AD_____

Award Number: DAMD17-01-1-0621

TITLE: Vision-Model-Based Image Enhancement for Digital Mammography

PRINCIPAL INVESTIGATOR: Jeffrey P. Johnson, Ph.D.
John Nafziger, Ph.D.
Elizabeth Krupinski, Ph.D.
Jeffrey Lubin, Ph.D.
Hans Roehrig, Ph.D.
Michael Brill, Ph.D.

CONTRACTING ORGANIZATION: Sarnoff Corporation
Princeton, New Jersey 08543

REPORT DATE: September 2002

TYPE OF REPORT: Final

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

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	September 2002	Final (1 Jun 01 - 31 Aug 02)

4. TITLE AND SUBTITLE
Vision-Model-Based Image Enhancement for Digital Mammography

5. FUNDING NUMBERS
DAMD17-01-1-0621

6. AUTHOR(S)
Jeffrey P. Johnson, Ph.D., John Nafziger, Ph.D.,
Elizabeth Krupinski, Ph.D., Jeffrey Lubin, Ph.D.
Hans Roehrig, Ph.D., Michael Brill, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Sarnoff Corporation
Princeton, New Jersey 08543

E-Mail: jjohnson@sarnoff.com

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

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

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 Words)***

The objective was to evaluate the utility of a visual discrimination model, JNDmetrix, for predicting/optimizing the effects of grayscale window and level (W/L) parameters on the detectability of breast lesions. Two observer performance studies were conducted to correlate detection performance with JNDmetrix predictions of lesion conspicuity. In the first set of experiments, 2AFC detection thresholds were measured with five W/L conditions applied to simulated mammographic backgrounds and "lesions" using nonmedical observers. The detectability of real microcalcification clusters in digitized mammograms was then evaluated for three W/L conditions in an ROC study with mammographers. For the simulated lesions/backgrounds, the model correctly predicted W/L conditions that minimized the detection thresholds, supporting our hypothesis that the model could be used to automate the selection of optimal W/L settings. Experimental and model results showed significant reductions in thresholds when W/L processing was applied locally near the lesion. ROC results with digitized mammograms read by radiologists failed to show enhanced detection of microcalcifications using a localized W/L frame. This was probably due to the nonuniform appearance of parenchymal tissue across the image. Future experiments will provide readers both the uniform contextual information of the full image and enhanced lesion contrast within a localized W/L frame.

14. SUBJECT TERMS	15. NUMBER OF PAGES
breast cancer	14
	16. PRICE CODE

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

Table of Contents

Introduction

Studies of breast-cancer screening errors have found that almost half of all lesions missed at initial screening were visible retrospectively in the images. Digital mammography offers potential for significantly reducing these perceptual errors by allowing image processing that enhances the conspicuity of abnormal features. The primary hypothesis of this study was that the detectability of breast lesions could be enhanced by image-processing techniques guided by a computational model of human visual perception. The objective was to evaluate the utility of a visual discrimination model, JNDmetrix, for predicting and optimizing the effects of grayscale window and level (W/L) parameters on the detectability of breast lesions. Two observer performance experiments were conducted to establish the correlation of detection performance with JNDmetrix predictions of lesion conspicuity. In the first experiment, detection thresholds were measured for a wide range of W/L conditions applied to a variety of *simulated* mammographic backgrounds and "lesions" using nonmedical observers in a two-alternative, forced choice test methodology. In the second experiment, the detectability of microcalcification clusters in digitized mammograms was evaluated for a select set of W/L conditions in a receiver operating characteristic (ROC) study with mammographers. The correlation of JNDmetrix model predictions with experimental detection thresholds and performance was determined as a function of W/L condition, signal (lesion) type, and image characteristics.

Body

Methods

Detection Threshold Experiments

Contrast-threshold experiments were conducted to measure the detectability of simulated breast lesions imbedded in synthetic mammograms for five different W/L conditions. The synthetic mammographic backgrounds consisted of 512x512 pixel filtered-noise images with a $1/f^3$ noise power spectrum. Two sets of background images were selected with central regions that were approximately 45% brighter or darker than the average pixel value of 2048 (out of 4096). The "lesion" signal was either a 2-D Gaussian function, representing a mass, or a small cluster of six blurred disks, or specks, representing microcalcifications. The diameter of the Gaussian signal was fixed at 50 pixels. The diameter of the blurred-disk specks was 8 pixels; the cluster radius was 20 pixels. Examples of these backgrounds and lesions are presented in Figure 1. The selected W/L condition was applied to either the full 512x512 pixel images or to only a central 170x170 pixel region (one-ninth of the image area). This resulted in eight distinct combinations of signal type, background luminance level, and W/L frame size, which were presented in eight separate test sessions. Examples of the test images are presented in Figure 2.

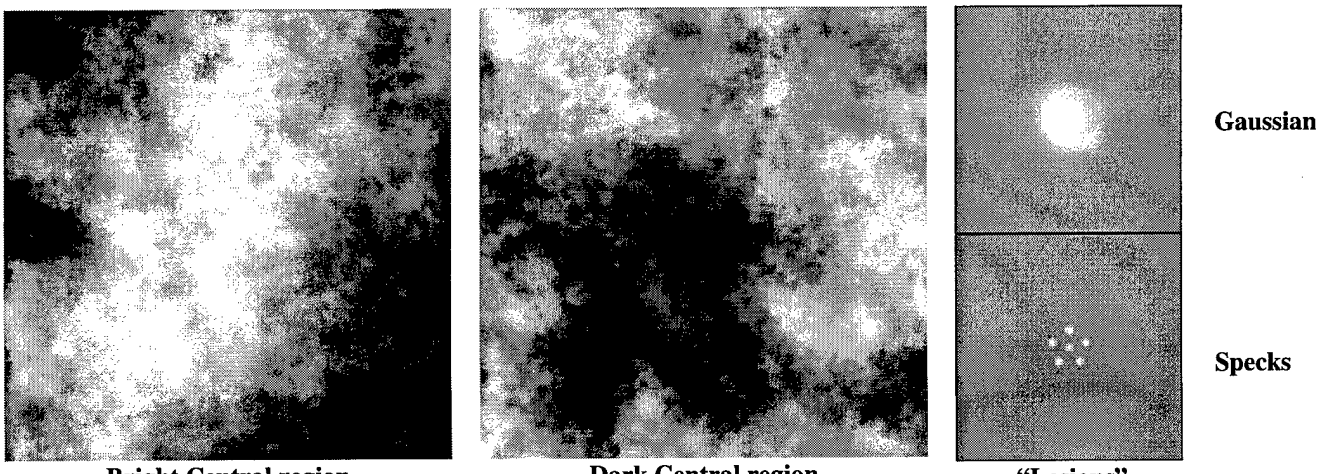


Figure 1. Examples of $1/f^3$ filtered noise images with bright or dark central regions. Simulated "lesions" consisted of a centrally located 2D Gaussian "mass" or a single cluster of six "specks" representing microcalcifications.

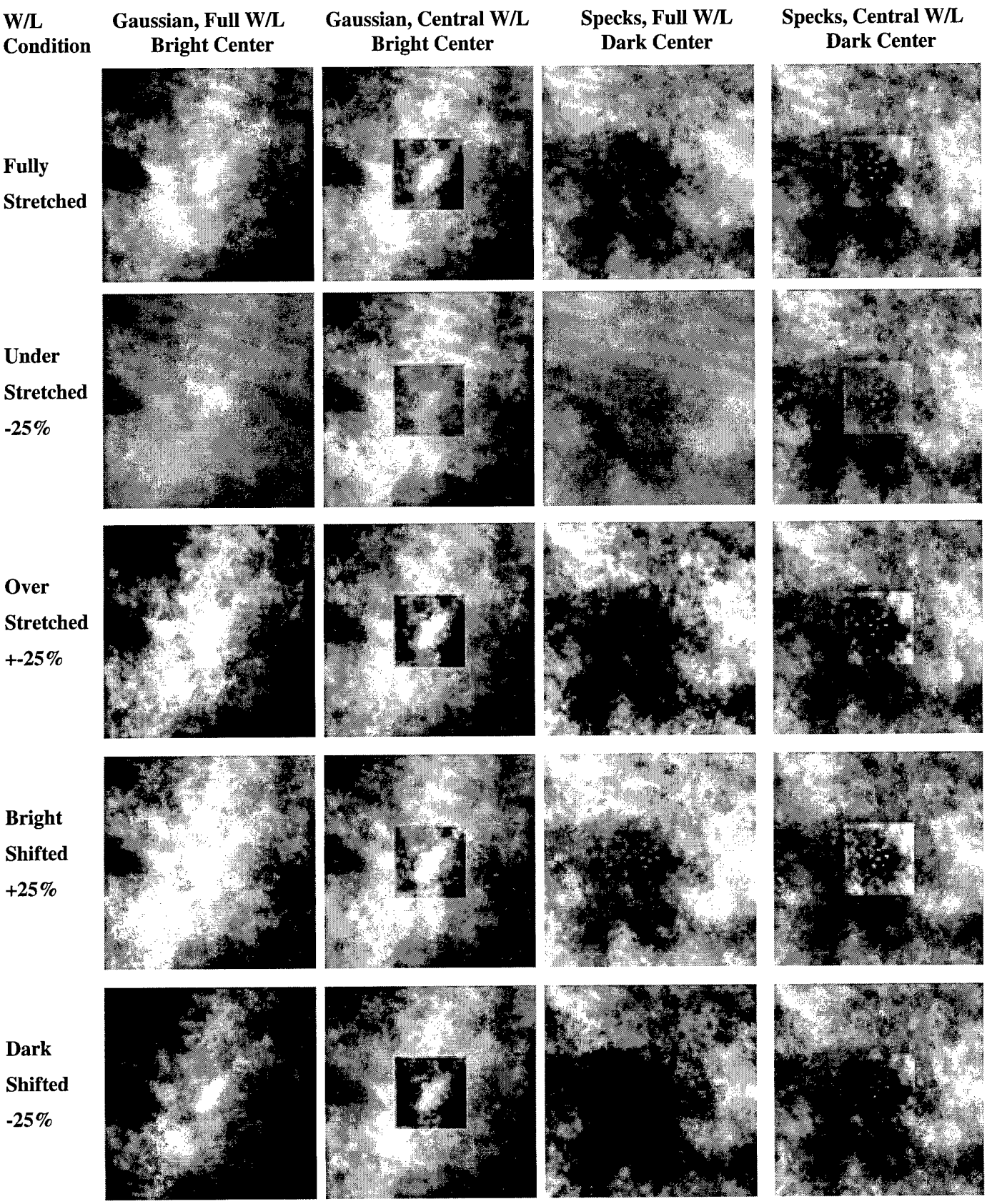


Figure 2. Examples of simulated breast "lesions" and backgrounds used in the threshold detection experiments. Five window/level conditions were applied to either the full images (Full W/L) or the central region only (Central W/L). These examples include a Gaussian mass or a cluster of specks on a $1/f^3$ filtered noise background with bright or dark central region, respectively.

The experimental paradigm interleaved five different W/L manipulations - fully stretched, under-stretched (-25%), over-stretched (+-25%), bright shifted (+25%), and dark shifted (-25%), - to measure the effects of W/L on signal detection. Figure 3 shows the W/L transformation for image pixel values, p.

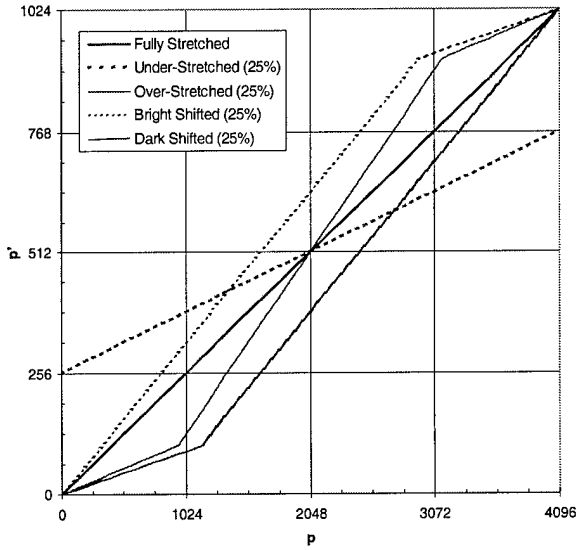


Figure 3. Window/level transformation used to generate test images for observer performance experiments and JNDmetrix model simulations.

Contrast thresholds were determined for three experienced, nonmedical observers corrected to normal myopes. Each observer was required to reach asymptotic performance prior to data collection. In a two-alternative, forced-choice (2AFC) procedure, the observer's task was to determine which of two side-by-side images contained a centrally located signal. The signal was also presented separately in a small mean background frame above the 2AFC images. The signal amplitude was varied in a transformed, 1-up/3-down staircase procedure with a threshold corresponding to 79% correct responses. Five W/L cases were interleaved in a single session to remove any sequential fluctuations in observer performance. A session was terminated after ten reversals of each staircase, where the final six reversals were used to compute the mean threshold amplitude. Thresholds were defined as peak signal amplitude divided by 2048, or one half of the 12-bit range of pixel values in these images.

The $1/f^3$ filtered noise images were generated by transforming a 512x512 uniformly distributed random matrix into the Fourier domain by a FFT. The random matrix was scaled in frequency space to be $1/f^3$ and transformed back to the spatial domain with the inverse FFT. The real component was extracted, discarding the imaginary part, and finally subtracting off the minimum. The resulting image was confirmed to have a $1/f^3$ noise power spectrum. In total, twenty images were generated and utilized in the experiments.

The stimuli were generated on a Windows NT PC with a 10-bit Barco display controller (1024 LUT entries, 12 bit DAC) and displayed on a Siemens 5-Mpixel CRT monitor. The monitor was perceptually linearized in conformance with the DICOM-14 grayscale standard display function. The monitor had a frame refresh rate of 71 Hz and a luminance range from 0.3 to 290 cd/m^2. The addressable pixel size on this monitor was 0.15 mm. The observers viewed the monitor binocularly in a darkened room. The viewing distance was fixed at 52 cm (60 pixels/degree) and head movements were stabilized with a chin-rest.

ROC Experiment

A receiver operating characteristic (ROC) experiment was conducted with six radiologists at the University of Arizona. Test images were acquired from digitized screen-film mammograms in the Database for Screening Mammography. Twenty-five cases were chosen with identified (biopsy proven) malignant (n = 13) and benign (n = 12) microcalcification clusters. A variety of breast densities and calcification types, distributions, and subtlety ratings were

included. The pixel resolution of the raw images ranged from 42 to 50 microns and 12-bit grayscale resolution. Regions of interest were extracted to form 512 x 512-pixel sub-images with centered microcalcification clusters (100% or original contrast). Individual calcifications were removed by digital processing using a median filter to generate "signal-absent" versions of the original "signal-present" images. A second set of 25 signal-present and signal-absent image pairs was created by superposing the extracted calcifications onto normal (lesion-free) regions of the full images. Additional images with calcifications at intermediate contrast levels (25%, 50%, and 75%) were created by weighted superposition of the signal-absent and signal-present pairs.

Test images were further processed by applying three W/L transformations – fully stretched, 15% under-stretched, and 15% over-stretched. The W/L conditions were applied to either the full images or to 200x200 pixel central regions. Half of the observers read the images with central W/L conditions during the first session and full W/L conditions in the second session. The other three observers did just the opposite for a counterbalanced presentation design. There was approximately two weeks between sessions to inhibit recall of cases. The confidence rating data were analyzed using the Multi-reader Multi-Case Receiver Operating Characteristic (MRMC ROC) technique, generating area under curve (A_z) values and comparing them with an Analysis of Variance (ANOVA).

Results

Detection Threshold Experiments

Experimental detection thresholds for the two signal types (Gaussian and specks), two central-background luminance levels (bright and dark), and two W/L frame sizes (full and central) are presented in Figures 4 and 5. The thresholds were averaged across all test sessions and observers; error bars show the 95% confidence intervals about the mean values. Significant variations in thresholds were observed across W/L conditions for all of the signal-background combinations. For the "Full W/L" cases, the lowest thresholds generally occurred for the fully-stretched condition and for dark-shifted bright backgrounds and bright-shifted dark backgrounds. For the specks, the under-stretched condition also yielded low thresholds. Most of the thresholds decreased significantly when the W/L condition was applied to only the central region of an image where the signal was located. For the Gaussian signals, the minimum thresholds overall were for the central W/L over-stretched and dark-shifted/bright and bright-shifted/dark conditions. For the specks on bright backgrounds, the lowest thresholds were for the under-stretched and dark-shifted central W/L conditions, while for dark backgrounds the central W/L threshold was slightly lower for the over-stretched condition.

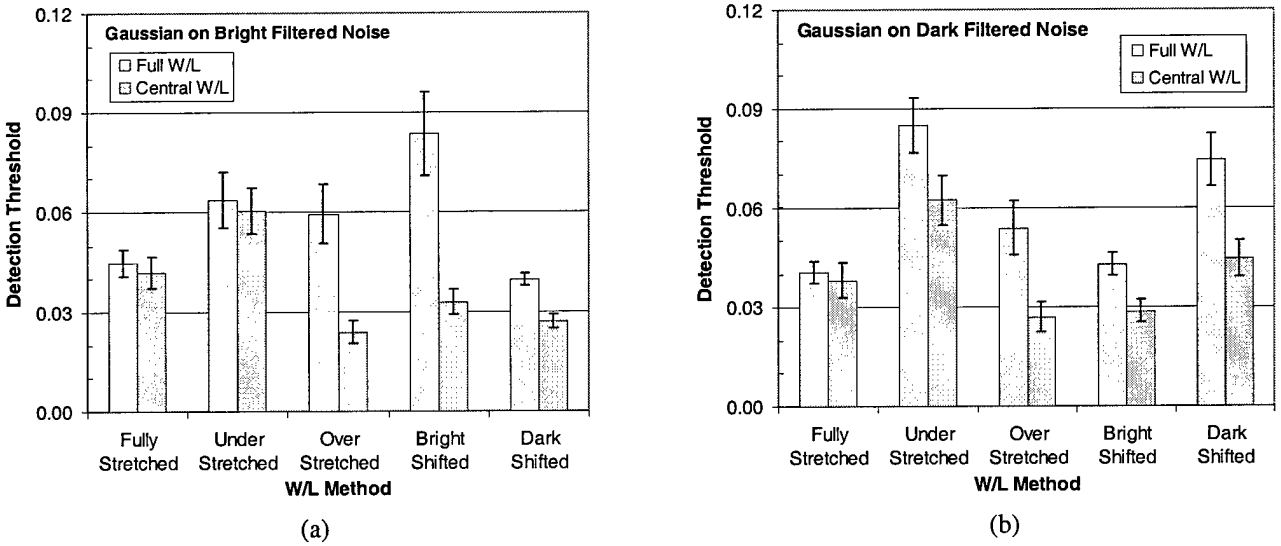


Figure 4. Mean experimental thresholds for the detection of 2D Gaussian "masses" at the center of filtered noise images with (a) bright or (b) dark central regions. Five W/L conditions were applied to either the full (512x512 pixel) or central (170x170 pixel) region of the images. Error bars show 95% confidence intervals.

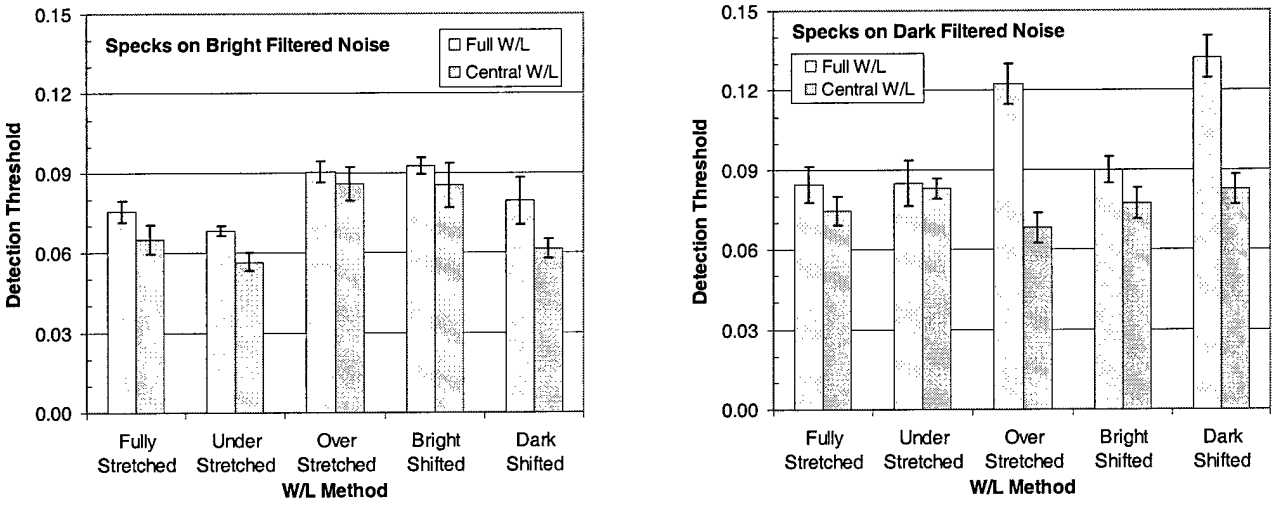
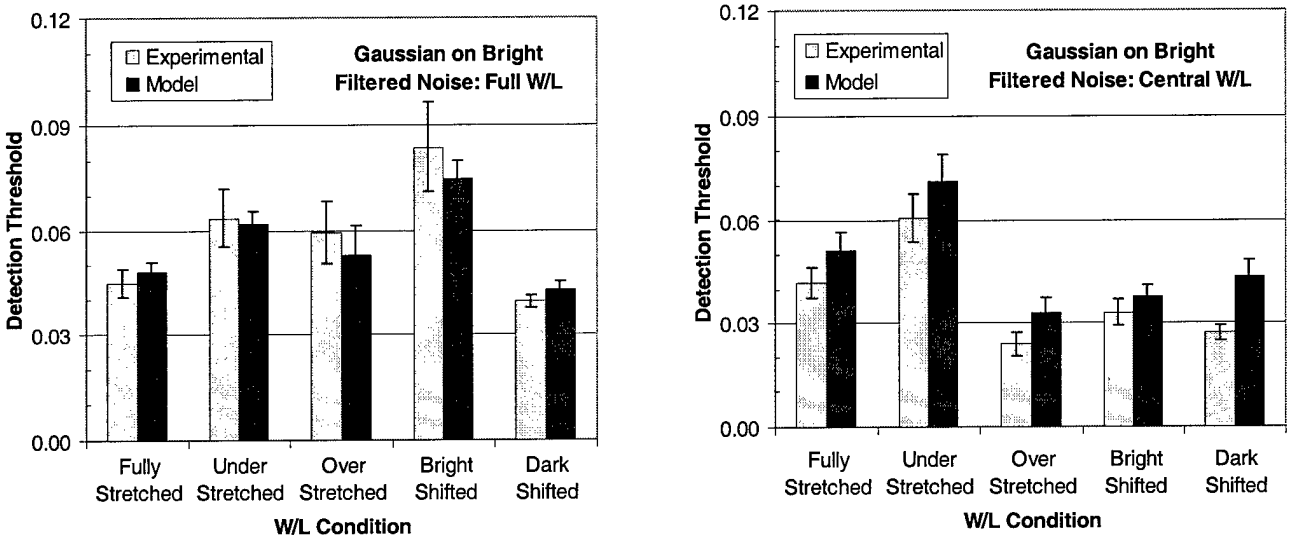


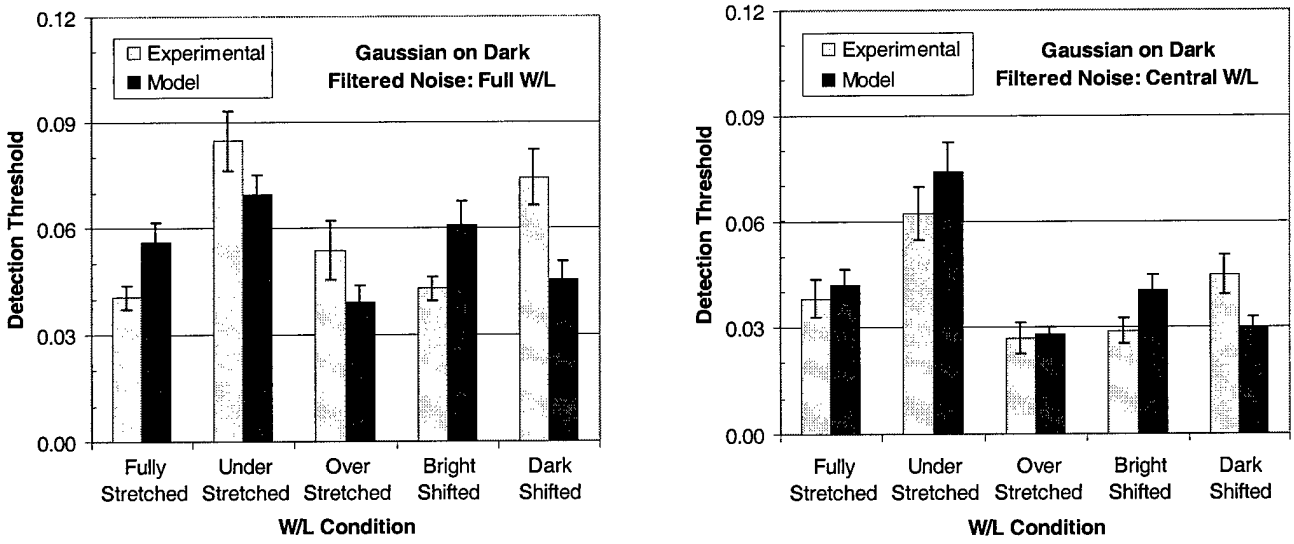
Figure 5. Mean experimental thresholds for the detection of a cluster of "specks" representing calcifications at the center of filtered noise images with (a) bright or (b) dark central regions. Five W/L conditions were applied to either the full (512x512 pixel) or central (170x170 pixel) region of the images. Error bars show 95% confidence intervals.

JNDmetrix model predictions of detection thresholds for the same signals, backgrounds, and W/L conditions used in the observer experiment are presented in Figures 6 to 9. In many cases, there was significant agreement between model and experimental thresholds across the W/L conditions. For Gaussian signals on bright backgrounds (Figure 6), the variations in measured thresholds were well matched by the predicted thresholds. For the dark backgrounds (Figure 7), the overall correlation was somewhat lower but many of observed variations in threshold were reflected in the model predictions. The most notable discrepancy for the Gaussians on dark backgrounds was in the predicted differences in threshold between dark and bright shifted conditions: the model predicted lower thresholds for the dark shifted images, which was contrary to the observed differences. For the specks, the correlation of model and experimental thresholds across W/L conditions was best for the bright-background/full-WL (Figure 8a) and dark-background/central-WL (Figure 9b) cases. For the two other cases (Figures 8b and 9a), the overall correlation was somewhat lower. It is worth noting, however, that the model correctly predicted the W/L condition that yielded the absolute minimum threshold for three out of the four signal-background combinations (central W/L, over-stretched condition for Gaussians on bright and dark backgrounds, and specks on dark backgrounds). For the fourth combination (specks on bright backgrounds), the model predicted an optimal W/L condition (dark stretched) for which the measured threshold did not differ significantly from the minimum (under-stretched) threshold.



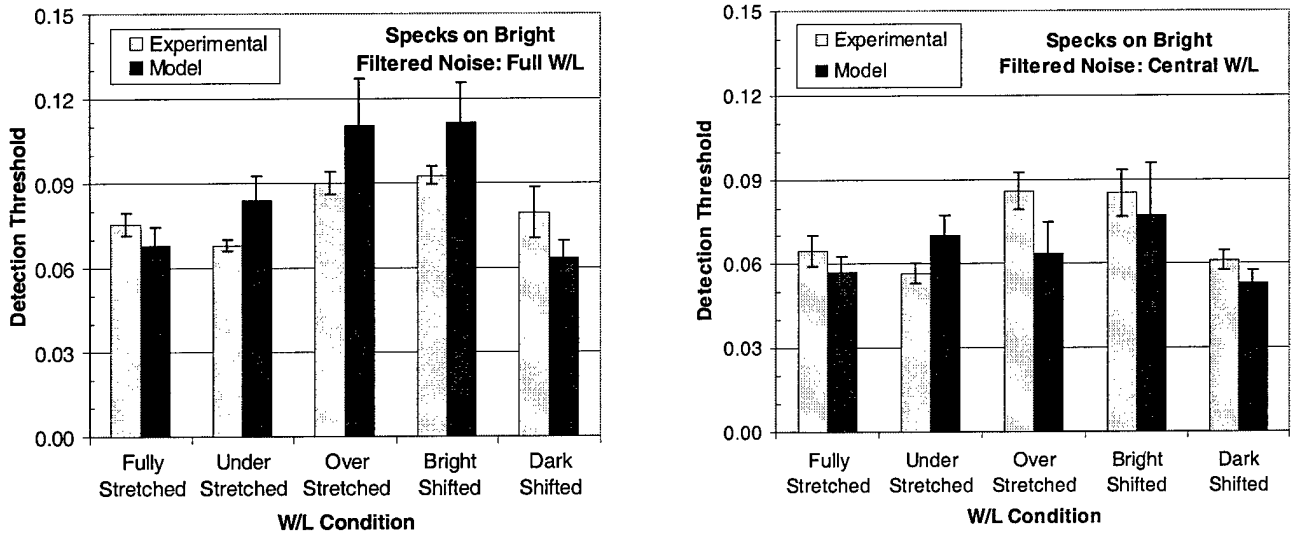
(a) (b)

Figure 6. Mean experimental and model thresholds for the detection of a 2D Gaussian "mass" at the center of filtered noise images with a bright central region. Five W/L conditions were applied to either the (a) full image or (b) central region. Error bars show 95% confidence intervals.



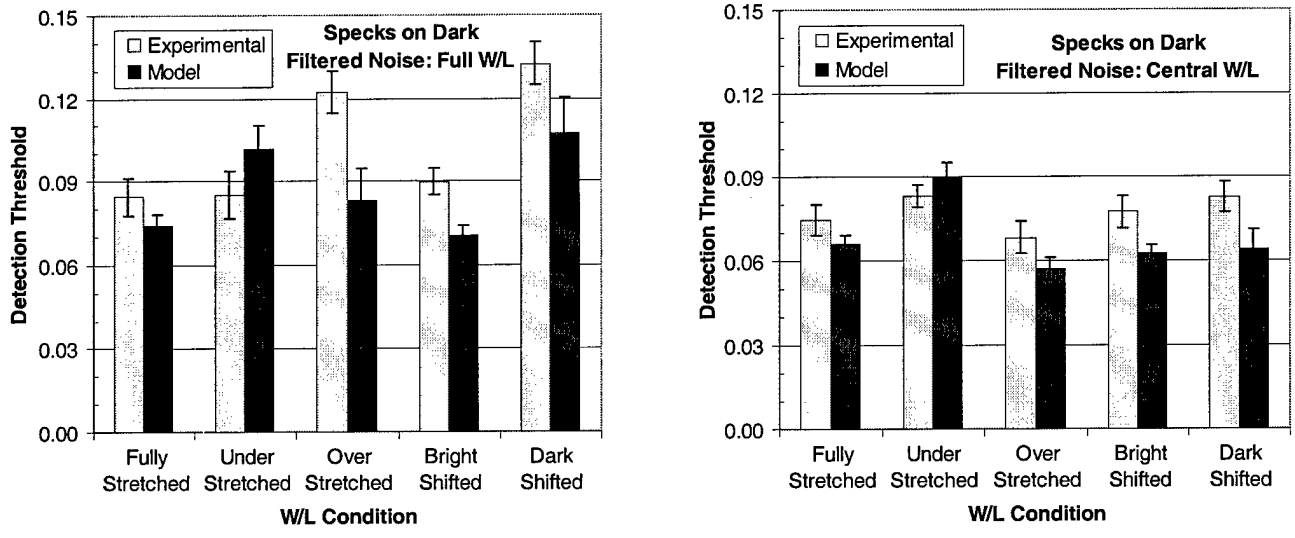
(a) (b)

Figure 7. Mean experimental and model thresholds for the detection of a 2D Gaussian "mass" at the center of filtered noise images with a dark central region. Five W/L conditions were applied to either the (a) full image or (b) central region. Error bars show 95% confidence intervals.



(a) (b)

Figure 8. Mean experimental and model thresholds for the detection of a cluster of specks at the center of filtered noise images with a bright central region. Five W/L conditions were applied to either the (a) full image or (b) central region. Error bars show 95% confidence intervals.



(a) (b)

Figure 9. Mean experimental and model thresholds for the detection of a cluster of specks at the center of filtered noise images with a dark central region. Five W/L conditions were applied to either the (a) full image or (b) central region. Error bars show 95% confidence intervals.

ROC Experiment: Detection of Microcalcifications

Combining all of the data and comparing central W/L vs. full W/L conditions (Table 1), there was a statistically significant difference ($F = 9.7691$, $p = 0.0261$) with the average A_z for the central W/L being 0.8399 (se = .0203) and 0.8809 (se = .0217) for the full-WL condition.

Table 1. All data combined comparing central versus full W/L presentation conditions.

Reader	Central W/L	Full W/L
1	.8458	.8940
2	.8313	.9198
3	.8293	.8301
4	.8252	.8890
5	.8233	.8489
6	.8848	.9038
Mean	**.8399**	**.8809**

The second comparison looked at each processing condition – represented below as f, o, and u for fully stretched, over-stretched, and under-stretched, respectively - and compared central vs. full W/L presentation mode. For each processing mode, performance was higher in the full W/L condition than in the central W/L condition. For the f condition the average A_z for the central W/L was 0.8333 (se = 0.045) and was 0.8974 (se = 0.0335) for the full W/L. There was no statistically significant difference ($F = 4.9664$, $p = 0.0538$). For the o condition the average A_z for the central W/L was 0.8555 (se = 0.0248) and was 0.8666 (se = 0.0305) for the full W/L. There was no statistically significant difference ($F = 0.1477$, $p = 0.7009$). For the u condition the average A_z for the central W/L was 0.8696 (se = 0.0353) and was 0.8715 (se = 0.0312) for the full W/L. There was no statistically significant difference ($F = 0.0023$, $p = 0.9638$).

Table 2. Central vs. full W/L presentation conditions compared for each of the 3 processing modes (f, o, u).

Reader	f-central W/L	f-full W/L	o-central W/L	o-full W/L	u-central W/L	u-full W/L
1	.8677	.9314	.8434	.9299	.8257	.8539
2	.7944	.9554	.8503	.9127	.9999	.8177
3	.8354	.8130	.8314	.7763	.8713	.8900
4	.8423	.8941	.8433	.9025	.8067	.9133
5	.7765	.8640	.8881	.8347	.8274	.8734
6	.8838	.9265	.8765	.8435	.8866	.8808
Mean	**.8333**	**.8974**	**.8555**	**.8666**	**.8696**	**.8715**

Combining all the data and comparing f vs o vs u for full W/L ($F = 0.682$, $p = 0.5206$) and central W/L ($F = 0.837$, $p = 0.4522$) presentations separately did not yield statistically significant differences overall between processing modes.

The third analysis compared central W/L vs full W/L presentation mode for each contrast level within each processing mode. For the f processing mode (Table 3) there were no statistically significant differences at 25% contrast (F =

2.6926, p = 0.1116), at 75% contrast (F = 1.5587, p = 0.2218) or at 100% contrast (F = 0.0659, p = 0.8076), but there was at 50% contrast (F = 6.8517, p = 0.0321) with the full W/L being higher than central W/L performance. Although the differences did not reach statistical significance, performance at each contrast level was higher for the full W/L than the central W/L presentation mode.

Table 3. Central vs. full W/L for each contrast level for the fully stretched W/L condition.

Reader	f-25 central W/L	f-25 full W/L	f-50 central W/L	f-50 full W/L	f-75 central W/L	f-75 full W/L	f-100 central W/L	f-100 full W/L
1	.6923	.7967	.8675	.9417	.9254	1.00	1.00	1.00
2	.5609	.8327	.7285	1.00	.9184	1.00	1.00	1.00
3	.6419	.5902	.7802	.8205	.9476	.9088	1.00	.9352
4	.6376	.6780	.8649	.9373	.9419	.9671	.9724	1.00
5	.5009	.6284	.7878	.7833	.8980	.9440	.9363	1.00
6	.6926	.7715	.8672	.9599	.9782	1.00	1.00	1.00
Mean	**.6210**	**.7162**	**.8160**	**.9221**	**.9349**	**.9700**	**.9848**	**.9892**

For the o processing mode (Table 4) there were no statistically significant differences at 25% contrast (F = 0.5376, p = 0.4646), at 50% contrast (F = 0.0892, p = 0.7657), at 75% contrast (F = 0.1245, p = 0.7247) or at 100% contrast (F = 0.3126, p = 0.6002). Although the differences did not reach statistical significance, performance at the 25% and 50% contrast levels was higher for the full W/L than central W/L presentation mode. At the 75% and 100% contrast levels performance was higher for the central W/L presentation mode.

Table 4. Central vs. full W/L for each contrast level for the over-stretched W/L condition.

Reader	o-25 central W/L	o-25 full W/L	o-50 central W/L	o-50 full W/L	o-75 central W/L	o-75 full W/L	o-100 central W/L	o-100 full W/L
1	.6273	.8667	.8367	.9229	.9222	.9441	.9741	.9692
2	.6846	.7711	.8255	.9324	.9418	.9806	.9591	1.00
3	.7034	.5866	.7409	.7451	.9204	.8894	.9866	.8506
4	.6416	.8046	.8635	.8901	.9116	.9090	.9684	1.00
5	.6826	.6482	.8894	.7867	.9999	.9562	.9729	.9540
6	.6307	.5735	.9151	.8709	.9693	.9359	1.00	1.00
Mean	**.6617**	**.7084**	**.8452**	**.8580**	**.9442**	**.9359**	**.9768**	**.9623**

For the u processing mode (Table 5) there were no statistically significant differences at 25% contrast (F = 1.1988, p = 0.2826), at 50% contrast (F = 3.1932, p = 0.0760), at 75% contrast (F = 2.0861, p = 0.2083) or at 100% contrast (F = 0.3577, p = 0.5507). Although the differences did not reach statistical significance, performance at each contrast level was higher for the full W/L than central W/L presentation mode.

Table 5. Central vs. full W/L for each contrast level for the under-stretched W/L condition.

Reader	u-25 central W/L	u-25 full W/L	u-50 central W/L	u-50 full W/L	u-75 central W/L	u-75 full W/L	u-100 central W/L	u-100 full W/L
1	.3638	.5900	.8402	.8413	1.00	1.00	1.00	1.00
2	.3709	.4652	.9966	.9665	.9998	1.00	1.00	1.00
3	.6720	.6691	.8593	.9097	.9732	.9682	1.00	1.00
4	.5723	.7242	.7601	.9378	.9089	1.00	.9688	1.00
5	.6423	.6346	.8004	.8986	.9145	.9706	.9512	1.00
6	.5595	.5085	.9266	1.00	1.00	1.00	1.00	1.00
Mean	**.5301**	**.5986**	**.8639**	**.9256**	**.9661**	**.9898**	**.9867**	**1.00**

Discussion

JNDmetrix model predictions of detection thresholds for the simulated breast lesions and backgrounds were in good quantitative agreement with the measured thresholds and reflected variations across W/L conditions in most but not all cases. The model correctly predicted W/L conditions that minimized the detection threshold for a given class of images, supporting our hypothesis that the model could be used to automate the selection of optimal W/L settings. The experimental and model results showed consistent reductions in detection thresholds when the W/L processing was applied locally in a region of interest that enclosed the lesion. This effect is reasonable when lesion contrast is relatively low and can be enhanced by local stretching of graylevels around the lesion to fill the available luminance range of the display.

Areas for improvement in the correlation between predicted and measured thresholds were identified. Refinements in the computations of low spatial-frequency contrast energy and crossmasking effects of spatial-frequency bands are in progress, and are expected to improve the degree and robustness of the correlation of model and experimental thresholds.

Results of the ROC study with mammographers were in some ways inconsistent with those of the detection threshold experiments. The most significant finding was the reduction in observer performance associated with the use of a localized ("central") W/L frame. One possible explanation is that when searching for microcalcifications (and other abnormalities) and deciding whether or not they are present, mammographers do not simply examine the area where the microcalcifications are located. They compare the tissue and features in that area to the entire image to get an impression of the overall appearance of the parenchymal tissue. In this experiment, the full W/L condition allowed them to do this since the appearance was uniform across the image, allowing better assessment of the characteristics of the surrounding parenchymal tissue, and thereby increasing their decision confidence. The center W/L condition did not allow this and the tissue inside the window looked very different from outside the window in most cases, which made it difficult to evaluate tissue appearance. The radiologists all agreed that the window presentation for the center mode was exactly like what they see when doing softcopy mammography reading (at least for zooming and some have seen or used a window that processes as well), so the window idea itself is not the issue. The difference seems to be that when reading softcopy on full images they read the whole image first without processing (or with whatever global preprocessing the viewing system provides). This gives them their impression of overall parenchymal patterns. If they see something questionable, then they use the image processing functions or the zoom. In future experiments, a modified procedure in which the observer is given the capability to turn the central W/L frame on and off during each trial could improve observer performance by providing both the uniform contextual information of the full image and the enhanced lesion contrast within the localized W/L frame.

Key Research Accomplishments

- Demonstrated good overall agreement between experimental lesion-detection thresholds and JNDmetrix model predictions for images with simulated breast lesions and backgrounds.
- Showed that the model correctly predicted W/L conditions that minimized the detection thresholds, supporting our hypothesis that the model could be used to automate the selection of optimal W/L settings.
- Identified potential for improved lesion detectability through local enhancement of contrast within a region-of-interest W/L frame.
- Determined that the applicability of W/L enhancements to mammography has to accommodate the radiologist's need for initial presentations with uniform tissue appearance across the image.

Reportable Outcomes

Submitted abstract for SPIE 2003 poster presentation.

Conclusions

Results of detection-threshold experiments with simulated breast lesions and backgrounds showed that the JNDmetrix model accurately predicted experimental thresholds and most of their variations across W/L conditions. The model correctly predicted W/L conditions that minimized the detection threshold for a given class of images, supporting our hypothesis that the model could be used to automate the selection of optimal W/L settings. The experimental and model results showed consistent, significant reductions in detection thresholds when the W/L processing was applied locally in a region of interest enclosing the lesion. This result suggested the potential for improved lesion detectability through local enhancement of contrast within a region-of-interest W/L frame.

Results of an ROC experiment with digitized mammograms read by radiologists failed to show enhanced detection of microcalcifications using a localized, region-of-interest W/L frame. This was due most likely to the nonuniform appearance of parenchymal tissue across the image in this experiment. In future experiments, a modified procedure in which the observer is given the capability to turn the central W/L frame on and off during each trial could improve observer performance by providing both the uniform contextual information of the full image and the enhanced lesion contrast within the localized W/L frame.

Areas for improvement in the correlation between model predictions and observer performance were identified in this study. Refinements in the computations of low spatial-frequency contrast energy and crossmasking effects of spatial-frequency bands are in progress, and are expected to improve the degree and robustness of the correlation of model and experimental detection thresholds and performance in detection tasks.