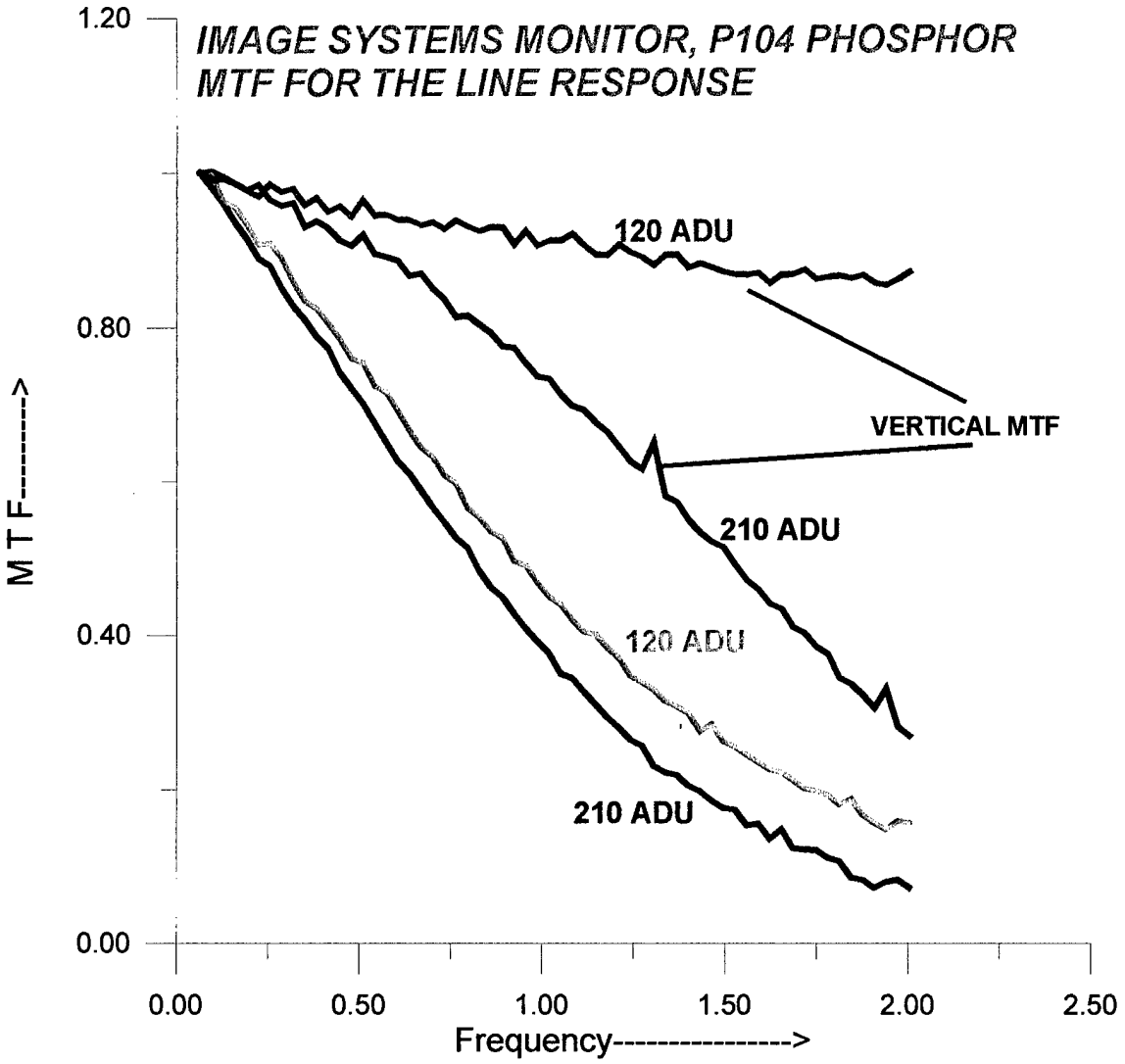


Fig 16b MTFs derived manually from the line response

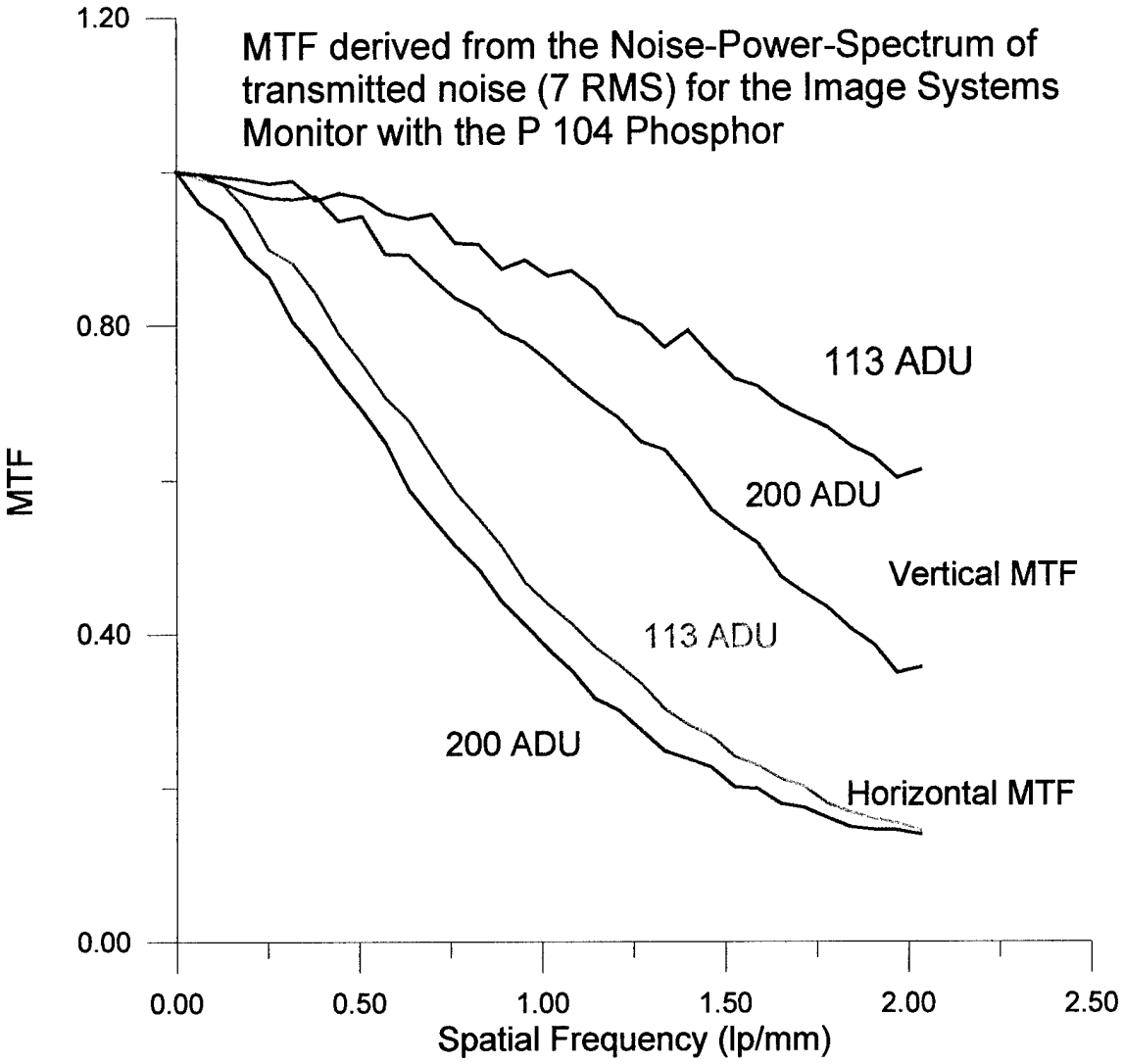


Fig 16c MTFs derived manually from the broadband response

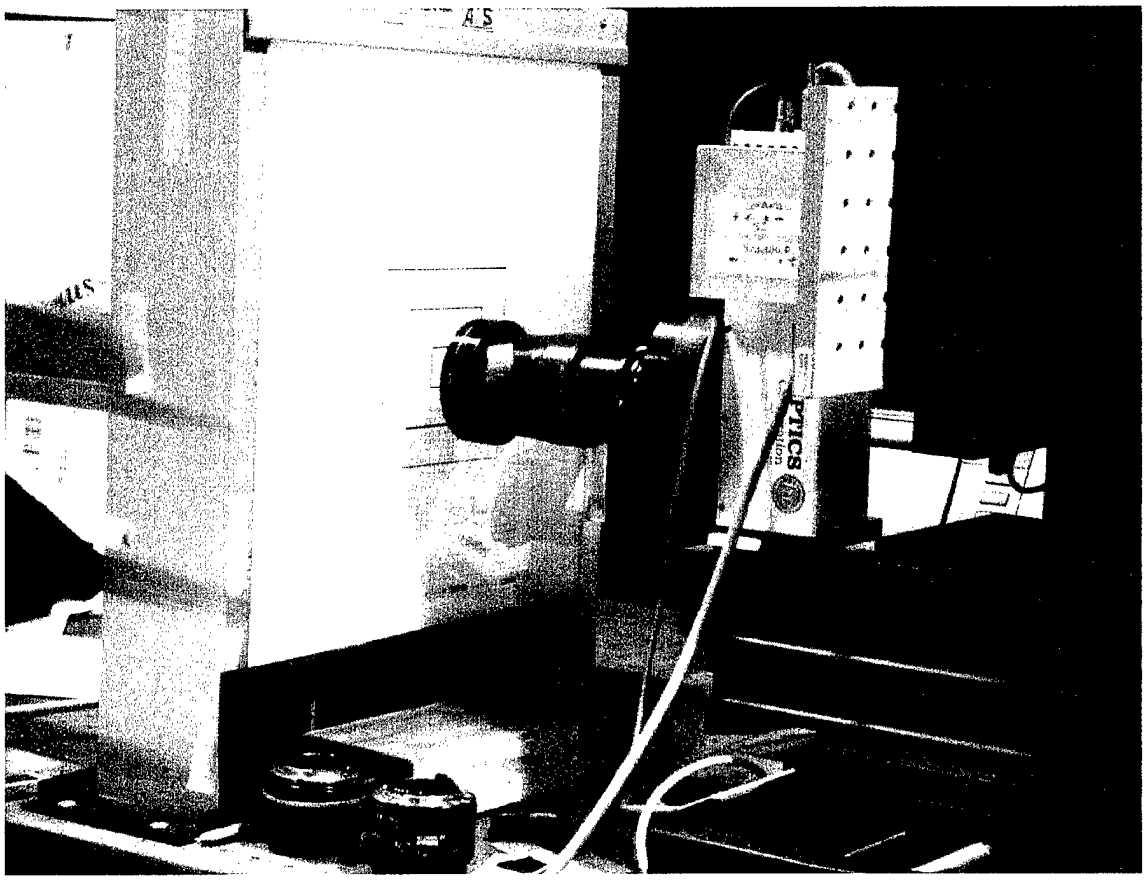


Fig 17 Photographs of the CCD camera, mounted on an x-y-z translation stage on top of an optical table and in front of a lightbox to investigate geometrical errors when imaging with a CCD camera

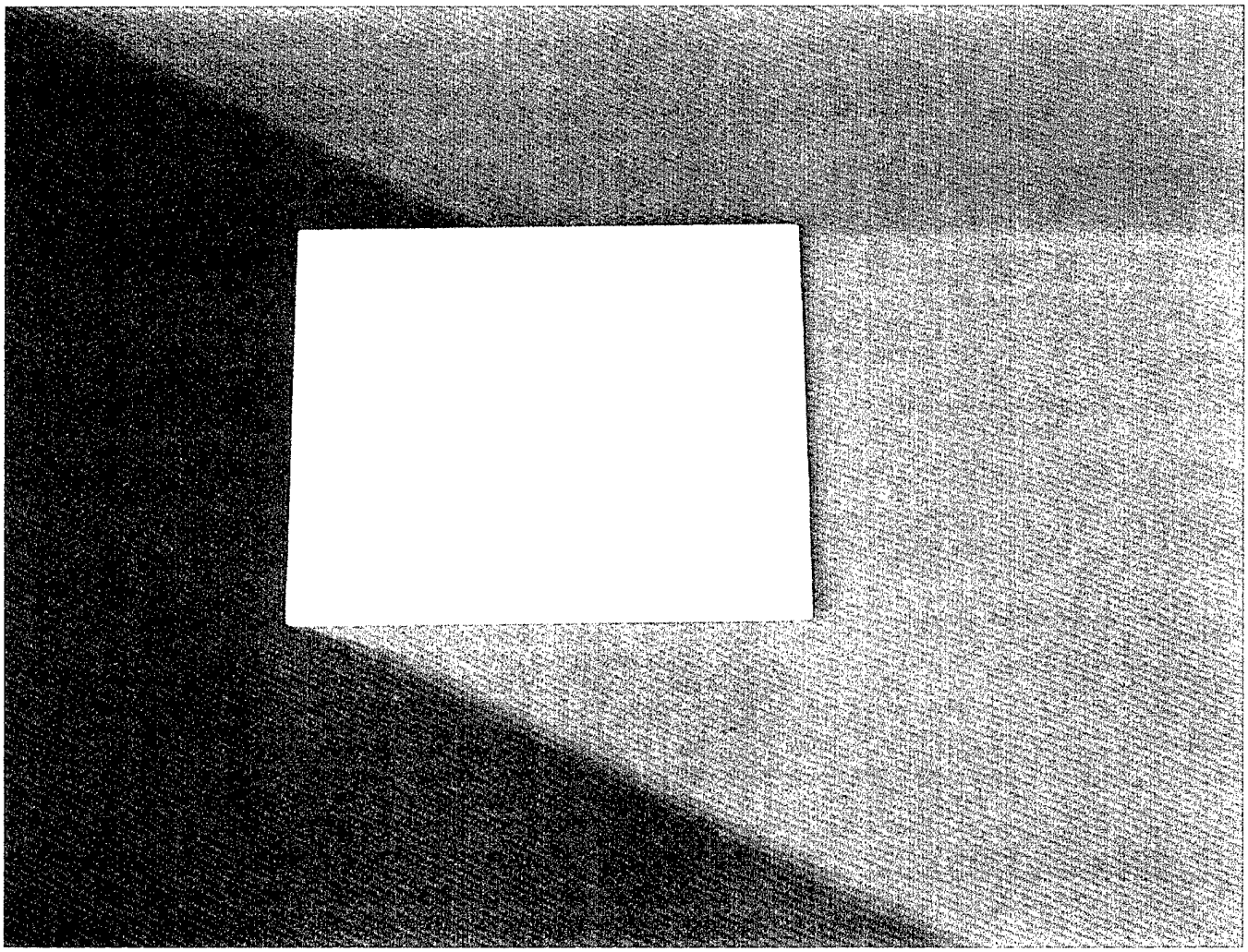


Fig 18 Photograph of an opal glass plate serving as an ideal diffuser

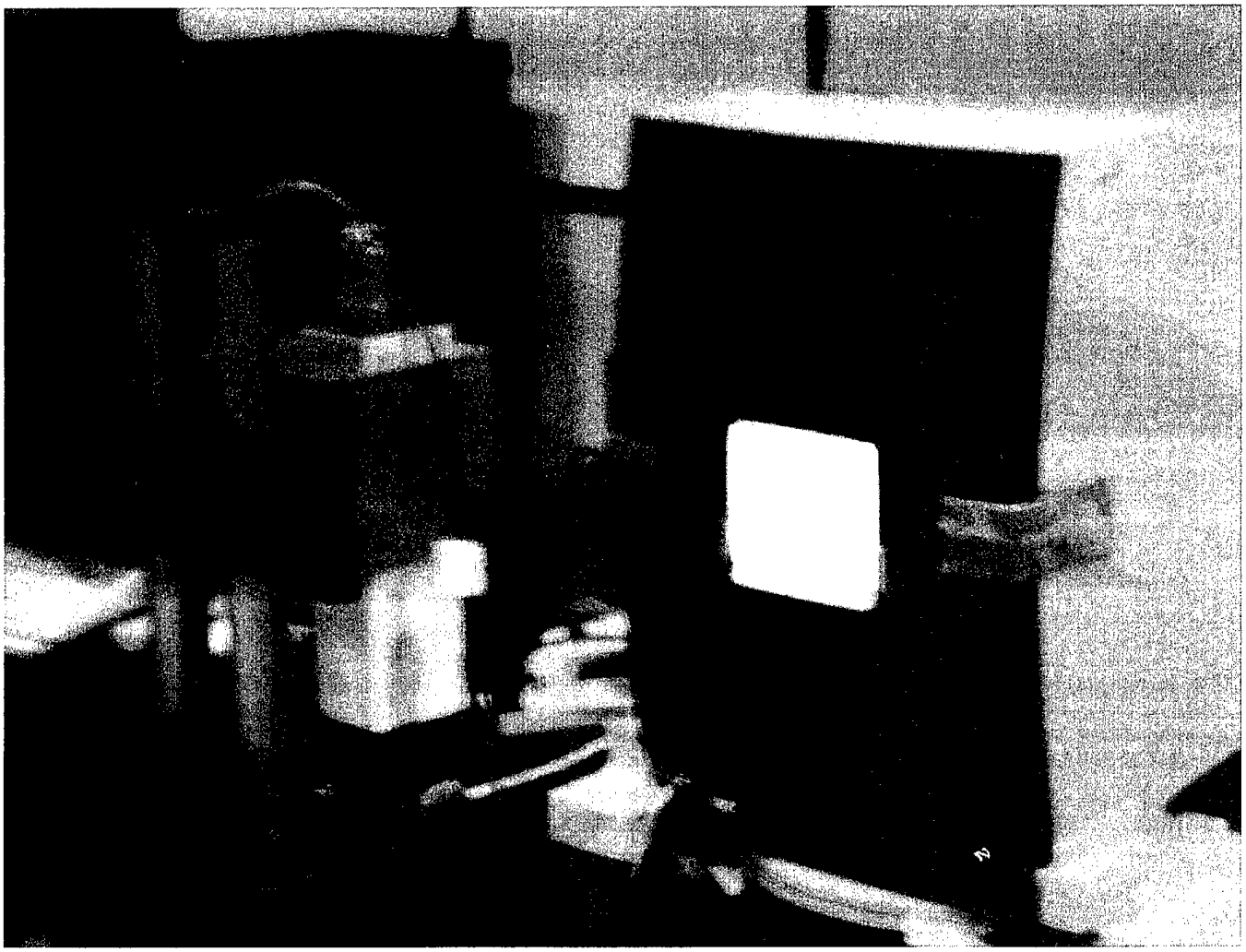


Fig 19 Photograph of the arrangement CCD camera and opal glass, mounted on a CRT faceplate, ready for the procedure to generate a "flatfield" for flatfield correction of CCD camera images.

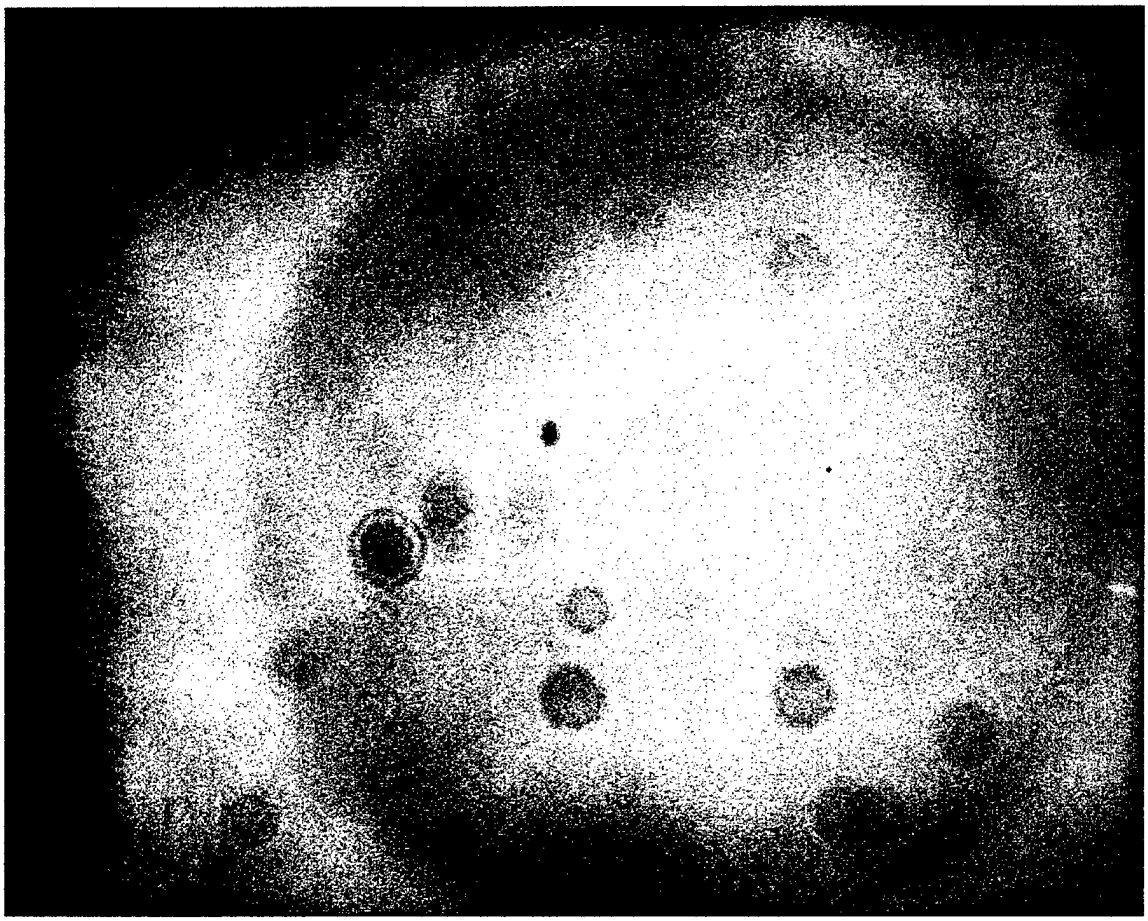


Fig 20 Flatfield for "flatfield correction" of the CCD camera images

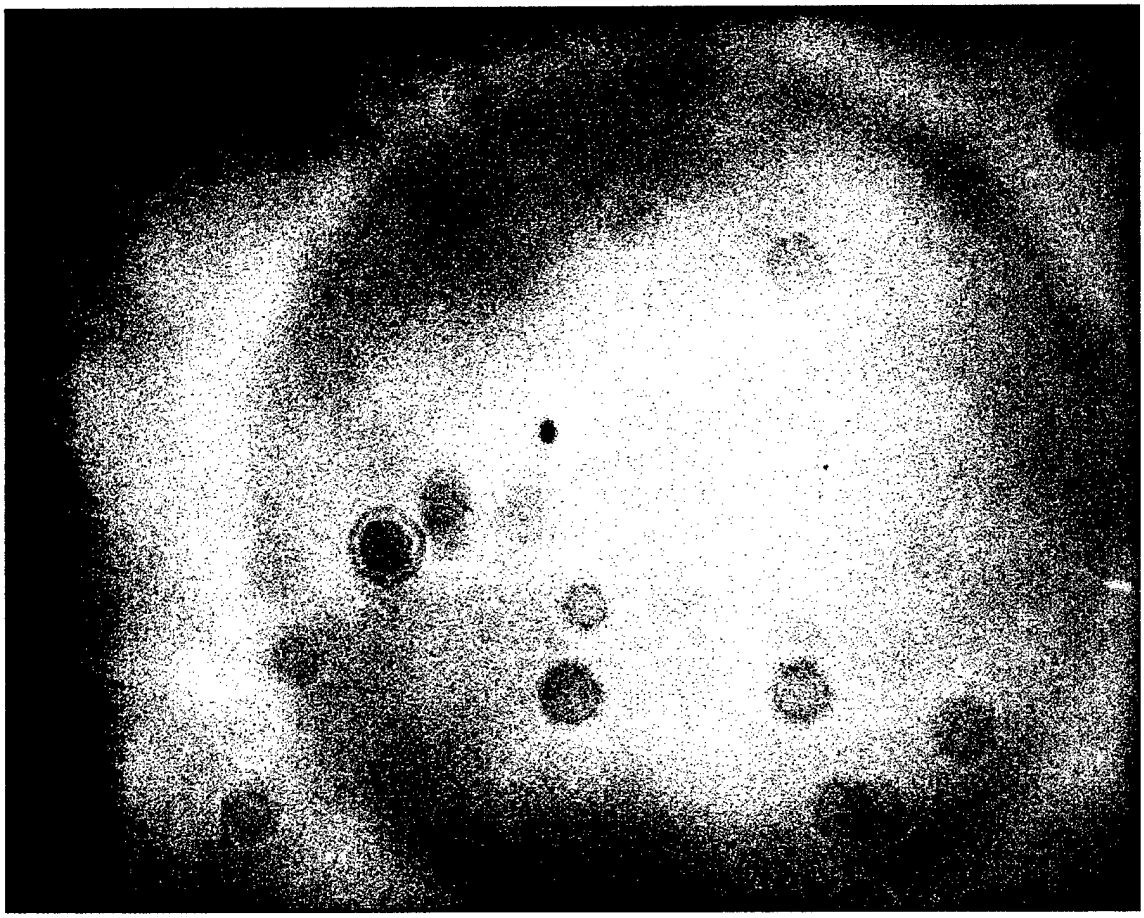


Fig 21 Raw CCD image of the uniform opal glass, showing non-uniformities due mainly to the lens of the CCD camera.

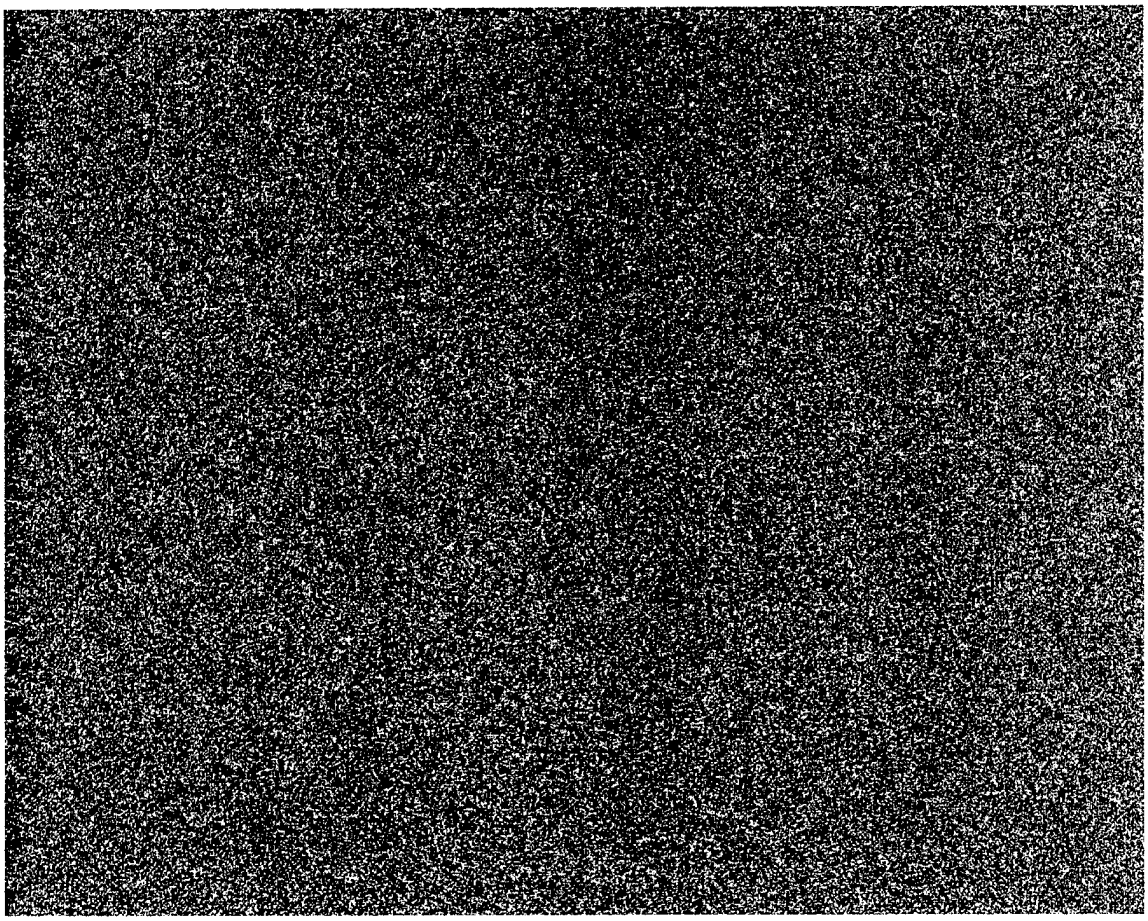


Fig 22 Flatfielded CCD image of the uniform opal glass. Flatfield shown in Fig. 10 was used to flatfield (correct) the raw image shown in Fig. 11.

MODIFICATIONS OF THE ORIGINALLY PLANNED METHODOLOGY

As outlined in the Year 1 report, we made a major improvement and simplification of the originally proposed method by switching from physical phantom images to test patterns, i.e. digital images with known pixel values. This is impossible in the context of x-ray image quality measurement, since one wishes to include x-ray factors, but is natural when optimizing CRT image quality. One regards the monitor/CCD camera combination as a transducer that transforms an input pixel distribution into an output pixel distribution. We apply signal detection theory directly (non-pre-whitening matched filter in a SKE task), which is also behind CAMPI. Including phantom noise (x-ray and other) would obfuscate the effect of the CRT, as we would not be able to separate x-ray and other degradations from that due to the CRT.

While not included in the original SOW, we had hoped to include measurements on structured backgrounds in Year 2. This idea has been dropped. This was due to the complexity of including a Hotelling observer along the lines of Barrett and colleagues – we continue to follow their advances with great interest and we are in communication with them to help us implement their algorithm. However, this cannot happen in the context of this project, and more immediate and interesting results need to be followed up on.

The details of the measurements are somewhat different from that originally outlined. For example, we have focused more on the final monitor comparison than on the details of how image quality changes when a given monitor is de-focused. This is because of the current interest in the P-45 vs. P-104 issue. We plan to conduct a few measurements on defocused monitors in the remaining year, but most of our effort will be focused on understanding the results and conducting confirmatory measurements.

PROBLEMS ENCOUNTERED

No major non-scientific problems or issues were encountered in this period. A laboratory move on the PI resulted in a minor disruption of activities.

KEY RESEARCH ACCOMPLISHMENTS

- Continuing improvements of basic CAMPI methodology to analyze the CCD images; high degree of automation and flexibility in the programs; these programs will run on any platform that supports IDL – i.e., practically all computer systems.
- Application of the measurement of image quality of two monitors. Over 500 images analyzed to date, including 414 in the monitor comparison study.
- Preliminary results showing the superiority of the P-45 design over the P-104 design.

REPORTABLE OUTCOMES

1. Chakraborty, Dev P.; Liu, Xiong; O'Shea, Michael; Toto, Lawrence C.: A quantitative method for visual phantom image quality evaluation. Proc. SPIE Vol. 3981, p. 24-33, Medical Imaging 2000: Image Perception and Performance, Elizabeth A. Krupinski; Ed.

2. Chakraborty DP, Kundel HL: Anomalous nodule visibility effects in mammographic images, Proc. SPIE Vol. 4324, p. 68-76, Medical Imaging 2001: Image Perception and Performance, Elizabeth A. Krupinski; Dev P. Chakraborty; Eds., Publication Date: 6/2001.

3. Roehrig,H, Fan JH: "CRTs for Medical Imaging: Operation, Performance and Calibration". Invited Talk presented at the Annual Meeting of the Japanese Radiological Society, April 7, 2001, Kobe, Japan.

4. Roehrig H, Fan JH:"Performance Evaluation of CRT for Medical Application". Invited Talk, presented on April 10, 2001at the weekly Seminar of Gifu-University, Gifu, Japan.

CONCLUSIONS

The complexity of digital mammography systems requires development of sophisticated methods to maintain their quality. This project addresses the critical display element of the mammographic imaging system. We have applied the CCD/CAMPI method to answering a critical question regarding phosphor types (P-104 vs. P-45) for mammography. Due to a late start, which was outside our control, we had to request an extension. However, we are very excited about the work performed this year and the work to be done in the final period.

REFERENCES

NA

APPENDICES

1. Chakraborty, Dev P.; Liu, Xiong; O'Shea, Michael; Toto, Lawrence C.: A quantitative method for visual phantom image quality evaluation. Proc. SPIE Vol. 3981, p. 24-33, Medical Imaging 2000: Image Perception and Performance, Elizabeth A. Krupinski; Ed.

2. Chakraborty DP, Kundel HL: Anomalous nodule visibility effects in mammographic images, Proc. SPIE Vol. 4324, p. 68-76, Medical Imaging 2001: Image Perception and Performance, Elizabeth A. Krupinski; Dev P. Chakraborty; Eds., Publication Date: 6/2001.

APPENDIX 1

PROGRESS IN BIOMEDICAL OPTICS AND IMAGING

Vol. 1, No. 26
ISSN 1605-7422

Reprinted from

Medical Imaging 2000

Image Perception and Performance

16–17 February 2000
San Diego, California

Proceedings of SPIE
Volume 3981

A quantitative method for visual phantom image quality evaluation

Dev Chakraborty, Xiong Liu, Michael O'Shea and Lawrence Toto

University of Pennsylvania, Department of Radiology, Philadelphia, PA 19104

ABSTRACT

This work presents an image quality evaluation technique for uniform-background target-object phantom images. The Degradation-Comparison-Threshold (DCT) method involves degrading the image quality of a target-containing region with a blocking processing and comparing the resulting image to a similarly degraded target-free region. The threshold degradation needed for 92% correct detection of the target region is the image quality measure of the target. Images of American College of Radiology (ACR) mammography accreditation program phantom were acquired under varying x-ray conditions on a digital mammography machine. Five observers performed ACR and DCT evaluations of the images. A figure-of-merit (FOM) of an evaluation method was defined which takes into account measurement noise and the change of the measure as a function of x-ray exposure to the phantom. The FOM of the DCT method was 4.1 times that of the ACR method for the specks, 2.7 times better for the fibers and 1. 4 times better for the masses. For the specks, inter-reader correlations on the same image set increased significantly from 87% for the ACR method to 97% for the DCT method. The viewing time per target for the DCT method was 3-5 minutes. The observed greater sensitivity of the DCT method could lead to more precise Quality Control (QC) testing of digital images, which should improve the sensitivity of the QC process to genuine image quality variations. Another benefit of the method is that it can measure the image quality of high detectability target objects, which is impractical by existing methods.

Keywords: Image Quality Assessment, Mammography, Methodology, Phantoms, Quality Control.

1. INTRODUCTION

Phantoms are widely used to assess and monitor the image quality of medical imaging devices. Currently, the most prevalent use of phantoms is in mammography, where they are used for accreditation and quality control (QC) purposes. Phantom image quality evaluation consists essentially of determining the number of target objects that are visible[1]. Often targets of different types are used, e.g., the fibers, specks, and masses in the American College of Radiology (ACR) accreditation phantom. Within each target type there are several targets with varying detectability, from easily visible to marginally visible. As image quality degrades the marginal targets are less easily visualized, and the total count of visualized targets decreases. The total count of a given target type is used as an image quality index. The ACR has set minimum standards, namely 4 fibers, 3 speck groups and 3 masses must be visualized to pass the phantom test.

While conventional phantom evaluation as described above is relatively inexpensive and easy to perform, it does have its limitations[2]. An individual reader can be variable on any given image on different reading occasions, intra-reader variability. This can be minimized by careful reader training, which is done for readers used by accreditation programs, but is generally impractical in the clinical QC setting. Additionally, there is substantial inter-reader variability in phantom scores, i.e., different readers produce variable scores on any given image. These problems are probably related to variability in reader visual performance characteristics, variability in decision criteria and lack of control for false positives. Consequently, while the ACR phantom reading procedure may be adequate for determining if an image passes or fails accreditation, it may not be as useful as a quantitative image quality indicator. Based on our professional experience the ACR phantom test is not very sensitive to subtle image quality changes – typically clinicians detect image quality degradation on clinical images before they are detected by the technician on the phantom images. This is, of course, defeating the purpose of the QC testing program. One way to assure increased sensitivity is to devise more finely graded phantoms, and such phantoms are available commercially (e.g., Computerized Imaging Reference Systems Inc., Norfolk, Virginia). This paper suggests an alternative technique applicable to direct digital images or digitized films of any uniform background target object phantom. It is termed the Degradation-Comparison-Threshold, or DCT method.

2. METHODS

As implied by the acronym, degradation, comparison and threshold degradation determination are the essential features of the DCT method. The method involves degrading the image quality of a target-containing region T with a *blocking* processing,

In *Medical Imaging 2000: Image Perception and Performance*, Elizabeth A. Krupinski, Editor, Proceedings of SPIE Vol. 3981 (2000) • 1605-7422/00/$15.00

and comparing the resulting image T' to a similarly degraded target-free region N'. We use the term "blocking" in the sense that the processing blocks or impedes the detection of the target. The blocking processing must be capable of being characterized by a variable L (for *level* of the processing) and the image quality must be a monotonic function of L. We define L_0 as the level corresponding to zero degradation. The threshold degradation is defined as that level L_t that yields 92% correct in the two-alternative forced choice (2AFC) task. If the image quality is an increasing (or decreasing) function of L, L_t (or $1/L_t$) is defined as the image quality measure of the DCT method. Note that the image quality depends on the blocking processing used. In the present study we used a random "white noise" field as the blocking processing, and a dilution process to control the level of the blocking processing. In addition we used the well-known QUEST procedure[3] to determine the threshold L_t.

For each target region T_{ik}, where i is the image index and k is the target index, a target-free region N_{ik} of minimum standard deviation was determined by exhaustive search of the image. The mean and minimum standard deviation of the noise region are denoted by μ_{ik} and σ_{ik}, respectively. A dilutor region D_{ik} with mean μ_{ik} and standard deviation σ_{ik} was synthesized using a gaussian random number generator. Two images T'_{ik} (diluted target) and N'_{ik} (diluted noise) were synthesized using the linear combinations shown below:

$$\left.\begin{aligned} T'_{ik} &= d\,T_{ik} + (1-d)\,D_{ik}\ , \\ N'_{ik} &= d\,N_{ik} + (1-d)\,D_{ik}\ , \end{aligned}\right\} \qquad (1)$$

where the *dilution* variable d satisfies $0 < d < 1$. Note that linearly combining the dilutor region corresponds to the blocking processing, d corresponds to the level variable L defined above, and d=1 corresponds to L_0, the level of the blocking processing that introduces no degradation. Window settings were determined from the N'_{ik} region by determining its minimum and maximum pixel values, which values was mapped to the available range of the display (0 to 255). The two images T'_{ik} and N'_{ik} were identically windowed using these settings and displayed side-by-side on the monitor. The observer was asked to choose the side most likely to contain the diluted target. His response (correct or incorrect) was supplied to the QUEST algorithm, which generated a new dilution value. If the observer is correct then QUEST decreases the dilution value so as to make the task more difficult, and conversely if the observer is incorrect, then QUEST will increase the dilution value. The procedure was iterated until the *error* of the maximum likelihood estimate (MLE) of the observer's threshold, as determined by the QUEST algorithm, fell below a preset value. Let MLE_{ik} be the final MLE of the threshold dilution. The image quality IQ_{ik} for image i and target k is defined as

$$IQ_{ik} = 1 / MLE_{ik}\ . \qquad (2)$$

We adopt the following notation in referring to the phantom targets. The first character is determined by the type of target, e.g., fiber = F, microcalcification = M, mass or nodule = N. The second character is the index number of the target within its type, with lower numbers referring to more visible targets. Using this notation, the least visible targets would be F6, M5 and N5. Five (5) observers participated in the experiments. They included two experienced and American Board of Radiology (ABR) certified medical physicists, an x-ray technician and two programmers. Each observer read 3 targets (F1, M2 and N4) on each of ten images of an accreditation phantom (GAMMEX RMI, Middleton, WI, Model RMI-156). The displayed ROIs were typically 170 pixels square, and each pixel was 0.1 x 0.1 mm^2 measured at the image receptor. To obtain variability data and to provide training, each observer repeated the targets on two images, a total of four times prior to beginning the main study. The total reading time was about 5 hrs per observer. Typically 70-80 iterations were needed to reach the termination criterion. The reading time per target was 3 – 5 minutes, with specks taking longer times.

Quest Implementation: We used the general algorithm described by Watson and Pelli[3], however, it had to be tailored to our needs. The signal strength parameter x in the QUEST algorithm is on a decibel scale while the dilution is not. Therefore, we defined the QUEST scale x variable by

$$x = 20\,A\,\log_{10}(d) + B\ , \qquad (3)$$

where d is the dilution parameter and A and B are constants defined, see below, to optimally span the quest variable space. We used N = 200 corresponding to an 801-dimensional array for the QUEST probability function, with the center index (i.e., 400) corresponding to x = 0. Denoting the starting dilution value by d_0, then setting

$$B = -20\,A\,\log_{10}(d_0) \qquad (4)$$

ensures that the initial dilution value will map to x = 0, i.e., the search begins in the middle of the QUEST scale. Defining f as the factor by which the initial estimate is uncertain, i.e., $d_0/f < d_0 < d_0 f$, one has $N \sim 20\,A\,\log_{10}(d_0 f) + B$, or

$$A \sim N/[40\,\log_{10}(f)], \qquad (5)$$

which assures optimal spanning of the QUEST scale. In our case, f = 8, corresponding to an assumed 6 dB uncertainty ($\pm\sigma$) in the initial estimate, and the additional factor of 2 is needed to obtain 95% confidence limits ($\pm 2\sigma$). With these values one has A = 11.1.

The convergence criterion was determined by testing the uncertainty in the MLE against a preset value. The uncertainty was determined by performing a quadratic fit to the likelihood function[3] in the region of the peak. The abscissa corresponding to the peak is equal to the MLE value. The width of the likelihood function at a value 2.51 units below the peak is the estimate of the uncertainty at the 97.5% confidence level. When this number fell below a preset value, see below, the algorithm terminated.

There are two possible sources of error of the MLE. Since x is not continuous, discretization or "binning" error is possible. This can be estimated from the equation

$$\Delta x = 8.69\ A\ \Delta d/d \qquad (6)$$

by setting $\Delta x = 1$. This yields $\Delta d/d \sim 1/(8.69\ A) \sim 0.010$, which is negligible. A larger error results from the finite width of the likelihood function. We set the terminal value of the width to 0.75 A. Applying Eqn. 6 one has $0.75\ A = 8.69\ A\ \Delta d/d$, or $\Delta d/d = 8.6\%$. Simulations of 40 trials under these conditions yielded 6.9% for the coefficient of variance. Four (4) observers repeated the DCT readings 4 times for each target in two images. The average coefficients of variance over 4 readers, all targets and both images, was 10%.

The QUEST implementation was tested using the model observer suggested by Watson and Pelli[3]. Multiple simulations were used to generate CV estimates. Based on the simulation study we decided to use a two-pass QUEST method, where the first pass determines an approximate MLE value (the stopping criterion was 1.25 A). This estimate is then used as the initial value of the second pass, which had a convergence criterion of 0.75 A. Problems were encountered with the algorithm due to instances where the likelihood function did not have a maximum. This could occur if the observer was correct in all trials preceding the convergence test. Therefore we do not perform any tests for convergence if the observer has yet to make an error. Reader inconsistency, particularly in the early trials, where the dilution adjustments are larger, can lead to convergence problems. We have found that it is useful to record the reader correct/incorrect responses in a file for subsequent off-line analysis of problem runs.

ACR method: Readers also read the images in the ACR specified manner[1]. The numbers of visualized target-objects were recorded, e.g., 4.5 masses, 4 specks and 5.5 fibers. Half scores and deductions for false positives were allowed. For the ACR readings the observers viewed the entire image, while for the DCT readings they only saw two targets. The ACR score for a given target type (e.g., specks) is the sum over all visualized targets of that type, while the DCT score was calculated for targetsF1, M2, and N4. Maximum scores for the ACR method are 6, 5 and 5 for the fibers, specks groups and masses respectively. All software was implemented in the IDL programming language (Interactive Data Language, Research Systems Incorporated, Boulder Co). All readings were performed using a BARCO MGD-6P monitor driven by a DOME card in a Dell computer running the NT operating system (4.0). To minimize glare effects the CRT illumination was about 2 lux.

Exposure: From measurements on the FFDM machine it was found that exposure had a nearly cubic dependence on kVp and the expected linear dependence on mAs. Therefore a quantity **E**, proportional to the true x-ray exposure incident on the phantom, was defined as

$$E = kVp^3\ mAs\ . \qquad (7)$$

The following 3-parameter sigmoid function was used to fit the image quality measures to the logarithm of the exposure values:

$$y = \frac{a}{1 + \exp(-\frac{x-c}{b})}\ , \qquad (8)$$

where $y = \log_{10}(IQ)$, $x = \log_{10}(E)$ and **a**, **b**, and **c** are constants. All fitting was done using a scientific software package (SigmaPlot for Windows, Version 4.0, SPSS Inc., Chicago, IL). The least-squares fit yielded the standard error of the predicted y-values. We identified this with the variability of y, denoted by σ_y. We defined a Figure of Merit (FOM) for an evaluation methodology, at a specified log exposure value x_0, as the local derivative of the fitted curve divided by the noise, $\sigma(y)$. It is seen that this definition ensures that a method with a larger FOM results in a larger change in image quality, in units of σ_y, for a given change in exposure. The FOM can be thought of as the *sensitivity* of the method.

$$FOM=\left(1/\sigma_y\right)\left.\frac{\partial y}{\partial x}\right|_{x_0} . \quad (9)$$

Images: Table 1 summarizes the x-ray techniques for the images used in this study. The images were acquired on a prototype General Electric Diagnostic-Molybdenum-Rhodium (GE-DMR) full-field digital mammography (FFDM) machine. All images were acquired at 100-micron resolution, with grid, Mo-Mo target /filter combination and consisted of 1800 x 2304 x 16 bit image files. Of the 16 bits, 14 bits represent useful gray-scale data. Each image was finally clipped to a 1000 x 1000 region that encompassed the ACR phantom image. The sizes of the displayed target – noise square regions were typically 161 x 161 pixels.

Table 1: This shows the tube kilovoltage (kVp), tube-charge (mAs) and the log(E) data for the images used in this study.

Image #	kVp	mAs	log(E)
1	32	50	6.21
2	32	25	5.91
3	32	200	6.82
4	32	100	6.52
5	30	50	6.13
6	30	25	5.83
7	30	200	6.73
8	30	100	6.43
9	30	4	5.03
10	28	25	5.74

3. RESULTS

Shown in Table 2 are FOM results for the ACR and the DCT methods at log (E) = 6.34, corresponding to a 28 kVp 100 mAs technique. Also shown in the last row of Table 2 are the ratios of DCT to ACR methods. Note that with the exception of reader A, who did better on the masses using the ACR method, every reader did better using the DCT method for every target. The largest improvement was observed for the specks (ratio = 4.12), followed by the fibers (ratio = 2.73), and the least improvement was observed for the masses (ratio = 1.39).

Table 2: Figure-of-merit (FOM) results for the ACR and the DCT methods for a 28 kVp, 100 mAs technique, corresponding to log(E) = 6.214. Shown in the last row are the ratios of DCT to ACR methods.

	DCT			ACR		
Readers	Fiber	Mass	Specks	Fiber	Mass	Specks
A	7.43	2.55	12.48	3.41	7.53	3.86
B	7.59	3.24	7.78	2.96	2.51	2.04
C	11.83	12.85	21.42	2.64	5.17	4.91
D	6.13	4.32	8.88	2.15	1.12	2.93
E	3.59	3.86	16.33	2.22	3.00	2.48
Average	7.31	5.37	13.38	2.68	3.87	3.24
DCT/ACR	2.73	1.39	4.12			

To quantify inter-reader variability, we calculated the correlations of the DCT scores between pairs of readers. These are listed in Table 3 for the 10 (i.e., 5x4/2) pairs of readers and the three targets. The last row lists the p-values for a 2-tailed t-test between the corresponding ACR and the DCT scores. Note that the correlations for the masses and fibers are not significantly different at the 5% level. On the other hand, for the specks the readers uniformly agreed better on the DCT scores (r = 97%) than on the ACR scores (r = 87%) and the difference was very significant ($p < 0.003$).

Table 3: Correlation coefficients of the DCT scores between pairs of readers.

	FIBER		MASS		SPECKS	
Pairing	DCT	ACR	DCT	ACR	DCT	ACR
A/B	0.90	0.77	0.88	0.72	0.96	0.84
A/C	0.93	0.97	0.83	0.98	0.98	0.93
A/D	0.93	0.73	0.89	0.71	0.97	0.85
A/E	0.89	0.84	0.75	0.92	0.97	0.91
B/C	0.81	0.95	0.73	0.81	0.95	0.80
B/D	0.87	0.81	0.86	0.92	0.98	0.93
B/E	0.85	0.97	0.69	0.90	0.98	0.88
C/D	0.85	0.77	0.73	0.64	0.98	0.68
C/E	0.98	0.99	0.83	0.91	0.97	0.92
D/E	0.83	0.83	0.91	0.88	0.95	0.92
Average	0.88	0.86	0.81	0.84	0.97	0.87
p-value	0.58		0.53		0.003	

Shown below in Table 4 are detailed results of the non-linear regression procedure used to determine the FOM. The readers A and E are experienced Medical Physicists, B and C are programmers (C has digital imaging experience) and D is an experienced technologist. The numbers in this table can be used to predict the variation of the FOM with log(E), see Fig. 1.

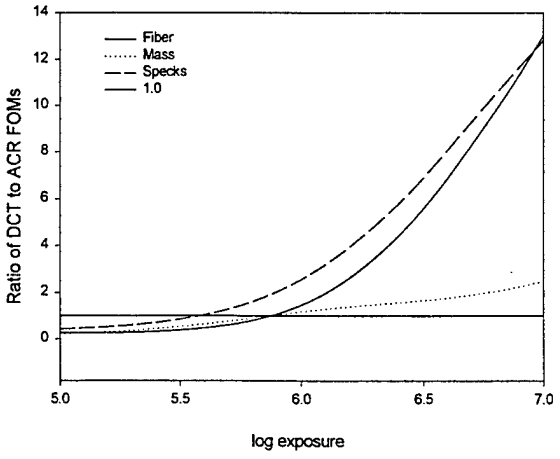


Fig. 1: Plot of ratio of DCT to ACR Figure-of-Merits for the three targets. Data have been averaged over all 5 readers.

Table 4: Results of the fitting procedure used to obtain the Figure-of-Merit (FOM). The parameters a, b, c, and σ_y are defined near Eqn. 7. The values indicated by ¶ had much larger uncertainties and the fitted curves were almost quadratic. See discussion.

Reader	Parameter	QUEST			ACR		
		Fiber	Mass	Specks	Fiber	Mass	Specks
A	a	2501¶	32	161	5.50	4.77	4.77
	b	0.651	0.716	0.841	0.31	0.71	0.43
	c	9.52	6.35	8.43	5.46	5.58	5.16
	σ_y	3.20	4.33	0.96	0.391	0.185	0.209
B	a	28	18	36	5.04	3.80	3.86
	b	0.367	0.399	0.600	0.27	0.41	0.34
	c	6.05	5.69	6.79	5.43	5.40	5.19
	σ_y	2.38	2.26	1.56	0.317	0.393	0.254
C	a	35	13	19	5.12	5.04	4.37
	b	0.542	0.306	0.551	0.27	0.50	0.61
	c	6.586	5.963	6.351	5.41	5.21	5.39
	σ_y	1.204	0.715	0.399	0.332	0.216	0.227
D	a	62	22	22	8.94	5.09	8.83
	b	0.648	0.674	0.506	3.16	0.77	3.17
	c	7.15	6.68	6.38	5.88	3.48	7.28
	σ_y	2.42	1.71	1.17	0.328	0.160	0.231
E	a	563¶	116	39	6.12	5.21	3.87
	b	0.807	0.767	0.626	1.30	0.47	0.33
	c	8.79	7.68	6.92	4.70	5.09	5.37
	σ_y	7.39	4.43	0.71	0.384	0.280	0.316

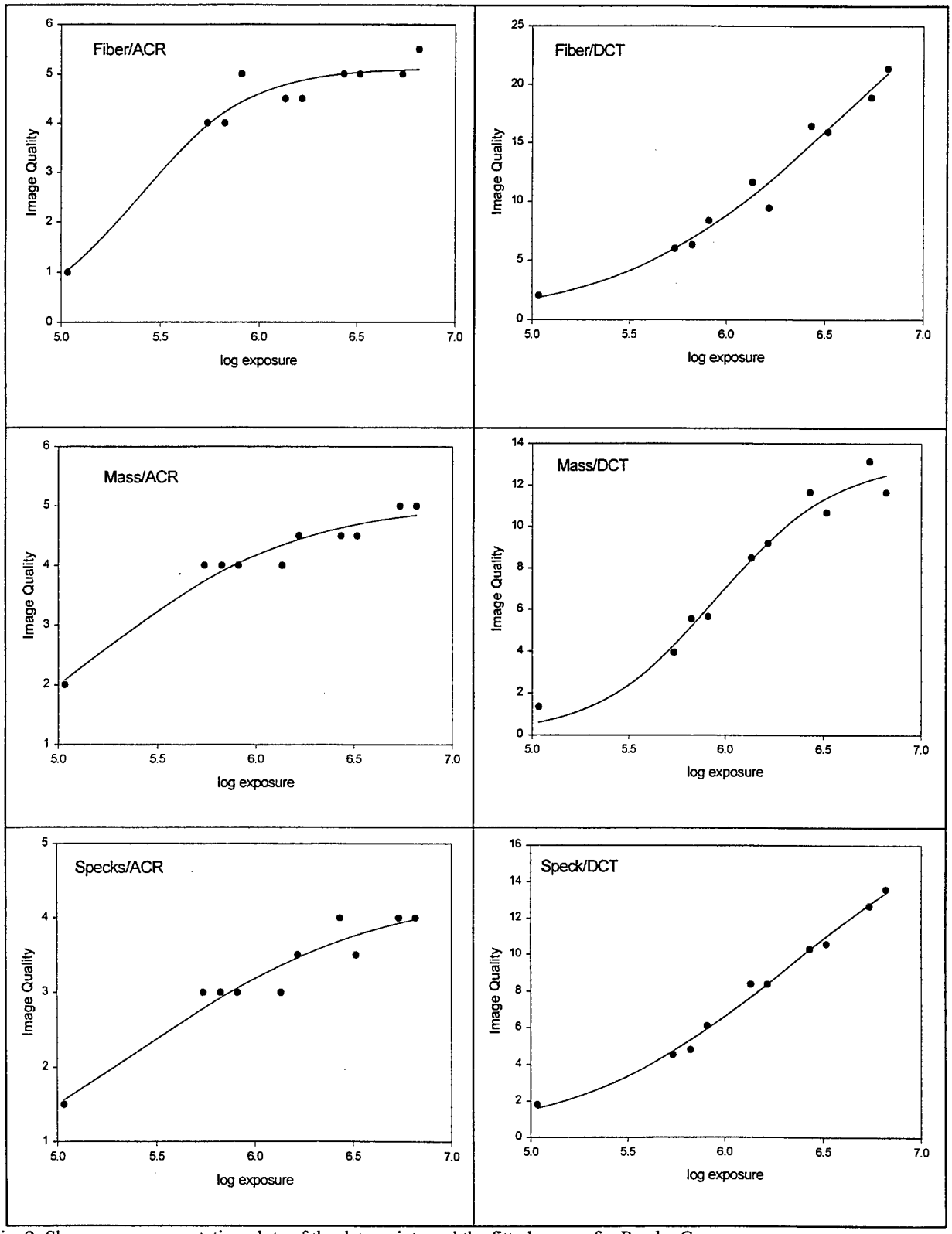


Fig. 2: Shown are representative plots of the data points and the fitted curves for Reader C.

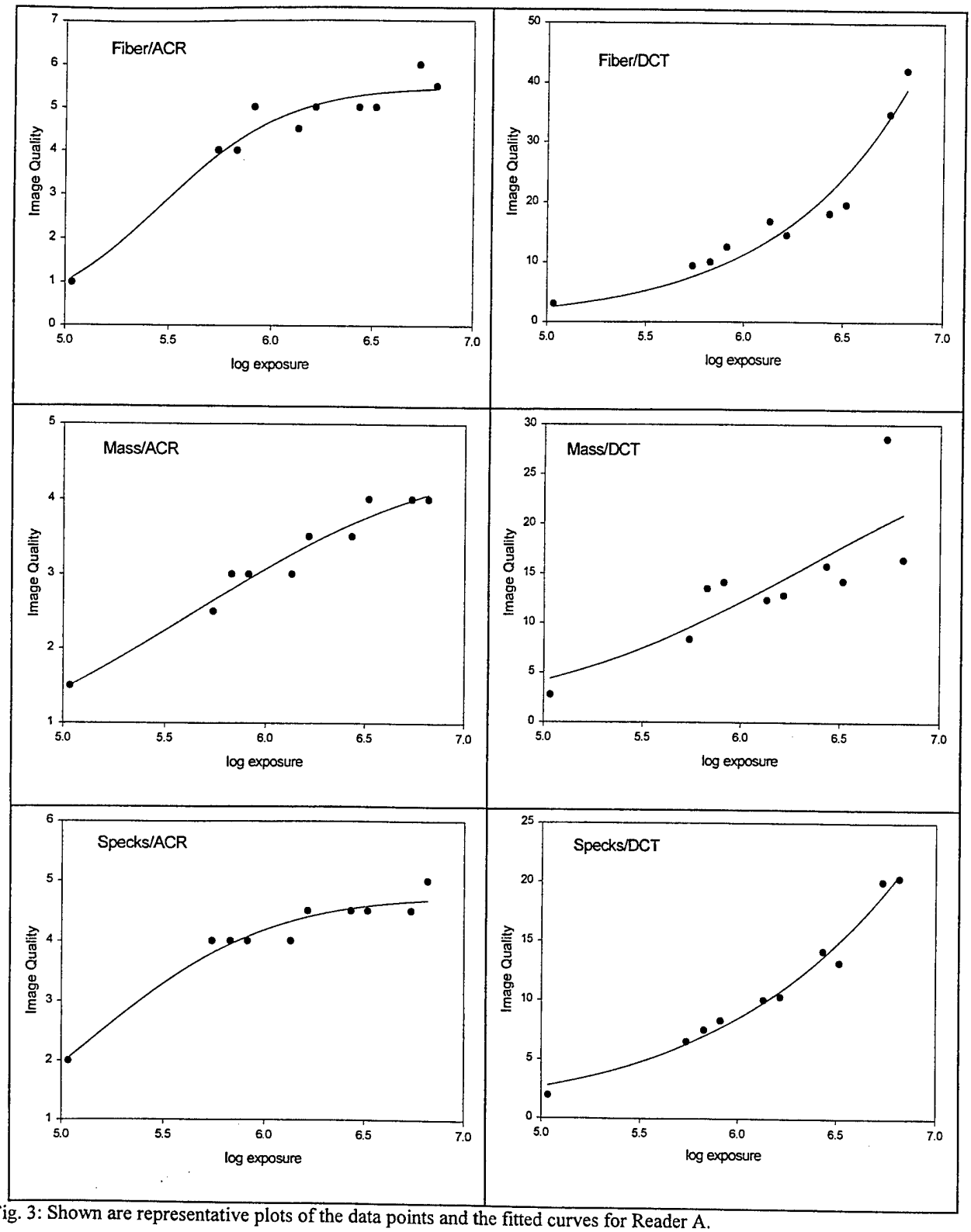


Fig. 3: Shown are representative plots of the data points and the fitted curves for Reader A.

4. DISCUSSION

It is evident from Figs. 2 and 3 that, as the exposure increases the ACR score increases in jumps in a "staircase" manner, i.e., a sequence of jumps each followed by a plateau. This is most evident in the middle-left panel in Fig. 2, where three such plateaus are seen. The plateaus are due to the fact that using the ACR method the observer cannot give finer gradations than 0.5 units. *On the plateaus the observer's score does not discriminate between the images*. The image quality could be deteriorating, but the target score would not reveal that fact. Examining Fig. 2 one sees a plateau in the range $5.74 < \log(E) < 6.1$, which is a factor of 2.3 in exposure. The same behavior is true for observer C's and observer A's ACR speck scores, and similar results were obtained for all readers. The FOM analysis presented above smoothens out this behavior by fitting a continuous curve to the data. *None of the DCT scores displayed staircase behavior*. Given only 5 specks groups, and allowing half-scores, there are at most 10 levels to a speck score. The DCT method has a higher limit and gives a more continuous response with larger dynamic range. It is our opinion that the current ACR phantom has a too-coarse gradation of targets, which moreover, are too easy. As seen in Figs. 2 and 3 most of the change in the ACR scores occur in the low exposure range, well below the clinical range. The scores tend to saturate in the clinical exposure range resulting in a low slope of the IQ vs. log (E) curves, which limits the FOM achievable by the ACR method. A solution to this would be to design a more difficult and more finely graded phantom. Then the region of saturation would be shifted to higher exposures, above the clinical range, and the observer would be able to indicate finer gradations of image quality.

The DCT method appears to allow superior discrimination over the ACR method for all target types but is most notable for the speck targets. There are a few possible reasons for the anomalous behavior for the fiber and mass targets. It could be related to their lower contrast relative to the specks, which, due to bit-truncation effects associated with the 8-bit display, could lead to artifacts at low dilutions. Another reason is that the speck group actually contains 6-specks, and target averaging is improving the stability of the speck scores. For the mass and fiber no such averaging occurs. The mass and fiber DCT measures apply to specific region, and do not take into account region-to-region variations over the entire field-of-view. This is undesirable and it is necessary to design a phantom specifically optimized for the DCT method. For a mass target this could be a phantom that contains several identical masses, rather than the graded masses, as in the ACR phantom. Including the additional identical masses in the paired-comparisons is expected to lead to more stable results.

A likely contributing factor to the anomalous behavior of the larger mass and fiber targets is that they are more affected by low-frequency noise (e.g., image non-uniformity) than white noise, the blocking processing used in this work. By contrast the specks are expected to be relatively insensitive to low-frequency noise. White noise may be quasi-optimal for specks but not for the larger targets. Note that the image quality depends on the blocking processing used and the choice of the blocking processing can affect how well the DCT method performs for a specific target type. Studies using model observers are needed to elucidate these effects further. We are currently using a non-pre-whitening zero suppressing matched filter observer[4], and in future intend to study more sophisticated model observers[5] for DCT method optimization. Measurement difficulties at low spatial frequencies are not unique to DCT. It is well known that Noise Power Spectrum is difficult to measure at low frequencies due to interference from deterministic noise effects, e.g., film-processor roller marks. We have experienced difficulties with mass image quality measurements using Computer Analysis of Mammography Phantom Images (CAMPI)[6,7]. We suspect that the problem is worse for the masses as they are close to the edge of the phantom. That fact, combined with the x-ray intensity variation near the anode end of the image (Heel effect) leads to visually greater non-uniformity near the mass targets than the fibers. We also performed a limited set of measurements on N1, but the results showed even greater anomalies.

The numbers in Table 2 can be used to predict the variation of the FOM with log(E), see Fig. 1. These curves are monotonic increasing and greater that 1.0 for techniques that satisfy $\log(E) > 5.9$. Also, note that the ratio increases with log (E). This is because the DCT measure is essentially measuring signal-to-noise-ratio of the target, which increases with log (E). On the other hand the ACR score saturates with increasing log (E). Note that the fitted noise σ_y of the ACR measures are frequently smaller than the corresponding DCT noise values. However, the DCT has larger FOM since the corresponding slopes are higher. Noise by itself is not an appropriate measure of image quality tracking ability. Reduced noise could be obtained simply by scaling the measure to smaller values. The FOM defined in Eqn. 8 is scale-factor independent. It is also independent of the units of E. Two outlier values are noted on the Table. These occurred when data only straddled the lower part of the sigmoid fitting function, as the regression had insufficient information to estimate the height of the sigmoid, the parameter a in Eqn. 8. Since we never use the fits outside of the data region, this is not expected to create problems.

The DCT method is potentially applicable to clinical images. Since the method requires about 75 comparisons for convergence, one needs approximately 75 normal and 75 abnormal images for 10% accuracy using a 2AFC design. The

comparisons would be performed between images from the normal group randomly paired with images from the abnormal group, with no image being shown more than once (to eliminate memory effects). Based on the observer's response the QUEST algorithm would determine the next level of the blocking processing, and the final MLE value is the desired image quality. An advantage of the DCT method is that one is no longer restricted to using only subtle cases, which are often difficult to acquire. Also, one has the flexibility of setting the threshold d' where optimal statistical accuracy is expected in the 2AFC task[8]. In the ROC case the statistical accuracy is pre-determined by the difficulty level of the cases. Finally, as noted by Watson and Pelli, the QUEST experiment does not preclude a rating response (say a 5-point scale such as used in Receiver Operating Characteristic, ROC, studies). Incorporating a rating type response in DCT could lead to further increases in efficiency.

5. ACKNOWLEDGEMENTS

We are grateful to technical assistance from Aparna Katakam of General Electric Medical Systems in transferring the image files to our workstation and supplying the information to obtain the pixel data from the image files. This work was partially supported by a grant from the Department of Health and Human Services, National Institutes of Health, National Cancer Institute, RO1-CA75145, and in part by a contract from the US Public Health Service's Office on Women's Health, Department of Health and Human Services, contract number RFP 282-97-0077. We are also grateful to Derek Suragh, RT, who served as a reader on this project.

6. REFERENCES

1. R.E. Hendrick, L. Bassett, M.A. Botsco et al, Mammography Quality Control Manual, 4th edition, American College of Radiology, Committee on Quality Assurance in Mammography, 1999.
2. D.P. Chakraborty, M.P. Eckert, "Quantitative versus Subjective Evaluation of Mammography Phantom Images", Medical Physics **22**: pp. 133-143, 1995.
3. Andrew B. Watson and Denis G. Pelli, "QUEST: A Bayesian adaptive psychometric method"; Perception and Psychophysics, **33**, 113-120, 1983.
4. M. J. Tapiovaara and R. F. Wagner, "SNR and noise measurements for medical imaging: I. A practical approach based on statistical decision theory," Phys. Med. Biol. **38**, pp. 71-92, 1993.
5. K. J. Myers and J. J. Barrett, "Addition of a channel mechanism to the ideal observer", J. Opt. Soc. Am. A **5**, pp. 2447-2457, 1987.
6. Chakraborty DP and Eckert MP: Quantitative versus Subjective Evaluation of Mammography Accreditation Phantom Images: Medical Physics 22 (2), 133-143, 1995.
7. Chakraborty DP: Computer Analysis of Mammography Phantom Images (CAMPI): an Application to the Measurement of Microcalcification Image Quality of Directly Acquired Digital Images, Medical Physics 24 (8) 1269-1277, 1997.
8. A.E. Burgess, "Comparison of ROC and forced choice observer performance", Medical Physics, **22**, pp. 643-655, 1995.

APPENDIX 2

PROGRESS IN BIOMEDICAL OPTICS AND IMAGING

Reprinted from

Medical Imaging 2001

Image Perception and Performance

21–22 February 2001
San Diego, USA

Proceedings of SPIE
Volume 4324

Anomalous Nodule Visibility Effects in Mammographic Images

Dev P. Chakraborty and Harold L. Kundel, University of Pennsylvania, Philadelphia

ABSTRACT

This study was undertaken to further investigate recent reports of unusual contrast-detail (CD) behavior in images with mammographic backgrounds, namely threshold contrast for detection increased with nodule size for Gaussian nodules. In this work we investigated the effects on the CD curve of allowing differently shaped nodules, in particular nodules with sharper edges. The following types of nodules/disks were studied: Gaussian shaped nodules and blurred disks, the latter characterized by a radius and an independent edge sharpness parameter. In a second type of disk the edge blur was held proportional to the disk radius. Ideal Observer detection thresholds were calculated for different nodule/disk radii ranging from 1.5 to 15 mm and observer performance studies were conducted. Noise power spectra (NPS) measurements confirmed the frequency dependence previously reported, NPS $\propto 1/f^{3.1}$. For the Gaussian nodules we confirmed the reported CD behavior, with threshold contrast $\propto$ radius$^{0.2}$. However, for the disk nodules with fixed blur edges we observed different behavior (larger objects required less contrast), with threshold proportional to radius $^{-0.28}$. For the disk nodules with variable blur the threshold contrast was almost independent of radius. In summary while we duplicated the reported CD diagrams for Gaussian nodules, different behavior was observed for nodules with edges. We conclude that in addition to considering the details of the noise, it is necessary to consider the signal properties in more detail.

Keywords: Mammography, nodules, detection threshold, ideal observer models.

1.INTRODUCTION

The threshold contrast for detection is defined as that contrast that yields a specified detection rate. Most imaging physicists are familiar with contrast-detail (CD) diagrams, which show the threshold contrast necessary for detection as a function of size of the object to be detected. In this paper we define the CD diagram to be a plot of the logarithm of threshold lesion amplitude vs. the logarithm of the radius. This plot typically has negative slope, as larger objects generally require less contrast to detect them. In fact, based on a Rose model [1] calculation for white (i.e., uncorrelated) noise one expects a slope of -1 on a log-log plot. The negative slope appears to be reasonable since one expects a larger object to produce a larger measured signal, thereby increasing the numerator of the signal-to-noise-ratio (SNR). Also the associated noise generally decreases with the size of the object to be detected, since more pixels are being averaged. It is less well known that the expected CD diagram depends critically on the nature of the background noise. It has been shown recently [2, 3] that if the noise power spectrum has an inverse power law dependence on spatial frequency, i.e., $1/f^{\beta}$ with $\beta > 2$, as is true for mammographic backgrounds, qualitatively different behavior was observed, namely the threshold contrast increased with lesion size.

Burgess' calculations [3] were made for a class of nodules termed 'scalable' in the sense that they could be characterized by functions of r/a, where **r** is the distance from the center and **a** is the nodule radius. For example, for Gaussian nodules the function is exp $(-r^2/2a^2)$. For scalable nodules he showed that the slope **m** of the CD diagram on a log-log plot was $m = (\beta-2)/2$. Note that with scalable nodules it is not possible to independently specify the sharpness of a nodule's edges. The scalability criterion represents an assumption that real nodules may not satisfy. An example is a nodule characterized by both a size parameter (the lateral extent of the nodule) and an independent edge-sharpness parameter, namely how rapidly the x-ray transmission changes near the edge. Since the human visual system tends to accentuate edges [4], the CD behavior could change significantly based on the presence or absence of such edges. The objective of this study was to test the effects of deviations from scalability, specifically to investigate the influence of including an independent edge parameter on nodule detection in mammographic backgrounds.

Presented in the following sections are details of the nodules and disks studied, the ideal observer methodology used to calculate threshold amplitudes and the noise power spectra measurement. These are followed by details of the observer experiments and the results.

Image Perception and Performance, Elizabeth A. Krupinski, Dev P. Chakraborty, Editors, Proceedings of SPIE Vol. 4324 (2001)

2.METHODS

2.1 ANALYTICAL

2.1.1 IDEAL OBSERVER MODEL: In order to understand human observer performance it has proven useful to study a variety of model observers, also called numerical observers, which are algorithms for calculating the SNR for specified signal and noise statistics [2, 5, 6]. The model observers differ from each other in how much use is made of information regarding the imaging task. In this work we used the Fisher Hotelling (FH) observer, sometimes called the pre-whitening matched filter, which takes into account knowledge of both the signal and the noise statistics, and the non-prewhitening matched filter with eye-filter (NPWE). Observer internal noise was neglected. Following Ref. 3 we describe a nodule profile by specifying its amplitude rather than its contrast, the two approaches being equivalent. Fourier methods were used to calculate the threshold amplitude for a given signal and noise [7] for the Fisher- Hotelling (FH) numerical observer. Two classes of cylindrically symmetric nodules were generated for this study: blurred disks (BD) and Gaussian nodules (GN) see Figure 1. The amplitude of the Gaussian nodule $S_{GN}(r, a)$ is defined by

$$S_{GN}(r, a) = A \exp\left[-\frac{r^2}{2a^2}\right], \quad \text{Eqn. 1}$$

where A is the peak nodule amplitude. The blurred disk amplitude function $S_{BD}(r, a, \sigma)$ is defined as a two-dimensional convolution of the Rectangle function $\Pi(r/2a)$ (= 1 if r < a and = 0 otherwise) and the Gaussian function $\phi(r, \sigma)$:

$$S_{BD}(r, a, \sigma) = A\, \Pi(r/2a) \otimes \phi(r, \sigma), \quad \text{Eqn. 2}$$

where σ is the edge-blur parameter and

$$\phi(r, \sigma) = \frac{1}{2\pi\sigma^2} \exp\left[-\frac{r^2}{2\sigma^2}\right]. \quad \text{Eqn. 3}$$

The Fourier Transform (FT) of a radially symmetric two-dimensional function is a Hankel transform. Using results from [8], the Hankel transform of the blurred disk amplitude function is given by

$$\tilde{S}_{BD}(f, a, \sigma) = A \frac{a J_1(2\pi a f)}{f} \times \exp[-2\pi^2 \sigma^2 f^2], \quad \text{Eqn. 4}$$

where J_1 is the Bessel function of order 1, and f is the radial frequency. The Hankel-transform of $S_{GN}(r, a)$ is given by

$$\tilde{S}_{GN}(f, a) = A\, e^{-2\pi a^2 f^2}. \quad \text{Eqn. 5}$$

In either case the SNR of the Fisher-Hotelling (FH) numerical observer is given by

$$SNR^2(A) = \int_0^\infty 2\pi f \frac{\tilde{S}^2(f, a)}{NPS(f)}\, df, \quad \text{Eqn. 6}$$

where NPS is the Noise Power Spectrum of the background and $\tilde{S}(f, a)$ is the relevant Hankel transform. The SNR is equivalent to the detectability index d' measured with a Two-Alternative Forced Choice (2AFC) experiment[9]. Since SNR is proportional to A, one has SNR(A) = A SNR(1). The threshold amplitude A_T, defined by $SNR(A_T) = 1.0$ is given by $A_T = 1/SNR(1)$. The integral in Eqn. 6 was evaluated numerically using 20th-order Romberg integration as implemented in the Interactive Data Language (IDL) development system (Research Systems Inc., Boulder, CO). All calculations were also repeated with the NPWE observer using the formula given in Ref. 3.

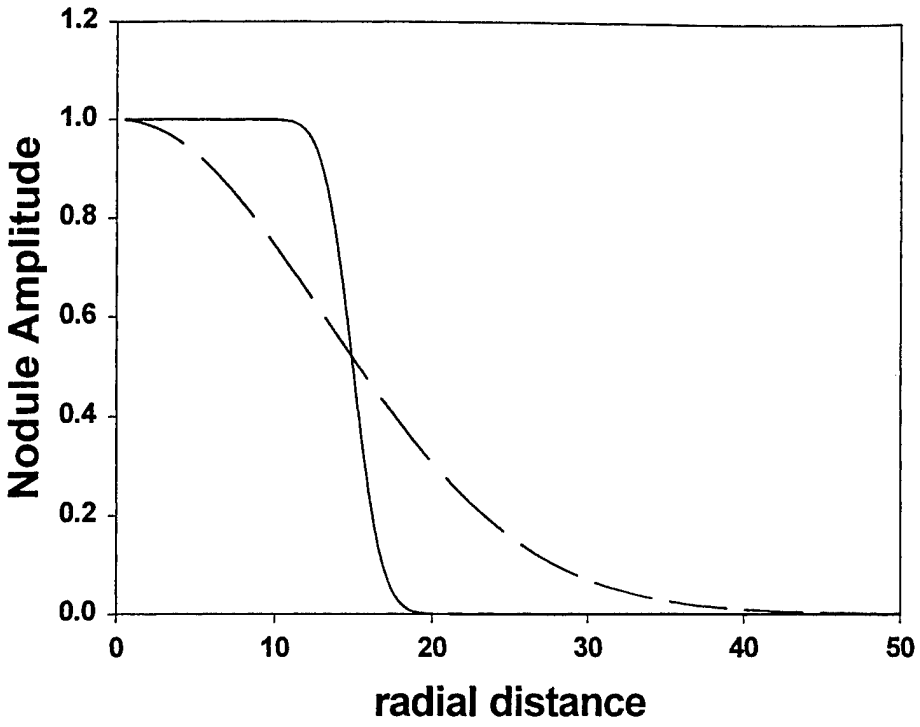


Figure 1: The solid line shows the radial profile of a blurred-disk nodule, with radius a = 15, edge blur parameter $\sigma = 1.5$, and the dashed line is for a Gaussian nodule with radius a = 13.

In addition to the blurred-disk model with fixed edge blur described above, we also studied nodules in which the edge-blur parameter was proportional to the nodule radius, $\sigma = 0.3$ a. These will be referred to as VB (variable blur) disks to distinguish them from the previously described FB (fixed blur) disks.

2.1.2 NPS MEASUREMENTS: The Noise Power Spectrum (NPS) of mammographic backgrounds was measured using 256 x 256 ROIs (regions-of-interest) randomly rotated, using nearest-neighbor interpolation, and extracted from digitized mammograms. The rotation served to average over any non-cylindrically symmetric asymmetry present in the images and to be consistent with the subsequent NPS analysis. A larger (by a factor of 1.414) square ROI was extracted around an operator chosen point in a relatively homogeneous region of the image. This ROI was randomly rotated around the center and the 256 x 256 ROI was then extracted. By choosing to rotate the larger ROI we eliminated the possibility that missing or invalid pixels would contaminate the smaller one.

We followed the Digital Fourier Transform (DFT) convention given in [10]. Before taking the DFT it is necessary to apply a window function to prevent leakage of power between frequency channels[11]. We used the Hanning window function H_{jk} ($0 < j,k < N-1$, where N is the number of pixels per square ROI-edge) defined by [12]:

$$H_{jk} = w_j \, w_k, \qquad \text{Eqn. 7}$$

and

$$w_j = \frac{1}{2} \left| 1 - \cos\left(\frac{2\pi j}{N-1} \right) \right|. \qquad \text{Eqn. 8}$$

The window normalization factor W_{ss} is defined by

$$W_{SS} = \sum_{j,k=0}^{N-1} w_j^2 \; w_k^2 \; . \qquad \text{Eqn. 9}$$

With these definitions the NPS, in units of mm^2, is given by

$$NPS = \frac{| \, DFT \, (H_{jk} \, I_{jk}) \, |^2 \, P^2}{W_{SS}} \; , \qquad \text{Eqn. 10}$$

where I is the average-subtracted ROI and P is the monitor pixel size, equal to 0.15 mm in our case. As a check on the normalization the sum of the NPS values divided by the area of the ROI, should yield the variance of the pixel values.

Finally a radial-frequency average was performed by averaging all pixels in the 2-dimensional NPS matrix at a fixed distance from the origin in frequency space. The measurement was repeated for 60 images, and the individual-image NPS functions were averaged. In this manner we determined the NPS (f) function of spatial frequency. The NPS function was fitted to the following empirically determined form (all logarithms are with respect to base 10):

$$\log (NPS \, (f)) = A + B \, f - \log (D + f^{\beta}) , \qquad \text{Eqn. 11}$$

where A, B and β were treated as fitting parameters and D was fixed at $(1/256)^3$ to prevent the NPS from going to infinity at zero frequency [3]. The NPS algorithm was also tested on simulated Gaussian noise images and images of an ACR phantom. In our experience the effect of the Hanning window was fairly significant. Without it the phantom NPS also showed unrealistic ~ $1/f^3$ frequency dependence, which disappeared on use of the filter.

2.2 OBSERVER STUDIES

2.2.1 NODULE GENERATION: Gaussian nodules of unit amplitude were generated using Eqn. 1. To generate blurred-disk nodules we convolved a Rectangle function of suitable radius with the Gaussian function defined in Equation 3. Six (6) nodule sizes with radii ranging from 1.5 mm < a < 15 mm were generated and the blur parameter σ was held at 5 pixels (0.75 mm) for the FB disks and 0.3 a for the VB disks. The display window (512 pixels square) was not large enough to accommodate the largest disk in the VB condition. We did not employ a look-up-table to linearize the display as recommended in Reference 3). Linearizing the display and physical imaging chain may be insufficient since a significant source of uncorrected non-linearity, the human observer, exists beyond it.

2.2.2 IMAGES: Fifty (50) screen-film mammograms were digitized (Lumisys LS-100) at 100-micron spot size for this study. Note that the digitizer pixel size is irrelevant to this study – it is the monitor pixel size (0.15 mm) that enters the formulae. Typically twenty to forty 512 x 512 ROIs were extracted from each image with the constraints that (1) the ROIs were separated by at least 100 monitor pixels (1.5 cm) in the horizontal or vertical direction, and (2) the ROIs were extracted from areas of reasonably uniform breast thickness. ROIs from different images were displayed in pairs on a 2048 x 2560 monitor (BARCO MGD-5, Clinton Electronics, Rockford, IL) driven by a DOME MD-5PCI display card (Dome Imaging Systems, Inc, Waltham, MA) using a Dell Optiplex computer running the Windows NT 4.0 operating system. A simulated nodule, as described above, was randomly superposed (i.e., added) at the center of one of the image-pairs. An algorithm using histogram analysis, applied separately to each image of the pair, was used to determine linear look-up-tables to display the images at optimal contrast and brightness. The region used to determine the histogram was confined to the central 256x256 region of the image.

2.2.3 READERS: Figure 2 shows the user interface for the human reader studies. The readers for this study were paid subjects (university students and staff) with good vision. They were told that a nodule was always present at the center of one of the image pairs. Their task was to indicate with a pointing device which side (left-or-right) was more likely to contain the superposed nodule. A modified QUEST algorithm [13] was used that utilized all preceding observer responses to

determine the amplitude of the nodule for the next presentation, to achieve a target detection rate of 92% (corresponding to a d' of 2.0). QUEST also calculated the maximum likelihood estimate (MLE) of the threshold amplitude and its uncertainty. The experiment was terminated after 512 image pairs had been viewed. The MLE value of the nodule amplitude at that point was defined as the threshold amplitude.

The readers were trained on 200 ROI pairs randomly drawn from the set described above, and these results were not scored. Following each response the software indicated if the reader's selection was correct or incorrect. The continuous feedback condition was maintained for the duration of the studies. A total of 6 readers performed the experiment for six nodule-radii and three nodule types (Gaussian, fixed and variable blur disks). The viewing distance was not restricted.

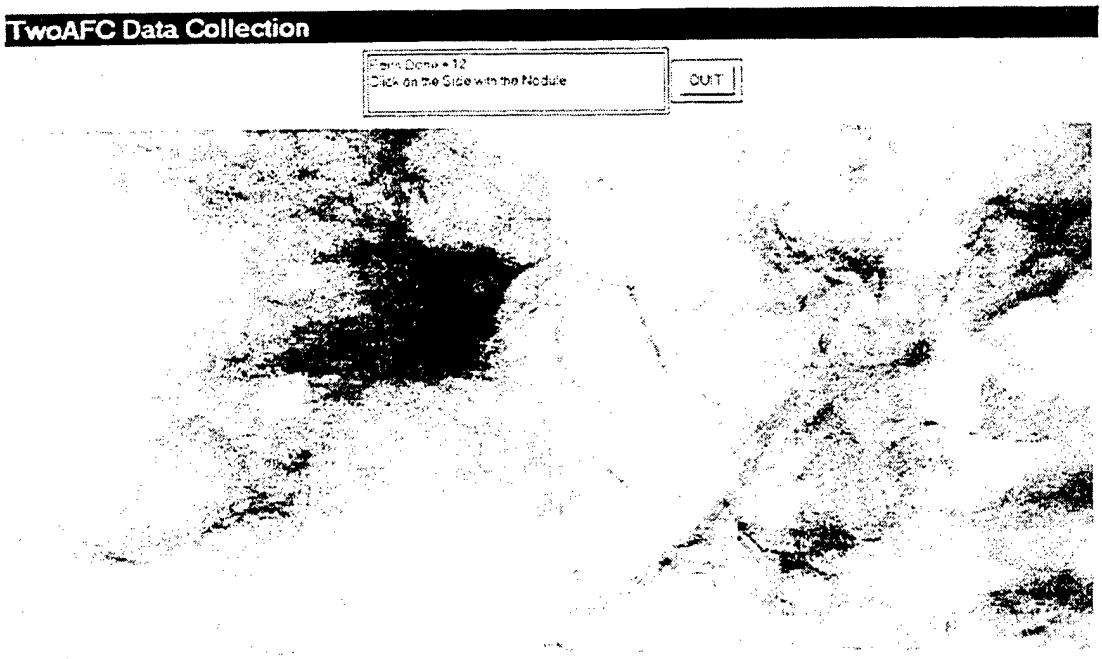


Figure 2: The user interface for the 2AFC human reader experiments. The two regions shown are 512 x 512 ROIs extracted from mammographic images. At the center of the left ROI is a 3 mm (20 pixel) radius blurred pixel disk with edge blur = 1 mm. Reference examples of the disks at three different contrasts were displayed below this window.

3.RESULTS

3.1 NPS Measurements

Figure 3 shows the NPS data plotted as $\log_{10}$(NPS) vs. frequency and the fitted function according to Eqn. 11. The fitting parameters were A = -2.0, B = 0.4 and β = 3.1. The parameter β is, of course, the exponent of the inverse power law dependence of NPS on frequency, i.e., if the constant D in Eqn. 11 is neglected, then for small frequencies NPS ~ $1/f^{\beta}$. All curve fitting was done using Sigma Plot software (SPSS, Chicago, IL). The fitted NPS function was used to calculate the threshold amplitudes of the nodules, as shown in Eqn. 6.

3.2 Analytical Threshold Results

3.2.1 Gaussian nodules: In Figure 4 the logarithm of the amplitude threshold is shown versus the logarithm of the radius in mm. This is seen to be similar to Figure 1 in Reference 3, the differences could be accounted for by the differing noise power spectra used. The peak threshold occurs at radius = 29.8 mm. Also shown is a straight line with slope 0.5, which is the limiting behavior at small radius. It is seen that for the Gaussian nodules our results confirmed the findings reported in the earlier study.

3.2.2 Blurred Disks: For the blurred disks with fixed edge blur of 0.75 mm, we found different behavior, as shown in Figure 5. For the range of nodule sizes studied, the threshold amplitude decreased as the radius increased, with an approximate $a^{-0.5}$ dependence. Finally, as shown in Figure 6, for the blurred disk with variable edges we found behavior similar to that for Gaussian nodules. The results shown here are for $\sigma = ka$, with $k = 0.3$. The peak occurs at radius $a = 49.5$ mm. For the NPWE numerical observer we obtained qualitatively similar results. The only substantive difference was that the thresholds were somewhat higher as compared to the FH observer. This is consistent with the fact that the NPWE observer is necessarily less efficient in detecting signals than the FH observer.

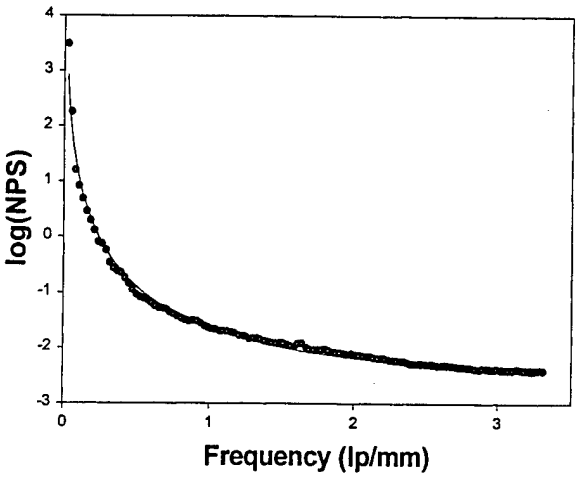


Figure 3: This shows Noise Power Spectra (NPS) results for mammographic backgrounds plotted as $\log_{10}$(NPS) vs. frequency. The solid line is the empirically determined fitting function described in Eqn. 11.

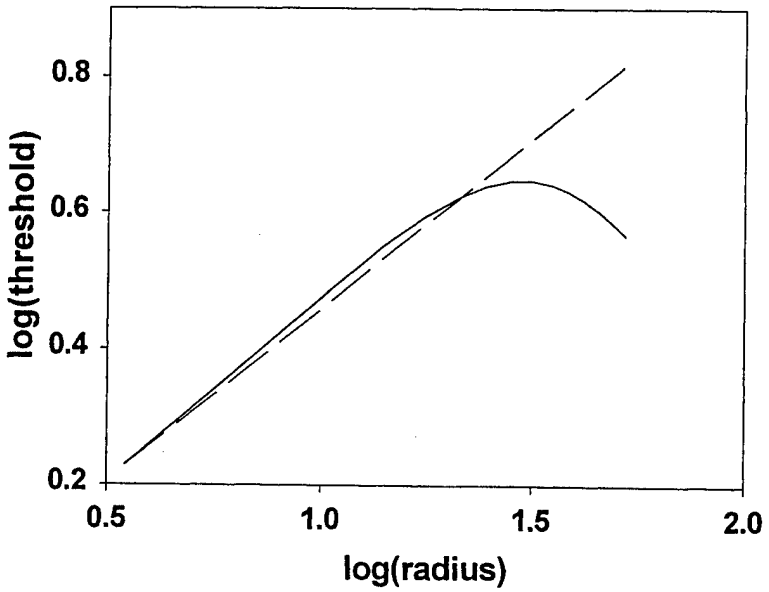


Figure 4: Contrast-detail plot for the Fisher-Hotelling numerical observer, showing the variation of threshold with radius (in mm) for Gaussian nodules. Note the increase in threshold with radius for radius less than 29.8 mm. Also shown is a straight line with slope 0.5, which is the predicted behavior at small radius.

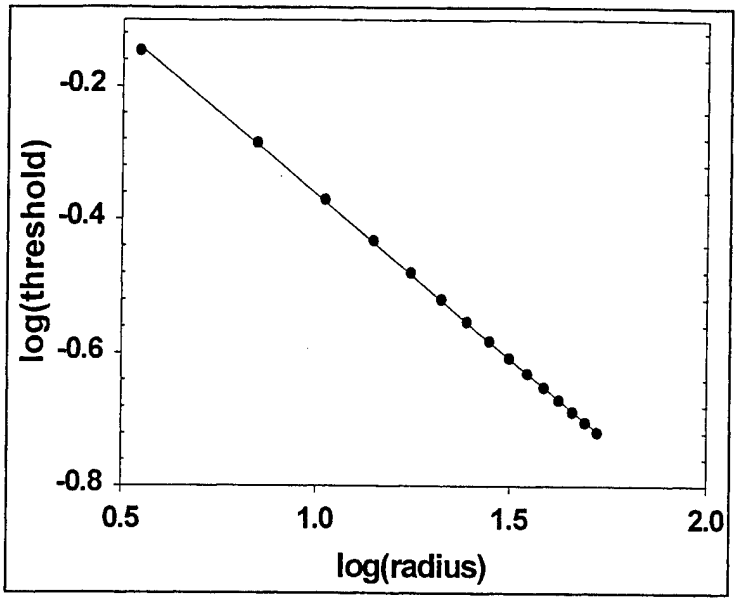


Figure 5: A contrast-detail plot for the Fisher-Hotelling numerical observer, showing the variation of threshold with radius (in mm) for fixed blur disks with $\sigma = 0.75$ mm. Note the uniform decrease of threshold with radius. The straight line is a regression fit to the data points with slope = -0.5, which is the predicted slope.

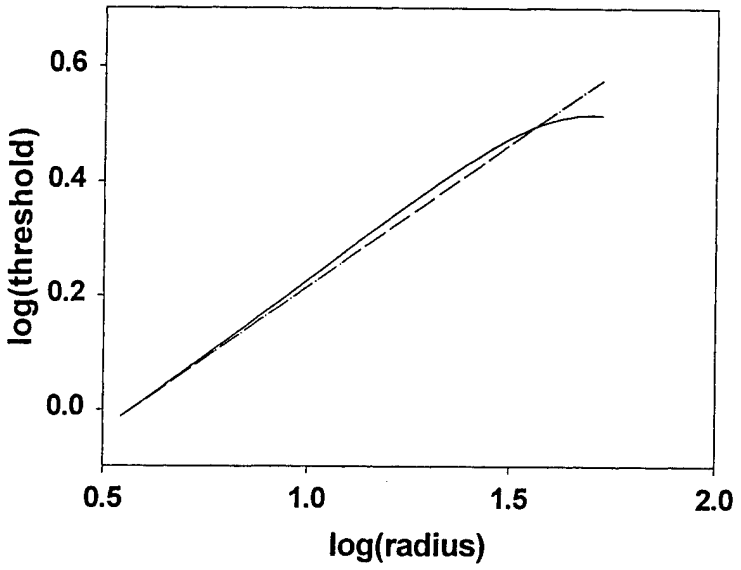


Figure 6: Contrast-detail plot for the Fisher-Hotelling numerical observer, showing the variation of threshold with radius (in mm) for variable blur disks with $\sigma = 0.3$ a. Note the resemblance of the behavior to that shown in Figure 4 for Gaussian nodules. The straight line, the limiting behavior at small radius, has slope = 0.5. The maximum occurs at a = 49.5 mm.

3.3 Human Observer Threshold Results

3.3.1 Gaussian nodules: Figure 7 shows the measured thresholds vs. radius, on a log-log plot, for the Gaussian nodules. The observed slope was 0.2 ± 0.09, which should be compared to the predicted value of 0.5. The standard deviation of each data point is about 10%, of which roughly 4% is the case sampling error, and the rest is inter-and intra- reader variability.

3.3.2 Disk nodules: Figure 8 shows the measured thresholds vs. radius for the FB disks with $\sigma = 0.75$ mm. The observed slope was -0.28 ± 0.06, which should be compared to the predicted value of 0.5. Similar results were observed for other values of σ in the range $0.075 < \sigma < 5.00$. As σ becomes smaller, the threshold amplitude drops. This is consistent with the notion that the increased edge information decreases the need for the observer to rely on amplitude to detect the nodule. For the variable blur disks the threshold data were observed to be constant to within the experimental uncertainty (5%), indicating that one is near the peak of the CD curve whose general form is shown in Fig. 4.

3.3.3 Efficiency: The human observer and ideal observer results can be combined to yield the efficiency of the human observer, defined as the square of the ratio of human and ideal observer d's[14]. This reduces to the square of the ratio of the ideal and human observer amplitude thresholds. Typically the efficiencies ranged from 10% to 20% for the FB disks, with the larger disks yielding smaller efficiencies, and 25%-80% for the Gaussian nodules, with larger nodules yielding greater efficiencies, and 15% to 50% for the variable blur disks, with the larger disks yielding greater efficiencies. The uncertainty in the efficiency numbers is considerable, about 30%.

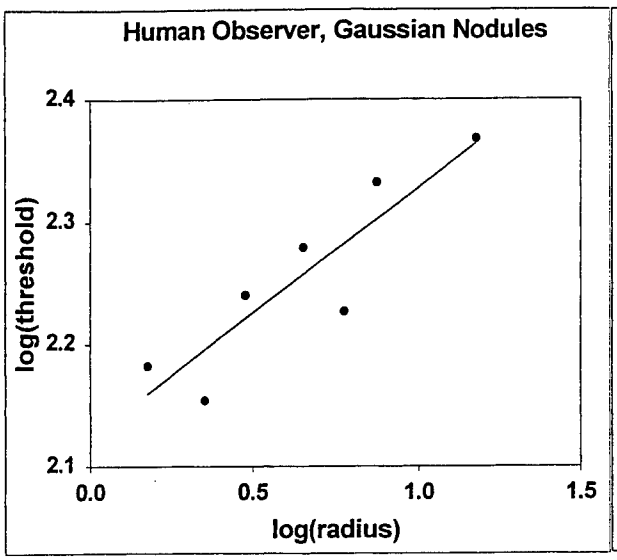


Figure 7: Contrast-detail plot for the average of six human observers showing the variation of threshold with radius for Gaussian nodules. The regression line has a slope of 0.20±0.09, which should be compared to the predicted value of 0.5 for the FH observer.

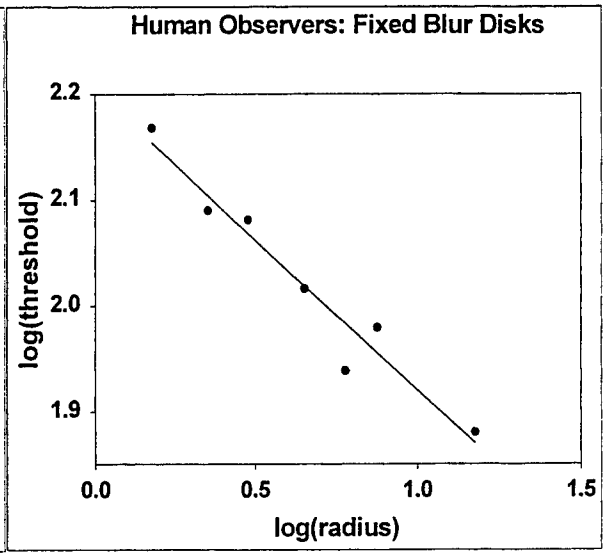


Figure 8: Variation of threshold with radius for the average of six human observers for fixed blur disks with σ = 0.75 mm. The regression line has a slope of -0.28 ± 0.06, which should be compared to the predicted value of -0.5 for the FH observer.

4.DISCUSSION

As indicated in the Introduction these experiments were undertaken to further investigate an unusual and significant result reported earlier in these proceedings. Our results confirmed the earlier findings for the Gaussian nodules. The NPS measurements, on an independent set of images, are in good agreement with the quoted range of exponents in the earlier studies. The fixed blur results can be understood by noting that as the disk becomes larger the edge information does *not* shift to lower spatial frequencies, as is true for Gaussian nodules. Therefore, the low frequency noise does not interfere with the edge information, leading to increased SNR and decreased threshold. While the shape of clinical nodules is open to question, one may get clues from the field of Computer Aided Detection (CAD) mammography. Researchers in CAD have found it necessary to include other independent features to characterize clinical masses, e.g., edge gradient, spiculation,

texture, etc. Each of these features could result in departures from the scalability assumption and a study of their effects is warranted. We have shown that the presence of an independent edge gradient can significantly change the CD diagram from that expected for scalable nodules, namely one can observe CD curves with negative, positive or near zero slopes as we have demonstrated. In spite of significant departures from the underlying assumptions (linearity and stationarity), the Fourier method appears to be able to qualitatively predict CD behavior (the value of the radius at the peak of the CD curve tends to be overestimated). Lastly, this study emphasizes the need to consider signal properties more accurately. The historical progression in human performance modeling has been from simple phantoms (i.e., flat backgrounds, white noise) to phantoms with more complex backgrounds (structured phantoms with power law noise) and a similar emphasis needs to be given to the complexity of the signal.

ACKNOWLEDGMENTS

We gratefully acknowledge several helpful comments by A. E. Burgess, PhD. This work was partially supported by a grant from the Department of Health and Human Services, National Institutes of Health, National Cancer Institute, RO1-CA75145.

REFERENCES

1. Rose, A., *Vision - Human and Electronic.* 1973, New York: Plenum Press.
2. Bochud, F.O., C.K. Abbey, and M.P. Eckstein, *Visual Signal Detection in structured backgrounds IV, Calculation of Figures of Merit for Model Observers in Non-Stationary Backgrounds.* J Opt Soc of Am A, 1999.
3. Burgess, A., F. Jacobson, and P. Judy, *On the detection of lesions in mammographic structure.* SPIE: The International Society for Optical Engineering, Medical Imaging Conference, 1999. **3663**: p. 304-315.
4. Barten, P.G., *Contrast Sensitivity of the Human Eye and its Effects on Image Quality.* 1999, Bellingham: SPIE Optical Engineering Press.
5. Wagner, R. and D. Brown, *Unified SNR analysis of medical imaging systems.* Physics Medicine Biology, 1985. **30**(6): p. 489-518.
6. Eckstein, M., Abbey CK, Bochud FO, *A Practical Guide to Model Observers for Visual Detection in Synthetic and Natural Noisy Images*, in *Handbook of Medical Imaging*, H.L. Kundel, J. Beutel, and R.L. Van-Metter, Editors. 2000, SPIE: Bellingham, Washington. p. 593-628.
7. Burgess, A., Li, X, Abbey CK, *Visual signal detectability with two noise components: anomalous masking effects.* Journal Opt. Soc. Am. A, 1997. **14**(9): p. 2420-2442.
8. Bracewell, R.N., *The Fourier Transform and Its Applications.* 2nd ed. 1986, New York: Mcgraw-Hill. 474.
9. Burgess, A., *Comparison of receiver operating characteristic and forced choice observer performance measurement methods.* Medical Physics, 1995. **22**(5): p. 643-655.
10. Cunningham, I.A., *Applied Linear-Systems Theory*, in *Handbook of Medical Imaging*, H.L. Kundel, J. Beutel, and R.L. Van-Metter, Editors. 2000, SPIE: Bellingham. p. 79-159.
11. Bendat, J.S. and A.G. Piersol, *Random Data: Analysis and Measurement Procedures.* second ed. 1986, New York: John Wiley & Sons. 566.
12. Press, W.H., *et al.*, *Numerical Recipes in C: The Art of Scientific Computing.* 1988, Cambridge: Cambridge University Press. 735.
13. Watson, A. and D. Pelli, *QUEST: A Bayesian adaptive psychometric method.* Perception and Psychophysics, 1983. **33**(2): p. 113-120.
14. Burgess, A.E., ed. *High level visual decision efficiencies.* X ed. Vision-Coding and Efficiency, ed. C. Blakemore. Vol. X. 1990, Cambridge University Press. 431-440.