Improving the Detection of Simulated Masses in Mammograms through Two Different Image-processing Techniques[1]

Bradley M. Hemminger, PhD, Shuquan Zong, MS, Keith E. Muller, PhD, Christopher S. Coffey, PhD
Marla C. DeLuca, MS, R. Eugene Johnston, PhD, Etta D. Pisano, MD

Rationale and Objectives. The purpose of this study was to determine whether contrast-limited adaptive histogram equalization (CLAHE) or histogram-based intensity windowing (HIW) improves the detection of simulated masses in dense mammograms.

Materials and Methods. Simulated masses were embedded in portions of mammograms of patients with dense breasts; the mammograms were digitized at 50 μm per pixel, 12 bits deep. In two different experiments, images were printed both with no processing applied and with related parameter settings of two image-processing methods. A simulated mass was embedded in a realistic background of dense breast tissue, with its position varied. The key variables in each trial included the position of the mass, the contrast levels of the mass relative to the background, and the selected parameter settings for the image-processing method.

Results. The success in detecting simulated masses on mammograms with dense backgrounds depended on the parameter settings of the algorithms used. The best HIW setting performed better than the best fixed-intensity window setting and better than no processing. Performance with the best CLAHE settings was no different from that with no processing. In the HIW experiment, there were no significant differences in observer performance between processing conditions for radiologists and nonradiologists.

Conclusion. HIW should be tested in clinical images to determine whether the detection of masses by radiologists can be improved. CLAHE processing will probably not improve the detection of masses on clinical mammograms.

Key Words. Breast neoplasms, diagnosis; breast radiography; diagnostic radiology, observer performance; images, processing.

Effective image display allows for an improvement in the clarity of structural details. Mammography, especially in patients with dense breasts, is a low-contrast examination that might benefit from increased contrast between malignant tissue and normal dense tissue. Image processing may enable improved visualization of details (1). Our overall aim is to improve the accuracy of mammography through image processing, because at least 10% of palpable breast cancers are not visible with standard mammographic techniques (2).

Image-processing methods accentuate or emphasize particular objects or structures in an image by manipulating the gray levels in the display. A predetermined transformation is imposed to amplify the contrast between structures and modify the recorded intensities, enhancing visualization of important features on the displayed image

Acad Radiol 2001; 8:845–855

[1] From the Departments of Radiology (B.M.H., R.E.J., E.D.P.), Biomedical Engineering (S.Z.), and Biostatistics (K.E.M.. M.C.D.) and the UNC-Lineberger Comprehensive Cancer Center (E.D.P.), University of North Carolina, Chapel Hill, NC 27599-7510; and the Department of Preventive Medicine, Division of Biostatistics, Vanderbilt University, Nashville, Tenn (C.S.C.). Received February 5, 2001; revision requested March 20; revision received May 2; accepted May 7. Supported by NIH PO1-CA 47982, NIH RO1-65583, and DOD DAMD17-94-J-4345. **Address correspondence to** B.M.H.

(3). These methods are designed not to increase or supplement the inherent structural information provided by the image but simply to improve the contrast and, theoretically, to enhance particular characteristics (4). Intensity windowing (IW) is an image-processing technique whereby new pixel intensities are determined by means of a linear transformation that maps a fixed subrange of intensity values for the image onto the available gray level range of the display system (4). Histogram-based IW (HIW) is an extension of IW that dynamically determines a selected subrange for IW based on the histogram of intensity values for an individual image. Contrast-limited adaptive histogram equalization (CLAHE) modifies the intensity values of the image in a nonlinear fashion to maximize contrast at all pixels on the recorded image.

Many investigators have studied the application of digital image-processing techniques to mammography. McSweeney et al (5) tried to enhance the visibility of calcifications by using edge detection for small objects, but they never reported any clinical results. Smathers et al (6) showed that intensity band filtering could increase the visibility of small objects. Chan et al (7) applied unsharp masking to mammograms in a digital form and found that the detectability of microcalcifications improved. Chan et al (8) noted that the improvements may have been greater had the observers been trained to make diagnoses from processed rather than unprocessed (normal) mammograms.

Hale et al (9) used Photoshop software (Adobe, San Jose, Calif) to apply nonspecific contrast and brightness adjustment to digitized mammograms and found that radiologists were better able to determine the likelihood of malignancy for mammographically apparent lesions. Yin et al (10,11) showed that nonlinear bilateral subtraction is useful in the computer detection of mammographic masses. Kheddache and Kvist (12), using their own qualitative visual grading analysis, found improvements in the grading of masslike structures when they compared digitally acquired, processed mammograms with conventional mammograms. Wiebringhaus et al used phantoms to evaluate four different acquisition techniques. They found that techniques that used digital image processing (IW, unsharp masking, and regulatable edge enhancement) had more favorable receiver operating characteristic curves (13). Muramatsu et al (14) and Nakata (15) have investigated the parameters of the computed radiography machines used for mammography and found that certain parameter combinations are more favorable than others for the depiction of masses and microcalcifications.

Previous work at the University of North Carolina (UNC), Chapel Hill, has explored the use of IW and the adaptive histogram equalization family of algorithms in mammography and computed tomography (16–21). We have previously described a laboratory-based method to test the efficacy of image-processing algorithms in improving the detection of masses in dense mammographic backgrounds (22,23). With that method, on which our current work is based, radiologists and nonradiologists exhibit similar trends in detection performance. While nonradiologists (who are easier to recruit for such studies) did not perform as well as radiologists overall, the two populations displayed parallel increases and decreases in performance when image-processing methods were applied. Thus, nonradiologists may be used as an effective surrogate for radiologists when the effects of image processing on feature detection is being evaluated. The experiments described in this article were performed to determine whether CLAHE or HIW could improve the detection of simulated masses in dense mammograms in a laboratory setting.

MATERIALS AND METHODS

Both the experimental design and the data analysis for the two experiments are the same, as in several previously published studies; further details can be found in earlier publications (19–22). This same experimental paradigm was used for both the HIW and CLAHE studies and allows for the laboratory testing of a range of parameter values. The experimental subject is shown a series of test images that include an area of a dense mammogram with a simulated mass embedded in one of the four quadrants. The observer's task is to determine in which quadrant the mass is located, resulting in a four-alternative forced-choice task. The test images are displayed with masses inserted in different locations multiple times for each combination of different processing parameters and contrast values. The contrast of the object is varied from easy to impossible to detect. Figure 1 shows an example image from the CLAHE experiment with an inserted mass of medium contrast.

HIW Study

The processing algorithm for HIW is one developed locally at UNC by two of us (B.M.H., S.Z.). The algorithm uses peak-seeking methods to identify the "hump" of breast tissue in the histogram. Other humps corresponding to nonbreast tissue (markers, labels, etc) are

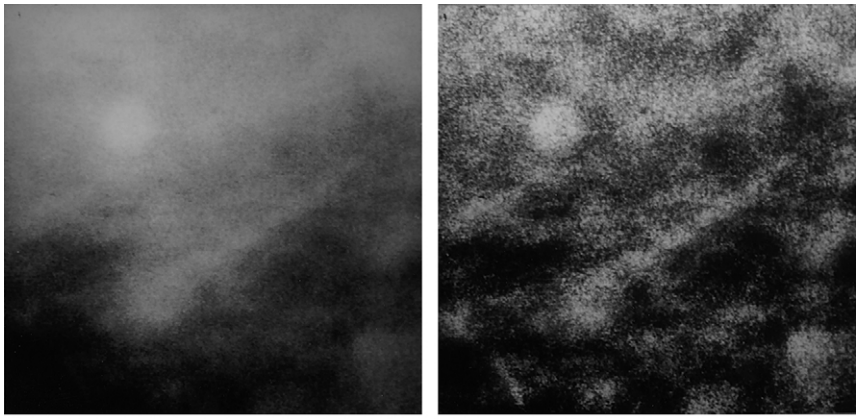


Figure 1. Example images from the CLAHE study with simulated mass inserted. **(a)** Unprocessed image. **(b)** CLAHE-processed image (clip level = 4, region size = 256 × 256 pixels). The mass is in the center of the upper left quadrant. The contrast of the mass is 90 digital driving levels (second smallest of four contrast levels).

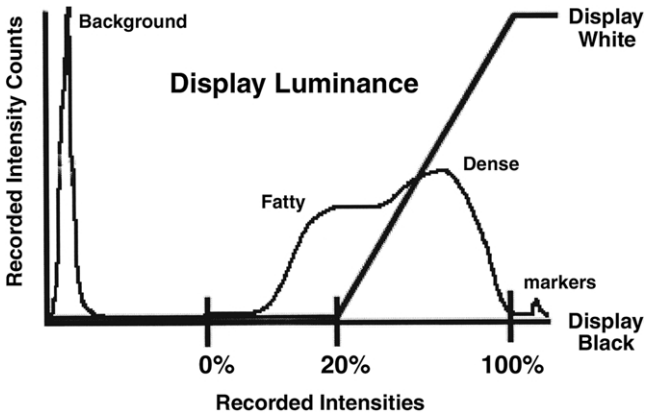


Figure 2. Histogram of recorded intensities for mammograms, showing percentiles for breast tissue and resulting IW mapping to display luminance (low-end HIW value is the 20th percentile).

discarded through heuristics. Once the hump of intensity values that matches breast tissue has been identified, percentiles are calculated across this region of the histogram. The resulting intensity window remapping is calculated on the basis of these percentiles. This mapping is shown in Figure 2.

The high end of the intensity window is always the 100th percentile, and the low end is chosen from four contending percentile values (ie, 20%, 35%, 50%, 65%). Only the single value of 100% for the high end of the intensity window range was considered because the pilot work showed no difference for values close to 100%. Values farther away from 100% were not considered because smaller values narrowed the IW range, resulting in images with too much contrast, while values larger than 100% resulted in images with too little contrast, becoming flatter and grayer. Unprocessed and IW-processed images were also evaluated in the HIW study. In this article, "IW-processed images" are images processed with a fixed (constant) intensity window range.

Figure 3 shows one of the study images processed under all six conditions. IW was evaluated as a positive control, since it has been previously shown to improve the detection of simulated masses on mammograms compared with unprocessed images, in a study that had the same experimental paradigm (20). A range of fixed intensity windows were evaluated in the prior study, and the one that performed best was chosen for evaluation in this study (20).

Forty background images (512 × 512 pixels each) were extracted from clinical screen-film mammograms digitized with a Lumisys digitizer (Lumisys, Sunnyvale, Calif) with a 50-μm pixel size and 12 bits (4,096 values) of contrast per pixel. The images were selected from craniocaudal or mediolateral oblique mammograms by a radiologist (E.D.P.) who is expert in breast imaging. Only areas that contained relatively uniformly dense tissue were included, with adjacent fatty areas specifically excluded. These areas were selected because they are most likely to hide soft-tissue masses in the clinical setting. The mammograms were known to be normal by virtue of at least 3 years of normal findings at clinical and mammographic follow-up.

Mammographic masses were simulated by blurring (via convolution with a Gaussian kernel with a spatial standard deviation of 6 pixels) a disk that is approximately 5 mm in diameter when printed on film (1.51° visual angle at a 38-cm viewing distance). Figure 1 shows a typical background image with the mass added to it. We used a simulated mass instead of real mass features so that we could control precisely the location of the

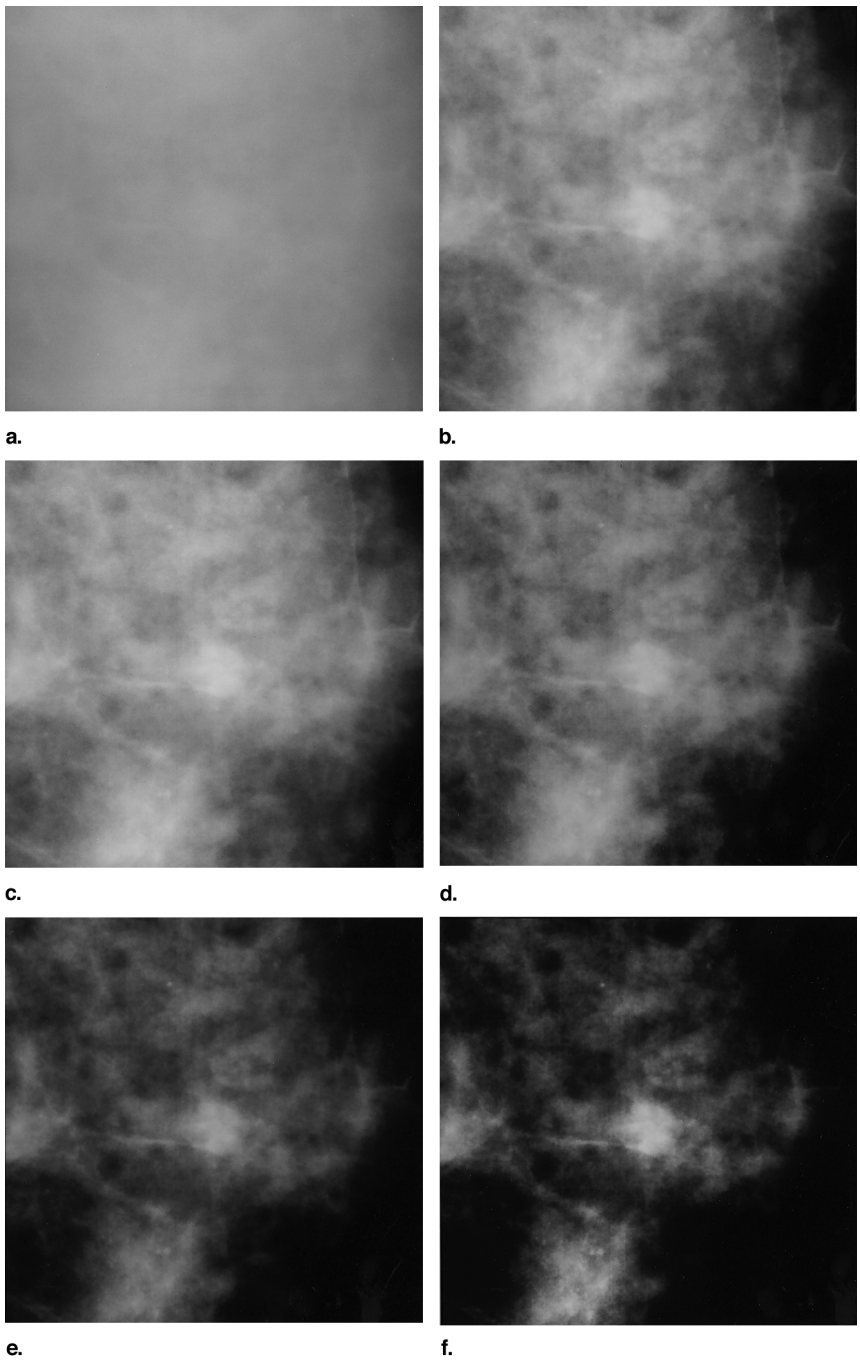


Figure 3. Example background from the HIW study, shown in all six processing conditions. Simulated mass is positioned in the center of each image (unlike in the experiment). **(a)** Unprocessed image has very low contrast since it is from a uniformly dense section of the mammogram. **(b)** IW processing based on the best results from the earlier UNC study of IW (20). **(c)** HIW processing with lower IW value set to 20%. **(d)** HIW processing with lower IW value set to 35%. **(e)** HIW processing with lower IW value set to 50%. **(f)** HIW processing with lower IW value set to 65%.

mass feature and the contrast between the feature and the background. The simulated mass target is identical to those in our previous experiments (18,20,22,23) to allow for comparisons. In this experimental paradigm the radiologists had indicated that while the simulated structures were not perfectly realistic, they had the same scale and spatial characteristics as actual masses typically found in mammography (18,22). Because the simulated masses are based on round disks, they are more representative of benign masses than malignant masses as seen clinically.

The masses were added at four fixed contrasts levels: 2.00, 2.88, 4.17, and 6.03. Contrast is defined as the change in luminance at the inserted feature location divided by the mean luminance of the background before insertion of the target ($\Delta L/L$), and thus the values have no units of measure. The contrast levels in this study

were determined through a separate pilot study. They were chosen to provide equal spacing on a logarithmic scale, covering the range of 30% detectable (just above 25% chance in a four-alternative forced-choice experimental paradigm) to 90% detectable (just above the 88% shoulder point on the probit curve) according to the predicted probit curves. The calculation for the HIW processing was based on the entire mammogram and then applied to the 512 × 512 pixel background with the mass target inserted.

Originally, there were 22 observers in the study, including 20 students and two radiologists from UNC. Because of concerns about differences in performance between radiologists and students, three radiologist observers were added after completion of the original study to verify the prior assumption of similar increases or decreases in performance for radiologists and nonradiologists (22). Because the inclusion of radiologists was not planned from the outset, there was the possibility of lack of power due to the imbalance between the two groups. Nevertheless, nonsignificant results for comparing radiologists and students would reinforce the assumption of parallel increases or decreases in performance for the two groups (22).

The digital images were printed onto standard 14 × 17-inch single-emulsion film (3M HNC laser film; 3M, St Paul, Minn) with a film printer (Lumicam; Lumisys). Each original 50-μm pixel was printed at a spot size of 80 μm, which produced film images of 4 × 4 cm (×1.6 enlargement). The background and target are magnified together. The radiologist observers in previous experiments with this same paradigm have reported that the presentation maintained the realistic appearance of the lesions and the mammographic backgrounds (18,22). Forty images were printed per sheet of film. The images were randomly ordered into an eight-by-five grid on each sheet of film. Both the film digitizer and film printer were calibrated, and the relationship between optical density on film and digital units on the computer was measured to generate transfer functions describing the digitizer and the film printer. To maintain a linear relationship between the optical densities on the original analog film and the digitally printed film, we calculated a standardization function that provided a linear matching between the digitizer and printer transfer functions (24). This standardization function was applied when the films were printed, to maintain consistency between the original optical densities of the original mammography film and those reproduced on the digitally printed films. The film printer produces films with a constant relationship between an optical density range of 3.35–0.13, corresponding to a digital input range of 0–4,095.

The order of the presentation of the stimuli was counterbalanced to eliminate any systematic effect of unimportant variables. Each observer scored 720 images, that is, 30 independent trials for each combination of 24 settings (six processing selections times four contrast levels). The experiments were conducted in our experimental laboratory, which is controlled for light, sound, and other distractions. The ambient light in the room was 4 lux. Film images were displayed on a standard mammography view box that was masked to exclude excess light. Observers were free to move and could use a standard mammography magnifying glass, if desired. The average viewing distance was 38 cm. Observers were dark adapted to the light levels of the experiment for 5 minutes before any readings. Observers were instructed to take breaks after each block of stimuli and more often if necessary. No time limits were imposed on the observers when they were viewing the test images.

CLAHE Study Materials and Methods

The CLAHE study used the same experimental setup, except for the following differences. There were 20 observers (all students). Ten processing settings were evaluated: unprocessed and nine CLAHE parameter selections (three-by-three grid). As with the HIW experiment, the calculation for the CLAHE processing was based on the entire image, with the result applied to the 512 × 512-pixel background image. The nine CLAHE settings were clip levels of 2, 4, and 16 combined with region sizes of 32 × 32, 64 × 64, and 128 × 128 pixels. Smaller clip levels result in less contrast enhancement, and larger values produce more enhancement (17). In this earlier experiment, the contrast of the mass stimulus was set to one of four fixed levels (60, 90, 135, and 200), defined in terms of digital driving level values of the display system instead of contrast to the underlying background, as in the HIW experiment. The display system was perceptually standardized so that the digital driving levels of the display corresponded to fractions of a just-noticeable difference. As in the HIW experiment, a pilot experiment was used to determine the four contrast levels. There were 10 processing choices combined with four contrast levels multiplied by 32 trials, resulting in 1,280 observations for each observer. The observers viewed 32 films of 40 images each.

Data Analysis for HIW Study

The relationship between $\log_{10}$ (contrast) and the probability that the observer will correctly identify the quadrant containing the mass can be described with a probit model. Fitting such a model requires the assumption that the relationship between $\log_{10}$ (contrast) and the probability of a correct response can be described by the cumulative Gaussian distribution. The location parameter, μ_{ij}, is the mean of the corresponding Gaussian distribution for the *i*th subject and *j*th processing condition. Processing conditions that improve detection will cause this parameter to be smaller, thus shifting the probit curve to the left. To make all values positive, we have added a value of 2 to the estimated values of μ_{ij}. Note that this is equivalent to multiplying the original contrast values by 100, which is consistent with the values used in the pilot study; this factor will cancel out when calculating differences of the θ values for two processing conditions. We assume a common spread parameter, σ_i, for all processing conditions on the same subject. The assumption of a common spread parameter makes sense, as it corresponds to an equal change in $\log_{10}$ (contrast) producing an equal change in perception, which is true for the display range of this experiment. Smaller values of σ_i correspond to steeper slopes, or greater increases in detection rates per unit change in $\log_{10}$ (contrast).

The model to be fit may be summarized as follows:

$$\Pr(\text{correct}) = ¼ + (1 - ¼)\Phi[(x_{ij} - \mu_{ij})\sigma_i^{-1}].$$

The formula gives the probability that a subject gets the correct answer. Here, *i* indexes subjects, and *j* indexes enhancements with x representing the $\log_{10}$ (contrast). The probit analysis will summarize the relationship between $\log_{10}$ (contrast) and proportion correct for each subject by processing condition. For comparisons among processing conditions, further analysis is required. For each observer and processing condition, $\theta_{ij} = \mu_{ij} + \sigma_i$ corresponds to that value of $\log_{10}$ (contrast) on the estimated curve for which the *i*th subject viewing the *j*th processing condition will achieve an 88% probability of giving a correct response. The difference between each subject's performance with unprocessed and processed images may be defined as follows:

$$\delta_{ij} = \theta_{iu} - \theta_{ij} = \mu_{iu} - \mu_{ij},$$

with larger (more positive) values of δ_{ij} indicating better performance under the processed condition. The difference scores for the five processing conditions were used as the outcomes in a repeated-measures analysis of variance with processing condition a between-subject factor. There were four hypothesis tests of primary interest: (*a*) test of no difference between the average differences across HIW conditions, (*b*) test for trend over HIW conditions compared with the unprocessed condition, (*c*) test of difference between each processed condition and the unprocessed condition (five tests), and (*d*) test of differences between the processed conditions (10 tests).

Each of the tests was conducted by using the univariate approach to repeated measures with the Geisser-Greenhouse test. To maintain an overall type I error rate of 0.10 for the experiment, the first two hypotheses were each tested at the 0.04 level. The five tests in the third set of hypotheses were tested at the 0.002 (0.01/5) level. Similarly, the 10 tests in the fourth set of hypotheses were tested at the 0.001 (0.01/10) level.

The study was not originally designed to include radiologists as observers. After the study was begun we decided to include radiologists to investigate the interaction between observer training (radiologists vs nonradiologists) and processing conditions. The motivation was to confirm our earlier results (21), which suggest that there is no interaction. To assess the effect of differences between radiologists and nonradiologists, the analysis plan was modified to include a test of interaction for each set of hypotheses. None of the tests showed significant differences, confirming that nonradiologist observers may be used as surrogates for radiologist observers in this and similar laboratory experimental paradigms. For this reason, results for radiologist and nonradiologist observers are not reported separately but are combined as a single observer pool for the remaining hypotheses.

Data Analysis for CLAHE Study

The CLAHE study used essentially the same methods of analysis to test a slightly different set of hypotheses: (*a*) test of interaction between CLAHE parameters of region size and clip level, (*b*) test of effect of region size on feature detection rate, (*c*) test of effect of clip level on feature detection rate, and (*d*) test of no difference in detection rate between CLAHE-processed and unprocessed images.

RESULTS

HIW Experiment

HIW hypothesis 1.—Our first hypothesis was that the average difference between unprocessed and processed

Table 1
Test of Trends across Choice of HIW Lower Percentile Values

Test	*F* Value	*df*	Den *df*	ε	*P* Value*
Average	12.24	1	24		.002
Trends	25.82			0.941	<.001
Linear	19.75	1	24		<.001
Quadratic	42.80	1	24		<.001
Cubic	19.87	1	24		<.001

*All differences were significant at the .04 nominal α level.

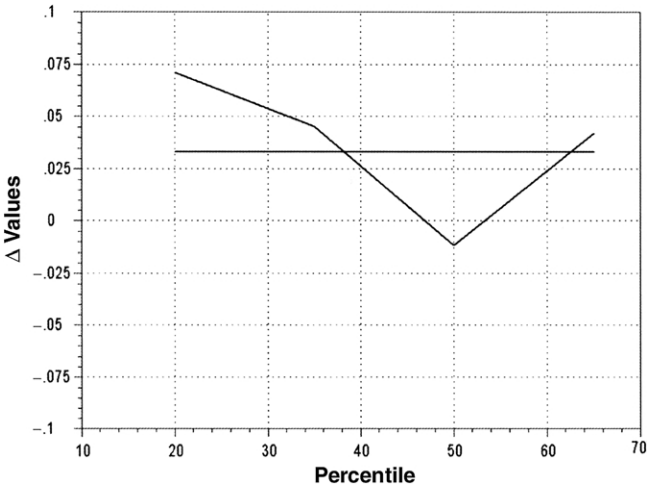


Figure 4. Estimated δ values (unprocessed − HIW-processed θ values) are shown for the four lower percentile choices of the HIW algorithm. Reference line corresponding to the *y* value of 0.033 indicates the performance of fixed IW, which was evaluated as a positive control. The *y* value of 0.0 corresponds to no difference compared with the unprocessed condition and is included for visual comparison.

conditions did not differ across HIW conditions. Table 1 displays the results from these tests. The test of average difference between processed HIW conditions and the unprocessed condition was significant at the .04 level ($P = .002$). Thus, observer detection performance was changed by the use of HIW processing.

HIW hypothesis 2.—This was a test for trend over HIW lower percentile values as compared with unprocessed images. We want to estimate which percentage ranges are most likely associated with best observer performance in feature detection. Results of an overall test for a trend associated with increasing levels of lower percentiles in HIW processing were significant ($P < .001$). Hence, a series of step-down tests was implemented to explore the trend further. Results of step-down tests for the linear, quadratic, and cubic trends were significant

Table 2
Mean Differences between Processed and Unprocessed θ Scores

Condition	Mean Difference ± Standard Deviation	*P* Value
Fixed	0.033 ± 0.013	.022
HIW		
20%	0.071 ± 0.010	<.001*
35%	0.045 ± 0.013	.0018*
50%	−0.012 ± 0.012	.306
65%	0.042 ± 0.013	.004

*These differences were significant at the .002 nominal α level.

($P < .001$ for all trends). Figure 4 presents these trends graphically. Performance was best at the 20% value, with the possibility of high performance in the ranges below 20% and above 65%.

HIW hypothesis 3.—This was a test of differences between each processed condition and the unprocessed condition (five tests). Table 2 displays two-sided *P* values for the five tests of no difference between processed and unprocessed conditions. When tested at the 0.002 α level, processing with HIW techniques at lower percentiles of 20% and 35% demonstrated significantly improved performance compared with the unprocessed condition ($P <$ .001, $=$ .0018, respectively). It should be noted, however, that the test for significant improvement for HIW with a lower percentile value of 65% (compared with the unprocessed condition) also yields very small *P* values. Furthermore, although it was not significantly different from 0, the parameter estimate for the difference between θ scores is negative for HIW processing with a lower percentile value of 50%, which corresponds to worse performance than with the unprocessed condition. The mean differences in θ scores for the processed conditions are illustrated graphically in Figure 4.

HIW hypothesis 4.—This was a test of differences between the processed conditions (10 pairwise tests). To determine which processing parameters to use clinically, we wanted to compare all the choices to see whether there was a single best choice or collection of best choices. Table 3 shows the two-sided *t* test *P* values for the 10 pairwise comparisons between the five processing conditions. The processing condition that uses HIW techniques with a lower percentile value of 50% is significantly worse than any of the other four processing conditions ($P < .001$, in all cases). HIW processing with a lower percentile value of 20% performed significantly

Table 3
Mean Differences between Processed and Unprocessed θ Scores

Comparison	Mean Difference ± Standard Deviation	*P* Value
IW vs HIW at 20%	−0.038 ± 0.010	.001*
IW vs HIW at 35%	−0.012 ± 0.009	.199
IW vs HIW at 50%	0.045 ± 0.007	<.001*
IW vs HIW at 65%	−0.009 ± 0.009	.324
HIW at 20% vs 35%	0.025 ± 0.008	.006
HIW at 20% vs 50%	0.083 ± 0.010	<.001*
HIW at 20% vs 65%	0.029 ± 0.010	.009
HIW at 35% vs 50%	0.058 ± 0.010	<.001*
HIW at 35% vs 65%	0.003 ± 0.009	.699
HIW at 50% vs 65%	−0.054 ± 0.010	<.001*

Note.—Larger differences correspond to better performance due to the processing condition on the left.
*These differences were significant at the .001 level.

better than fixed windowing ($P = .001$). There were no other significant differences among the other four processing conditions, although the comparisons between the HIW techniques with a lower percentile of 20% and the other HIW processing conditions all yielded small *P* values, as shown in Table 3.

CLAHE Results

CLAHE hypothesis 1.—This was a test of interaction between the CLAHE parameters of region size and clip level. The first question was whether region size and clip level were independent and could be studied separately. Table 4 includes tests for the effects of $\log_2$ (region size), $\log_2$ (clip level), and their interaction based on the univariate approach to repeated measures. The results of the test of no interaction were significant at the 0.04 level ($P < .0001$). Hence, a series of step-down tests was implemented to investigate the nature of the interaction. The test of a quadratic-by-quadratic trend had significant results ($P < .0011$), as did the test of linear-by-linear interaction ($P < .0001$). Thus, choice of region size depends on clip level, and vice versa. Because of the significant interaction between region size and clip level parameters in CLAHE, no further investigation of region size and clip level individually was performed (ie, hypotheses 2 and 3 were not tested).

CLAHE hypothesis 4.—This was a test of no difference between CLAHE processing and the unprocessed condition. To find out which of the processing conditions performed best, all of the conditions were compared with the unprocessed condition. Table 5 includes two-sided *t* test *P* values for the nine tests of no difference between processed and unprocessed conditions. Five enhancement conditions were not significantly different from the unprocessed condition; four were significantly different, demonstrating poorer performance.

Table 4
Mean Difference between CLAHE Processed and Unprocessed θ Scores by Region Size and Clip Level

$\log_2$ (region size)	$\log_2$ (clip level) 1	2	4
5	−0.014	−0.079	−0.155
6	−0.048	−0.052	−0.112
7	−0.0005	−0.028	−0.053

Table 5
Mean Differences between CLAHE Processed and Unprocessed θ Scores

Region Size (pixels)	Clip Level	Mean Difference ± Standard Deviation	*P* Value
32 × 32	2	−0.014 ± 0.056	.2730
64 × 64	2	−0.048 ± 0.059	.0018
128 × 128	2	−0.005 ± 0.052	.6640
32 × 32	4	−0.079 ± 0.047	.0001*
64 × 64	4	−0.052 ± 0.056	.0005*
128 × 128	4	−0.028 ± 0.060	.0516
32 × 32	16	−0.155 ± 0.069	.0001*
64 × 64	16	−0.112 ± 0.054	.0001*
128 × 128	16	−0.053 ± 0.072	.0035

Note.—Larger differences correspond to better performance. Negative values correspond to worse performance.
*These differences were significant at the .0011 level (ie, these processed conditions were significantly different from the unprocessed conditions).

DISCUSSION

The results of these analyses combined with the previous laboratory work that used the same experimental paradigm suggest several conclusions. First, for the detection of simulated masses on mammograms and similar lesion detection tasks, the choice of parameters for a particular image-processing algorithm can make the difference between improvement and degradation in detection performance. In the HIW and CLAHE studies reported here, as well as the previous CLAHE and IW studies (16,18–22), some parameter choices perform better than the unpro-

cessed condition for each processing method, while others perform worse. These evaluations are of the "best" parameter choices, that is, many less desirable ones have already been discarded through comparison in pilot studies. Thus, it is very important to verify the proper choice of parameters for image-processing algorithms applied to medical images.

Second, we considered which algorithm of those tested performs best in the detection of simulated masses on mammograms. Some results were consistent across the current HIW and CLAHE experiments, our previous IW experiments (19–21), and the prior CLAHE experiment (22). The best results from each of the processing methods, suggest the following relationship:

$$\text{HIW} > \text{IW} \geq \text{unprocessed} = \text{CLAHE}.$$

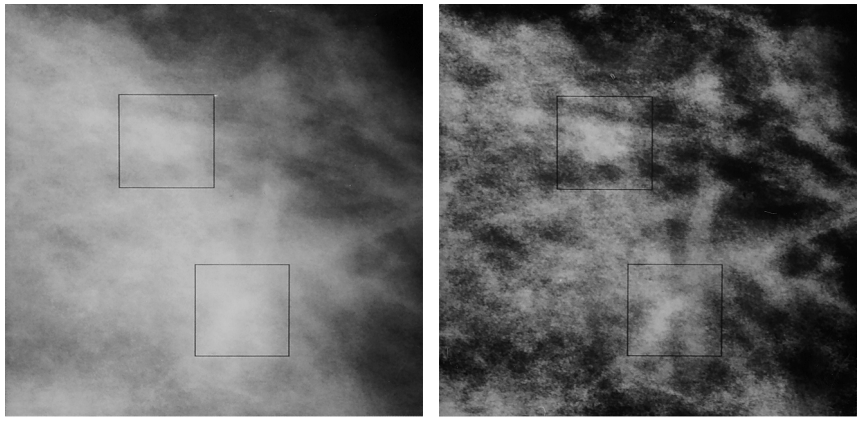
In the prior CLAHE study, the best CLAHE performance was the same as that for unprocessed images when the images were displayed on video (one parameter combination was slightly better but not significantly so). In this study, which used film image display, results were similar, with five parameter settings performing the same as the unprocessed condition and four performing worse. In the prior IW study we found IW to be significantly better than unprocessed images (20). In the current study, in which we used an improved definition of contrast, we again found IW to perform better, but not significantly better. We did find an HIW condition that performed significantly better than both IW and unprocessed imaging.

The results of these studies suggest that processing techniques such as HIW have the greatest potential to improve detection rates, but as with all image-processing techniques, the benefit of HIW is not uniform and is strongly dependent on the choice of the lower percentile value. These results suggest that the benefits of this HIW technique are at a minimum somewhere in the middle of the conditions considered here (ie, around the 50% percentile choice for the low end of the intensity window range). The best results were observed with the 20% lower percentile for the low end of the intensity window range. The trend analysis found trends through cubic order in these data, suggesting the possibility of further maximums at the extreme values (close to 0% and 100%) for the low-end percentile. Our practical experience and the pilot work for this experiment suggest that a maximum benefit occurs around 20%–40%, with minimum benefit at the extremes (0% and 100%). We have found that low-end percentile values near 0% caused a flatter, lower-contrast image, while low-end percentile values near 100% result in a black-and-white image with little to no contrast in the midrange. Therefore, while the data from this experiment suggest that choices of percentile values below 20% or above 65% could further improve detection ability, on the basis of the pilot work and our other experiences with HIW, we do not expect much improvement in those areas.

We also found no statistically significant interaction between observer type (radiologists vs nonradiologists) and laboratory feature detection, under any of the processing conditions, individually or as a whole. This further confirms a result previously described by Puff et al (22). The conclusion is strengthened by the fact that we reached the same results with a different display medium (film vs video) and a different processing method (HIW vs CLAHE). This finding is important because the use of radiologist observers is a limiting factor in many medical imaging studies. The ability to use nonradiologist observers as surrogates for radiologist observers in laboratory feature detection experiments considerably expands the scope of possible experiments. An especially important application of nonradiologist observers is to reduce a large set of processing methods and parameters choices to a small set as the preliminary step for a clinical evaluation with radiologists.

Finally, no combination of the CLAHE parameters studied resulted in improved detection of masses on mammograms, which was consistent with the findings of Puff et al (22). The most likely explanation is that any improvement in mass conspicuity was generally negated or outweighed by the increase in false-positive results caused by other structures that were processed with CLAHE and mistaken for positive masses. An example is seen in Figure 5, which shows one of the backgrounds with the highest false-positive rate when processed with CLAHE. Highlighted in boxes are the inserted simulated mass (lower right), which is not easily discernible, and the structure that with CLAHE enhancement has a masslike appearance (upper left).

For CLAHE parameters, there is an interaction between region size and clip level, making it difficult to isolate their effects, but there seemed to be an effect with each parameter. First, lower amounts of contrast enhancement (ie, smaller clip levels) performed better. The best performance was for a clip level of 2. A clip limit of 1 is the smallest amount of enhancement, with a value of less than 1 equivalent to not changing the image. Since none



a. b.

Figure 5. **(a, b)** Images from the CLAHE study. Boxes at upper left contain the "false-positive" structure that was commonly mistaken for a mass. Boxes at lower right contain the real mass target at the lowest level of contrast. **(a)** Unprocessed image. **(b)** CLAHE-processed image (clip level = 4, region size = 8 × 8 pixels).

of the CLAHE processing combinations outperformed the unprocessed condition and since the best-performing combinations were with a clip level of 2, it may be that detection with the CLAHE-processed images was generally inferior and was only close in performance when the clip level was set so low that the processed image was minimally different from the unprocessed image.

For region size, it was expected that smaller regions, approaching the size of the mass, would narrowly encompass the mass and optimize the enhancement of the mass feature against its background. The best-performing region sizes, however, were the largest evaluated (128 × 128 pixels), suggesting that larger regions were more appropriate, perhaps because the smaller regions did not enclose the masses and immediate surround well, or (as with clip level) because the observers performed better when the CLAHE-processed images were most similar to the unprocessed images (as region size increases in CLAHE the difference between processed and unprocessed images decreases).

In conclusion, we believe that of the techniques compared (HIW, IW, CLAHE), HIW would be the most valuable for the mass detection task in mammography and should be tested clinically to determine whether its use improves radiologists' diagnostic performance. The best parameter choices for our HIW implementation are the percentile values of 20% for the low end and 100% for the high end of the intensity window (calculated just over the breast tissue portion of the histogram). Recent work in our laboratory has evaluated HIW in clinical settings, with use of the best-performing parameter settings identified in this study. A multicenter clinical trial found that HIW had the best overall performance of the nine processing methods tested with digital mammograms for radiologist preference and performance evaluations, when averaged across scanner type and feature type (25,26). These results suggest that HIW may be effective for presenting other mammographic features (microcalcifications and spiculations) and that it may be as effective for digital mammograms as it is for digitized screen-film mammograms. In the HIW experiment, no significant interaction was found between observer type (radiologist or nonradiologist) and feature detection performance, confirming our earlier result that nonradiologist observers can be used as surrogates for radiologist observers in laboratory feature detection tasks. With CLAHE processing, the detection of masses on mammograms was not improved; the best results were achieved with low amounts of enhancement (small clip levels) and larger regions, both of which make the processed image more similar to the original unprocessed image. Further clinical study of CLAHE processing was not indicated.

REFERENCES

1. Rosenman J, Roe CA, Cromartie R, et al. Portal film enhancement: technique and clinical utility. Int J Radiat Oncol Biol Physics 1993; 25: 333–338.
2. Homer MJ. Mammographic interpretation: a practical approach. New York, NY: McGraw-Hill, 1991; 4–5.
3. Pizer SM. Psychovisual issues in the display of medical images. In: Hoehne KH, ed. Pictorial information systems in medicine. Berlin, Germany: Springer-Verlag, 1985; 211–234.
4. Jain AK. Fundamentals of digital image processing. Englewood Cliff, NJ: Prentice Hall, 1989.
5. McSweeney MB, Sprawls P, Egan RL. Enhanced image mammography. AJR Am J Roentgenol 1983; 140:9–14.
6. Smathers RL, Bush E, Drace J, et al. Mammographic microcalcifications: detection with xerography, screen-film, and digitized film display. Radiology 1986; 159:673–677.
7. Chan HP, Doi K, Galhorta S, et al. Image feature analysis and computer-aided diagnosis in digital radiography. I. Automated detection of microcalcifications in mammography. Med Phys 1987; 14:538–547.

8. Chan HP, Vyborny CJ, MacMahon H, et al. Digital mammography ROC studies of the effects of pixel size and unsharp-mask filtering on the detection of subtle microcalcifications. Invest Radiol 1987; 22:581–589.
9. Hale DA, Cook JF, Baniqued Z, et al. Selective digital enhancement of conventional film mammography. J Surg Oncol 1994; 55:42–46.
10. Yin F, Giger ML, Vyborny CJ, et al. Comparison of bilateral-subtraction and single-image processing techniques in the computerized detection of mammographic masses. Invest Radiol 1993; 28:473–481.
11. Yin F, Giger M, Doi K, et al. Computerized detection of masses in digital mammograms: analysis of bilateral subtraction images. Med Phys 1991; 18:955–963.
12. Kheddache S, Kvist H. Digital mammography using storage phosphor plate technique-optimizing image processing parameters for visibility of lesions and anatomy. Eur J Radiol 1997; 24:237–244.
13. Wiebringhaus R, John V, Muller RD, Hirche H, Voss M, Callies R. ROC analysis of image quality in digital luminescence radiography in comparison with current film-screen systems in mammography. Aktuelle Radiologie 1995; 5:263–267.
14. Muramatsu Y, Nawano S, Anan M, et al. A study of image processing in CR mammography; gradation processing. Jpn J Clin Radiol 1990; 23:271–276.
15. Nakata M. A study of image processing condition of mammography using FCR. Nippon Acta Radiol 1989; 49:454–467.
16. Hemminger BM, Johnston RE, Muller KE, et al. Comparison of clinical findings between intensity-windowed versus CLAHE presentation of chest CT images. SPIE Medical Imaging VI: Image Capture, Formatting, and Display 1992; 1653:164–175.
17. Pizer SM, Zimmerman JB, Staab EV. Adaptive grey level assignment in CT scan display. J Comput Assist Tomogr 1984; 8:300–305.
18. Puff DT, Cromartie R, Pisano ED. Evaluation and optimization of contrast enhancement methods for medical images. Proc SPIE Visualization in Biomedical Computing Conference 1992; 1808:336–346.
19. Pisano ED, Zong S, Hemminger BM, et al. Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms. J Digit Imaging 1998; 1:193–200.
20. Pisano ED, Chandramouli J, Hemminger BM, et al. The effect of intensity windowing on the detection of simulated masses embedded in dense portions of digitized mammograms in a laboratory setting. J Digit Imaging 1997; 10:174–182.
21. Pisano ED, Chandramouli J, Hemminger BM, et al. Does intensity windowing improve the detection of simulated calcifications in dense mammograms? J Digit Imaging 1997; 10:79–84.
22. Puff DT, Pisano ED, Muller KE, et al. A method for determination of optimal image enhancement for the detection of mammographic abnormalities. J Digit Imaging 1994; 7:161–171.
23. Hemminger BM, Dillon AW, Johnston RE, et al. Effect of display luminance on the feature detection rates of masses in mammograms. Med Phys 1999; 26:2266–2272.
24. Hemminger, BM, Johnston RE, Rolland JR, Muller KE. Introduction to perceptual linearization for video display systems for medical image presentation. J Digit Imaging 1995; 8:21–34.
25. Pisano ED, Cole EB, Hemminger BM, Yaffe MJ, et al. Image processing algorithms for digital mammography. RadioGraphics 2000; 20: 1479–1491.
26. Pisano ED, Cole EB, Major S, et al. Radiologists' references for digital mammographic display. Radiology 2000; 216:820–830.