25

Journal of Digital Imaging

VOL 7, NO 4 NOVEMBER 1994

A Method for Determination of Optimal Image Enhancement for the Detection of Mammographic Abnormalities

Derek T. Puff, Etta D. Pisano, Keith E. Muller, R. Eugene Johnston, Bradley M. Hemminger, Christina A. Burbeck, Robert McLelland, and Stephen M. Pizer

We present a paradigm for empirical evaluation of digital image enhancement algorithms for mammography that uses psychophysical methods for implementation and analysis of a clinically relevant detection task. In the experiment, the observer is asked to detect and assign to a quadrant, or indicate the absence of, a simulated mammographic structure characteristic of cancer embedded in a background image of normal breast tissue. Responses are indicated interactively on a computer workstation. The parameter values for the enhancement applied to the composite image may be varied on each trial, and structure detection performance is estimated for each enhancement condition. Preliminary investigations have provided insight into an appropriate viewing duration, and furthermore, suggest that nonradiologists may be used under this methodology for the tasks investigated thus far, for predicting parameter values for clinical investigation. We are presently using this method in evaluating several contrast enhancement algorithms of possible benefit in mammography. These methods enable an objective, clinically relevant evaluation, for the purpose of optimal parameter determination or performance assessment, of digital image-processing methods potentially used in mammography.

KEY WORDS: digital image display, image evaluation, contrast enhancement.

THE EMERGENCE OF digital mammography as a breast-imaging modality necessitates the development of methods of optimal electronic image display. For many years extensive research has been devoted to the optimization of film/screen mammography, a characteristically high-resolution imaging modality. Likewise, for digital mammography to be effective clinically, the methods by which digital mammograms are processed and displayed electronically must be evaluated to allow the human observer to acquire the greatest amount of clinically important information from the resulting images.

Image-enhancement algorithms applied to the digital mammogram before its display may enable a significant improvement in detection of structures indicative of breast cancer. For example, the results from an earlier pilot study conducted in our laboratory suggest that the application of a contrast enhancement method may be helpful in enhancing the features of breast cancers.[1] Several subtle cancers missed in the original interpretation were more visually obvious after the application of the algorithm contrast-limited adaptive histogram equalization (CLAHE).[2]

The task for any investigator or clinician who attempts to use or show a benefit from image-enhancement methods is to optimize the performance of the methods. Inevitably, all methods possess various parameters, and successful application of any technique relies on optimal settings of those parameters. In our attempt to evaluate the potential benefit in mammography of several contrast enhancement algorithms, we have developed a methodology for determining the optimal parameter values for these algorithms for enhancing the detectability of mammographic abnormalities, such as masses, calcifications, and spiculations.

In our initial attempts at determining optimal parameter values for CLAHE, which possesses

From the Departments of Biomedical Engineering, Radiology, Biostatistics, Psychology, and Computer Science, University of North Carolina-Chapel Hill.

Supported by National Institutes of Health Grants No. PO1-CA47982 and RO1-CA60193-01.

Address reprint requests to Etta D. Pisano, MD, Department of Radiology, CB No. 7510, 503 Old Infirmary Bldg, University of North Carolina Hospital, Chapel Hill, NC 27599-7510.

0897-1889/94/0704-0002$3.00/0

two parameters, radiologists were presented with a 5 × 5 grid displayed on a computer workstation. Each square in the grid contained a single simulated mammographic structure and surrounding background cropped from a digitized image of the ACR phantom (American College of Radiology Phantom, RM-161) and enhanced with CLAHE according to the parameter values (which spanned the entire range of relevant values) on the corresponding axes (Fig 1). The radiologists were asked to rate the visibility of the structures in the images. Their ratings were extremely diverse and subjective, apparently corresponding to their aesthetic judgments about the equally-enhanced noise in the images. Whereas this approach was useful for visualizing trends because parameter values were varied, an objective determination of the optimal position in the grid was difficult.

Because simple rating schemes were insufficient, we determined that a more rigorous preliminary investigation must be conducted to determine the optimal parameter values. This objective preliminary determination is important for several reasons. First, the extensive computation time required in the generation of an enhanced image, for most of the existent methods, prohibits an interactive adjustment of the relevant parameters. Second, because these algorithms enhance both the signal and noise in the image, an interactive determination of enhancement parameter values might be influenced by aesthetic judgments about the image and not dictated instead by optimal performance at a specified task. Thus, the determination of parameter settings calculated from the performance of several observers offers an objective, experimentally-proven assignment of the optimal values with respect to the task.

We have developed a method, described below, which uses psychophysical methods for implementation and analysis of a five-alternative forced choice (5-AFC) detection task, to achieve an objective and clinically realistic determination of the optimal parameter values for the enhancement algorithms under investiga-

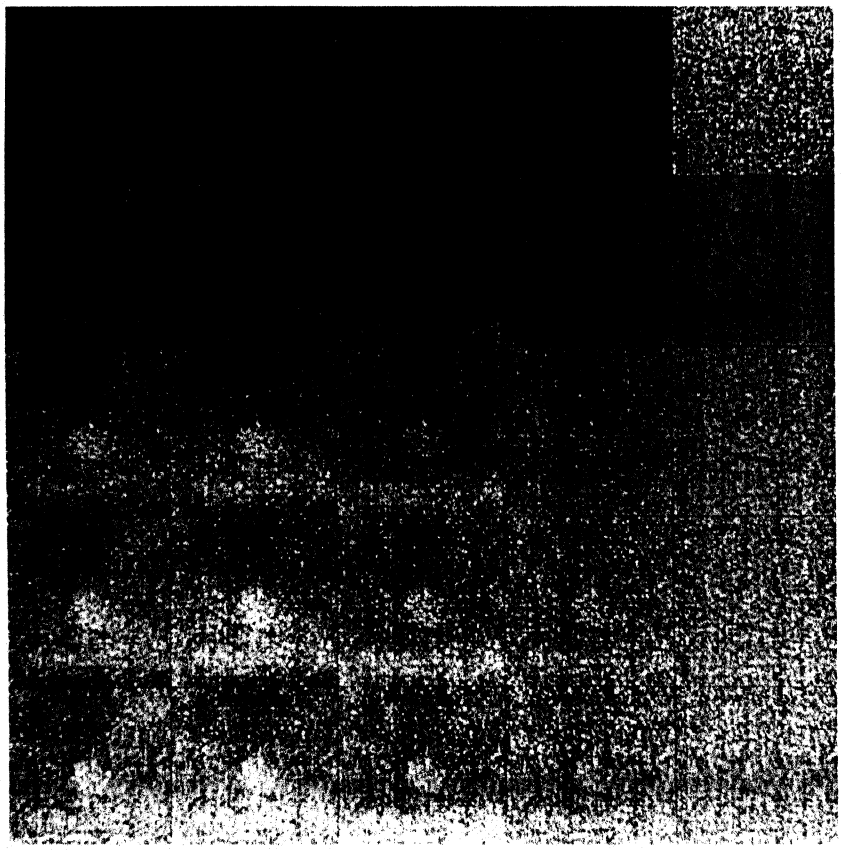

Fig 1. A 5 × 5 grid of CLAHE enhancement parameter combinations as applied to a mass and surrounding background cropped from a digitized image of the ACR phantom. The squares in the grid are enhanced according to the CLAHE parameter values on the axes; values to the right along the horizontal axis reflect decreasing contextual region size, whereas values toward the bottom along the vertical axis reflect decreasing contrast limitation.

tion. This approach incorporates the following emphases. First, it is important to make a determination of the optimal parameter values with respect to the visual task conducted in the interpretation of the image. The task of detection of abnormal mammographic structures may be improved by the application of the enhancement algorithms we are investigating, and thus, we have attempted to determine the parameter values to optimize detection. Second, real mammographic images are used as background to determine parameter values that are clinically applicable. Finally, it is also important to recognize that the optimal enhancement parameter values may be different for structures of different sizes and shapes; we have attempted to develop this method, and suggest structure simulations, for determining optimal parameter values for the enhancement of several different mammographic structures (masses, calcifications, and spiculations) of sizes typical of earlier breast cancer.

We report here a description of the proposed methods, along with several results integral to the construction of this methodology. Specifically, we conducted preliminary investigations to determine the appropriate viewing duration and acceptable observer population for use with these methods. We also measured possible training effects that may have occurred throughout the experiments.

In determining a viewing duration, we originally believed that a brief duration for the display of the images on each trial would be necessary to acquire an objective assessment of the effect of the application of contrast enhancement in making this detection. Furthermore, we believed the brief duration would allow us to measure the saliency of the effect by maximizing any differences among contrast-enhancement conditions. However, although mammograms are often viewed rapidly in busy screening contexts, a limited viewing duration is never explicitly imposed upon radiologists in the clinic. Thus, we endeavored to measure detection performance as the viewing duration was varied.

Secondly, recognizing that recruiting radiologists for observer studies is difficult and time consuming, we conducted an experiment to determine whether nonradiologists were predictors of the performance of radiologists as the enhancement conditions were varied. We believe that it is not important if nonradiologists perform worse in general than radiologists at this task. However, if the relative distributions of the two populations' scores are sufficiently similar, nonradiologists might be used as observers under this methodology for the purpose of determining enhancement parameters that might later be investigated clinically. Thus, we report the results of an experiment in which detection performance was measured for both radiologists and nonradiologists as the CLAHE enhancement parameters were varied.

Finally, we report data gathered using this methodology in a preliminary evaluation of a subset of the parameter space for the contrast-enhancement method CLAHE.

Any repeated-measures human-subjects experiment, particularly one that measures the effect of novel image presentations, is susceptible to training effects. The differences between conditions may shift because of practice with images that were initially less familiar or standard in appearance. Such effects are inevitable, and the best that any investigator can do is provide sufficient practice beforehand so that observers are well along the learning curve at the onset of the experiment. We report an analysis of training effects for radiologists and nonradiologists in one of the CLAHE-enhanced mass-detection experiments.

Whereas the results presented in this paper were obtained from an investigation of CLAHE, we believe that the methodology proposed here will be applicable in the evaluation of virtually any image-processing algorithm. Therefore, we do not make any claims about the efficacy of CLAHE or any other algorithm: we present what we believe to be a useful, relevant method for systematically investigating such image enhancements.

MATERIALS AND METHODS

Physical Specifications

The experiments were conducted on a Sun4 computer workstation (Sun Microsystems, Inc, Mountain View, CA). The observer was seated with his or her head positioned in a chin rest, such that the viewing distance was 38 cm. The images were presented and observer judgments were indicated with a mouse using an interactive windowing system (X Window System; Massachusetts Institute of Technology.

Cambridge, MA). The images presented on each trial occupied 11.57 degrees of visual angle, and were 256 × 256 pixels. The experiments were conducted in a darkened room. The overall luminance of the screen was 2.96 fL, and the images were displayed with 246 gray levels ranging from 0.001 to 11.91 fL. Perceptual linearization of the display was conducted essentially as described by Pizer.[3]

Observers

We conducted preliminary investigations with two observer populations: board-certified radiologists who are specialists in breast imaging, and nonradiologists. Exemption from human-subjects review was obtained from the Institutional Review Board of the University of North Carolina (UNC) Medical School. The observer was instructed to make his or her best guess as to the presence and location of the structure, and to indicate (with the appropriate button) that the structure was absent only when he or she was certain that the structure was not present in the image. Preliminary training was performed with a minimum of 25 practice trials before the experiment. This allowed us to acquaint observers, particularly radiologists, with these nonstandard images.

Images

The background images were 256- × 256-pixel images cropped from mammograms digitized at 50-μm spot size, 12-bit resolution (Lumisys laser film digitizer, Lumisys Inc, Sunnyvale, CA). The images contained relatively dense breast parenchyma. They were selected from mammograms known to be normal by virtue of a lack of mammographic or clinical evidence of malignancy for a 3-year period.

Mammographic masses were simulated by blurring (via convolution with a Gaussian kernel with a standard deviation of 2.0 pixels) a circle ~5 mm in diameter (1.51-degree visual angle at the 38-cm viewing distance). The masses were embedded by a pixelwise addition of the structure and background images (Fig 2).

The contrast of the mass then is defined to be the maximum gray level at the center of the mass before addition with the background. Normally, contrast is a unitless measure, formed by a ratio of foreground and background luminances or gray levels. However, because we perform pixelwise addition of the mass, which is nonuniform in intensity because of the blurring, with a background that is variable because of structure and noise, such a ratio would be difficult to formulate. Furthermore, our definition allows a dependent variable with discrete levels for use in the threshold determinations in the analysis.

Whereas we do not present results here of experiments using other structures, we have produced reasonable simulations of calcifications and spiculations. Calcifications were generated by creating a cluster of five 2- × 2-pixel rectangles positioned randomly within a 9-mm^2 (2.7-degree) region. Spiculations were generated with a 1-pixel–wide line 11 mm (3.32 degrees) in length, positioned in one of four orientations (0, 45, 90, 135 degrees). In all cases, although the simulated structures are not entirely realistic, they do possess the same scale and spatial characteristics of the actual mammographic structures for which we are trying to optimize digital enhancement and display.

Task Sequence

The observer initiated the trial by clicking with the mouse in the window. The observer was then presented with an image on a black background. His or her task was to indicate the presence of the embedded structure by clicking in one of the four imaginary quadrants in the image, or to indicate its absence by clicking in a window (with an indicative label) next to the image (Fig 3). Because the observer must select from one of five response options, this is a 5-AFC experiment. The image was presented for an interval specified by the designated viewing duration for the experiment (0.5, 2.0, 5.0 seconds, or unlimited). For experiments with a limited viewing duration, if the observer had not responded within the display duration, the image was erased and the subject had an unlimited time to make a decision on the blank screen (in which the quadrants were indicated with lines).

Feedback was provided on each trial in the form of a circle surrounding the structure (or an appropriate visual indication if the structure was not present), as well as with an audible signal. We believed it was important to always provide feedback for several reasons. First, in a threshold detection task, observers rarely see fully or clearly the target they are searching for. By circling the target briefly after each trial that it was present, the observer is continually

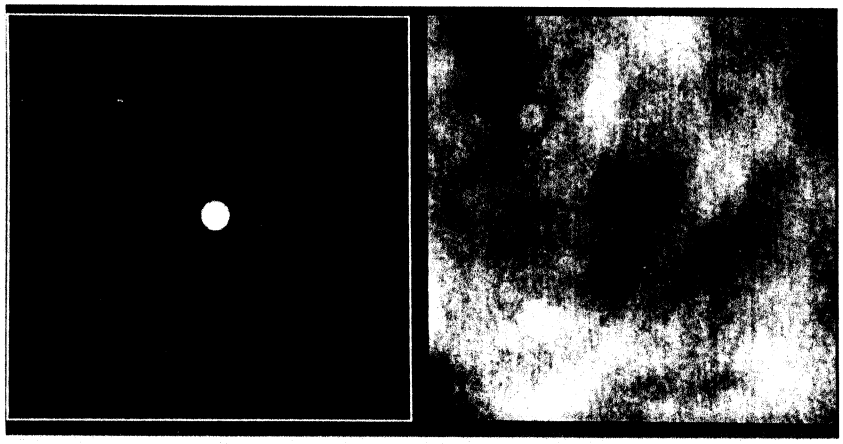

Fig 2. A simulated mass before (left) and after (right) insertion, in the upper-left quadrant, in a background image of real mammographic tissue.

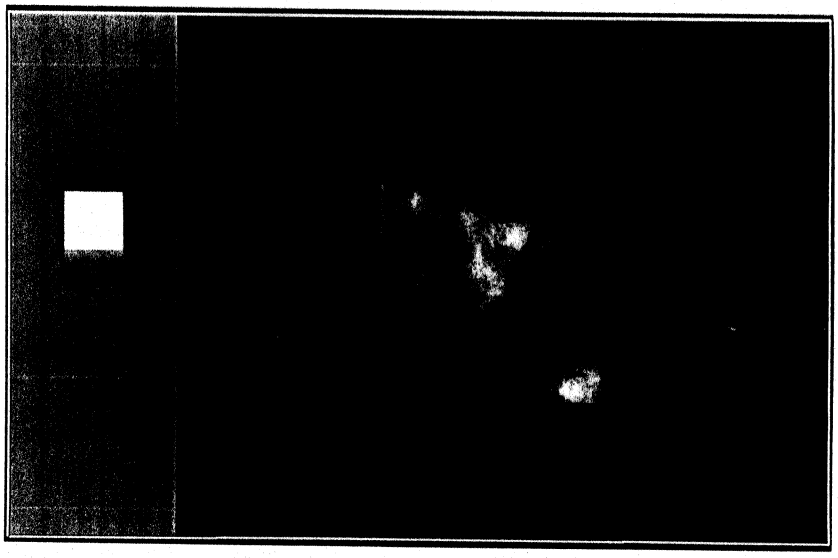

Fig 3. Experiment window on a computer workstation. The image (256^2 pixels, 246 gray levels) is presented on a black background in the center of the window. The simulated mammographic structure may be in one of the quadrants in the image. The button on the left containing the box icon is used to indicate the absence of the structure.

reminded of the appearance of what they are searching for. Similarly, observers can learn when they are repeatedly incorrectly selecting features present in the repeating backgrounds. Secondly, feedback provides motivation for the observers; without it they become uncomfortable about their performance or work at the task with less effort. Any training effects that might arise from continual feedback should occur in the practice sessions before the experiment.

An arrow cursor was used in the experiment window. Whereas a nondirectional, symmetric cursor probably should have been used, the zone around the borders of the quadrants in which masses could not be located should have prevented unintended responses arising from minor errors in mouse positioning. Furthermore, any misconceptions related to the directional mouse should have been eliminated through extensive training with feedback.

Design

For each trial, the background image was chosen randomly from among a list of 25. In those trials in which the structure was in fact present, the structure was embedded randomly in one of the four quadrants. To provide realistic variability in mass locations, the mass was also positioned within the chosen quadrant in a random manner. However, to eliminate ambiguity in the selection of quadrants, the structure was positioned in such a way that it was never directly adjacent to the quadrant borders. For each trial, the contrast of the structure was chosen randomly from among a list of predetermined contrast values chosen for the experiment.

Determination of a viewing duration. Using the described protocol, an experiment was conducted to determine the most appropriate viewing duration for this methodology. Nine contrast values ranging from 15 to 125 gray levels above background were chosen in a preliminary experiment as the simulated mass contrast values for use in this experiment. Mass-detection performance for two radiologists and four nonradiologists was measured in separate experiments using display durations of 0.5, 2.0, and 5.0 seconds. Four CLAHE contrast enhancement conditions were applied throughout the experiments in addition to unenhanced images. Later, an additional experiment using 7 radiologists and 16 nonradiologists was conducted with an unlimited viewing duration. An average of 10 observations per contrast per enhancement condition were collected for each observer.

Determination of the observer population. With the same protocol, an experiment was conducted to measure the performance of two observer populations as the CLAHE enhancement parameters were varied. The nine mass-contrast values chosen for the viewing-duration experiment were used here as well. Mass-detection performance for 7 radiologists and 16 nonradiologists was measured in a single experiment using four CLAHE contrast-enhancement conditions (as well as unenhanced) and an unlimited viewing duration. Again, an average of 10 observations were collected from each observer for each contrast in each enhancement condition.

CLAHE investigation. For the experiments determining the optimal parameter values for CLAHE, five contrast values were determined in a preliminary study, and are the values at the threshold contrast (60% correct) and 0.5 and 1.0 standard deviations in either direction, extracted from a fitted curve of the psychometric function for several observers (see Analysis for an explication of the relevant statistical methods in making this determination).

Five values of each of CLAHE's two parameters were studied. The parameter values were chosen to have logarithmic spacing and to span the entire range of possible values. A single block in the experiment consisted of 125 trials: each of the structure contrast levels was presented at each of the 25 enhancement parameter combinations. Furthermore, the structure was presented once in each of the five positions (four quadrants and absent) in each of the 25 background images. Each new block contained a newly randomized order for background image presentation, structure contrast and placement, and enhancement application. A complete experiment consists of over 20 blocks of data.

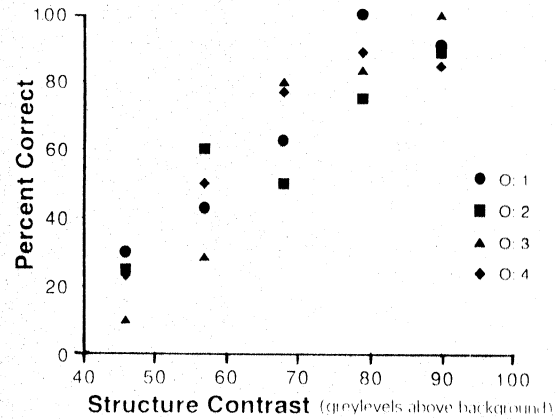

Fig 4. Psychometric function for four subjects depicting detection performance (percent correct) as a function of five structure contrasts. The data reflect detection performance for a simulated mass for a single CLAHE parameter combination. Structure contrasts are gray levels of the simulated structure above background before addition with a mammographic image.

Using this design, a preliminary experiment was conducted measuring simulated mass-detection performance for three nonradiologists for a 2-second viewing duration. For this experiment, only five blocks were run for most observers, giving five observations per contrast per enhancement.

Analysis

This design systematically varies structure contrast to derive a detection threshold, a measure of performance, for each enhancement condition. As we will show, the detection threshold is defined as that structure contrast corresponding to a chosen performance (percent correct) level. Comparisons of different processing parameters can be made by assessing the shifts in detection thresholds that they cause. An alternative to this approach might be to study detection performance for a single, predetermined contrast level. Whereas this would require fewer trials because only a single contrast was examined, it would be very difficult to determine, given the variability in both background image content and human performance, what single contrast to use.

From an experiment with a single observer, data were obtained in the form of a psychometric function relating detection performance (percent correct) as structure contrast was varied along a number of levels. Figure 4 shows example data collected from four observers, for a single enhancement parameter combination, for the experiment in which they were asked to detect simulated masses embedded in images enhanced with CLAHE. Classical sensory discrimination theory would predict that, because contrast values were varied from virtually imperceptible to highly apparent, a typical S-shaped curve would be expected to describe the data.[4] Specifically, detection performance for structure contrasts well-below threshold asymptotically approaches 20%: in a 5-AFC paradigm, observers will, by chance alone, correctly detect the structure one out of every five trials on the average. Likewise, performance for easily-detected contrasts asymptotically approaches 100%. The intermediate values, along which the theoretical detection threshold lies, represent the monotonically increasing transition between the two extremes. Data not resembling such a form indicate the presence of additional, uncontrolled factors in the experiment: the observer's task may not be well-defined or he or she may be adopting different or inconsistent strategies for detecting stimuli. Using the described method, with sufficient data we indeed consistently obtained psychometric functions possessing this standard, expected shape.

The data from each observer was described using probit analysis,[5] which models a proportion outcome (percent correct divided by 100) as a function of a continuous predictor (in this case, structure contrast). In probit analysis, one assumes that the function follows the cumulative Gaussian (normal) distribution, and the analysis yields estimates for the two parameters (location and shape), which completely determine a function that characterizes the data. The location parameter (labeled μ in subsequent discussion) indicates the inflection point of the sigmoidal probit curve and is effectively the detection threshold. The location parameter is in contrast units; as the value for this parameter increases, performance accuracy decreases. Similarly, large values for the probit shape parameter (σ) indicate a small (shallow) slope of the function. Fig 5 shows the probit curves for each subject's data from Fig 4.

Extensive exploratory analysis has led to a statistical model with the following characteristics. First, *log* contrast is the appropriate metric for the dependent variable, because the log data is linear in probit space, allowing the most accurate fit of the original data. Furthermore, log units are often the most appropriate metric for psychophysical observations. Secondly, whereas a distinct probit curve location parameter was estimated for each subject in each of the enhancement conditions, a single value (for all enhancements) for the shape parameter was estimated for each subject. A stable numerical solution was not possible in the statistical analysis of the data when attempting to fit a distinct σ for each enhancement and subject.

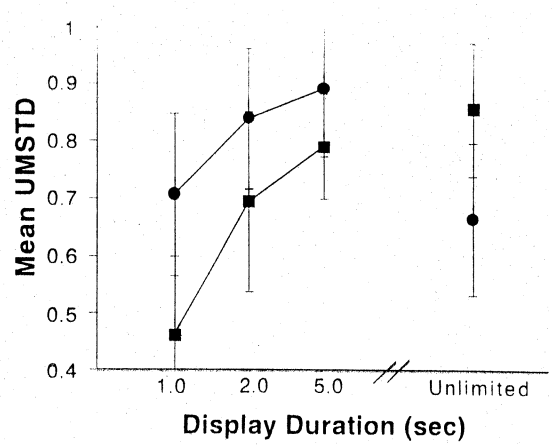

Fig 5. Probit curves corresponding to individual data from Fig 4. (■), radiologists; (●), nonradiologists.

This strategy has lead to the definition of the following response variable, a standardized inverse location measure we have labeled *umstd*.

$$\text{umstd}_{i,j} = (2 - \mu_{i,j})/\sigma_i,$$

where i is an index for the subject, and j is an index for the enhancement condition, μ is the probit curve location parameter, and σ is the probit curve shape parameter. Subtraction of μ from 2 inverts the function, enabling the intuitive representation of increased detection accuracy for larger values of *umstd*. μ is in units of log contrast and was always contained within 0 and 2, thus making 2 an appropriate value for inverting the function. The probit curve location values are standardized by the estimate of the subject's shape parameter, σ. This standardization was necessary because each subject possessed a different σ and the amount by which detection accuracy increases with an increase in structure contrast may be different for each subject. The purpose of these experiments is to measure the effect of contrast enhancement, and contrast enhancement will undoubtedly cause a shift in some direction in the structure-contrast threshold. Because of the different slopes, this shift from contrast enhancement may result in different accuracies for different subjects. Standardizing the probit curve location values by the subject's probit curve slope allows a common scale for the comparison of detection accuracy among the different contrast enhancement conditions. Thus, the *umstd* variable serves as a measure of detection accuracy.

Variance estimates and confidence intervals can easily be computed with the designs used here. All but one of the study designs involved at least ten subjects. Furthermore, considering the set of *umstd* scores from each enhancement separately simplifies the problem to the traditional one for independent observations. The combination of adequate sample size and independence imply that the usual formula for independent observations applies, and allows computing sturdy estimates of the variance of *umstd* separately for each enhancement. Heterogeneity of variance often occurs within subjects in behavioral studies. This motivated our choice of statistical tests and also makes trying to define and compute a within-subject variance problematic. Hence, in the results, we report only the between-subject standard deviations.

Repeated-measures analysis of variance (ANOVA) on the *umstd* scores was chosen to test differences between enhancements and populations. Enhancement constituted a within-subject factor (because each subject experienced all enhancements). Observer population constituted a between-subject factor (because each subject can be a member of only one population, radiologists or nonradiologists). Finally, viewing duration constituted a within-subject factor. (See Maxwell and Delaney[6] or Kirk[7] for thorough treatment of repeated measures ANOVA.) Such an analysis, a parametric analysis of continuous data, can be expected to be much more sensitive than any other approach for the same design, as it results in higher statistical power, and thus requires fewer subjects.

ANOVA provides, in addition to main effects (indications of the effect of manipulation of single variables), appropriate interactions. An interaction indicates the extent to which the relationship between the dependent variable and one independent variable are dependent upon a second (or more) independent variables. In attempting to determine whether nonradiologists are effective predictors of the performance of radiologists, it is essential to show that there is no interaction of the performance of the two populations as the enhancement is varied. If the results for the two observer populations exhibit the same relative distribution, where nonradiologists and radiologists are performing optimally at the same enhancement parameters and the results for the two populations vary similarly as the enhancement parameters are varied, then the two populations do not interact.

If we wish to show the equivalence of these performance distributions, we are burdened with showing adequate power, the probability of correctly detecting the alternative that there are in fact differences between the populations.[8] Consider plotting the average *umstd* value for each population separately, as a function of enhancement. Power calculations were performed assuming an X-shaped interaction (a pure interaction between subject type and enhancement), with a maximum *umstd* difference (at the endpoints of the X) corresponding to a 5% difference in proportion, correct between subject populations. This provides a worst-case scenario in that as one population improves performance, the other deteriorates. The data were analyzed with repeated-measures ANOVA techniques. Hence, the methods described by Muller et al,[10] were used. Those methods allow computing power for exactly the repeated-measures hypothesis of interest. Readers desiring a technical tutorial of the methods, or a copy of free software which implements them, should consult the reference.

Whereas we feel the described statistical approach is the most sensitive single analysis, it may be important to consider the extent to which a parameter combination allows false-positive selections (a response in one of the quadrants when the structure was not present). Additional discrimination amongst enhancements can be obtained with ROC analysis, the standard approach for such data.[9] In addition, the reader should recognize that other metrics may be preferred to the *umstd* variable in some settings. For example, $\mu + \sigma$ (from the probit curve) provides the

Table 1. Mean *Umstd* Scores for Radiologists and Nonradiologists for Three Different Viewing Durations

	0.5 s	2.0 s	5.0 s
Radiologists	0.462 (0.274)	0.695 (0.315)	0.791 (0.182)
Nonradiologists	0.707 (0.282)	0.841 (0.247)	0.892 (0.238)

Scores (and standard deviations combined across four CLAHE enhancement conditions, as well as unenhanced images) for three different viewing durations.

Table 2. Mean *Umstd* Scores for Radiologists and Nonradiologists for an Unlimited Viewing Duration

	Unlimited
Radiologists	0.857 (0.234)
Nonradiologists	0.747 (0.255)

Scores (and standard deviations combined across four CLAHE enhancement conditions, as well as unenhanced images) for an unlimited viewing duration.

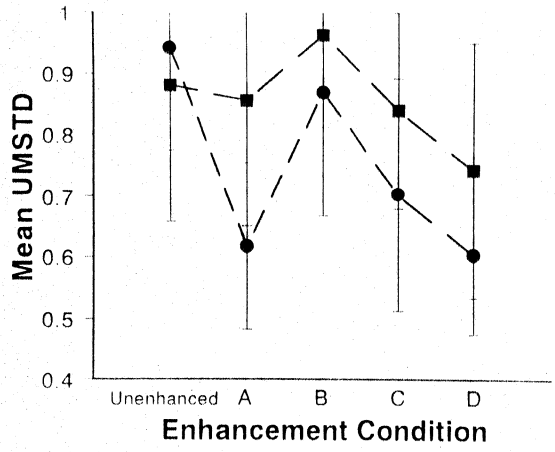

Fig 6. Mean *umstd* scores for radiologists and nonradiologists at three viewing durations and unlimited viewing duration. (■), radiologists; (●), nonradiologists.

stimulus value corresponding to ~87% correct in a five-alternative task. This, or some other higher percentile, may be preferred when interest lies in assessing performance at higher levels of detection, rather than in characterizing threshold performance.

A difficulty with the investigation of any enhancement method lies in systematically studying the large number of enhancements that may arise from the possible combinations of the method's parameter settings. A standard statistical solution involves study designs such as fractional factorials.[7] Such a design sacrifices the ability to test the highest order interactions among the parameters to reduce the number of treatment conditions that must be examined. In addition, assigning one (or more) treatment parameters as between subject factors, can dramatically ease the burden on each subject and simultaneously enhance the statistical sensitivity of the study. The range of possible enhancement methods as well as the alternative designs for evaluating each of them precludes detailed recommendations in any compact form.

RESULTS

Determination of a Viewing Duration

Repeated-measures ANOVA with viewing duration as the only within-subject factor was conducted on the *umstd* scores combined across five enhancement conditions for radiologist and nonradiologist observers. As mentioned in the Design section, two display duration experiments were conducted: data from the experiment examining display durations of 0.5, 2.0, and 5.0 seconds are presented in Table 1, whereas data from the separate experiment examining the unlimited display duration are presented in Table 2. The data are depicted graphically in Figure 6.

Notably, mean *umstd* scores for radiologists are poorer than those for nonradiologists, though not significantly so ($F = 1.02$, $P = .369$, at the 0.5-second viewing duration where the difference is greatest). With increased viewing time (from 0.5 to 5.0 seconds), scores for both populations improved, without interaction ($F = 0.82$, G-G $P = .4441$). The Greenhouse-Geisser epsilon (G-G P), a more conservative probability estimate applied in the analysis where appropriate, is a measure of the probability of rejecting the null hypothesis of equal means that incorporates a correction for unequal variances between the treatment conditions in a within-subjects design. At the unlimited viewing duration, there was also no significant difference between the mean for the two populations ($F = 0.94$, $P = .343$). Based on the trend noted at the first three brief viewing durations, as well as the comments of the radiologist participants regarding their standard viewing practices, we have decided upon an unlimited viewing duration for future experiments.

Determination of the Observer Population

Repeated-measures ANOVA with enhancement as the within-subject factor was conducted on the *umstd* scores for radiologists and nonradiologists. The data, in Table 3 and Fig 7, show mass-detection thresholds for the two observer populations at five CLAHE enhancement conditions (including unenhanced) with an unlimited viewing duration. Whereas the means for radiologists are greater than those for nonradiologists in four of the five enhancement conditions, pairwise comparisons indicated that there were no significant differences between the two populations in any of the enhancement conditions (see Table 4). The data indicate a similar

Table 3. Mean *Umstd* Scores for Radiologists and Nonradiologists for Four CLAHE Enhancement Conditions

	Unenhanced	CLAHE A	CLAHE B	CLAHE C	CLAHE D
Radiologists	0.881 (0.46)	0.856 (0.41)	0.964 (0.30)	0.840 (0.32)	0.743 (0.42)
Nonradiologists	0.942 (0.34)	0.618 (0.27)	0.869 (0.40)	0.703 (0.38)	0.604 (0.26)

Standard deviations are shown in parentheses.

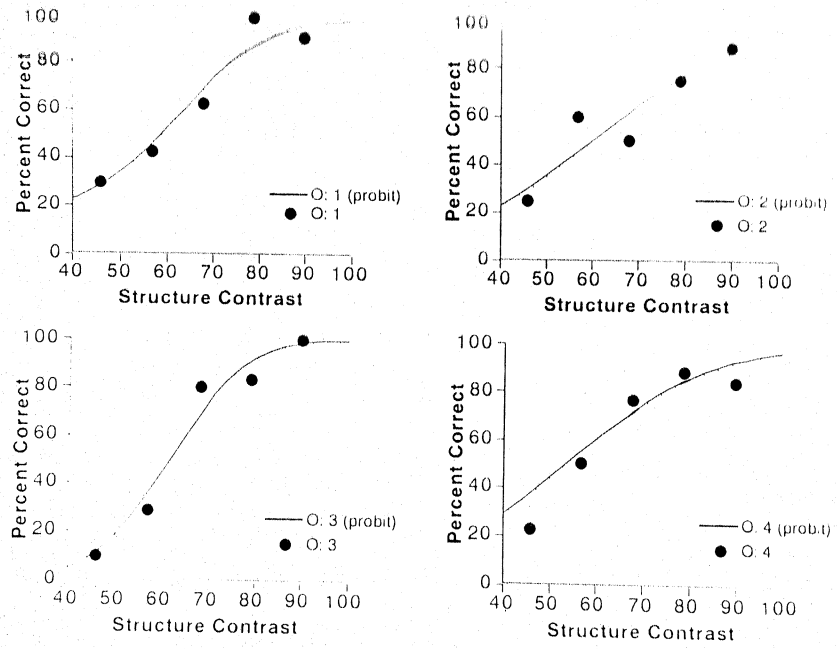

Fig 7. Mean *umstd* scores for radiologists and nonradiologists for four CLAHE parameter combinations and unenhanced images.

variation in the performance of the two populations as the enhancement condition is varied, and statistical calculations for the data indicate there is no significant interaction between the two populations ($F = 0.77$, G-G $P = .5248$). Power calculations, computed as described in the Analysis section, indicated a power of 83% to detect a subject population by enhancement interaction.

Training Effects

The data from the experiment in the previous section was analyzed in two sequential blocks to test if either subject population exhibited differential improvement in the enhancement conditions as a function of experiment block. For instance, one hypothetical scenario is that radiologists, less familiar with enhanced images than standard images, would show an improvement in detection performance over the course of the experiment for enhanced images. The data, represented as mean difference in *umstd* scores from block one to block two, are shown in Table 5. These almost uniformly positive difference scores indicate an improvement for both groups for almost all enhancement conditions. Figure 8 shows the mean *umstd* scores for the two observer types for the first and second experiment blocks, and shows again the increase in scores. Our analysis indicates that improvements in scores were not statistically significant. There was no main effect of experiment block; that is, there was no overall difference between the scores from block one and those from block two ($F = 1.2$, $P = .286$). Furthermore, we found no three-way interaction among experiment block, enhancement condition, and observer type ($F = 1.36$, G-G $P = .2594$), suggesting that the two observer populations were not exhibiting different kinds of training effects for different enhancement conditions. We should point out that whereas we have quite a few observers for these experiments, we actually had relatively few observations per condition for each observer. The means clearly indicate improvement for both groups during the experiment. With more observations, there might have been significant training effects.

Table 4. Pairwise Comparisons of Mean *Umstd* Scores for Radiologists *v* Nonradiologists

	Unenhanced	CLAHE A	CLAHE B	CLAHE C	CLAHE D
Radiologists – Nonradiologists	$F = 0.13, P = .72$	$F = 2.73, P = .11$	$F = 0.31, P = .58$	$F = 0.69, P = .41$	$F = 0.95, P = .34$

Scores are F and *P* values for four CLAHE enhancement conditions as well as unenhanced images.

Table 5. Mean Difference in *Umstd* Scores for Radiologists and Nonradiologists

	Unenhanced	CLAHE A	CLAHE B	CLAHE C	CLAHE D
Radiologists	0.08 (0.34)	0.36 (0.55)	0.40 (0.94)	−0.03 (0.48)	0.19 (0.32)
Nonradiologists	0.05 (0.78)	0.02 (0.60)	0.15 (0.70)	0.27 (0.56)	0.25 (0.67)

Values are the difference (block 2 − block 1) in *umstd* scores for four CLAHE enhancement conditions and unenhanced images. Standard deviations are shown in parentheses.

CLAHE Investigation

In an example application of this methodology, mean *umstd* scores were calculated for three nonradiologists across an entire grid of CLAHE parameter combinations. Whereas the data were collected for a viewing duration we have since eliminated, the results provide an indication of the methods of analysis and representation of data obtained with the methodology. An attractive mechanism for visualizing the detection performance data, at least for the instance of a two parameter algorithm like CLAHE, is a three-dimensional plot. Each of the points in the plot represents the detection performance corresponding to the parameter settings on the axes. Thus, the highest regions on the performance surface are, in theory, the parameter combinations producing superior detection performance (see Fig 9).

DISCUSSION

The ultimate determination of the efficacy of a particular image-processing algorithm is best made after the implementation of a clinical study using radiologists conducting real detection tasks. Because of practical constraints as well as observer biases, the parameter values for each of the enhancement algorithms investigated in such a study must be determined beforehand. We have proposed here a method for making that preliminary determination for mammographic images that uses a clinically relevant detection task, and eliminates the subjective judgments characteristic of other preference or ranking methods.

We investigated an appropriate viewing duration and observer population for the proposed methodology. We concluded that an unlimited viewing duration ought to be used. We also showed that nonradiologists were not significantly worse at the mass-detection task, and showed good power against any interaction among the two observer populations as the enhancement condition was varied. We believe that we have shown conclusively that nonradiologists may be used in determining the optimal parameter settings for image enhancement methods aimed at optimizing detection of structures in the image. For interpretation tasks that might benefit from the application of image processing, nonradiologists might not be useful for optimizing the parameters for using those methods. Furthermore, we do not suggest that nonradiologists could be used in the interpretation of real mammograms. Rather, our results suggest that preliminary parameter optimization experiments can be performed using nonradiologists. It is our intention to fully investigate with radiologist observers the efficacy of the algorithms computed with parameter settings

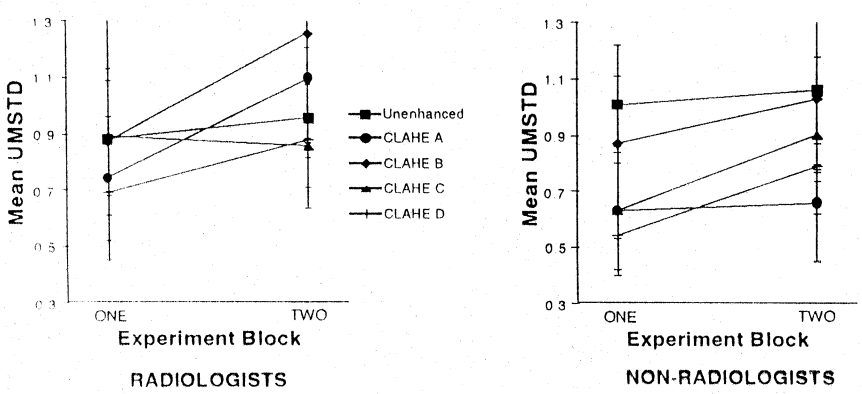

Fig 8. Mean *umstd* scores for radiologists and nonradiologists for five enhancement conditions as a function of experiment block.

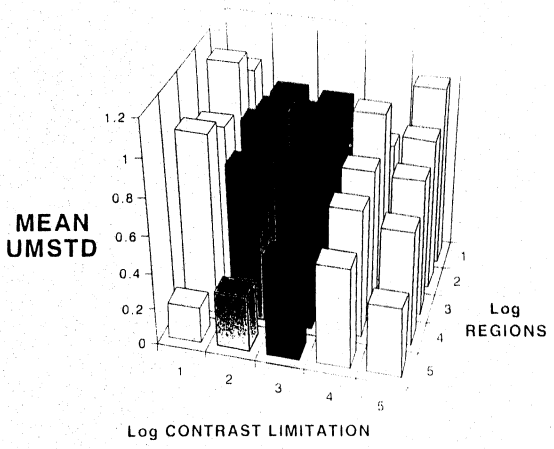

Fig 9. Mass detection performance (mean *umstd* scores) as function of combinations of the two CLAHE contrast enhancement parameters, contrast limitation and regions.

determined by these initial experiments. Given the number of potential parameter combinations that may exist for a multiparameter method, and the cost, in time and money, of using radiologists for this work, this initial exploration of parameter settings with nonradiologists may be particularly efficient.

Our measurement of potential training effects in the experiments highlights an important principle for any observer methodology. Conclusions drawn from the evaluation of proposed, novel image enhancements must be made with the confidence that the observers were sufficiently practiced with the new techniques so that substantive improvements did not occur in their performance during the experiment. It is only with extensive prior training that this phenomenon may be prevented.

We advocate the proposed methodology as a practical, thorough approach to image quality evaluation, and continue to apply it in our investigation of a number of enhancement algorithms of potential efficacy in mammography. This method does not require the acquisition of difficult cases, and subsequent verification of truth that image evaluation methods using real clinical cases entail. Realistic simulation of anomalous structures allows random, known placement of structures in an otherwise normal background. Similarly, to the extent that abnormal structures that must be detected in other regions of the body can be sufficiently simulated, this methodology may be applicable to other proposed image-processing algorithms of potential benefit in other imaging tasks.

The image enhancement methods developed recently, as well as those that will most certainly be developed in the future, offer great promise for enhancing digital image display and interpretation. The experimental paradigm described in this paper enables experimental determination of the parameter settings for these enhancement methods to ensure optimal performance in attempting to measure their clinical utility.

ACKNOWLEDGMENT

We wish to thank Douglas J. Taylor and Cheryl A. Roe, who contributed to the implementation of the statistical analysis. We also wish to thank the radiologist and nonradiologist observers who participated in the experiments.

REFERENCES

1. Pisano ED, Johnston RE, Pizer SM, et al: Computer enhancement of digitized mammograms. RSNA Annual Meeting Scientific Session, December 5, 1992

2. Pizer SM, Amburn EP, Austin JD, et al: Adaptive histogram equalization and its variations. Computer Vision, Graphics, and Image Processing 39:355-368, 1987

3. Pizer SM: Intensity mappings to linearize display devices. Computer Graphics and Image Processing 17:262-268, 1981

4. Corso JF: The Experimental Psychology of Sensory Behavior. New York, NY, Holt Rinehart, 1967

5. Finney DJ: Probit Analysis. Cambridge, UK, Cambridge University, 1971

6. Maxwell SE, Delaney HD: Designing Experiments and Analyzing Data. Belmont, CA, Wadsworth, 1990

7. Kirk RE: Experimental Design: Procedures for the Behavioral Sciences (ed 2). Belmont, CA, Brooks/Cole, 1982

8. Muller KE, Benignus VA: Increasing Scientific Power With Statistical Power. Neurotoxicol teratol 14:211-219, 1992

9. Metz CE: Basic principles in ROC analysis. Semin Nucl Med 8:283-298, 1978

10. Muller KE, LaVange LM, Ramey SL, et al: Power calculations for general linear multivariate models including repeated measures applications. J Am Stat Assoc 87:1209-1226, 1992