12-1-1993

Pre-Clinical ROC Studies of Digital Stereomammography

Jean Hsu
Purdue University School of Electrical Engineering

Charles F. Babbs
Purdue University Biomedical Engineering Center

David M. Chelberg
Purdue University School of Electrical Engineering

Zygmunt Pizlo
Purdue University Department of Psychological Sciences

Edward J. Delp
Purdue University School of Electrical Engineering

PRE-CLINICAL ROC STUDIES OF DIGITAL STEREOMAMMOGRAPHY

JEAN HSU
CHARLES F. BABBS
DAVID M. CHELBERG
ZYGMUNT PIZLO
EDWARD J. DELP

TR-EE 93-47
DECEMBER 1993

SCHOOL OF ELECTRICAL ENGINEERING
PURDUE UNIVERSITY
WEST LAFAYETTE, INDIANA 47907-1285

Pre-Clinical ROC Studies of Digital Stereomammography

Jean Hsu†, Charles F. Babbs‡, David M. Chelberg†,
Zygmunt Pizlo††, and Edward J. Delp†

†School of Electrical Engineering
‡Biomedical Engineering Center
††Department of Psychological Sciences
Purdue University
West Lafayette, Indiana

This work was supported in part by the Purdue University TRASK fund and by a Digital Equipment Corporation Faculty Incentives for Excellence Grant.

ABSTRACT

This paper reports the diagnostic performance of observers in detecting abnormalities in computer generated mammogram-like images. A mathematical model of the human breast is defined in which breast tissues are simulated by spheres of different sizes and densities. Images are generated by casting rays from a specified source, through the model, and onto an image plane. Observer performance with two viewing modalities (stereo versus mono) is compared. In the stereo viewing mode, left and right images are presented to the observer (wearing liquid crystal shuttle glasses), such that the left eye sees the left image only and the right eye sees the right image only. In this way, the images can be fused by the observer to obtain a sense of depth. In the mono viewing mode, left and right images are presented side by side and the observer can see both images at the same time. Observer response data are evaluated using receiver operating characteristic (ROC) analysis to characterize any difference in detectability of abnormalities (in either the density or arrangement of simulated tissue densities) using the two viewing modes. The results indicate the clear superiority of stereo viewing for detection of arrangement abnormalities. For detection of density abnormalities, the performance of the two viewing modes is similar. These preliminary results suggest that stereomammography may permit easier detection of certain tissue abnormalities, perhaps providing a route to earlier tumor detection in cases of breast cancer.

Index Terms : diagnostic radiology; digital mammography; receiver operating characteristic (ROC) analysis; stereo imaging;

I INTRODUCTION

The breast is the most frequent site of incidence of cancer in American women, accounting for 32% of incident cancers [1]. The disease is now projected to affect one woman in nine [2] and has been targeted by the National Institutes of Health (NIH) and by society in general for intensive research [3]. Prior to metastatic spread, breast cancer is a regional disease that is often cured by surgery or radiation. After metastatic spread, however, it becomes a generalized disease that is resistant to aggressive regimens of chemotherapy. The probability of metastasis is directly related to the size of the primary lesion. Hence, a highly effective means to diminish breast cancer mortality is earlier diagnosis [4], leading to a reduction in average tumor size at initial treatment. The Health Insurance Plan (HIP) study which began in 1963, after eighteen years of follow-up, has clearly demonstrated that the screening of presumably well women, with the possibility of early tumor detection, can result in a substantial (25%) reduction in mortality from breast cancer [5]. Early tumor detection has also been identified by the National Cancer Institute as a major priority for the decade of the 1990's, and recent NIH announcements have pointed out the need for improved screening technologies for the detection of breast cancer [3].

Mammography is the standard for diagnosis of localized breast cancer. It is well known to be more effective than physical examination, sonography, thermography and diaphanography [5, 6, 7]. Despite their utility, mammographic images are complex. Abnormalities in mammograms, when present, may be small or subtle. Any diagnostic technique that improves the sensitivity or specificity of breast cancer detection would be highly valued. Given that the estimated number of new breast cancer cases in the United States for 1993 is 183,000 [1], even a slight improvement in the diagnostic accuracy of mammography would benefit thousands of women.

In this paper, we report a study of stereo perception as an adjunct to mamrnographic screening. The work is based on the hypothesis that recognition of subtle abnormalities in a complex three-dimensional object, such as the matrix of glandular and fatty tissues of the breast, can be enhanced when the scene is viewed in stereo. The depth information provided

by stereo display may allow better radiographic definition of abnormal masses from similar surrounding normal tissues, increasing the observer's ability to distinguish and characterize abnormal masses. Experiments have been designed to investigate, in a systematic way, the effectiveness of stereo imaging in aiding the detection of abnormalities in simulated mammograms. A preliminary study involving fewer subjects has been reported [8]. We have since then extended our work to include more subjects.

Section II gives an introduction to stereo perception and discusses the potential benefits of stereo viewing for mammographic screening. Section III presents an overview of the experiments and describes the image generation process. The advantages of using computer simulated images, our mathematical model of the human breast and the types of abnormalities that are defined are also described in section III. Section IV contains a description of the experimental design and equipment. In section V, the basic concepts of ROC analysis are introduced. Section VI is the results section. A discussion of the results and the conclusion can be found in sections VII ancl VIII.

II STEREO PERCEPTION

Stereo perception, or stereopsis, refers to the impression of visual depth created by binocular parallax of the images cast upon the left and right retinas [9]. The two dissimilar retinal images are fused in the visual centers of the brain to obtain a three dimensional appreciation of depth. The fact that one can see well with only one eye indicates that monocular cues such as linear perspective, occultation, shading, shadow, and texture can also provide a sense of depth [10, 11, 12, 13]. However, in the domain of x-ray imaging, radiologists have minimal monocular depth cues. In mammography, in particular, the observer may not know the exact shape of a possible tumor. Binocular stereo vision could well be of great benefit in mammography, since it is the most powerful depth cue for scenes viewed close at hand and it is sufficient for perception of objects even in the absence of other cues such as color and contour[14]. Stereo vision can help in resolving ambiguities by revealing position, form and structure of objects. Intriguingly, as indicated in [15], stereopsis may access specialized

brain centers that allow an observer to look through the clutter of insignificant depth planes to concentrate visual attention on a triangulated target. Visual information processing in the brain is organized with relatively large areas devoted to binocular stereopsis and the analysis of depth cues. By using stereo viewing, observers may utilized these areas of the brain to achieve better accuracy in mammographic screening.

Stereo techniques have been available to radiologists for decades [16]. In 1898, just three years after the discovery of x-rays, J. Mackenzie Davidson studied stereo x-ray images[17]. However, the poor quality of stereo display devices in the past, and more recently, the availability and interest in computed tomography (CT) and magnetic resonance imaging (MRI), have led to the neglect of stereo x-ray research. The particular clinical requirements of mammographic screening for breast cancer, however, create a special case in which CT and MRI may not be appropriate for routine use. The most recent American Cancer Society guidelines call for baseline mammography in all women 35-40 years of age, and yearly mammography in all women over 50 years of age [18]. The large number of women who must be screened on a repeated basis, the relatively high radiation dose of CT and the relatively high cost of MRI (about $1000 for MRI versus $60 for mammography), make these imaging approaches inappropriiate for widespread application to breast cancer screening. Furthermore, recent improvements in technology have made realistic high-quality three-dimensional x-ray imaging possible. For initial diagnostic screening, stereoradiographic approaches may well fill a technological niche of considerable public health importance.

III OVERVIEW OF THE EXPERIMENT AND IMAGE GENERATION

To determine if stereo viewing has any effect on the accuracy of abnormality detection, we have designed experiments to systematically investigate the effectiveness of stereo viewing for detection of abnormalities in the density and arrangement of simulated tissue densities. Twenty two subjects have participated in the experiments. Subjects include the authors, some graduate students and some undergraduate students. The perceptual tasks that are required in the experiments are analogous to those required of radiologists; in the diagnosis

of breast cancer, but do not require radiological training. The use of lay subjects for the experiments is therefore appropriate. The accuracy and precision with which lay subjects can detect abnormalities using stereo versus mono viewing mode will define in a fundamental sense the potential diagnostic benefits of depth information provided by digital stereo displays.

Computational models are used to create computer simulated images. In this way, fundamental questions relating to the virtue of stereo displays as an aid to human perception can be answered without exposing any human subjects to radiation. Also, the cost of obtaining images for analysis is minimal. Other advantages of using computer simulated images are that

1. the "ground truth" about the images is known exactly, since the abnormalities are deliberately created and mathematically defined;
2. the number of possible abnormalities is unlimited, and the nature, background, and context of the abnormalities can be systematically varied to determine under what circumstances perception and diagnostic performance are most and least influenced by stereo display techniques;
3. a coniputational model of the breast can be made more anatomically realistic and more complex than physical models (phantoms), such as those constructed from resin;
4. a conlputational model is exactly reproducible; and
5. full control over the image formation process is possible.

Our mathematical model of a breast consists of a large, truncated hemisphere (radius of 8 cm) with a small sphere at its apex as the nipple. "Truncation" is used to simulate flattening of the breast between compression plates during mammographic examination. Approximately 70 spherical densities are distributed beneath the "skin" of the breast-like hemisphere in place of the glandular and connective tissue elements of the breast. The mean size of the embedded spheres is 0.6 cm radius and the mean density is set to a level

sufficient to provide image contrast similar to that in clinical mammogramis. The embedded spheres have a Gaussian distribution of size, density and center coordinates within the model. Limited random variations in sphere size and density are included to mimic normal biological variations, from which "abnormal" features have to be distinguished. A diagram illustrating our mathematical breast model is shown in Figure 1.

Three types of images have been generated for the experiments:

Images with no abnormality. These are control images with aforementioned normal variations only. Figure 2 shows an example of such an image.

Images with abnormal density. In these images, there is increased density of one sphere in the population, relative to the remaining normal ones. Figure 3(a) shows an example of a density test image. The abnormal sphere is highlighted in Figure 3(b). The mean abnormal density is 3 standard deviations above the mean density of normal spheres. The task of the observer is to identify if an image contains a relatively denser sphere. This perceptual task is considered psychophysically analogous to the task of locating abnormal densities that may be associated with breast cancers in clinical mammograms.

Images with an abnormal arrangement pattern. We define "daisy rings" as abnormal formittions of six spheres tangential to one another and surrounding a seventh central sphere, all lying in the same plane. These non-random groupings :resemble a daisy, the central sphere being the heart of the daisy and the six outer spheres being the petals. An example image is shown in Figure 4. The diagnostic task of the observer is to detect if a randomly oriented abnormal daisy grouping is present in an image. This perceptual task is considered psychophysically analogous to the task of locating architectural distortion of normal tissue densities that may be associated with signs of malignancy in clinical mammograms.

Simulated images are computed for x-rays originating from a source, passing through the simulated tissue volume containing the breast model, and striking an image plane. The image plane is described by a rectangular grid of the desired resolution. For each pixel in

the image plane, a ray is cast from the x-ray source, through the scene into the center of the pixel. This ray is tested for intersection with each of the objects in the breast model. For each successive object the ray intersects, the amount of attenuation is computed using Beer's law for absorption of photons by radiodense materials[16] :

$$N = N_0 e^{-\mu x}$$

where

N = *number of transmitted photons*

N_0 = *number of incident photons*

μ = *linear attentuation coefficient (object density)*

x = *object thickness.*

Object densities are specified in an input file. Object thickness is given by the distance that a ray passes through an object. It is computed by finding the intersection points of the ray with all spheres along its path and then obtaining the distances between the intersection points in sequence. Details of computing the intersection points between a ray and a sphere can be found in [19]. The intensity of the emerging ray when it reaches the image plane is recorded. Since a mammogram is a negative image, the inverse of intensities collectively form the simulated mammogram.

The left and right images which are necessary for stereo viewing can be generated by irradiating the breast from two different perspectives, in succession, corresponding to positions of the left and right eyes when viewing an object at arm's length. Using the ray tracing technique as described above, stereo images are easily generated by moving the x-ray source. Columniation of x-rays can be simulated by moving the point source relatively far from the image plane to create parallel rays. If desired, magnification views can be simulated by moving the point source closer to the tissue and the image plane farther from the tissue. Thus, many features of actual radiographs are present in the computer generated images.

IV EXPERIMENTAL DESIGN

Each subject participates in four, approximately one hour sessions:

- Arrangement abnormalities (Stereo)
- Arrangement abnormalities (Mono)
- Density abnormalities (Stereo)
- Density abnormalities (Mono)

Use of separate stereo and mono sessions makes the perceptual tasks less complicated and easier to learn. It also helps subjects to keep response criteria constant throughout each session. To balance reading order and learning effects, alternate subjects start with the stereo session.

The CrystalEyes system from StereoGraphics Corporation is used to display images. It consists of a pair of glasses, an infra-red emitter and a graphics display controller (GDC3). Figure 5 shows the experimental apparatus. There is a bypass switch on the GDC controller which allows the selection of stereo or mono viewing mode. For both the stereo and the mono sessions of the experiments, the stereo mode of the GDC controller is used and the subject wears active liquid-crystal display (LCD) glasses. This ensures that the viewing condition is kept constant between the stereo and mono sessions.

A stereo image displayed in mono mode is shown in Figure 6(a). In the stereo mode, the GDC controller doubles the refresh rate of the screen so that the left image (top image) and the right image (bottom image) are alternately displayed on the full screen. Due to the fast screen refresh rate, the perception of an image when viewed without the LCD glasses is as shown in Figure 6(b). A vivid three-dimensional stereoscopic depth effect can be achieved by presenting each eye with its own perspective view of the scene. To achieve this effect, the observer wears a pair of wireless, infra-red controlled LCD glasses. Each lens is electrically controlled to be opaque or transparent so that the right eye sees only the right image (the

left eye is blocked by the opaque lens). For the next video frame, the right eye is blocked and the left eye sees only the left image. The switching rate is 144/sec, providing flicker-free perception of the scene.

For the mono sessions, the images are formatted in such a way that when the GDC controller is in the rnono mode, the left image is displayed on the top and bottom of the left side of the screen. Similarly, the right image is displayed on the top and bottom of the right side of the screen. Since the images on the top and bottom are the same, doubling the refresh rate simply means a vertical stretching of the image. The left and right images are simultaneously visible to each eye, hence no depth can be perceived. Figure 7 shows a sitmple image that subjects see during the mono session.

Images of the same computational models are presented in both the stereo and mono sessions. This control is to ensure that if a case sample is atypically simple or atypically difficult, it will be so for both modalities. In this way, the performance difference between the two modalities will be an accurate measure [20]. The order of presentation of the images is randomized for each experimental session so that subjects cannot derive clues from the order of the images. No image is shown more than once during a session.

Arrangement and density abnormalities are tested in separate experimental sessions. In this way, the performance of stereo versus mono viewing for detecting each type of abnormality can be independently analyzed. There is at most one abnormality present in each image.

The experimental session is fully automated to minimize any subtle effects of investigator interactions with subjects. Standardized lighting and viewing distance are: also maintained. An introduction to the experiment and instructions for responding are presented on-screen. Training images, giving examples of the types of abnormalities to be searched for, are displayed as part of the subject training sequence. These training images serve to acquaint the subject with the visual display and the nature of abnormalities (either densities or daisy rings) to be identified in a given session. In the training session, feedback is provided to the subject, as recommended by Straub et al [21]. The system highlights the abnormal object (see Figure 3(b) and Figure 4(b)) when the subject presses a key to indicate that he or she

is ready to see the answer.

The actual experimental session begins immediately after the training session is completed. In the experimental session, test images are displayed and keystrokes, indicating the subjects' responses, are recorded by the computer system. Depending on whether the subject is participating in the stereo or mono session, the appropriately formatted images are shown, one at a time, on the video display. The subject is allowed to view each image for a maximum of 30 seconds. The time limit is set to simulate the fact that radiologists have limited time to spend on each mammogram. It also limits that maximum time that an experimental session can last. If desired, the subject can enter the response in less than 30 seconds by pressing the n-key for "next".

After viewing each image, subjects are asked to rate the image for abnormality on a graded scale of 1 to 5.

A response of 1 indicates definitely no abnormality is present.

A response of 2 indicates probably no abnormality is present.

A response of 3 is an equivocal response i.e. possible abnormality.

A response of 4 indicates an abnormality is probably present.

A response of 5 indicates an abnormality is definitely present.

The rating scale is displayed whenever responses are expected from the subject. In this way, the subject does not have to memorize the scale. This display also helps to reduce errors due to misunderstanding of the meaning of the numeric scale. Subjects are given feedback after each response. A tone is automatically sounded if the image that has just been rated contains an abnormality. In each session the subject evaluates 60 images, with the knowledge that exactly half of the images contained a particular abnormality.

V DATA ANALYSIS

Receiver operating characteristic (ROC) analysis has been accepted as the most rigorous and objective means of comparing diagnostic imaging modalities in radiology [20, 22]. It is often used for contrasting the technical potential of one modality with that of others. Many studies utilizing ROC analysis in radiology are referenced in [23]. In mamirnography-related research, ROC analysis has been used to characterize the accuracy of mammography [24], to compare the performance of mammography and palpation [7] and to characterize the spatial resolution requirement and the effect of unsharp-mask filtering on the detectability of subtle micro-calcifications in digital mammography [25]. ROC analysis has also been utilized in a study on the effect of attention-cueing on breast cancer detection performance [26, 27].

For our experiments, observer response data are evaluated by ROC analysis using standard techniques for five category ratings as described by Metz [22]. ROC curves are constructed in which the correct detection rate of a particular abnormality is plotted as a function of the false alarm rate.

The sensitivity of a diagnostic procedure, or the true positive fraction (TPF) refers to the fraction of patients actually having the disease that are correctly diagnosed as positive, i.e.

$$Sensitivity = TPF = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

The specificity of a diagnostic procedure or true negative fraction (TNF) refers to the fraction of patients actually without the disease that is correctly diagnosed as negative, i.e.

$$Specificity = TNF = \frac{True\ Negatives}{True\ Negatives + False\ Positives}$$

Additionally, in ROC analysis it is convenient to define the false positive fraction,

$$FPF = 1 - TNF = 1 - Specificity$$

as an objective measure of the response bias or confidence criterion of an individual observer.

Unless the discrimination capacity of a diagnostic image is perfect (Sensitivity = Specificity = 100%), there will always be some overlap in image characteristics of those with and without the disease. This results in difficult and ambiguous cases which make the sensitivity and specificity of diagnostic procedures less than 100%.

According to the ROC model (Figure 8), a radiologist or an observer decides to render a positive or negative diagnosis by comparing his or her confidence concerning each image with an internal confidence criterion. If confidence in a positive diagnosis exceeds this confidence criterion, the image is read as positive and vice versa. If the observer is keen to make a positive diagnosis and desires to minimize false negative readings, then the false positive fraction will be increased. This mind-set is appropriate in breast cancer screening, in which a radiologist generally prefers high sensitivity (TPF), even at the expense of a high false positive fraction, because of the great importance of detecting early breast cancers. The penalty to the patient for a false positive diagnosis (a negative biopsy) is much less than the penalty for a false negative one (a continuing malignancy). In current clinical mammography, the false positive fraction is in the range of 30% to 60% [7, 24].

By nature, human observers may vary greatly in terms of individual confidence criterion or tendency to under-read or over-read. ROC analysis, however, allows comparison of the diagnostic accuracies of imaging systems, despite variability in the confidence criteria. Diagnostic performance data from single or multiple observers are analyzed to create a curve of TPF (sensitivity) as a function of FPF (a measure of positive response bias). As FPF increases, TPF increases in a curvilinear fashion from the lower left to the upper right quadrant of the unit square (Figure 9). Points on the curve in the upper right quadrant indicate less strict confidence criteria and less specificity. Points on the curve in the lower left quadrant indicate more strict confidence criteria and more specificity. The entire curve describes both the sensitivity and specificity of the observations and represents all of the trade-offs between sensitivity and specificity that can be achieved by a diagnostic system as the confidence criterion is varied. A curve describing the performance of a perfectly discriminating observer will indicate 100% TPF, even when the FPF is vanishingly small. In this case the area under the ROC curve, usually denoted A_z, will completely fill the unit square. For inherently im-

perfect observations, the area under the ROC curve is a hybrid summary index, describing the performance of the observer using the particular technology under consideration.

To compare the performance of observers using a modified or innovative technology with the performance of the same observers using conventional technology, it is necessary to gather performance data for a series of images in which diagnostic truth is known and then to construct the ROC curves [20, 28]. In the domain of medical imaging, ROC curves are most commonly assumed to have the binormal functional form [22]. The two adjustable parameters of binormal ROC curves can be fitted from the ROC data by using the maximum likelihood parameter estimation scheme [23]. If for the same observers, the curve for the new technology lies above the curve for the conventional technology, there is objective evidence that the new technology permits a greater fraction of correct diagnoses, regardless of variability in observers' bias for or against making a false positive diagnosis.

A problem that may occur with ROC curve fitting is a degenerate data set. The most common degeneracy occurs when the observer does not distribute his or her responses more or less uniformly over all the possible rating categories. One of the subjects in our experiment did not use categories 1 and 3 in his responses. Another subject used category 3 most of the time. These idiosyncracies resulted in degenerate data sets. The data sets had to be discarded, since salvage of degenerate data sets is not recommended [20]. To minimize occurrences of degenerate data sets, subjects are instructed to use all categories and to distribute their responses uniformly over the rating scale.

In our experiments, a maximum likelihood curve fit for data from each subject is computed. For non-degenerate data sets, stereo and mono ROC curves for individual subjects are generated. Student-t tests of paired differences in the stereo and mono summary index, A_z, are performed to determine significance. Paired t-tests are performed instead of non-paired t-tests to minimize the effects of inter-observer differences such as differences in observer's skill and experience. For example, the significance of the results may be obscured by the tendency of better skilled and experienced observers to perform better and the tendency of others to perform poorly. This difference in performance increases the apparent variance and thus decreases the significance of any difference in the performance of the two modalities

if a non-paired t-test is used. In order to visualize the overall performance of all subjects, combined ROC curves are plotted using the average parameter values for all subjects. This is the recommended method to use when a heterogeneous observer population is studied [23].

VI RESULTS

A Arrangement Experiment

Of the 22 result sets from the arrangement experiment, there is a degenerate data set that had to be discarded (see discussion in section V). Figure 9 shows the overall performance of 21 subjects who participated in the arrangement experiment. The shape of mono ROC curve is typical for a visual detection experiment in radiology [22, 25] indicating that an appropriate perceptual task is required of the subjects, neither too trivial nor too difficult. In comparing the two modalities, the TPF for any given FPF in Figure 9 is clearly greater for stereo than for mono views. The difference in the A, index (describing the area under the ROC curve) is statistically significant (t=4.128447, p=0.000521) using a 2-tailed paired t-test. Hence, there is clear and significant benefit in using the stereo modality, even though exactly the same visual information is presented to the subjects in mono and stereo formats – except for the creation of stereoscopic depth.

B Density Experiment

Of the 22 :participants of the experiment, two had degenerate data sets, leaving 20 analyzable data sets. The overall performance of the 20 subjects is shown in Figure 10.. There is a slight difference between the ROC curves for the two modalities, with the stereo ROC curve lying above the mono ROC curve. Using a 2-tailed paired t-test, the difference in area under the ROC curve, A_z, is not statistically significant (t=1.392338, p=0.179903).

VII DISCUSSION

The purpose of the present experiments is to determine which viewing modality (stereo or mono) leads to better diagnostic performance. Our aim is not to measure the absolute detectabilities of abnormalities using the two viewing modes. In order to achieve our purpose, real-world viewing conditions are simulated as closely as possible in the design of the experiments. To simulate time constraint of radiologists, a maximum tinne limit is set for viewing images. Subjects were not asked to report any visual problems that they may have had. Subjects with different vision level are included in the experiment so that the potential utility in actual radiologic practice can be accessed.

The choice of which images should be included in an ROC experiment is not easy, but an appropriate choice is very important. If the abnormalities to be detected are too conspicuous, subjects will perform well regardless of the modality used. Similarly, if the abnormalities are chosen to be too subtle, poor performance will be recorded for all modalities. A rule of thumb that has been suggested for determining the most appropriate level of case difficulty is that tht: average A_z of two modalities under consideration should lie near the range 0.75 to 0.8 [20]. For our study, the average A_z index for the arrangement experiment is 0.78 and that for t'he density experiment is 0.73, indicating that the experiments have appropriate difficulty levels.

The results of the experiments allow us to conclude confidently that stereo viewing provides higher detectability of arrangement abnormalities. This finding is consistent with expectations since stereo viewing increases detectability of shapes and structures. Also, after the experimental sessions, when asked about their preferences of the two modalities, all subjects except one have preference for the stereo sessions over the mono sessions. The subject who preferred the mono sessions mentioned that the stereo sessions are blurry. It should be noted that this subject has astigmatism which is known to affect stereo vision if it has not been properly corrected.

Our experimental results also show that stereo viewing has an advantage over mono viewing

on the observers' ability to detect density differences. However, the performance difference is not statistically significant. There is great variation in the performances of the subjects. In our earlier work with fewer subjects [8], a low significant difference (at 4% significance level) is found in the performance of the two modalities – with stereo mode out-performing mono mode.

The following are some points that should be noted about the experiments. A basic assumption that we make in the study is that any given observer is equally skilled with the two imaging modalities in question. This may not be true in general but any differences in skill of stereo versus mono modalities are minimized by the training session which each observer goes through before the actual experiment begins. Some subjects indicated that the training sessions are too short. However, a compromise has to be made between shorter training session and longer overall experimental session. As it is, each subject takes an average of 45 minutes to an hour to complete an experimental session. Another factor that may affect detection performance is the difference in overall image intensities for images displayed in the stereo session and the mono session. For the stereo sessions, the left and right images are being alternately displayed in quick succession. For the mono sessions, the left and right images are displayed side by side for the entire viewing time. Hence the overall image intensity for mono images is higher than that for the stereo images. This difference may have an effect on the performance of the observers in detecting abnormalities. Stereoacuity has been found to increase as the retinal illuminance increases, until at high intensities the curve approaches asymptotically a limiting value [29]. The screen phosphor decay and retinal sensitivity to the persisting image can result in some ghosting when stereo viewing is used. Since the problems with intensity levels and ghosting affect the stereo sessions only, if new stereo display technology that overcomes these problems becomes available,,the performance using stereo viewing may be even better.

VIII CONCLUSION

We have applied modern display technology which allows comfortable .viewing of three-dimensional stereo images to diagnostic mammography. We have also developed the novel concept of creating simulated x-ray medical images by ray-tracing. To our knowledge this is the first scientific evaluation of stereo versus mono medical imaging using ROC curve analysis. This paper reports fundamental results on the contribution to abnormality detection provided by stereo viewing and demonstrates the potential of stereo mammography for early detection of breast cancer.

References

[1] C. C. Boring, T. S. Squires, and T. Tong, "Cancer statistics, 1993," CA: a Cancer *Journal* for Clinicians, vol. 43, no. 1, pp. 7–26, 1993.

[2] "Cancer facts and figures." American Cancer Society, 1992.

[3] "Digital mammography development group (Naional Cancer Institute) PA-92-57," NIH Guide, vol. 21, no. 12, March 27 1992.

[4] E. R. Fisher and J. D. Paulson, "Identification of early changes in breast cancer," in *Early* Diagnosis of Breast Cancer (E. Grundmann and L. Beck, eds.), pp. 65–73, New York: Fischer, 1982.

[5] P. Strax, Make Sure You Do Not *Have Breast Cancer.* New York: St. Martin's Press, 1989.

[6] T. B. Hunter and L. L. Fajardo, "Digital genitoruniary, gastrointestinal, and breast radiology," in Digital *Imaging in Diagnostic Radiology* (J. D. Newell and C. A. Kelsey, eds.), pp. 43–70, New York: Churchill Livingstone, 1990.

[7] J. K. Gohagan, E. L. Spitznagel, M. McCrate, and T. B. Frank, "ROC analysis of mammography and palpation for breast screening," *Investigative* Radiology, vol. 19, pp. 587–592, 1984.

[8] J. Hsu, C. F. Babbs, D. M. Chelberg, Z. Pizlo, and E. J. Delp, "A study of the effectiveness of stereo imaging with applications in mammography," Proceedings of the SPIE Conference on Human *Vision, Visual Processing, and* Digital Display IV, vol. 1913, Feburary 1993. To appear.

[9] P. C. Dodwell, "Binocular vision and pattern coding," in Visual Pattern Recognition, pp. 120–136, New York: Holt, Rinehart and Winston, 1970.

[10] M. Livingstone and D. Hubel, "Segregation of form, color, movement, and depth: anatomy, physiology, and perception," *Science*, vol. 240, pp. 740–749, May 1988.

[11] J. T. Todd and R. A. Akerstrom, "Perception of three-dimensional form from patterns of optical texture," Journal of *Experimental* Psychology, vol. 13, no. 2, pp. 242–255, 1987.

[12] T. Troscianko, R. Montagnon, J. LeClerc, E. Malbert, and P.-L. Chanteau, "The role of colour as a monocular depth cue," *Vision* Research, vol. 31, no. 11, pp. 1923–1930, 1991.

[13] Z. Pizlo and A. Rosenfeld, "Recognition of planar shapes from perspective images using contour-based invariants," *Computer Vision,* Graphics and Image Processing : Image Understanding, vol. 56, no. 3, pp. 330–350, November 1992.

[14] B. Julesz, Foundations of *Cyclopean* Perception. Chicago: University of Chicago Press, 1971.

[15] W. Carter, "The advantage of single lens stereopsis," Proceedings of the SPIE Conference on Stereoscopic Displays and Applications III, vol. 1669, Feburary 1992, pp. 204–214.

[16] E. E. Christensen, T. S. Curry, and J. Nunnally, An Introduction to the Physics of Diagnostic Radiology. Philadelphia: Lee & Febiger, 1972.

[17] J. M. Davidson, "Remarks on the value of stereoscopic photography and skiagraphy," British Medical Journal, December 1898.

[18] G. D. Dodd, "American cancer society guidelines on screening for breast cancer: an overview," CA: a Cancer *Journal* for *Clinicians*, vol. 42, no. 3, pp. 177–180, 1992.

[19] A. S. Glassner, An *Introduction* to *Ray Tracing*. London; San Diego: Academic Press, 1989.

[20] C. E. Metz, "Some practical issues of experimental design and data analysis in radiological ROC studies," *Investigative Radiology*, vol. 24, pp. 234–245, 1989.

[21] W. H. Straub, H. Rockette, J. L. King, N. A. Obuchowski, W. F. Good, J. H. Feist, B. C. Good, and C. E. Metz, "Training observers for receiver operating characteristic (ROC:) studies," Proceedings of the SPIE *Conference* on Medical *Imaging* IV: PACS System Design and *Evaluation*, vol. 1234, 1990, pp. 126–130.

[22] C. E. Metz, "ROC methodology in radiologic imaging," Investigative *Radiology*, vol. 21, pp. 720–733, 1986.

[23] J. A. Swets and R. M. Pickett, *Evaluation* of *Diagnostic* Systems : Methods from Signal Detection Theory. Academic Press, 1982.

[24] J. E. Goin, J. D. Haberman, M. K. Linder, and P. A. Lambird, "Analysis of mammography: A blind interpretation of BCDDP radiographs," Radiology, vol. 148, pp. 393–396, August 1983.

[25] H. P. Chan, C. J. Vyborny, H. MacMahon, C. E. Metz, K. Doi, and E. A. Sickles, "Digital mammography: ROC studies of the effects of pixel size and unsharp mask filtering on detection of subtle microcalcifications," *Investigative* Radiology, vol. 22, no. 581–589, 1987.

[26] S. M. Astley and C. J. Taylor, "Combining cues for mammographic abnormalities," Proceedings of the First British *Machine* Vision *Conference*, September 1990, Oxford, pp. 253–258.

[27] S. Astley, I. Hutt, S. Adamson, P. Miller, P. Rose, C. Boggis, C. Taylor, T. Valentine, J. Davies, and J. Armstrong, "Automation in mammography : Computer vision and human perception," *Proceedings* of *the* SPIE *Conference* on Biomedical *Image* Processing, vol. 1'905, 1993. To appear.

[28] J. A. Swets, "ROC analysis applied to the evaluation of medical imaging techniques," Investigative Radiology, vol. 14, no. 2, pp. 109–121, 1979.

[29] H. Davson, ed., The Eye, vol. 4, ch. 15 (Spatial Localization Through Binocular Vision), pp. 2'71–324. Academic Press, 1962.

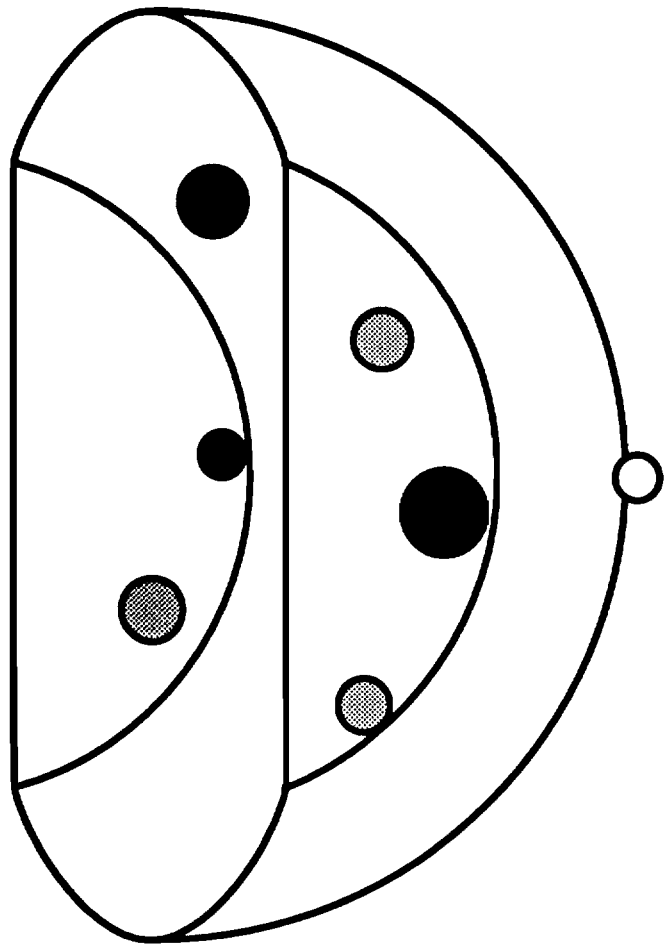

Figure 1: Mathematical model of a human breast. It consists of a hemisphere, truncated on both sides, with smaller spheres embedded within it.

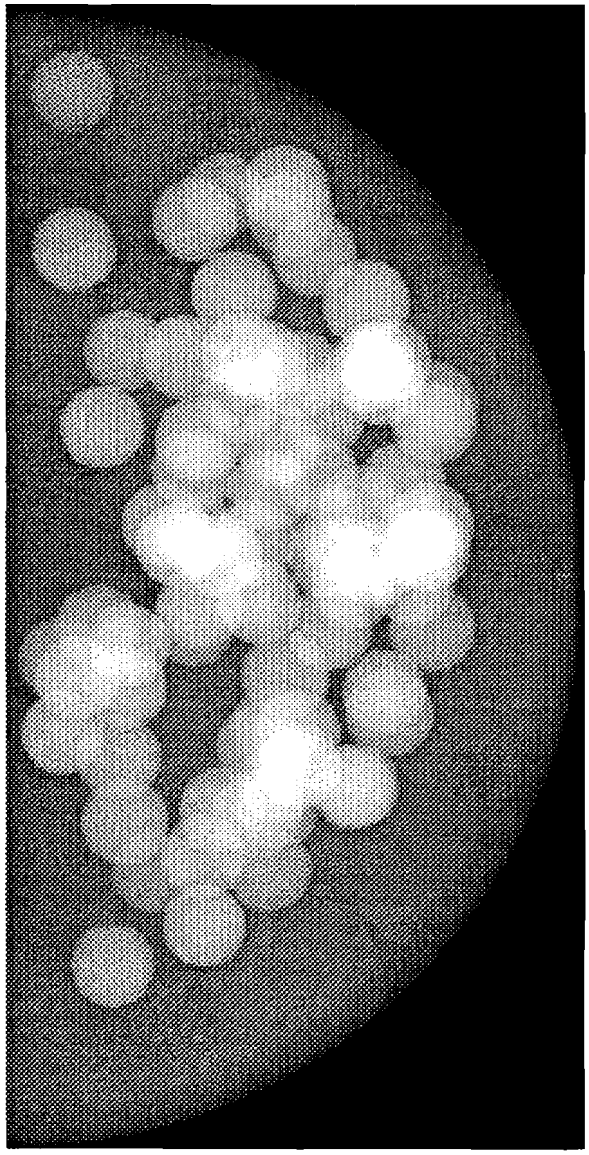

Figure **2:** An example of a test image with no abnormalities.

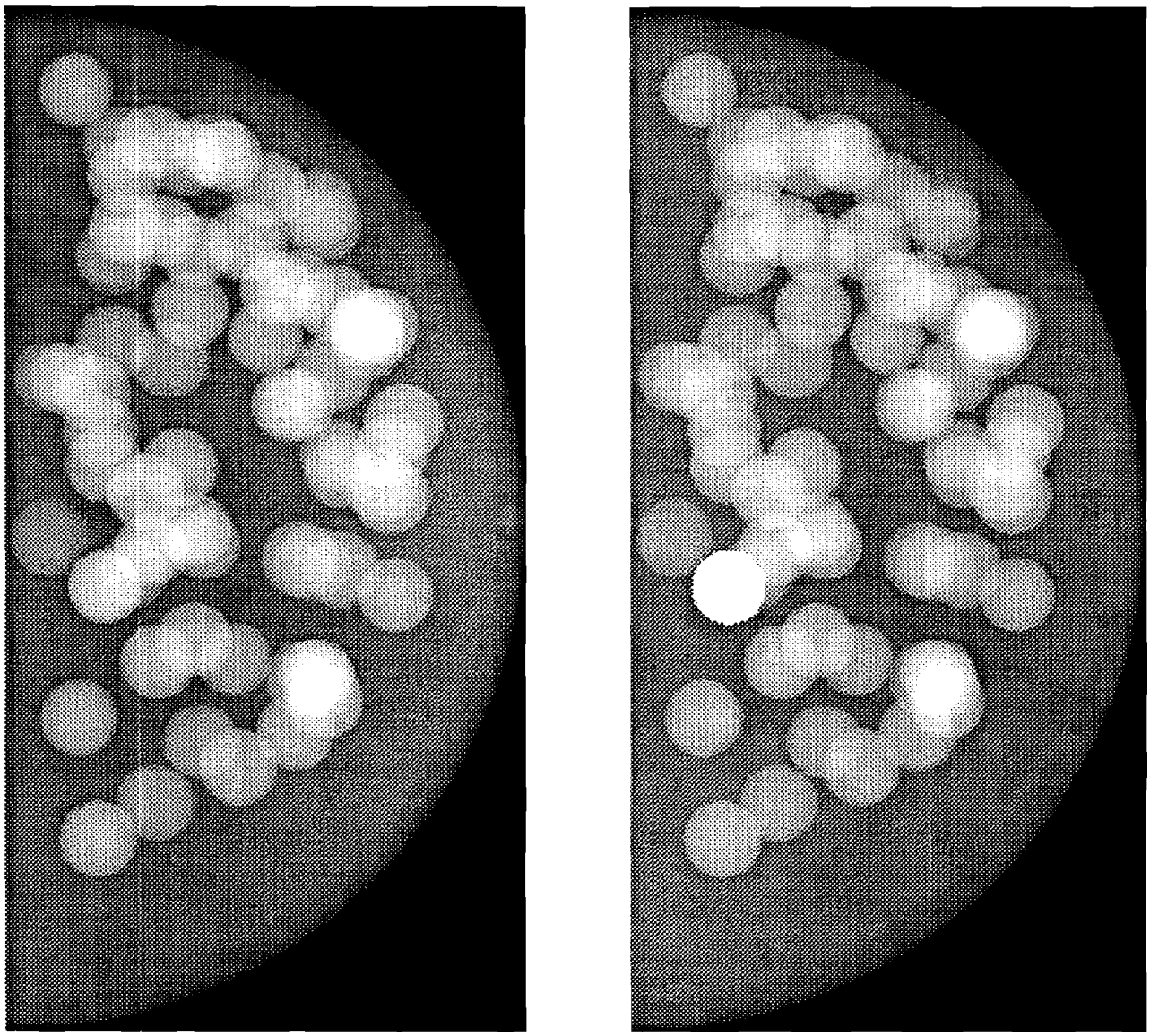

Figure 3: The left image is an example of a test image with density abnormality. On the right is the same image with the abnormality highlighted.

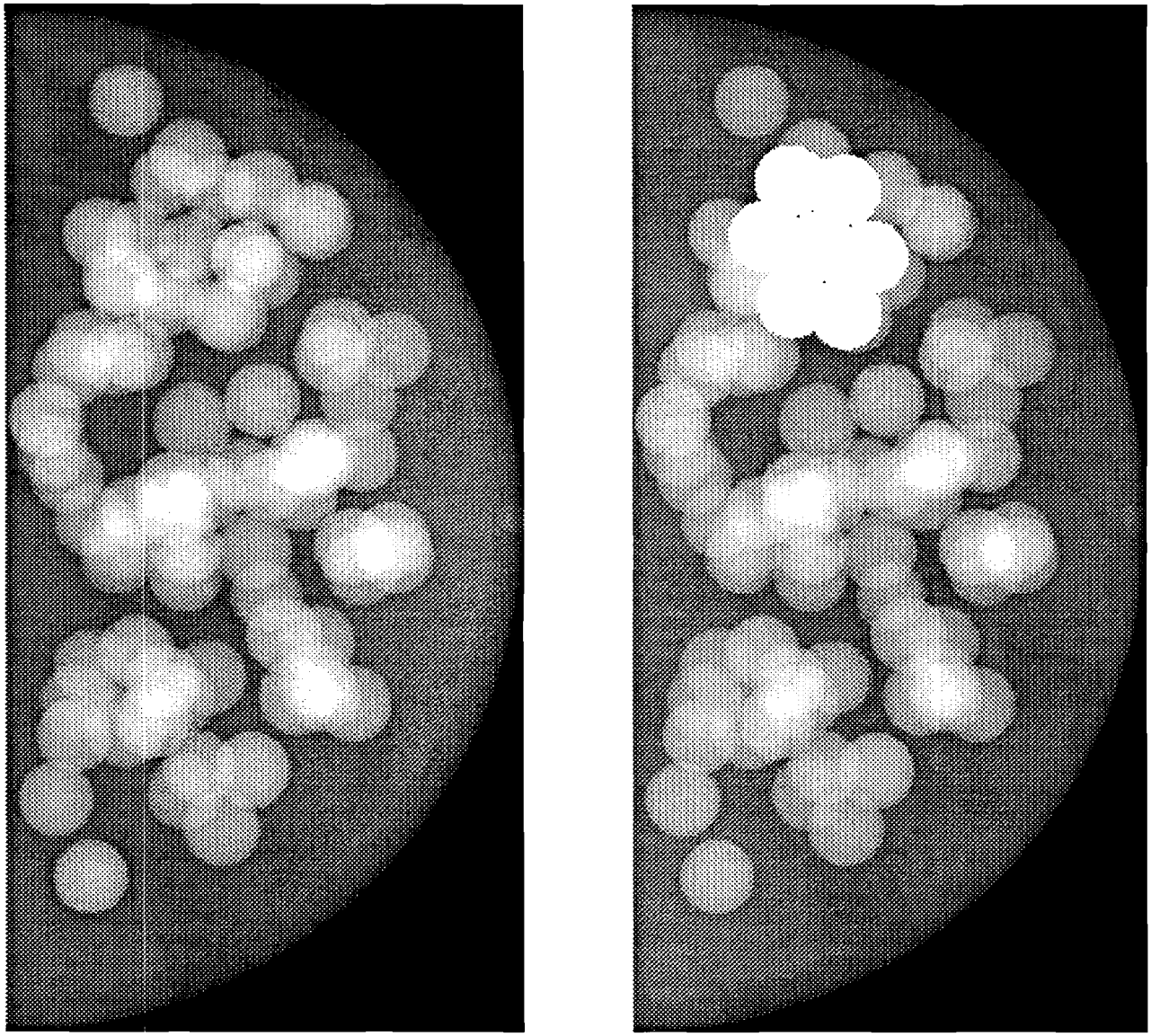

Figure 4: The left image is an example of a test image with arrangement abnormality. On the right is the same image with the abnormality highlighted.

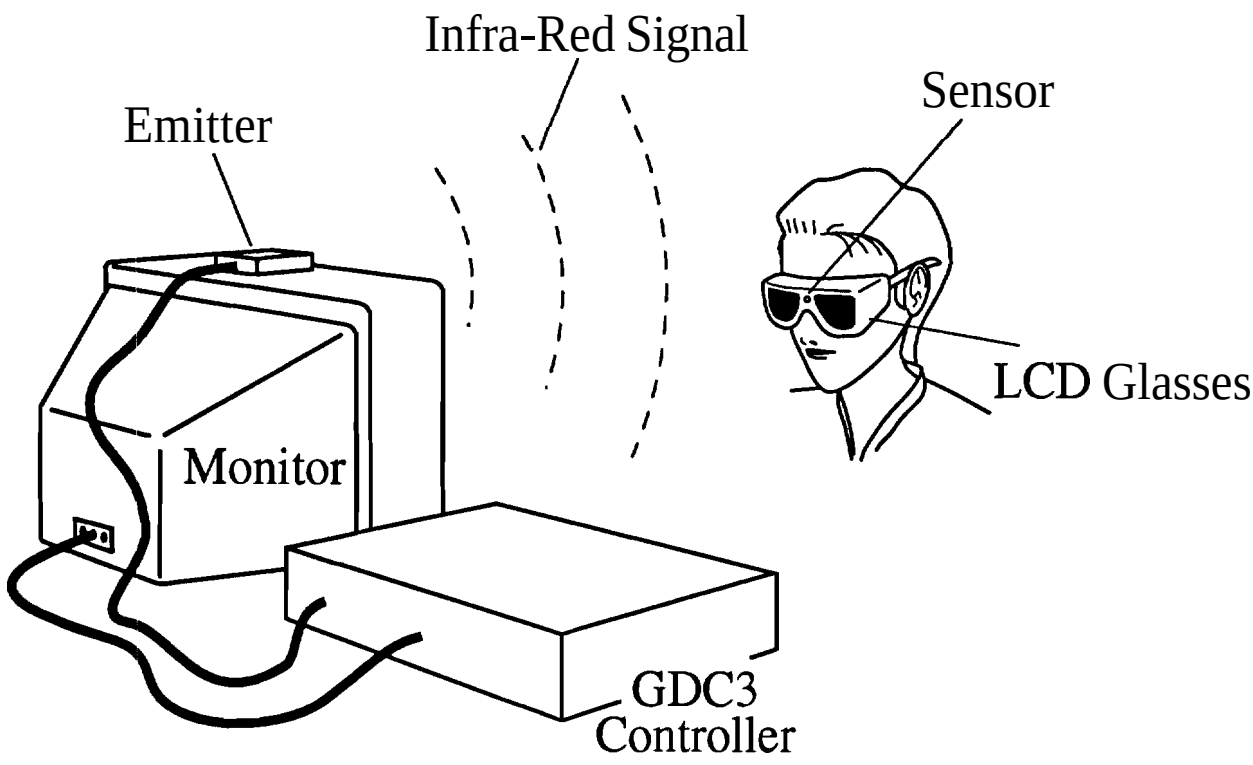

Figure 5: Experimental Apparatus

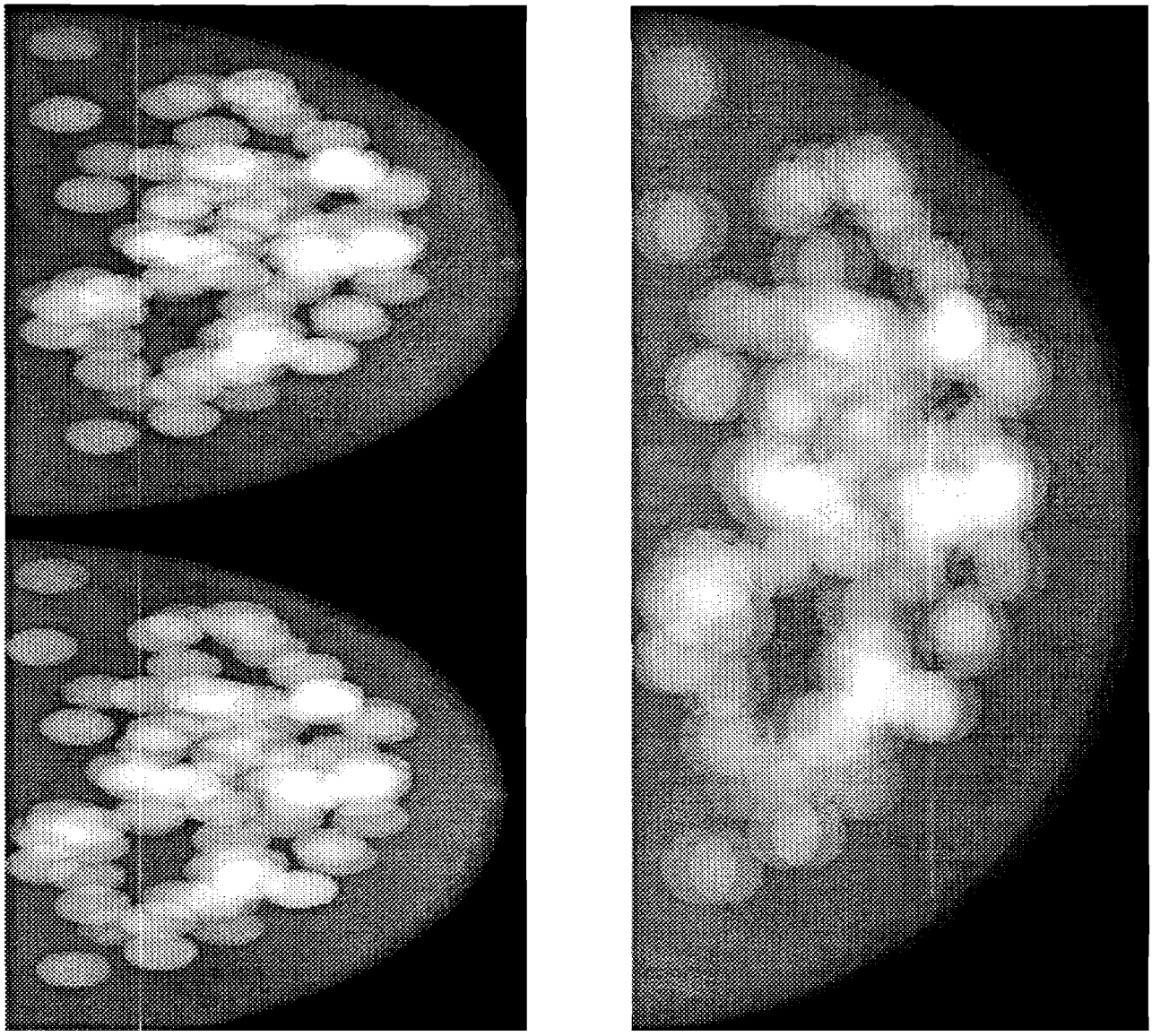

Figure 6: The left shows a stereo image displayed in mono mode and the right shows a stereo image displayed in stereo mode. LCD glasses have to be worn in order to see the right image in three-dimension.

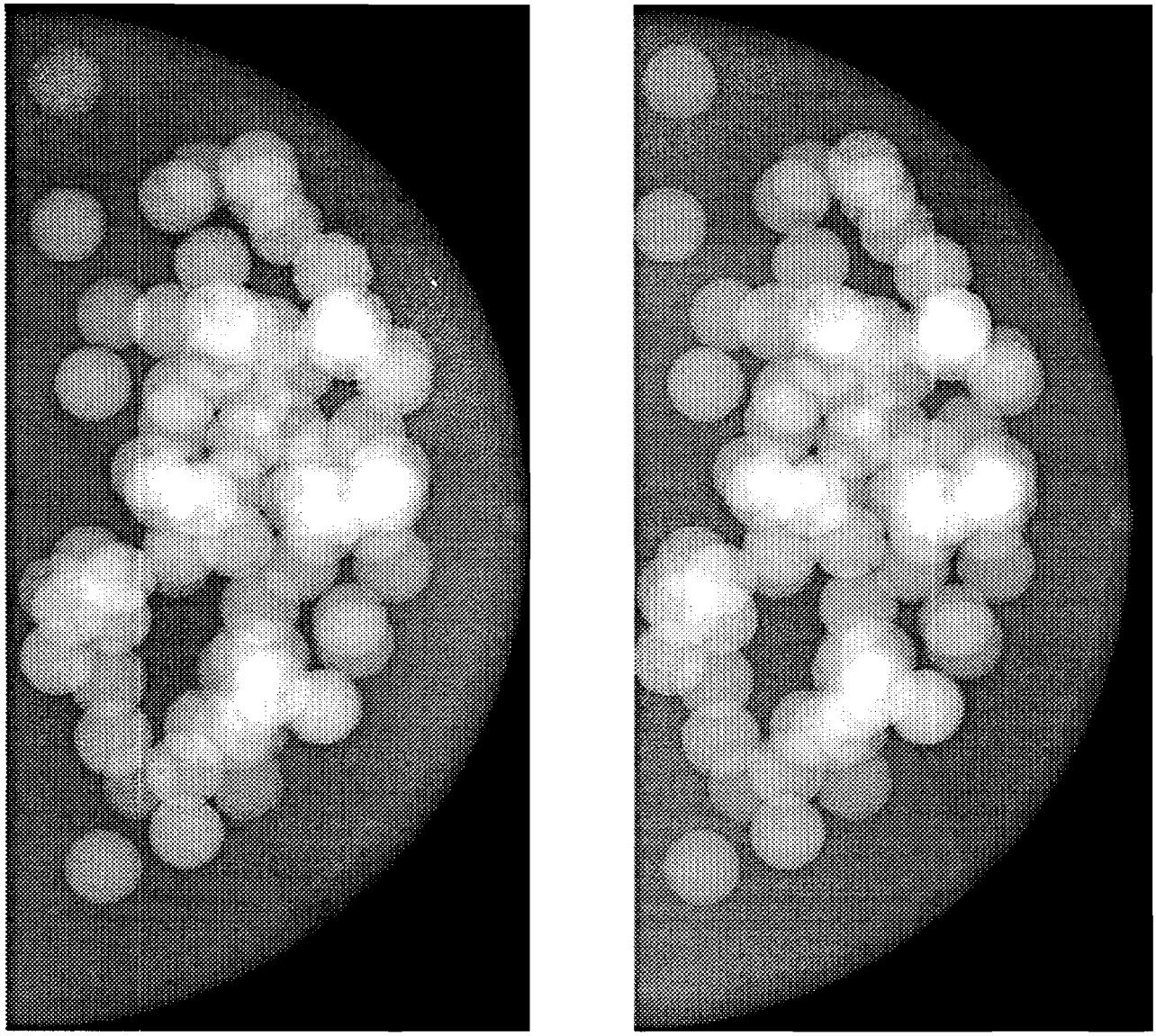

Figure 7: An example image from the mono session. Both the left and right images are displayed at the same time.

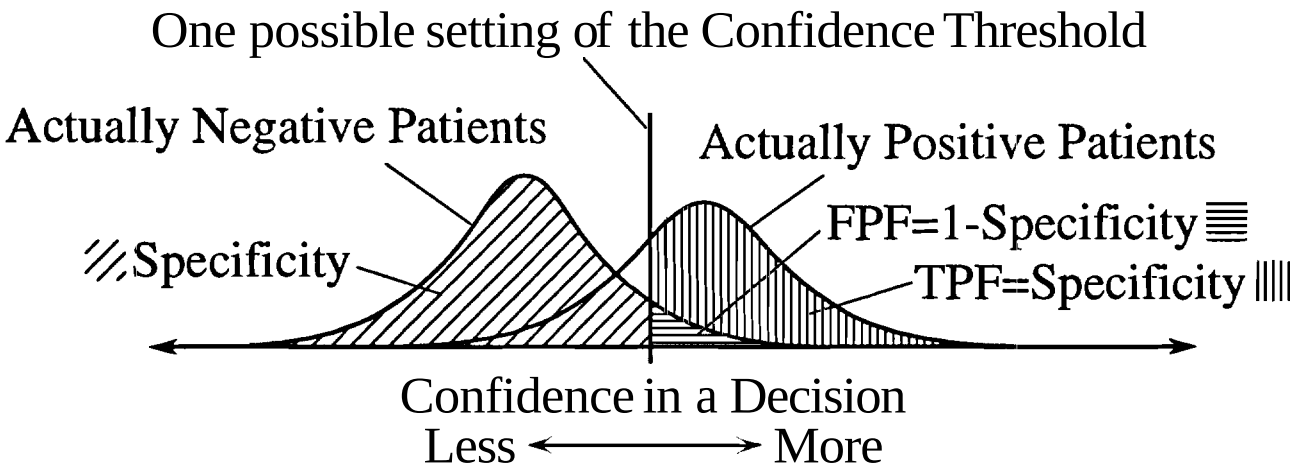

Figure 8: The ROC analysis model

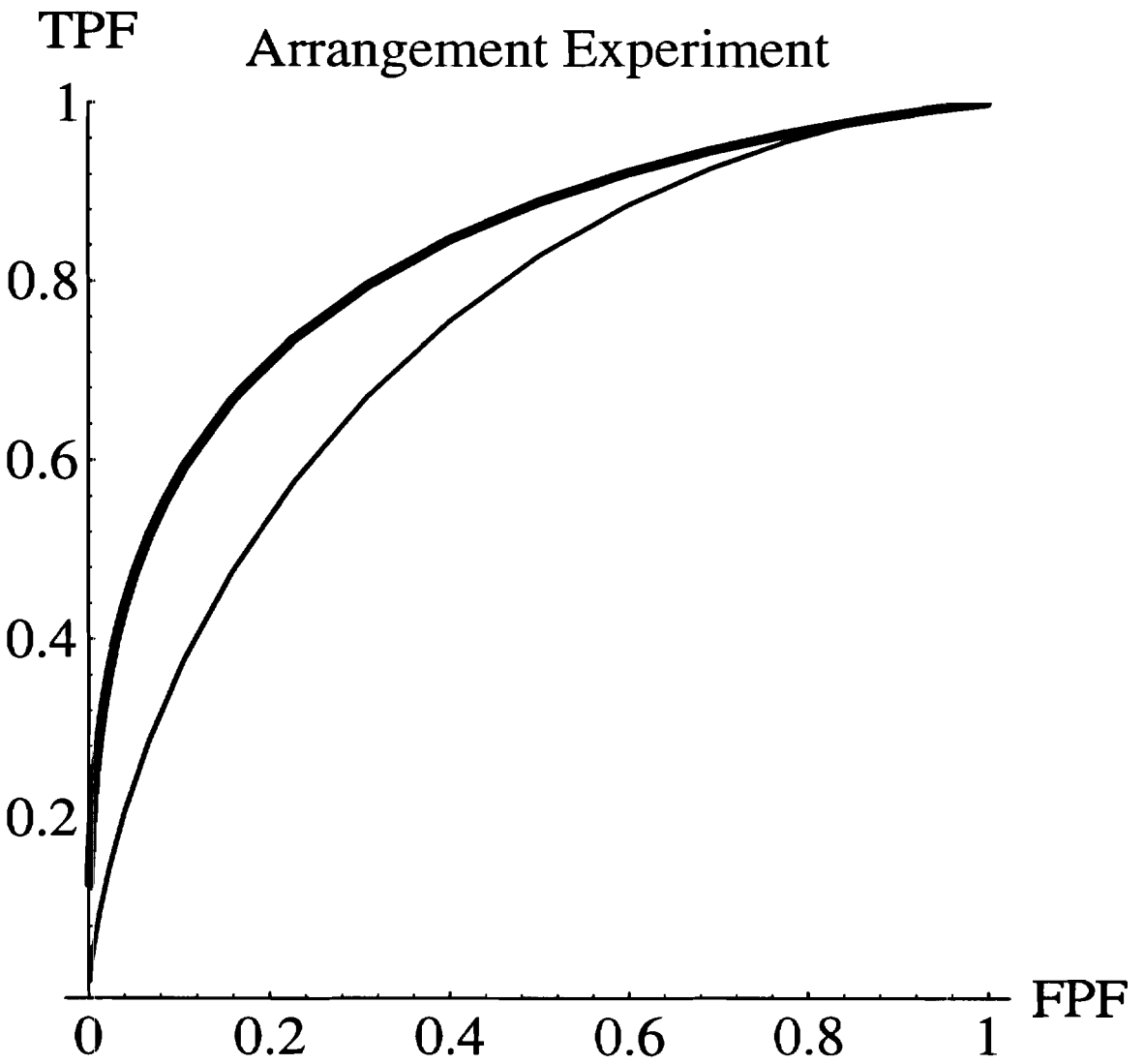

Figure 9: Combined ROC curve for Arrangement Experiment. Legend : thick lines (stereo viewing); thin lines (mono viewing).

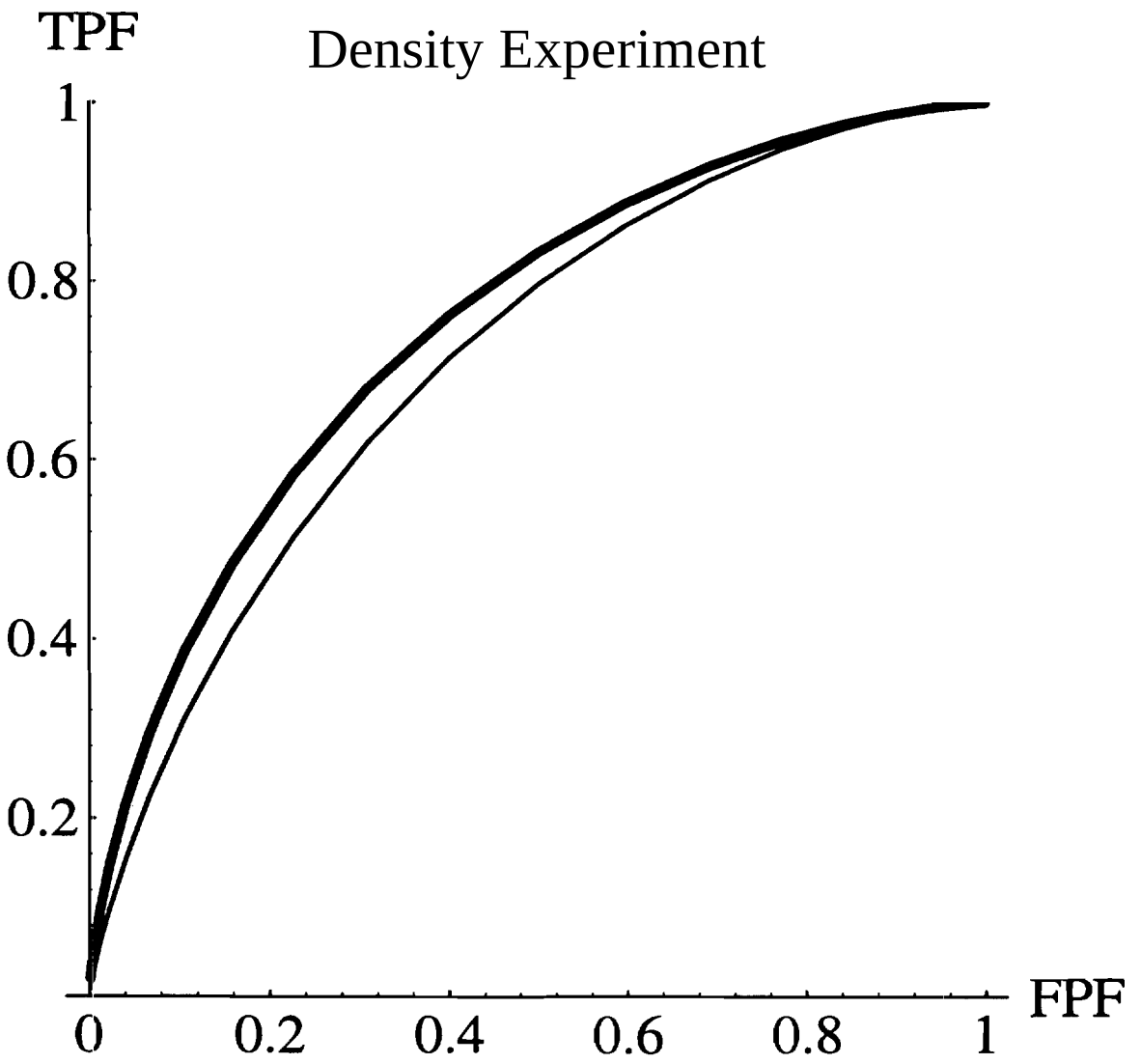

Figure 10: Combined ROC curve for Density Experiment. Legend : thick lines (stereo viewing); thin lines (mono viewing).