Does Intensity Windowing Improve the Detection of Simulated Calcifications in Dense Mammograms?

Etta D. Pisano, Jayanthi Chandramouli, Bradley M. Hemminger, Marla DeLuca, Deb Glueck, R. Eugene Johnston, Keith Muller, M. Patricia Braeuning, and Stephen Pizer

This study attempts to determine whether intensity windowing (IW) improves detection of simulated calcifications in dense mammograms. Clusters of five simulated calcifications were embedded in dense mammograms digitized at 50-μm pixels, 12 bits deep. Film images with no windowing applied were compared with film images with nine different window widths and levels applied. A simulated cluster was embedded in a realistic background of dense breast tissue, with the position of the cluster varied. The key variables involved in each trial included the position of the cluster, contrast level of the cluster, and the IW settings applied to the image. Combining the ten IW conditions, four contrast levels and four quadrant positions gave 160 combinations. The trials were constructed by pairing 160 combinations of key variables with 160 backgrounds. The entire experiment consisted of 800 trials. Twenty student observers were asked to detect the quadrant of the image in which the mass was located. There was a statistically significant improvement in detection performance for clusters of calcifications when the window width was set at 1024 with a level of 3328, and when the window width was set at 1024 with a level of 3456. The selected IW settings should be tested in the clinic with digital mammograms to determine whether calcification detection performance can be improved.

KEY WORDS: mammography, image processing, intensity windowing, observer studies, calcifications, computers, radiology.

MAMMOGRAPHY, especially in women with dense breasts, is not perfectly sensitive to all cancers. Approximately 10% to 15% of palpable malignancies are not visible mammographically.[1] There is some reason to believe that digital mammography might allow for greater contrast and improved detection of small and early tumors over standard film screen technology, especially if image processing is used to improve image contrast.[2,3]

There are many potentially useful image processing algorithms, and each algorithm has a number of parameters that can be systematically varied to improve or worsen lesion detectability. Radiologists cannot and should not evaluate these algorithms in the clinic with real patients. Such a task would be overwhelming and potentially could cause much unnecessary patient anxiety. Ideally, a test set of image phantoms with simulated lesions in known locations should be used to test each potentially useful algorithm and its attendant parameters in the laboratory setting before any patient's images are interpreted using these algorithms. We have developed such a laboratory method for evaluation of image processing algorithms.[4] In previous work, we have shown that detection performance with the application of contrast limited adaptive equalization (CLAHE) to digitized mammograms is parallel for radiologists and student observers.[4] Using the same experimental paradigm, we report here on whether intensity windowing (IW) can improve the detection of calcifications in dense mammograms in a laboratory setting. We have previously reported elsewhere that IW improves the detectability of masses in dense mammograms.[5]

Many investigators have studied the use of image processing techniques in digitized mammograms. McSweeney attempted to improve the visibility of calcifications by using edge detection for small objects, but gave no clinical results.[6] Smathers improved the visibility of small objects in images by intensity band-filtering.[7] Chan used unsharp-masking to reduce image noise to improve detection of clustered calcifications.[8] Chan, Hale, and Yin have tested other image processing methods on digitized mammograms with variable results.[9-12]

Contrast enhancement methods are not designed to increase or supplement the inherent structural information in an image, but rather to improve the image contrast and theoretically to enhance particular characteristics. IW is an image processing technique that involves the determination of new

From the Departments of Radiology, Computer Science, Biomedical Engineering, and Biostatistics, The University of North Carolina-Chapel Hill, School of Medicine, School of Public Health and College of Arts and Sciences, Chapel Hill, NC.

Supported by NIH PO1-CA 47982, NIH RO1-65583 and DOD DAMD 17-94-J-4345.

Address reprint requests to Etta D. Pisano, MD, CB 7510, Room 503 Old Infirmary Building, Dept of Radiology, UNC School of Medicine, Chapel Hill, NC 27599-7510.

0897-1889/97/1002-0002$3.00/0

pixel intensities by a linear transformation that maps a selected band of pixel values onto the available gray level range of the display device.[13]

The experiments described in this article were performed to determine whether IW could improve the detection of simulated clusters of calcifications in dense mammograms in a laboratory setting. Although the scope of this article is limited to the evaluation of observer performance with respect to the contrast of the simulated microcalcification to background using our established experimental paradigm, it may be interesting for follow-up work to evaluate these results with respect to measures proposed by other investigators, such as the conspicuity measure proposed by Revesz and Kundel.[14-16]

MATERIALS AND METHODS

The experimental paradigm used here is based on the model we have previously described and allows for the laboratory testing of a range of parameter values (in this case, window width and level).[4] The experimental subject is shown a series of test images that consist of an area of a dense mammogram with a simulated cluster of calcifications embedded in the image in one of four quadrants. The observer's task is to determine in which quadrant the cluster of calcifications is located. The test images are displayed in both the processed and unprocessed format, and the contrast of the object against the background is varied from quite easy to detect to impossible to detect.

A computer program randomly selected one of 40 background images and rotated that background to one of four orientations. The 40 backgrounds images of 256 × 256 pixels each were taken from actual mammograms that had been digitized using a Lumiscan digitizer (Lumisys, Inc, Sunnyvale, CA) with a 50 μm sample size and 12 bits of intensity data per sample. The images were selected from relatively dense parts of the mammograms that were known to be normal by virtue of 3 years of clinical and mammographic follow-up. They were selected by a radiologist expert in breast imaging from digitized film screen craniocaudal or mediolateral oblique mammograms. Fig 1 shows one of the backgrounds.

The gray scale values for the mammographic backgrounds are assigned the values recorded by the Lumisys digitizer. The digitizer assigns digital values in the range 495 to 4095 representing an optical density range of 3.43 to 0.08. The digitizer produces digitized gray values that map one to one with optical density (OD) values, ie, the same OD value on film will produce the same gray level.

The 40 images and four orientations provided 160 different dense backgrounds. The program then added a phantom feature, the simulated cluster of five calcifications into the background. The image was then processed with IW to yield the test stimulus.

Mammographic calcifications were simulated using a locally developed program. A cluster of five calcifications was generated. Each individual calcification was a square measuring 1 pixel by 1 pixel in size. Simulated clusters were used instead of real features so that we could have precise control over the structure location, orientation, and structure to background contrast of the calcifications. To more realistically simulate

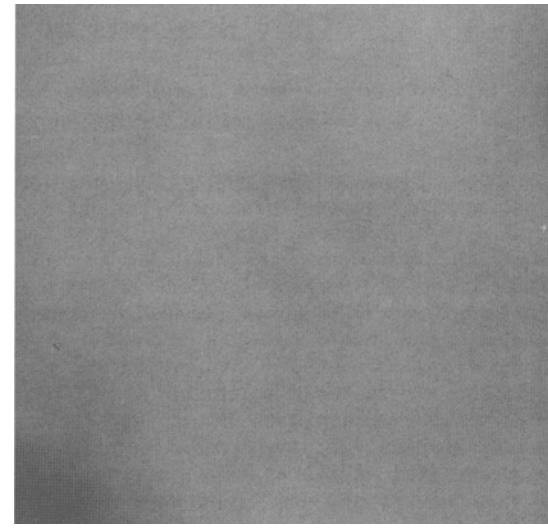

Fig 1. An example of a dense normal background taken from a patient's mammogram and used in the experiment.

microcalcifications would have required using multiple pixels per microcalcification, for instance a 2 × 2 or 3 × 3 matrix. Because the smallest spot size available to use at the time for printing films was 160 μm per pixel, the use of a 2 × 2 or 3 × 3 microcalcification would have unrealistically enlarged the simulated microcalcification. Thus we limited our simulated calcifications to single pixel areas, and varied only the contrast of the calcification. As a result, the simulated calcifications were not entirely realistic, but they did possess the same scale and similar spatial characteristics to actual calcifications seen at mammography.

The intensity difference of the calcifications from background was defined as the gray level of the digital microcalcifications before addition to the background. The calcifications were then embedded at four different intensity levels equally spaced in perceived brightness relative to background by pixel-wise addition of the structure and background images. Fig 2 shows an example of a simulated cluster of calcifications. Figure 3A shows a typical background with the cluster embedded in it without windowing applied. Figure 3B shows the same image with intensity windowing, with the window width of 1024 and a level of 3328. The images in Figs 2 through 3 were photographed from a video monitor with a larger pixel spot size.

A 3 × 3 grid of appropriate window and level parameter settings was selected based on the results of pilot preference

Fig 2. An example of a simulated cluster of calcifications. The actual size of the cluster used in the experiment was only 5 mm in diameter.

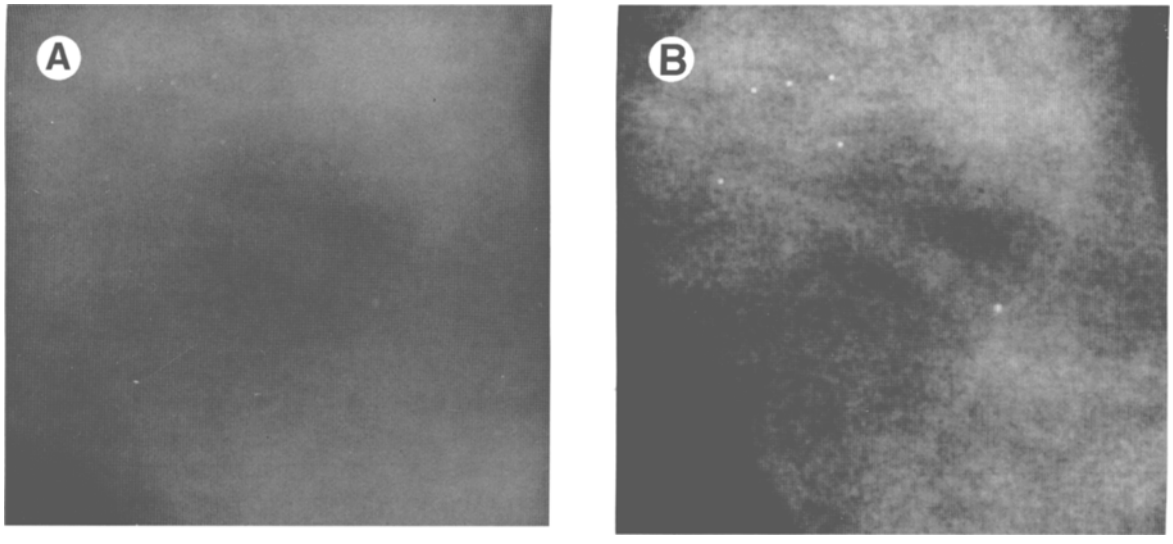

Fig 3. (A) A dense background with a simulated cluster of calcifications embedded in it in the left upper quadrant. The image is enlarged so that the calcifications are readily visualized. (B) The same image as shown in 3A with IW applied. Note how much more obvious the cluster of calcifications appears. The real breast calcification in the right lower quadrant also appears much more obvious with this window.

studies done with two radiologists who specialize in breast imaging (E.D.P. and M.P.B.). In these pilot studies, the two radiologists reviewed dense mammograms with real clinical lesions that were judged to be difficult to visualize using standard screen film mammography. There were seven cases of this type reviewed with 70 combinations of window width and level applied. The radiologists scored each combination of values as showing no change over standard image, improved visibility of the lesion, or worsened visibility of the lesion.

The grid of IW values tested spanned all the likely optimal settings as determined by the pilot work. The IW settings tested were the following: window width 256 with levels 3328, 2456 and 3584; window width 512 with levels 3328, 3456, and 3584; and window width 1024 with levels 3328, 3456, and 3584. The default or unprocessed settings were window width (WW) = 4096, with Level = 2048. There were thus a total of 10 IW settings tested in this experiment.

The digital images were printed onto standard 14 × 17-inch single-emulsion film (3M HNC Laser Film; 3M, St Paul, MN) using a Lumisys Lumicam film printer (Lumisys Inc, Sunnyvale, CA). Each original 50-μm pixel was printed at a spot size of 160 μm, which produced film images 4 × 4 cm, resulting in an enlargement by a factor of three. The radiologist observers in the pilot experiment reported that the magnification did not make the backgrounds unrealistic. Forty images were printed per sheet of film. The images were randomly ordered into an 8 × 5 grid on each sheet of film. Both the film digitizer and film printer were calibrated, and measurements of the relationship between optical density on film and digital units on the computer were determined to generate transfer functions describing the digitizer and 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 digital and printer transfer functions. This standardization function was applied when printing the films 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 OD to 0.13 OD, corresponding to a digital input range of 0 to 4095, respectively.

There were 20 observers for the experiment. They were medical students and graduate students from the biomedical engineering and computer science departments. Performance bonus pay was provided. Observers selected the quadrant of the image that they thought contained the cluster of calcifications. All images contained a simulated cluster of calcifications, for a four alternate-forced choice design. Observers were instructed to make their best guess if they could not tell where the simulated lesion was located in the image.

Films were displayed in a dark room on a standard mammography viewbox that was masked to exclude excess light. Observers could move closer to the image, and could use a magnifying glass, if desired. The observers were trained for the task through the use of two sets of images with instructive feedback before actually starting the experiment.

The order of presentation of stimuli was counterbalanced so as to eliminate any effects of learning and fatigue. All 160 possible combinations of processing conditions (10 IW combinations of WW and level), contrast level (four contrasts) and location of the simulated cluster (four quadrants) were used in the experiment. The experiment was designed to have five blocks, in which all 160 combinations appeared. Each observer saw all combinations in each block. All observers completed the experiment. There were 40 backgrounds and four possible rotations of each background, for 160 possible background patterns. For each block, a different background was uniquely assigned to each of the 160 possible processing condition combinations. The assignment was different for each block.

Each observer examined 800 images, for a total of 16,000 stimuli for the whole experiment.

Observers took breaks after each block of images, and more often if necessary. No time limit was imposed on the observation of the images. Typically, the experiment took 2 hours for each observer, divided into two sessions of 60 minutes each. The two sessions were always scheduled on two different days within a week of each other.

DATA ANALYSIS OVERVIEW

Probit models were fit for each subject and enhancement condition using log10 contrast as the predictor. The probability that a subject gets a correct answer is given by the following equation.

$$\Pr(\text{correct}) = 1/4 + (1 - 1/4)\,\phi\,[(x - \mu_{ij})\sigma_i^{-1}]$$

where i indexes subject, and j indexes IW settings. Here ϕ indicates the cumulative Gaussian distribution function. For each subject, this gave a separate location parameter estimate for each IW setting, and a common spread parameter estimate. Assuming a common spread parameter makes sense biologically, as it corresponds to an equal change in log contrast producing an equal change in perception, throughout the visual range. Also, the ¼ arises from the four-choice task.

The location parameter, μ_{ij}, is the mean of the corresponding Gaussian distribution for the ith subject and jth IW setting. Processing conditions that improve detection will cause this parameter to be smaller, and the curve will shift to the left. This occurs because lower contrast levels are required to spot the object. When the processing of the image makes detection harder, higher contrast levels are needed to locate the calcification, and the curve shifts to the right. The values of σ_i, the spread parameter for the ith subject correspond to the slope of the curve. Larger values of σ_i correspond to steep slopes, or greater increase in detection rates per log contrast.

To compare the processing conditions and to examine the effect of window width and level, further analysis was needed. We defined an overall measure to be $\theta_{ij} = \mu_{ij} + \sigma_i$, which corresponds to the log contrast level at which the ith subject viewing the jth IW condition scored 88% correct. We measured the "success" of a processing condition by calculating the difference between the θ score for the unprocessed image and the θ score for the condition for each subject, say $\delta_j = \theta u - \theta_j$, where u is unprocessed. A large positive δ_j score reflects improved performance. It indicates better detection with processed images than with unprocessed images.

Two analyses were performed using this outcome measure. To keep an overall nominal experiment-wise type 1 error rate of .05, a repeated measures analysis of variance was done at the .04 level, with a set nine of *t*-tests at a .01/9 nominal level for each, and hence a .01 level for the whole set.

Repeated measures analysis of variance (ANOVA) allows one to examine the effect of processing conditions and the interactions between window width and level, while accounting for the dependence of measurements taken on the same observer. The repeated measures ANOVA model was fitted, with the δ_j scores as the outcome, and window width and level as the predictors.

RESULTS

The repeated measures ANOVA showed that the interaction between window width and window level was significant at the .04 level (*P* value $< .0001$, G-G = .729). To examine the nature of this interaction, a series of step-down tests was planned. There was significant interaction between a quadratic trend in window width, and a quadratic trend in window level. Because the quadratic by quadratic interaction was significant, no further tests were examined. A quadratic by quadratic trend means that the surface was curved with respect to both window level and width, and that the shape of the curve differed for fixed values of window width and level (Fig 4).

At the nominal level of .01/9 = .0011, the differences between the default unprocessed condition and the IW conditions were examined. Two settings of intensity windowing processing conditions made finding the calcifications significantly harder, six made the task significantly easier, and one made no significant difference. The settings that made detection easier were window width 1024 with window levels 3328 and 3456 (Table 1, Fig 4).

Average μ_{ij} and σ_i parameters from the best processing condition and the unprocessed condition were used to calculate a typical probit curve. At most, on average, IW processing with settings of window width 1024 and window level 3328 increased the correct detection of calcifications by a maximum of 9%. This is shown in Fig 5.

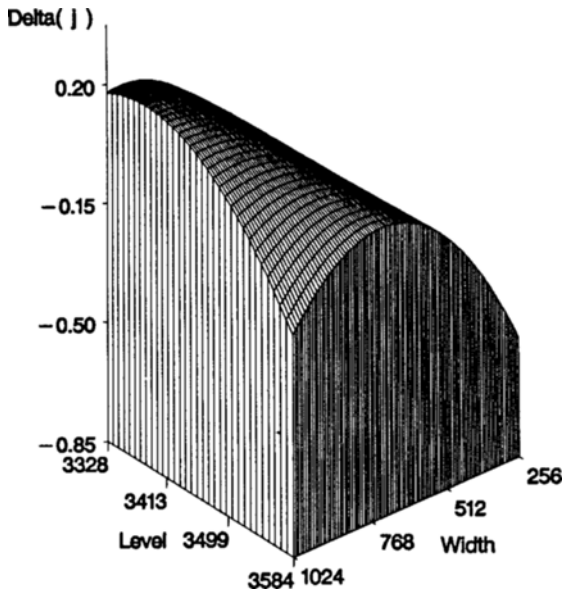

Fig 4. Interpolated predicted values from repeated measures ANOVA: difference in θ value versus window width and level. The peak shows the improved performance due to window width 1024 with window level 3328.

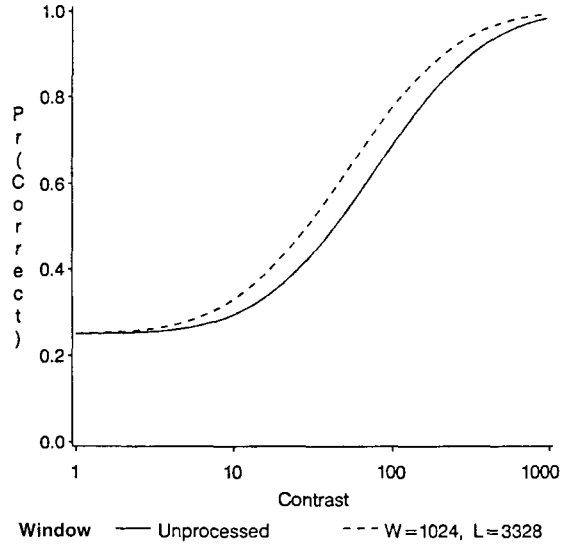

Fig 5. Estimated detection probability for WW of 1024 and level of 3328. The shift in the curve to the left for the processed image reflects improved detection.

DISCUSSION

These results suggest that IW can improve the detection of clustered calcifications on dense mammographic backgrounds, if used properly. Our results also indicate that significant lesion visibility degradation can occur if the window widths and levels are not chosen carefully. We believe that it is important to select the parameters to be applied in the testing of this tool in the clinic based on these types of careful analyses of laboratory studies. Preset intensity windows might then be selected to apply to printed digital mammograms or to mammographic work stations where radiologists might interpret images on line.

Table 1. Mean θ Scores, Difference Scores, and *P* Values for *T* Tests of No Difference

Window Width	Window Level	Mean θ Score	Difference Score	SD	*P* Value
4096	2048	2.46			
256	3328	3.27	−.814	.23	.0001*
256	3456	3.00	−.538	.16	.0001*
256	3584	2.96	−.504	.12	.0001*
512	3328	2.67	−.214	.12	.0001*
512	3456	2.60	−.137	.16	.0012*
512	3584	2.59	−.135	.13	.0002*
1024	3328	2.28	.177	.14	.0001*
1024	3456	2.33	.124	.11	.0001*
1024	3584	2.70	−.246	.10	.0001*

Note: Larger positive difference scores correspond to better performance.

*Significant at the .0011 level.

This work may not predict how this tool will function in a clinical setting. Specifically, graduate student observers and the use of simulated lesions might incorrectly predict the performance of radiologists in detecting real clusters of calcifications in real patients. We have demonstrated previously that graduate student performance at this task parallels the performance of experienced mammographers.[4] The signal-to-noise ratio and the type of image noise present in digital images might vary substantially from digitized mammograms when real full-field digital images are used as the stimuli. Because we have used real clinical images and we have simulated lesions using relatively realistic stimuli, we are optimistic that this image processing algorithm will improve clinical performance. If so, radiologists will be using IW to help them determine whether mammograms of women with dense breasts really do contain calcifications.

Digital mammography is coming to the clinic very soon. It is highly likely that radiologists will want to apply image processing in an attempt to improve their performance in interpreting mammograms. A simple approach to deciding how to view mammograms would be to test every single available algorithm in the clinic on real patients. That would be an expensive and time-consuming process that might have an impact on the care of real

women. It would be preferable, cheaper, and less time-consuming to test this technology in the laboratory before it is tested clinically. The work reported here is intended to help radiologists narrow their choices regarding what might be clinically helpful before expensive clinical tests are undertaken. This project was intended to be a more rigorous exploration of the window widths and levels that might be used clinically in the most challenging areas in the breast, namely the dense parts.

Furthermore, specific IW values depend on the calibration of the instrumentation used for digitization or acquisition, and the patient being imaged. IW values are not standardized and therefore may not directly translate from system to system. That is, the IW values reported on here may not be the correct ones for a different system. However, this experiment showed that there are IW values that can significantly improve detectability of calcifications as well as IW values that substantially degrade lesion visibility. With the advent of full-field digital mammography, and with the standardization of data acquisition, IW values could also be standardized across systems.

This experiment does not address how IW would affect the appearance of fatty areas of the breast, and the detection of calcifications in those parts. We would not want to view a mammogram solely with an algorithm applied that degrades performance in areas where sensitivity is currently quite high. If this algorithm is useful in dense areas, it could potentially be applied selectively to only the dense parts of the breast. Alternatively, it could be used as an adjunct with the image viewed in a standard format, and then with the calcification window width and level applied.

Our experiments to date cannot estimate the frequency of false positives when IW would be used clinically. Many alternate forced choice tests yield proportion correct as the primary outcome. Macmillan and Creelman describe methods for converting proportion correct in this setting to a value for d′, the sensitivity parameter of an ROC analysis.[17] Given the characteristics of the study design, subjects, and training, we believe that superior proportion correct will translate into superior d′. Of course, this must be proven in a true clinical setting with ROC analysis.

REFERENCES

1. Homer MJ: Mammographic Interpretation: A practical approach. New York, NY, McGraw Hill, 1991, pp 4-5
2. Rosenman J, Roe CA, Cromartie R, et al: Portal Film enhancement: Technique and clinical utility. Int J Radiat Oncol Biol Physics 25:333-338, 1993
3. Shtern F: Digital mammography and related technologies: A perspective from the National Cancer Institute. Radiology 183:629-630, 1992
4. Puff DT, Pisano ED, Muller KE, et al: A method for determination of optimal image enhancement for the detection of mammographic abnormalities. J Dig Imaging 7:161-171, 1994
5. Pisano ED, Chandramouli J, Hemminger BM, et al: Utility of Intensity Windowing in Improved Detection of Simulated Masses in Mammograms of Dense Breasts. Presented at the Radiologic Society of North America Meeting. Chicago, IL, November 27, 1995
6. McSweeney MB, Sprawls P, Egan RL: Enhanced Image Mammography. AJR 140:9-14, 1983
7. Smathers RL, Bush E, Drace J, et al: Mammographic microcalcifications: Detection with xerography, screen film, and digitized film display. Radiology 159:673-677, 1986
8. Chan HP, Doi K, Galhorta S, et al: Image Feature analysis and computer-aided diagnosis in digital radiography: I. Automated detected of microcalcifications in mammography. Med Phys 14:538-547, 1987
9. 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. Investigative Radiol 22:581-589, 1987
10. Hale DA, Cook JF, Baniqued Z, et al: Selective Digital Enhancement of Conventional Film Mammography. J Surg Oncol 55:42-46, 1994
11. Yin F, Giger ML, Vyborny CJ, et al: Comparison of Bilateral-Subtraction and Single-Image Processing Techniques in the Computerized Detection of Mammographic Masses. Investigative Radiol 28:473-781, 1993
12. Yin F, Giger M, Doi K, et al: Computerized detection of masses in digital mammograms: Analysis of Bilateral Subtraction Images. Med Phys 18:955-963, 1991
13. Pizer SM: Psychovisual issues in the display of medical images. KH Hoehne, ed, Pictorial Information Systems in Medicine. Berlin, Springer-Verlag, 1985, pp 211-234
14. Revesz G, Kundel HL, Graber MD: The influence of structured noise on the detection of radiologic abnormalities. Investigative Radiol 9:479-486, 1974
15. Kundel HL, Reversz G: Lesion conspicuity, structured noise and fil reader error. AJR 126:1233-1238, 1976
16. Revesz G, Kundel HL: Psychophysical studies of detection errors in chest radiology. Radiology 128:559-562, 1977
17. MacMillan NA, Creelman CD: Detection theory: A user guide. Cambridge, England, Cambridge, 1991, pp 135-136