Breast Texture Synthesis and Estimation of the Role of the Anatomy and Tumor Shape in the Radiological Detection Process: from Digital Mammography to Breast Tomosynthesis

THÈSE N^{O} 4347 (2009)

PRÉSENTÉE LE 20 MARS 2009
À LA FACULTÉ SCIENCES DE LA VIE
LABORATOIRE DE PHYSIQUE DES HAUTES ÉNERGIES
PROGRAMME DOCTORAL EN PHYSIQUE

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Cyril CASTELLA

acceptée sur proposition du jury:

Prof. O. Schneider, président du jury
Prof. A. Bay, Prof. F. O. Bochud, directeurs de thèse
Prof. C. Depeursinge, rapporteur
Prof. M. P. Eckstein, rapporteur
Dr J. O. Prior, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2009

Table of contents

Résumé 1
Abstract 2
List of acronyms 3
1. Introduction and motivation 5
    Breast cancer: a challenge in diagnostic radiology 5
    Psychophysical studies 6
    History of psychophysical studies in medical imaging 7
    Intuitive approach to model observers 9
    Breast tomosynthesis: the answer to tissue overlap issues? 12
    Goals of the study 14
2. PhD milestones 15
    Statistically understanding mammographic texture 15
    Creating visually and statistically realistic breast texture 16
    Experimenting with a simple spherical signal 17
    Enhancing realism with benign and malignant signals 18
    Looking further: towards objective evaluation of breast tomosynthesis potential 19
3. Papers 21
    Semiautomatic Mammographic Parenchymal Patterns Classification Using Multiple Statistical Features 23
    Mammographic texture synthesis: second-generation clustered lumpy backgrounds using a genetic algorithm 37
    Human linear template with mammographic backgrounds estimated with a genetic algorithm 51
    Mass detection on mammograms: influence of signal shape uncertainty on human and model observers 63
    Masses detection in breast tomosynthesis and digital mammography: a model observer study 75
4. Conclusions and perspectives 85
Acknowledgments 87
References 89
Appendix: Screening Mammography from a practical point of view 97
Curriculum Vitae 99

Résumé

Le cancer du sein est le plus répandu et la principale cause de décès par cancer parmi les femmes dans le monde. Détecté suffisamment tôt, il peut cependant être traité de manière efficace, dans le sens où il est alors possible d'éviter des traitements lourds et de réduire la morbidité et la mortalité. C'est dans ce but qu'ont été instaurés depuis les années 60 des essais cliniques randomisés, puis des programmes de dépistage systématique du cancer du sein par mammographie.

Le processus de détection du cancer du sein en mammographie est complexe, et sa compréhension offre de nombreux défis aux radiologues et aux physiciens médicaux. Une manière d'appréhender cette problématique est de modéliser pas à pas le processus en réalisant des expériences psychophysiques avec des images anatomiques ou de synthèse. Dans cette approche, le contenu en information des images est contrôlé. Depuis les premières expériences avec des images de synthèse constituées de bruit blanc et de simples signaux géométriques, de nombreux progrès techniques et informatiques ont permis de s'approcher peu à peu de la réalité clinique pour étudier le mécanisme de la perception d'un signal présent sur une image radiologique.

Le présent travail étend la liste des outils utilisés jusqu'ici dans les expériences psychophysiques en mammographie. Il propose une analyse statistique détaillée des images anatomiques, à partir de laquelle des algorithmes de classification de la densité mammaire et de synthèse d'images réalistes sont développés. Dans une seconde phase, diverses expériences psychophysiques utilisant des signaux simples ainsi que des masses bénignes ou malignes superposées aux fonds anatomiques et synthétiques sont présentées. La performance d'observateurs humains est analysée en fonction de paramètres tels que le type de fonds, de signal, ou l'incertitude à propos de la taille ou de la forme du signal. Ces résultats sont comparés à ceux de modèles existants ou adaptés de la littérature, et ces derniers sont évalués dans leur aptitude à prédire la performance des observateurs humains pour la détection de lésions dans ces conditions.

Pour chacune des étapes de ce projet, l'accent a été mis sur le côté objectif et reproductible de l'évaluation des images ou de la performance des observateurs. Des conditions à la fois contrôlées et réalistes assurent la robustesse des résultats, ainsi que leur adaptabilité clinique. Parmi les principaux résultats, des images synthétiques de texture mammaire ont été générées et validées. Celles-ci fournissent une base de données virtuellement inépuisable d'images au réalisme visuel et statistique démontré. Concernant l'analyse de la performance des observateurs humains, ce travail montre, entre autres, que ceux-ci sont sensibles à une incertitude quant à la taille du signal, mais pas quant à sa forme, qu'ils utilisent des stratégies similaires avec les images réelles ou synthétiques, et qu'ils sont principalement sensibles aux fluctuations anatomiques dans la proximité immédiate du signal. Ces effets, ainsi que le niveau de performance des observateurs humains pour les diverses tâches de détection, ont pu être reproduits par des modèles prenant en compte des caractéristiques du système visuel humain.

Les conclusions de ce travail pourront guider de futures études dans le domaine de la détection en mammographie ou en tomosynthèse. Cette technique d'imagerie permet une analyse tomographique de la glande mammaire, et offre tout comme la mammographie un grand nombre de défis en vue de la compréhension et de la caractérisation objective de sa performance. Dans ce but, les études via des modèles d'observateurs validés pour des tâches de détection en imagerie médicale offrent une excellente approche en termes de temps et de coûts.

Mots-clés: Mammographie, vision et perception visuelle, imagerie médicale, évaluation de la performance, modèles d'observateurs

Abstract

Breast cancer is the most common, and the number one cause of death by cancer among women. However, when it is sufficiently early detected, heavy treatments can be avoided, and morbidity and mortality can be reduced. This is the reason why randomized clinical trials were started in the 60s, followed in the last decades by screening mammography national programs.

The process of breast cancer detection in mammography is complex. Its understanding offers numerous challenges to radiologists and medical physicists. One way to apprehend it is to model this process step-by-step by performing psychophysical experiments with anatomical or synthetic images. In this approach, the images information content is controlled. Since the first experiments with synthetic images created with white noise and simple geometric signals, technical and computational improvements allowed to get ever closer to clinical realism for studying the mechanism of perception of a signal on a radiological image.

The present work extends the list of tools that have been used until now in psychophysical experiments in mammography. It proposes a detailed statistical analysis of anatomical images, from which algorithms for breast density classification and realistic breast texture synthesis are developed. In a second phase, psychophysical experiments with simple signals and benign or malignant masses combined with anatomical or synthetic backgrounds are presented. The performance of human observers is analyzed as a function of parameters such as background type, signal type, or uncertainty about the size or the shape of the signal. These results are compared to that of existing or adapted models from the literature, and the different models are evaluated in their ability to predict the performance of human observers for the detection of lesions in such conditions.

Each step of this project focused on the objective and reproducible aspect of the image evaluation or of the observer performance. Controlled yet realistic conditions ensure the robustness of the results, as well as their clinical adaptability. Among the main results, synthetic mammographic texture images have been generated and validated. They provide a virtually unlimited database of images with demonstrated visual and statistical realism. Concerning the analysis of the human observers' performance, this work shows that they are sensitive to uncertainty about the signal size but not about its shape, that they use similar detection strategies with real and synthetic images, and that they are mainly sensitive to the anatomical fluctuations in the immediate area around the signal location. These effects, as well as the performance level of human observers for the detection tasks under consideration, could be reproduced by models taking into account human visual system characteristics.

The conclusions of this work will be able to guide future studies in the field of detection tasks in mammography or tomosynthesis. This 3D breast imaging technique presents, like mammography, numerous challenges in order to understand and to objectively characterize its clinical potential. Studies with model observers specifically validated for detection tasks in medical imaging provide an excellent alternative in terms of time and costs for answering these questions.

Keywords: Mammography, vision and visual perception, medical imaging, performance evaluation, model observers

List of acronyms

ANOVA	Analysis of Variance
AUC	Area Under the ROC Curve
BI-RADS	Breast Imaging Reporting Data System
BT	Breast Tomosynthesis
CC	Cranio-Caudal
CH(O)	Channelized Hotelling (Observer)
CLB	Clustered Lumpy Backgrounds
CSF	Contrast Sensitivity Function
DDOG	Dense DOG
DDSM	Digital Database for Screening Mammography
DF	Degrees of Freedom
DM	Digital Mammography
DOG	Difference-of-Gaussians
FFT	Fast Fourier Transform
FP	False-positive
GA	Genetic Algorithm
GL	Gray Level
GLCM	Gray Level Co-occurrence Matrices
HLT	Human Linear Template
HSD	Honestly Significant Difference
LKE	Location Known Exactly
(M-)AFC	(M-)Alternative Forced-Choice
ML	Maximum-Likelihood
MLO	Mediolateral Oblique
MRMC	Multireader Multicase
NGTDM	Neighborhood Gray-Tone Difference Matrix
NPS	Noise Power Spectrum
NPW	Non-prewhitening matched-filter
NPWE	NPW with an Eye-filter
PC	Percent Correct
PDF	Probability Density Function
PM	Primitive Matrix
RMSE	Root-Mean-Square Error
ROC	Receiver Operating Characteristic
ROI	Region of Interest
SDOG	Sparse DOG
SF	Screen-film
SKE	Signal Known Exactly
SKEV	Signal Known Exactly but Variable
SKS	Signal Known Statistically
SQR	Square
TP	True-positive
(W)RMSD	(Weighted) Root-Mean-Square Difference

1. Introduction and motivation

Breast cancer: a challenge in diagnostic radiology

In high-income countries, about one woman out of 10 will develop breast cancer in her lifetime [Curado, 2007]. This makes breast cancer the most common cancer [Ferlay, 2001], and the leading cause of mortality by cancer among women [Ferlay, 2007]. Its incidence is the highest in the age group between 50 and 70, where about 10% of the deaths are due to breast cancer.

Most known risk factors of breast cancer cannot be directly influenced by women. They include age, family history of breast cancer, genetic predispositions, hormonal and reproductive history, or breast density [Kelsey, 1993; Boyd, 1995; Colditz, 1995; Van Gils, 1999; Fitzgibbons, 2000; Heine, 2001; Ziv, 2003]. This renders primary prevention difficult, and explains why the best prevention method radiologists currently use against breast cancer is to try to detect it as early as possible. Early detection indeed usually allows more efficient and less heavy treatments (surgery, chemotherapy, radiation therapy), thus reducing the financial as well as psychological costs.

For this reason, screening mammography was introduced for women above 50, first as controlled randomized trials launched between the 60s and the 90s in the USA, Sweden, Scotland, and Canada. Several meta-analyses of these randomized trials have shown that breast cancer mortality could be reduced by 15 to 30% in the target group with the introduction of screening mammography [IARC, 2002; Humphrey, 2002; Deck, 2006]. Although contested by some researchers because of imperfections in the randomization [Gotzsche, 2000; Olsen, 2001], there is now a general agreement that these figures can be accepted as a basis for justifying organized screening mammography [Boyle, 2003; Green, 2003; de Koning, 2003; Fletcher, 2003].

Since then, 18 European countries have started such screening programs, including Switzerland. The mortality drop observed in these countries is consistent with the controlled randomized trials.

However, early detection of breast cancer is not a trivial task for the radiologists. Recent studies, based on the randomized trials and/or the screening programs, have given estimates for mammography sensitivity (percentage of women actually having breast cancer and diagnosed as such) and specificity (percentage of women without lesion and diagnosed as such). For women over 50, sensitivity has been found to range from 68% to over 90%, and specificity from 82% to 97% [IARC, 2002]. In a recent and comprehensive analysis of the so-called Million Women Study data with adjustments for potential confounding factors, overall sensitivity was found to be 87%, and specificity 97% [Banks, 2004].

Even with the progresses in medical imaging, like the recent screen-film to digital transition that has been reported to improve diagnostic performance [Pisano, 2005; Skaane, 2005], mammography still suffers from diagnosis errors. Most of them come from the fact that the lesions of interest to be detected on mammograms are very low-contrasted signals, which visibility can be reduced by the inherent nature of x-ray projection imaging. On the mammograms, the superposition of the breast structures or the high density of the glandular areas sometimes hide the lesions or make them difficult to assess (false-negative). The opposite can also happen, and superimposed structures on the mammogram can be mistaken as actual lesions (false-positive).

In order to improve the current performance of screening mammography, it is desirable to be able to assess image quality not only with purely physical criteria (e.g. image resolution, system modulation transfer function, radiation dose to the breast), but rather through an optimization approach based on the tasks to be performed with such images [Barrett, 2004]. This way, the full potential of screening programs could be reached in the future, because such methodology would guarantee that radiologists are working with the best images today's technology can offer.

Psychophysical studies

While randomized clinical studies provide ultimate tools for evaluating image quality in a task-based framework, they are often not practical for researchers specialized medical imaging science. Many confounding factors have to be taken into account, especially when several observers, imaging units, imaging protocols, diagnostic centers, and lesion types are involved. Moreover, the usually low incidence in clinical studies tends to make these confounding factors even more difficult to analyze and separate.

For these reasons, a growing interest in psychophysical studies has been shown since Tanner and Birdsall’s paper about human efficiency in the ‘50s [Tanner, 1958]. Formally, psychophysics can be described as *the analysis of perceptual processes by studying the effect on a subject’s experience or behavior of systematically varying the properties of a stimulus along one or more physical dimensions* [Bruce, 2003]. Psychophysical experiments thus provide a convenient approach, since a total control over most experimental parameters is kept, while testing the effect of one or several others.

In medical imaging and detection tasks in particular, psychophysics approach is typically implemented using clinical, synthetic, or hybrid images obtained from a single acquisition method, with controlled signals, a controlled reference classification (also known as gold standard), and reproducible viewing conditions. These four key points, together with the concept of varying the properties of the task along a limited number of physical properties, significantly reduce the inherent uncertainty of purely clinical studies, and characterize the strength of psychophysics framework.

Depending on the experimenters’ goals, two psychophysical studies conditions have been developed and used: Receiver Operating Characteristic (ROC) [Barrett, 2004; ICRU, 2008] and M-alternative forced-choice (M-AFC) [Burgess, 1995; Eckstein, 2000; Barrett, 2004; Gallas, 2007]. In ROC studies, a single image is presented to the observer, who is asked to use a discrete or continuous scale representing his confidence level that an abnormality is present in the image. From the observer’s ratings and the gold standard, the ROC curve can be computed. This curve is used to compute the observer’s sensitivity and specificity depending on the threshold on the rating scale, as well as the overall performance. In this work however, we mostly focused on M-AFC approach. For this kind of task, M images are presented simultaneously to the observer. Out of these M images, one and only one contains a signal of interest, whereas the M-1 others are background-only images. The observer has to select among the M images the one that most likely contains the signal. Burgess has shown that each approach has its own practical advantages and disadvantages, but that the information about the observer’s performance they provide is essentially equivalent [Burgess, 1995]. We chose M-AFC with M equal to 2 or 4, because this kind of task is particularly fast and efficient for collecting data and is not sensitive to intra- or inter-observer variability in the use of a rating scale like ROC.

Another interesting fact concerning M-AFC tasks is that trained non-physician observers have a performance level very similar to that of radiologists in such simplified conditions [Brettle, 2007].

History of psychophysical studies in medical imaging

During the last decades, psychophysical studies in the medical imaging field have undergone a significant evolution. Current digital imaging units offer ready-to-use images, and computers often easily handle on-the-fly 2D signals embedding, a situation which is far from the first experiments with TV screens or digitized films. The aim of this section is to present an overview of some historical and recent works in psychophysics applied to medical imaging, with an emphasis on mammography.

The complexity of a hypothetical complete end-to-end model, from imaging device to the final physician's decision, makes it out of the scope of past and current studies. A complete diagnostic decision process typically involves many steps: choice of the imaging modality and conditions, acquisition of the data, choice of the visualization procedure, search of lesion candidates in the images, analysis of the suspicious areas, and characterization of those identified as lesions. Over the years, psychophysicists in medical imaging have inherited tools from various domains like computer science, ionizing and non-ionizing radiation physics, semi-conductor physics, vision science, and signal processing, and they have been able to investigate and optimize these decision steps, starting from simple conditions to get ever closer to actual clinical tasks.

Early psychophysical experiments were performed with simple geometric patterns on uncorrelated noise. In a series of experiments, Burgess *et al.* investigated detection and discrimination of sine waves [Burgess, 1981], projected spheres [Burgess, 1984], and Hadamard (rectangle-based) patterns [Burgess, 1985], embedded on Gaussian white noise. The analytically traceable statistics of these experiments allowed to estimate the efficiency of human visual system for such signals. Burgess *et al.* also reported a performance drop when the observers were not given information about the exact pattern they had to look for, prefiguring subsequent studies of Signal Known Exactly (SKE) versus Signal Known Statistically (SKS) tasks.

But white noise, although providing relatively simple statistical analysis, was found to be a too simplified vision of actual clinical backgrounds. Researchers switched to correlated noise, which was meant to better represent the influence of projected anatomical structures in radiographic images. Myers *et al.* [Myers, 1985], Bochud *et al.* [Bochud, 1999a], and Burgess *et al.* [Burgess, 2001] showed that performance in detection tasks was dependent on noise correlation level, and used the noise frequency power spectrum (NPS) as a parameter to match computer-generated backgrounds and clinical images statistics. In a recent study, Burgess and Judy [Burgess, 2007] used correlated backgrounds with NPS content given by $NPS(f) \sim f^{-\beta}$, and linked the performance of human observers in a nodule detection task to the slope β of a log-log power spectrum plot against frequency f. For clinical images ($\beta \approx 3$), this implied that the correlations cause the observer's performance to decrease as the lesion size increases, as opposed to white noise ($\beta = 0$). Similar conclusions have been drawn by Judy *et al.* [Judy, 1997] with disks and Gaussian signals in white and correlated noise. Based on NPS match between synthesized and clinical backgrounds, Rolland and Barrett [Rolland, 1992] developed a method they coined *lumpy backgrounds* in order to generate correlated noise with adjustable statistical properties by summing bright blobs, or lumps, assumed to mimic x-ray attenuation processes in heterogeneous matter. Later, Bochud *et al.* adapted the method specifically to mammography and improved visual and statistical realism of the synthetic images [Bochud, 1999b].

Besides projection radiography, correlated noise has been used to simulate tomographic reconstructed images in nuclear medicine: Abbey and Barrett, for example, showed that human observers were able to detect more subtle lesions when exposure time was increased, and that their performance was degraded as anatomical variability extended into higher spatial frequencies [Abbey, 2001b].

Anatomical noise has indeed been shown to be the main limiting factor in mammography: in a study with digital mammograms and filtered white noise ($NPS(f) \sim f^{-3}$), Burgess *et al.* [Burgess, 2001] observed the same positive contrast-detail slope effect for both kinds of backgrounds, and thus showed that the breast structure could not be considered as purely random noise. With an anthropomorphic breast phantom study, Huda *et al.* [Huda, 2006] also showed that the detection of millimeter-sized lesions was mainly limited by anatomical noise. Bochud *et al.* [Bochud, 2000] showed that the assumption of stationarity (statistical properties independent of the location in the image) did not hold for mammograms, as opposed to computer-generated backgrounds that are generally stationary by construction. While the nonstationarities are usually not critical and stationarity within the boundaries of the images can often be assumed, Zhang *et al.* [Zhang, 2006], showed that human observers seemed to be able to adapt their detection strategy in presence of very strong local nonstationarities, and perform better than on stationary backgrounds.

In order to increase the realism of the tasks, psychophysicists also tried using actual clinical backgrounds in their studies. For example, the influence of anatomical structure on spherical nodules detection was studied in lung radiography [Samei, 1998; Baydush, 2001], IRM and bone imaging [Brettle, 2007], and mammography. As breast masses are usually roughly spherical, projected spheres [Bochud, 2000; Abbey, 2002] or Gaussian signals [Abbey, 2001a] have often been used as signals for studying human detection performance with mammographic backgrounds. This method provided so-called hybrid images (real backgrounds + synthetic signals) that represented conditions much closer to actual clinical tasks than the first simplified studies. Recently, advances have been made in the breast lesion simulation area, allowing psychophysicists to synthesize highly realistic signals [Ruschin, 2005; Saunders, 2006]. Based on the analysis of the properties of actual lesions, these signals offer a virtually unlimited range of shapes and sizes, while keeping visually realistic properties.

While traceability of analytical (sine waves, spheres, Gaussians) signals and computer-generated backgrounds statistics had allowed psychophysicists to focus on fundamental understanding of detection tasks, studies with realistic hybrid images were usually more clinically oriented. Compared to actual clinical data, the hybrid images were convenient for keeping a total control over the reference classification (signal absent/present), and the lesion location and profile.

Recently, Chawla *et al.* [Chawla, 2007], Ruschin *et al.* [Ruschin, 2007b], and Samei *et al.* [Samei, 2007] used such images to examine the potential of dose reduction in screening mammography. Through ROC studies, they all concluded that, compared to current clinical practice, a dose reduction of about 50% would lead to an overall performance reduction (mainly due to a degradation of microcalcification detection and masses discrimination), but that masses detection would only be marginally altered. Such findings again emphasize that low-contrast signals detection is limited mainly by anatomical noise rather than quantum noise, and definitely leave room for improvement between radiation protection and medical imaging concerns. They also question the way current mammography protocols, optimized for homogeneous phantoms without anatomical structures, are developed.

Psychophysical studies have also found applications in the past and much debated topic of screen-film (SF) to digital systems transition. Lai *et al.* inserted microcalcifications on anthropomorphic breast phantom images, comparing SF, flat panels, and charged-coupled devices systems, and demonstrated the improvement of microcalcifications visibility when magnification was used with digital systems [Lai, 2005].

Finally, hybrid images and psychophysical studies offered a handy framework for studying digital image compression: Suryanarayanan *et al.* [Suryanarayanan, 2005] inserted simulated masses and microcalcifications into digital mammograms, and compared the detection performance across a wide range of data compression ratios. Masses detection was found not to be altered, even for compression ratios up to 30:1, but microcalcifications were significantly less detectable for ratios of more than 15:1.

To summarize, psychophysics have found a wide range of areas of interest and purposes. With analytical signals such as sine waves, disks, projected spheres, or realistic masses combined to white noise, correlated noise, phantom images or actual clinical backgrounds, detection tasks have been conducted in a various medical imaging modalities: angiography, chest imaging, nuclear medicine, MRI, or mammography, to cite only but a few. With absolute control over the backgrounds and/or the signals, researchers have been able to analyze most steps of the diagnostic processes, from image acquisition to final decision and data archiving.

Intuitive approach to model observers

In many psychophysical studies, experimenting with human observers was not the only goal. Gathering sufficient data and statistics from radiologists or medical physicists may be costly and time consuming, and directly generalizing the conclusions to other experimental conditions is complex. Moreover, human observers are known to be prone to intra- and inter-observer variability, coming from internal (weariness, tension, motivation, attention) or external (light, noise) perturbations sources, which add uncertainty to the intrinsic experience level of the observers.

Ultimately, a model of human visual system and information processing could solve all these issues: once correctly calibrated, one could compute the average human observer performance for a task given several conditions, and optimize all the imagery chain in order to maximize the performance level. Unfortunately, coupling our current knowledge of visual perception to that of decision-making processes is so far too complex. For that reason, researchers have used the same approach with models as with human observers over the years, starting from simple patterns on white noise to current experiments with clinical backgrounds and signals.

Over the years, two model observers approaches have made their way in medical imaging: the ideal and the human-like model observer. The ideal observer extracts all information available in the image and maximizes the performance, while the latter attempts to reproduce or predict human decisions. For the present work, we have focused on the second approach and an objective, task-based definition of image quality [Barrett, 2004], and we have tried to constantly link human and mathematical models decisions and performance levels.

In this section, we present an intuitive approach to the most commonly used model observers from the literature, in order to introduce the reader to the more mathematical descriptions that can be found in the papers in section 3.

Quite surprisingly, most models of human vision used in medical imaging are linear. For such models, the response λ_i to an image $\mathbf{g}_i$ is simply given by:

$$\lambda_i = \mathbf{w}^T \mathbf{g}_i + \varepsilon \quad (1)$$

$\mathbf{w}$ in Eq.(1) is called the model template, and both $\mathbf{w}$ and $\mathbf{g}_i$ are expressed as 1D vectors before computing the dot product (superscript ^{T} is for transposition operator). ε is an optional internal noise term that may be used to model decision variability given the same image $\mathbf{g}_i$.

Once computed from Eq.(1), the observer's response λ_i is used as a decision variable for the task under study. In an M-AFC task, where the signal is known to be present at one of the M locations, and absent at the (M-1) others, the model will compute λ_i for the *i=1, 2, …, M* possible locations, and select the one that obtains the highest response. In a ROC task, where the observer is requested to rate his degree of confidence that the signal is effectively present in an image $\mathbf{g}_i$ on a continuous or discrete scale, λ_i can directly be used, after having been mapped to the discrete scale if necessary.

In the absence of internal noise, the differences between the model observers reside only in their templates $\mathbf{w}$. Intuitively, the templates may be defined as the way the models "perceive" a given signal in a given background. Mathematical expressions or empirical parameters set the rules according to which the signals are processed to obtain the templates. Visual examples are given in Fig. 1, for a signal corresponding to a projected sphere.

The most basic model observer is the region of interest (ROI) observer. Based on the fact that the signal presence (bright area) implies local high pixel values, the ROI observer simply integrates them over the area covered by the signal. This ROI observer is the ideal observer for constant amplitude signals in white noise, but this naïve approach is of course suboptimal in presence of more clinically realistic signals or backgrounds.

Another simple model is the non-prewhitening matched-filter (NPW). This model uses the signal 2D profile itself as template in Eq.(1). The NPW model thus uses full knowledge of the signal to be detected, but no information about the backgrounds. For any kind of signal, it corresponds to the ideal observer in white noise but, again, fails to maintain a good performance level with more complex backgrounds.

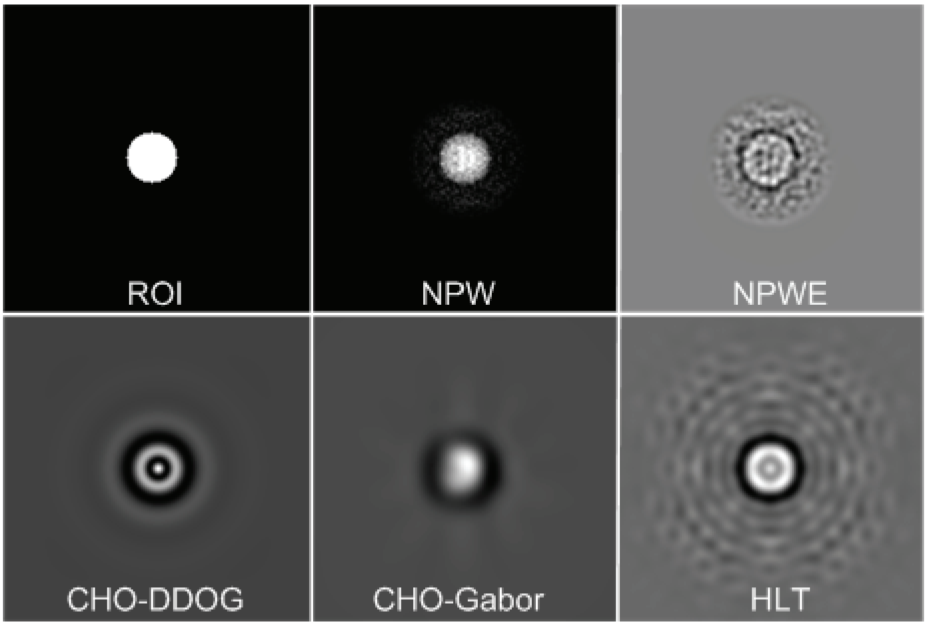


Fig. 1. Examples of 2D model observers' templates in the spatial domain (adapted from Castella *et al.* [Castella, 2007b]). The region of interest (ROI) template is a simple binary image corresponding to the area covered by the signal. The non-prewhitening matched-filter (NPW) observer template corresponds to the actual signal. The NPW observer with an eye-filter (NPWE) template filters the signal with an experimental human eye contrast sensitivity function. The channelized Hotelling (CHO) with dense differences of Gaussians (DDOG) or Gabor channels models human vision as a response to a limited number of frequency or orientation channels. The human linear template (HLT) is derived *a posteriori* from human observers' responses.

The next step in model observer science was to incorporate information about human visual system and its features into the model observers. Burgess [Burgess, 1994], on the basis of the early results with sine waves, added an empirical model of human eye contrast sensitivity function (CSF) into the NPW model. The CSF basically describes human eye's ability to perceive a signal as a function of its frequency content, and had been found to peak at middle frequencies (about 4 cycles per degree of visual angle) and decreasing rapidly towards low and high frequencies. This translates the fact that human observers are not efficient for detecting very large or very small signals in noise. In the NPW with an eye-filter (NPWE) model, the signal is thus filtered by the CSF, selectively suppressing or enhancing its frequency content in order to model human eye perception. This model has been successfully applied to predict human results for detecting synthetic thrombus embedded on x-ray coronary angiograms [Eckstein, 1998], spherical nodules [Abbey, 2002] or digitized masses [Burgess, 2001] embedded on mammograms, as well as for studying medical images compression algorithms [Suryanarayanan, 2005; Zhang, 2004a, b].

As discussed before, correlation in the noise patterns have a critical effect on the performance of the simple NPW model. A way to overcome this problem is to incorporate knowledge about the backgrounds statistics and to decorrelate the noise prior to matched filtering. This is exactly the aim of the Hotelling observer, which performs a prewhitening operation using the covariance matrix of the backgrounds, in order to derive the template corresponding to a given signal. When the statistics are Gaussian, this observer is the best linear observer [Barrett, 1998, 2004; Eckstein, 1998, 2000;

Rolland, 1992]. In presence of white noise, it simply reduces to the NPW model, since the only nonzero elements of the backgrounds covariance matrix are on its diagonal.

However, the Hotelling observer suffers two disadvantages when used to model clinically relevant tasks. First, obtaining a reasonably stable estimate of the backgrounds covariance matrix is needed for computing the observer template, which requires a large amount of images. The rule of thumb in such cases is a number of images equal to about ten times the number of pixels per image, which is tremendous even for reasonably small 128 by 128 pixel regions of interest. Second, this observer does not take into account any human visual system features and its limited ability to decorrelate the noise. In most cases, the Hotelling observer performance is thus much higher than that of human observers, and comparing model and human results may be challenging.

Fortunately, these two issues can be solved at once in a rather simple and mathematically elegant way by constraining the Hotelling observer to a limited set of N_c basis functions, or channels [Myers, 1987; Gallas, 2003]. The number of elements of the covariance matrix as seen through these channels is reduced to N_c^2, greatly reducing the computing time and complexity. In addition, the channel basis can be chosen in order to reflect human visual system processes. For example, it is believed that cells in the visual cortex preferentially respond to visual stimuli with a specific spatial frequency and/or orientation [Movshon, 1978; Marcelja, 1980]. For this reason, channelized Hotelling observers (CHO) using basis functions like square band-pass radial frequency filters [Myers, 1987; Abbey, 2001b, 2002], differences of Gaussians [Abbey, 2001b,2002] or Mesa filters [Burgess, 1997], or Gabor functions [Eckstein, 1998, 1999; Zhang, 2004a, b, 2005, 2007], have been used with various kinds of backgrounds. Other basis functions like Laguerre-Gauss channels [Barrett, 1998; Burgess, 2001; Suryanarayanan, 2005], which are not related to human vision features and were used in some studies solely to reduce the dimensionality of the images, were not considered in the present work.

Breast tomosynthesis: the answer to tissue overlap issues?

All studies about detection performance presented in the previous sections were conducted in the framework of conventional, planar radiography. One general conclusion from these studies is that structured or anatomical noise has a major degrading effect on human performance [Bochud, 1999a; Burgess, 2001]. In mammography, these tissue superposition effects can mislead the radiologists and lower both sensitivity and specificity. Therefore, a method that could reduce this noisy component directly during image acquisition by providing three-dimensional information about the breast instead of two-dimensional projections would be desirable.

Breast tomosynthesis (BT) is a refinement of the conventional tomography method, and allows to reconstruct an arbitrary number of in-focus planes retrospectively from a sequence of typically 7 to 11 low-dose projection radiographs acquired during a single motion of a conventional x-ray tube [Dobbins, 2003; Smith, 2005]. Although still relatively new and much less investigated by the scientific community than planar mammography, BT potential benefits include 3D tissue localization at a dose comparable to one conventional mammographic exposure, better-contrasted images and better depiction of masses borders even with dense breasts, reduction in recall rates, and higher positive predictive value [Park, 2007]. BT technique is still currently under development, but several studies have already shown its potential for improving accuracy compared to mammography [Suryanarayanan 2000, 2001; Chan, 2005; Gong, 2006; Ruschin, 2007a].

Suryanarayanan *et al.* [Suryanarayanan, 2000, 2001] investigated the potential gain on detection performance BT could offer. Their experiments comparing different tuned-aperture tomosynthesis reconstruction methods to digital mammography for imaging composite phantoms showed that threshold contrast characteristics were significantly better for all tomosynthesis methods than those with planar mammography. Significant differences between these two imaging modalities in a disk detection experiment were observed, in favor of BT, in a study involving board-certified radiologists. The authors concluded that breast tomosynthesis could improve visualization of valuable diagnostic information.

Another application of better-contrasted tomosynthesis images in dense breasts was investigated by Chan *et al.* [Chan, 2005], whose preliminary results on a computer-aided detection applied to 3D localization of breast masses led to excellent results.

Recently, Gong *et al.* studied the detection of a spherical signal embedded into a synthetic breast model, comparing mammography to BT. Five physicists participated to these experiments. The authors showed that the 5-millimeter lesion was statistically significantly better detected with BT than in digital mammography [Gong, 2006].

Ruschin *et al.* used hybrid images with synthetic breast masses and anatomical backgrounds to show that, using the same acquisition dose in mammography and BT, lesion projected intensity in BT could be reduced by about four times compared to mammography, while keeping the same human observers' performance level [Ruschin, 2007a].

Most of BT evaluation studies have been conducted with human observers only. The model observers presented in the previous sections, which have been developed and tested with success in various medical imaging fields, are indeed only in the early stage of development in tomosynthesis. A study by Gifford *et al.* investigated a scanning noiseless channelized Hotelling Observer and compared different number of projections and angular span combinations [Gifford, 2008]. Reiser *et al.* compared filtered-backprojection and iterative maximum-likelihood expectation maximization reconstruction methods with a prewhitening observer in a simplified detection task with a spherical signal in a homogeneous phantom [Reiser, 2008]. In another study, Pineda *et al.* used a channelized Hotelling and non-prewhitening model observers with and without eye-filters for optimizing a tomosynthesis system for the detection of lung nodules [Pineda, 2006].

To summarize, BT currently offers challenging perspectives in various research fields. They include acquisition and display techniques optimization [Dobbins, 2003; Carton, 2006; Heberhard, 2006; Maidment, 2006], as well as reconstruction and filtering methods [Wu, 2004; Chen, 2006; Mertelmeier, 2006], and patient dose comparison with other imaging modalities. Finally, detection strategies, human or model observers' performances in the particular framework of BT images, and comparison with other imaging techniques are far from being understood and characterized.

Goals of the study

The aim of this work was to provide a step-by-step and objective approach to a better understanding and modeling of clinically realistic detection tasks in mammography, starting with an in-depth study of the statistical properties of mammographic backgrounds, before conducting various psychophysical studies with different backgrounds and signals.

The project was developed towards the following goals:

- To study the statistical properties of mammographic textures depending on the breast density and to develop an objective breast density classification algorithm based on these properties.
- To optimize the *Clustered Lumpy Backgrounds* technique in order to generate synthetic mammographic backgrounds, while maximizing both visual and statistical realism, and to validate these images.
- To perform detection experiments of a simple spherical signal modeling a tumor with real and synthetic mammographic backgrounds and to compare the performance of human observers for this Signal Known Exactly (SKE) task to that of existing or adapted models.
- To conduct detection tasks with realistic benign and malignant signals varying throughout the experiment (Signal Known Statistically, SKS) and to study the influence of the signal size, shape, and variability, on human and model observers’ performance.
- To objectively compare detection performance of realistic masses with clinical backgrounds in mammography and tomosynthesis.

2. PhD milestones

The thesis core is composed by the articles in section 3, which represent the five main milestones of the project: understanding, creating, experimenting, enhancing realism, and looking towards the future (see Fig. 2).

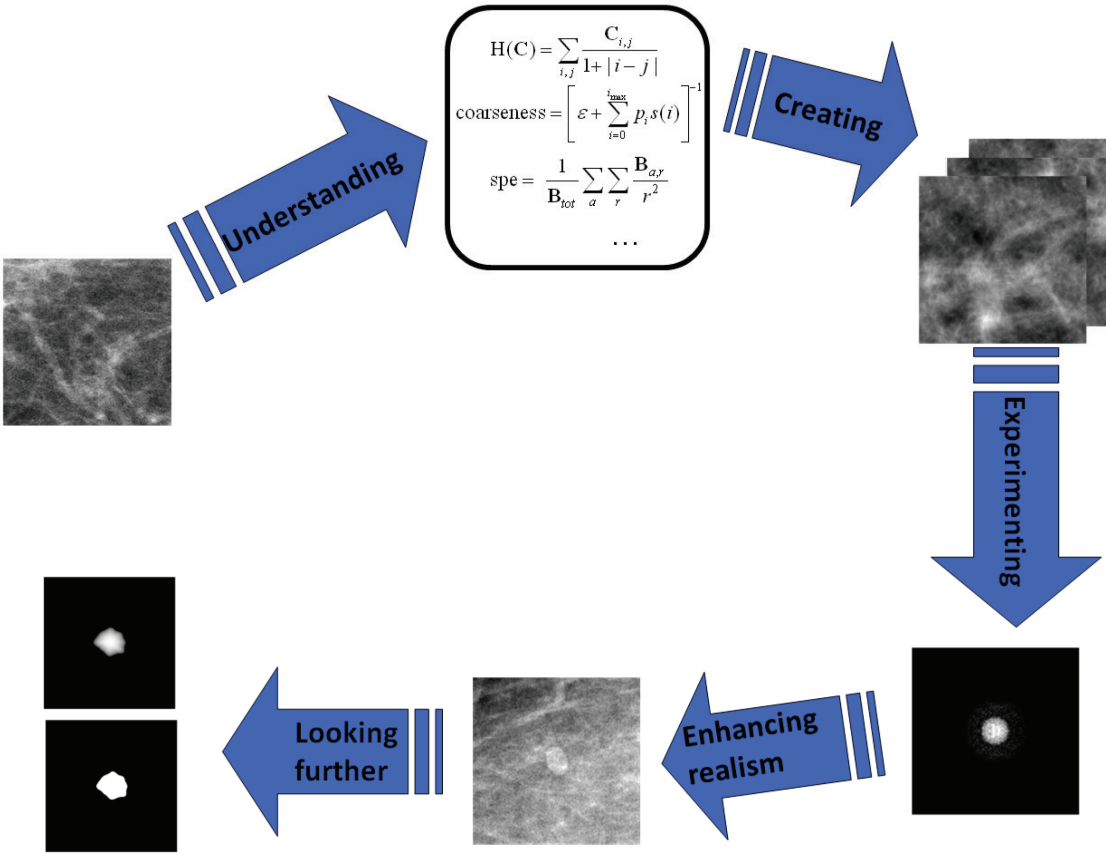
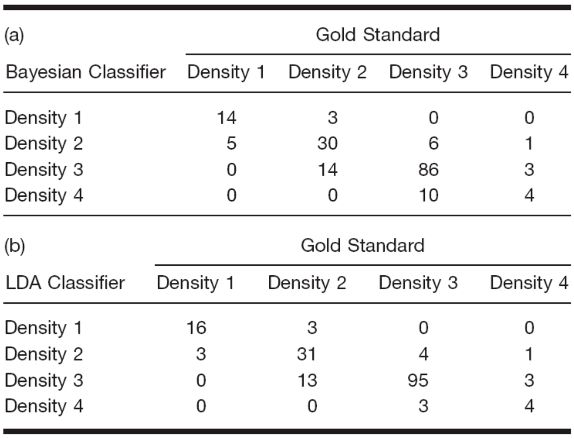

Understanding
$H(C)=\sum_{i,j}\frac{C_{i,j}}{1+|i-j|}$
$\text{coarseness}=\left[\varepsilon+\sum_{i=0}^{i_{\max}}p_i s(i)\right]^{-1}$
$\text{spe}=\frac{1}{B_{tot}}\sum_{a}\sum_{r}\frac{B_{a,r}}{r^2}$
...
Creating
Experimenting
Enhancing realism
Looking further

Fig. 2. Schematic representation of the PhD milestones.

Statistically understanding mammographic texture

This paper [Castella, 2007a] deals with the statistical properties of digital mammograms. Complementary texture analysis methods were used in order to characterize mammograms regions of interest (ROI) with mathematical features. Using these features, semi-automated classifiers for assessing breast density according to the BI-RADS scale [Obenauer, 2005] were developed. The classifiers used Bayesian rules or Linear Discriminant Analysis in order to determine which of the four classes of the BI-RADS scale should be attributed to single regions of interest, whole breast, or breast pair.



(a)

Bayesian Classifier	Gold Standard Density 1	Density 2	Density 3	Density 4
Density 1	14	3	0	0
Density 2	5	30	6	1
Density 3	0	14	86	3
Density 4	0	0	10	4

(b)

LDA Classifier	Gold Standard Density 1	Density 2	Density 3	Density 4
Density 1	16	3	0	0
Density 2	3	31	4	1
Density 3	0	13	95	3
Density 4	0	0	3	4

Fig. 3. Confusion matrix obtained for the Bayesian classifier (a), and the Linear Discriminant Analysis (LDA) classifier (b) [Castella, 2007a]. The Gold Standard is the reference classification established by radiologists. In this example, the regions of interest of both breasts are used for determining the breast density. For comparison, the exact agreement upper limit between two radiologists is estimated to be around 80% [Karssemeijer, 1998].

With these three analysis levels and the information provided by the statistical features, an excellent agreement with the reference classification established by three radiologists was obtained, as illustrated in Fig. 3. The classifiers thus provided an objective and reproducible method for assessing breast BI-RADS density, which is an indicator for breast cancer risk.

Creating visually and statistically realistic breast texture

This paper [Castella, 2008] describes the improvement of the Clustered Lumpy Backgrounds (CLB) technique using a genetic algorithm. CLB are images produced by the random superposition of blobs of various shapes. The free parameters of the image generation technique can be tuned in order to produce a wide range of textures with different visual and mathematical properties [Rolland, 1992; Bochud, 1999b].

Using the statistical features described in the previous section, a Mahalanobis distance in the features space was defined. This distance was used as a cost function to evolve a genetic algorithm, in order to improve the original CLB parameters and to produce images closer to real, clinical mammograms ROI. Several optimized variations of the CLB model were then evaluated through psychophysical studies involving radiologists and radiographers (see examples in Fig. 4). This showed that the optimized model improved both visual and statistical realism of the synthetic images. This “second-generation” CLB technique thus allowed to generate any amount of images for psychophysical studies, while guaranteeing their statistical traceability and their realism. The proposed approach could readily be adapted to other kinds of images.

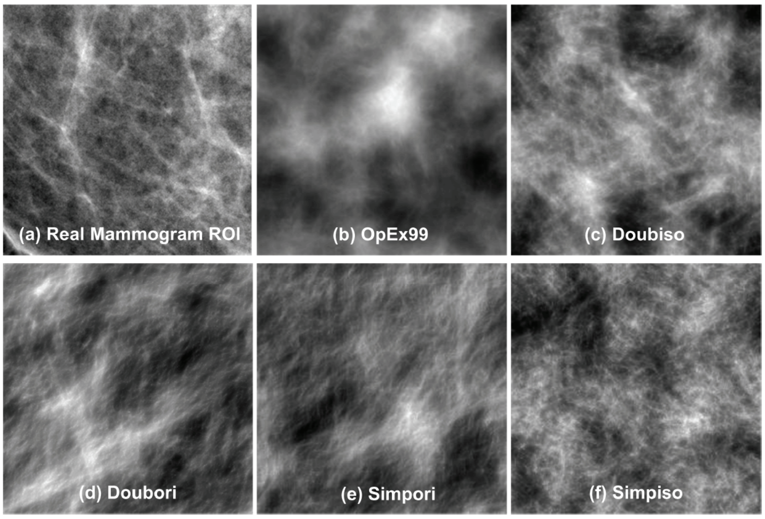


Fig. 4. Examples of realizations for the different types of Clustered Lumpy Background (CLB) variations [Castella, 2008]. The 2-layer variations are created with blobs having two different sizes. (a) Region of interest selected from a real mammogram; (b) original 1-layer CLB, Opex99 parameters [Bochud, 1999b]; (c) 2-layer CLB, isotropic orientation of the clusters; (d) 2-layer CLB, favored orientation of the clusters; (e) 1-layer CLB, favored orientation of the clusters; (f) 1-layer CLB, optimized version of (b).

Experimenting with a simple spherical signal

The next step of the project was to use the CLB and real mammographic backgrounds in psychophysical detection experiments with a simple spherical signal [Castella, 2007b]. The performance of human observers was analyzed and compared to linear model observers adapted from the literature. Additionally, the human observers' linear templates (HLT) were estimated from their responses (Fig. 5). This provided a way to evaluate not only their general detection performance, but also their strategies.

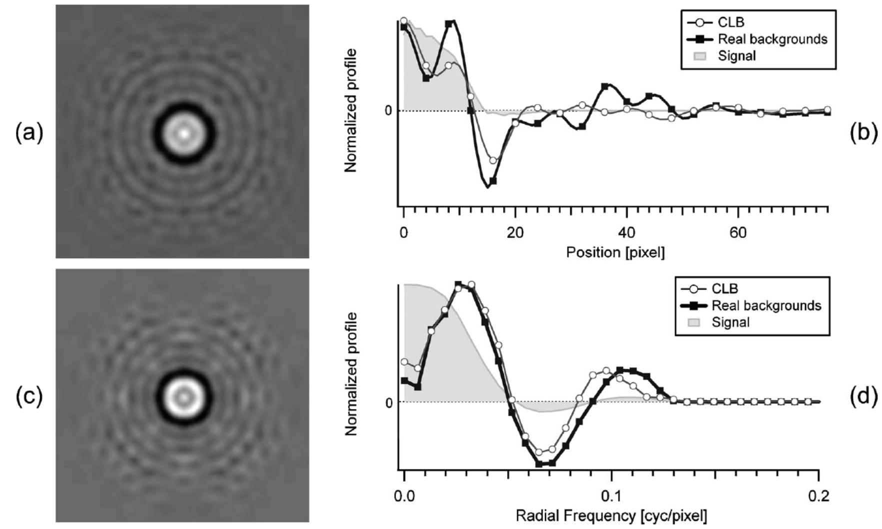


Fig. 5. Human observer linear template obtained for the spherical signal considered in Fig. 1. [Castella, 2007b] (a) Spatial domain 2D template with Clustered Lumpy Backgrounds (CLB). (c) Same with mammographic backgrounds. Profiles of both templates in (b) the spatial domain and in (d) the Fourier domain.

The main results of this study were that the HLT derived for real and synthetic backgrounds were not significantly different, and that they could reproduce the human observers' performance within 5% in terms of percent of correct answers. This study also emphasized the significance of local statistics of the backgrounds, since the detection performance was found to be significantly higher with the real backgrounds, which have a smaller local variance when the overall variance is matched with that of CLB. In other words, human observers processed the real and synthetic backgrounds the same way, but were too sensitive to local noise to obtain the same performance with both kinds of images.

Enhancing realism with benign and malignant signals

The goal of the fourth step was to get closer to clinical tasks, by introducing realistic signals into the experimental schemes [Castella, 2009a]. Simulated benign and malignant breast lesions mimicking real masses, designed by Saunders *et al.* [Saunders, 2006], were used. For this series of psychophysical experiments, another clinically relevant characteristic was also introduced: the detection performance when the signal was the same throughout a given task (Signal Known Exactly, SKE) was compared to the cases where the observers were not given exact information about the signal shape and/or size (Signal Know Statistically, SKS). Again, the performance of the human observers was compared to that of models (Fig. 6).

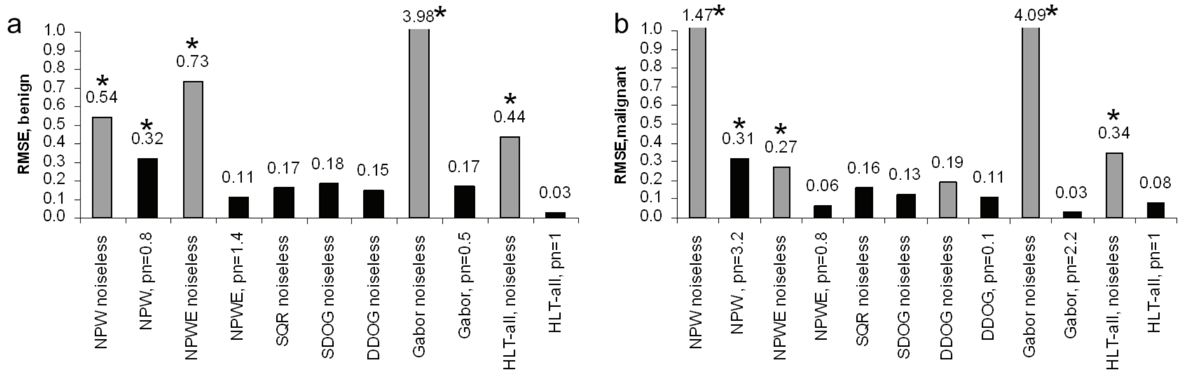


Fig. 6. Root mean square error (RMSE) in *d'* units between the generic human observer and the different model observers for Signal Known Exactly (SKE) tasks for noiseless (grey) and noise level-optimized (black) models with benign (a) and malignant (b) simulated masses. Stars indicate performance levels that are significantly different from humans (F-test, p<.05). Models acronyms are detailed in section 1.

As for the simple spherical signal task, human data could be fitted with some models taking into account human visual system properties, and with the HLT. Quite surprisingly, human and model observers obtained very similar performances in SKE and SKS tasks, as long as the lesion size was kept constant. A performance level drop, however, was observed when the information about the exact signal size was not given to the observers. These results suggest that evaluating the performance in detection tasks through human and/or model observers studies could be done adequately even with a small set of signals covering the size range of clinical interest.

Looking further: towards objective evaluation of breast tomosynthesis potential

The last part of the project consisted in applying the knowledge about model observers acquired in digital mammography (DM), to the emerging breast tomosynthesis (BT) modality [Castella, 2009b]. Three-dimensional breast reconstruction in BT offers exciting insights for early detection of breast cancer, with the opportunity of removing the superposition effect that lowers sensitivity and specificity in DM. As it is still a technique at a prototype stage, much improvement is expected within the next years, concerning the acquisition parameters like scan angle span and number of projections, or reconstruction algorithms and image processing.

Although being more and more used in mammography, model observers are still in their early stages in BT. In this work, the images and results from a human observer study by Ruschin *et al.* [Ruschin, 2007a] were used to compare DM and BT with the same set of matched hybrid images. These were generated by adding realistic mass signals to clinical images of patients who underwent both screening mammography and tomosynthesis exams.

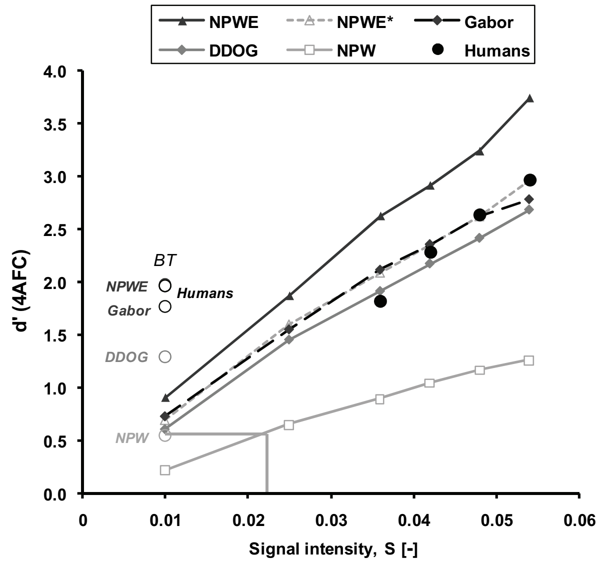


Fig. 7. 4-Alternative Forced-Choice digital mammography task performance of the model observers as a function of signal intensity S [Castella, 2009b]. For comparison, human results are indicated with filled circles, and performance in breast tomosynthesis (BT) task with open circles. NPWE* values correspond to the NPWE model with internal noise intensity level that minimizes the error over the DM tasks

Again, a good agreement was obtained between human observers and models using human visual system properties (see Fig. 7). In particular, it was shown that BT potential for detecting low-contrast lesions could be evaluated through model observers studies. This opens the way to testing future BT developments with such models, providing an objective and reproducible means for assessing diagnostic performance in BT, or comparing it to DM.

3. Papers

The following papers are inserted in this Section. They correspond to the five milestones listed in Section 2.

Statistically understanding mammographic texture

[Castella, 2007a] C. Castella, K. Kinkel, M. P. Eckstein, P.-E. Sottas, F. R. Verdun, and F. O. Bochud, "Semiautomatic Mammographic Parenchymal Patterns Classification Using Multiple Statistical Features," Academic Radiology **14**, 1486-1499 (2007).

Creating visually and statistically realistic breast texture

[Castella, 2008] C. Castella, K. Kinkel, F. Descombes, M. P. Eckstein, P. Sottas, F. R. Verdun, and F. O. Bochud, "Mammographic texture synthesis: second-generation clustered lumpy backgrounds using a genetic algorithm," Opt. Express **16**, 7595-7607 (2008).

Experimenting with a simple spherical signal

[Castella, 2007b] C. Castella, C. K. Abbey, M. P. Eckstein, F. R. Verdun, K. Kinkel, and F. O. Bochud, "Human linear template with mammographic backgrounds estimated with a genetic algorithm," J. Opt. Soc. Am. A **24**, B1-B12 (2007).

Enhancing realism with benign and malignant signals

[Castella, 2009a] C. Castella, M. P. Eckstein, C. K. Abbey, K. Kinkel, F. R. Verdun, R. S. Saunders, E. Samei, and F. O. Bochud, "Mass detection on mammograms: influence of signal shape uncertainty on human and model observers," J. Opt. Soc. Am. A **26**, 425-436 (2009).

Looking further: towards objective evaluation of breast tomosynthesis potential

[Castella, 2009b] C. Castella, M. Ruschin, M. P. Eckstein, C. K. Abbey, K. Kinkel, F. R. Verdun, A. Tingberg, and F. O. Bochud, "Masses detection in breast tomosynthesis and digital mammography: a model observer study," *to appear in Proc. SPIE Medical Imaging (2009).*

Post-publication comments:

The reader should be aware of the following typos and updates about the published papers.

First, the transpose operator in the covariance matrix formula, Eq. 3 in [Castella, 2007a] and Eq. 6 in [Castella, 2008] was not correctly placed. The correct expression, which was used in the computations, is:

$$\mathbf{K} = \sum_{i=1}^{n} \frac{1}{n-1} (\mathbf{v} - \mu)(\mathbf{v} - \mu)^T$$

Second, in the last paragraph of Section 2.G in [Castella, 2007b], ROC obviously does not stand for *radius of curvature*. It is the acronym of Receiver Operating Characteristic.

Finally, the match between human results and the NPWE model in the study with the phantom mass [Castella, 2007b] can be highly improved with the same internal noise addition mechanism as in the subsequent papers. In the phantom mass study, internal noise was added to the NPWE decision variable as a uniformly distributed random variable, which is quite a naïve approach.
If the internal noise is described by a zero-mean, Gaussian distributed random variable instead, with a variance chosen in order to match the performance between the NPWE model and the human observers with CLB, the AUC for the NPWE model becomes 0.74 ± 0.02 with CLB (human observers: 0.73 ± 0.02), and 0.81 ± 0.02 with real backgrounds (human observers: 0.84 ± 0.02). These updated results confirm the excellent potential of the NPWE model for matching human observers' performance level.

Note concerning the following papers: [Castella, 2007], [Castella, 2008], and [Castella, 2009a]

These papers were published in J. Opt. Soc. Am. A and Optics Express and are made available as electronic reprints with the permission of OSA. One print or electric copy may be made for personal use only. Systemic or multiple reproduction, distribution to multiple locations via electronic or other means, duplication or any material in these papers for a fee or for commercial purposes, or modification of the content of the papers are prohibited.

Copyright OSA (http://www.osa.org/pubs/osajournals.org).

Semiautomatic Mammographic Parenchymal Patterns Classification Using Multiple Statistical Features[1]

Cyril Castella, MSc, Karen Kinkel, MD, Miguel P. Eckstein, PhD, Pierre-Edouard Sottas, PhD
Francis R. Verdun, PhD, François O. Bochud, PhD

Rationale and Objectives. Our project was to investigate a complete methodology for the semiautomatic assessment of digital mammograms according to their density, an indicator known to be correlated to breast cancer risk. The BI-RADS four-grade density scale is usually employed by radiologists for reporting breast density, but it allows for a certain degree of subjective input, and an objective qualification of density has therefore often been reported hard to assess. The goal of this study was to design an objective technique for determining breast BI-RADS density.

Materials and Methods. The proposed semiautomatic method makes use of complementary pattern recognition techniques to describe manually selected regions of interest (ROIs) in the breast with 36 statistical features. Three different classifiers based on a linear discriminant analysis or Bayesian theories were designed and tested on a database consisting of 1408 ROIs from 88 patients, using a leave-one-ROI-out technique. Classifications in optimal feature subspaces with lower dimensionality and reduction to a two-class problem were studied as well.

Results. Comparison with a reference established by the classifications of three radiologists shows excellent performance of the classifiers, even though extremely dense breasts continue to remain more difficult to classify accurately. For the two best classifiers, the exact agreement percentages are 76% and above, and weighted κ values are 0.78 and 0.83. Furthermore, classification in lower dimensional spaces and two-class problems give excellent results.

Conclusion. The proposed semiautomatic classifiers method provides an objective and reproducible method for characterizing breast density, especially for the two-class case. It represents a simple and valuable tool that could be used in screening programs, training, education, or for optimizing image processing in diagnostic tasks.

Key Words. Image analysis; pattern recognition; feature extraction; mammography.

© AUR, 2007

While the etiology of breast cancer remains unclear, many studies have demonstrated a correlation between cancer risk and factors such as age, breastfeeding and pregnancy history, family history of breast cancer, hormonal treatments, genetic factors, and breast density (1–7). Breast density as a factor of risk was first investigated by Wolfe (8), who defined a four-grade density scale on the basis of the patterns and textures observed on mammograms. Later, the BI-RADS (Breast Imaging Reporting Data System) density scale was developed by the American College of Radiology to standardize mammography reporting terminology and assessment and recommendation categories (9,10). The BI-RADS density classification was created to inform referring physicians about the decline in sensitivity of mammography with increasing breast density. BI-RADS defines breast density 1 as almost entirely fatty, density 2 as scattered fibroglandular tissue, density 3 as heterogeneously dense tissue and density 4 as extremely dense

---

Acad Radiol 2007; 14:1486–1499

[1] From the University Institute for Radiation Physics, Centre Hospitalier Universitaire Vaudois (CHUV), and University of Lausanne, Grand-Pré 1, CH-1007 Lausanne, Switzerland (C.C., P.-E.S., F.R.V., F.O.B.); Clinique des Grangettes, Chemin des Grangettes 7, CH–1224 Chêne-Bougeries, Switzerland (K.K.); and the Department of Psychology, University of California, Santa Barbara, CA 93106-9660 (M.P.E.). Received May 15, 2007; accepted Jul 18, 2007. This work was supported by Swiss National Science Foundation FN 3252BO-104273. **Address correspondence to:** F.O.B. e-mail: francois.bochud@chuv.ch

© AUR, 2007

doi:10.1016/j.acra.2007.07.014

tissues. It was not intended to serve as a method of measuring breast density percentage, although as per Wolfe's scale (11), correlations with this more objective factor do exist (12). In clinical American and European conditions, the breast density of a given patient is typically evaluated and reported by a radiologist using BI-RADS on the basis of the simultaneous display of two mammograms per breast.

However, one of the difficulties for correctly assessing breast density is that the BI-RADS density scale definitions are rather subjective. A certain interpretational freedom prevents perfect interobserver and even intraobserver reproducibility (13,14). On the other hand, numerous pattern recognition and classification techniques have been developed and can be directly applied to this task (15). This is why different statistical approaches have been explored in the last few years in order to develop an objective classifier of mammograms according to Wolfe or the BI-RADS scale. These techniques have made use of various pattern recognition parameters to statistically describe the whole breast or part of it: fractal dimension (16–18), gray level histogram properties (19,20), moments (17,18,21), gray level variations matrices (17,20), or maximum response filters (22). These descriptions have been combined with several general classification algorithms: Bayesian classification (16,17), linear discriminant analysis (LDA) (20), nearest neighbor rules (21), neural networks, and textons (22).

The goal of this study was to develop a semiautomatic method for assessing the BI-RADS density category using features extracted on mammograms. For this purpose, we combined a large number of statistical features computed from manually selected regions of interest (ROIs) with LDA and Bayesian predictors. Special care was applied in order to assess the robustness of the three distinct classifiers we developed, and the validation of their individual performance. In contrast to most previous studies, we worked on multiple ROIs per mammogram. Homogeneity in both size and emplacement was retained in order to facilitate the interpatient comparisons of the statistical features without bias due to different breast sizes and shapes.

Each classifier was trained and tested using the leave-one-out technique to classify a set of 1408 ROIs extracted from 88 patients, on the basis of all computed features. Additionally, we averaged the individual ROI results over multiple ROIs from the same breast and/or patient. Finally, optimal subsets of features were computed and the classifiers ran the same processes. The results were then compared to a reference classification established upon a consensus of three radiologists through weighted κ statistics.

The developed semiautomatic classifiers may have valuable applications in screening exam procedures, to help radiologists objectively determine breast density in a reproducible way. Patients with higher density breast tissue may thus receive special attention and specific image display optimization, because pathologies tend to be hidden by dense backgrounds. The field of potential usefulness of such classifiers extends to training and education as well.

MATERIALS AND METHODS

Mammogram Database

The image database consisted of a set of 352 digital mammograms collected at the Clinique des Grangettes, Geneva, Switzerland, from patients who underwent screening exams. For each of the 88 patients, one craniocaudal (CC) and one mediolateral oblique (MLO) view mammogram per breast was considered. All mammograms were obtained using automatic exposure control (27- to 32-kV voltage) on a GE Senograph 2000D full-field digital detector (23–25). This means that not only the tube loading, but also the anode/filter combination and tube potential were selected automatically in a process involving a preexposure, depending on the thickness and density of the compressed breast, in order to control the dose delivered in the central breast region (26). Mammograms were outputted as 12-bit processed images, with 0.1 × 0.1 mm pixel size. All mammograms showing any sign of abnormal mammographic features such as masses, architectural distortion, or clusters of microcalcification were excluded from this study.

Selection of Regions of Interest

The first step consisted of the manual choice of four ROIs per mammogram. The ROIs were 256 × 256 pixel square regions chosen in the central breast region, about half way between the nipple and the chest wall. One example case is shown in Figure 1. The location choices were made under the control of the radiologists involved in the study and allowed us to obtain four nonoverlapping ROIs per mammogram, while covering most of the breast density. This location also ensured that we performed our analysis using only breast tissue, without bias introduced by the pectoral muscle or imaging artefacts.

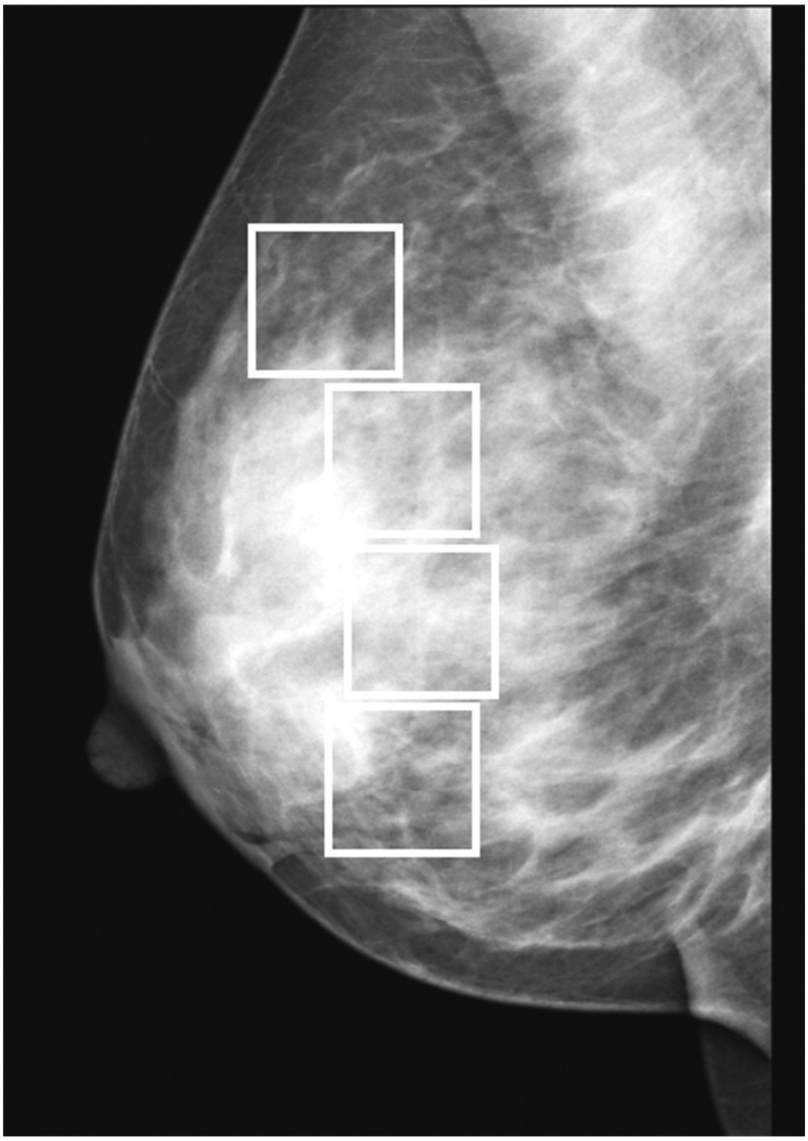


Figure 1. Digital mammogram and corresponding manually defined regions of interest.

Statistical Description

All ROIs were then characterized by the statistical quantities defined below. Unlike a global analysis of the whole breast projection, the square and uniform shape of all ROIs greatly simplifies the computation and interpatient comparison of these features.

In order to capture as much information as possible, we extracted 18 different and complementary statistical quantities from each ROI. Due to the diversity of definitions found in the literature for a given quantity, all expressions used in this work are presented explicitly in the Appendix. They involve quantities derived from the gray level histogram like the standard deviation, skewness, and kurtosis but also balance (15,27). Gray level co-occurrence matrices (GLCMs) provided quantities like energy, entropy, cmax, contrast, and homogeneity (28–30). From the primitive matrix (PM), we derived the short primitive emphasis (spe), the long primitive emphasis (lpe), as well as gray level uniformity (glu) and primitive length uniformity (plu) (28). The fractal dimension was calculated by a box-counting method (16,17,31). Finally, the neighborhood gray-tone difference matrix (NGTDM) provided the coarseness, contrast, complexity, and strength (32).

Features derived from the gray level histogram characterize the distribution of gray levels in a comprehensive way, in particular, its shape and its symmetry. Balance is closely related to skewness and describes the asymmetry of the gray level histogram.

GLCMs are a powerful tool for obtaining information about the spatial relationships of gray levels in structural patterns. The ROIs were linearly rescaled from 12 to 4 bits (16 gray levels), reducing the computing time by a factor of 65,536 and ensuring that the GLCM elements were essentially non-zero. Following, for each ROI, 20 co-occurrence matrices were computed, using directions of 0°, 45°, 90°, and 135° and distances of 1, 3, 5, 7, and 9 pixels. These directions correspond to the four natural directions for a square image, and the corresponding distances describe structures from the millimeter to the centimeter range, which are typical for the breast texture. Finally, five scalar features (energy, entropy, maximum, contrast, and homogeneity) were averaged on these 20 matrices.

Primitives matrices or acquisition length parameters characterize the shape and the size of the textural patterns in an image. GLCM features are four scalars extracted from a matrix **B**, where each element **B**(a,r) is the number of primitives of length r and gray level a, a primitive being a contiguous set of pixels with the same value. In our case, **B** was computed from the rescaled ROI as a 16×256 matrix.

Fractal dimension was calculated using the method described in detail by Caldwell (16) and Byng (17). The pixel value was seen as z-coordinate (x and y being its position in the ROI), and ruler sizes ε of 1 to 10 pixels were used to plot the log of the exposed surface $A(\varepsilon)$ versus $\log(\varepsilon)$. From this plot, the fractal dimension was computed using Equation 26 given in the Appendix. This feature indicates the degree of complexity in the textural patterns, a low fractal dimension denoting a rather simple and homogeneous structure.

Finally, we used the textural features described by Amadasun and King (32) to obtain four additional statistical parameters from the NGTDM. These features provide a mathematical description of the texture and are supposed to characterize texture properties like coarseness or complexity in the same way as human observers would do. ROIs were rescaled to 8 bits for the same reasons as for the GLCM and PMs.

The statistical characterization was also performed at another scale on the same ROIs. For this, all ROIs were averaged on square blocks of 8×8 pixels (thus leading to 32×32 pixels images). All the 18 above-mentioned quantities were then computed again on these images, and this provided a description of the texture at another scale, one order of magnitude higher than the first one. This step was inspired by the fact that the structures visible on mammograms are typically in the submillimeter to centimeter range. The total number of statistical features was thus 36, corresponding by definition to the dimension N of the classification process. Table 1 summarizes the whole set of 18 statistical features that were computed for each of the two scales, making a total of 36 features.

Table 1
Summary of the Texture Analysis Methods and the Corresponding Features

Analysis Method	Statistical Features
Gray level histogram	standard deviation
	skewness
	kurtosis
	balance
Gray level co-occurrence matrices	energy
	entropy
	cmax
	contrast
	homogeneity
Primitives matrices	short primitive emphasis
	long primitive emphasis
	gray level uniformity
	primitive length uniformity
Fractal analysis	fractal dimension
Neighbourhood gray-tone difference matrix	coarseness
	contrast′
	complexity
	strength

The 18 parameters in this table were computed for two scales as described in the text, making a total of 36 features.

Definition of Gold Standard From Radiologists' Ratings

In order to get a reliable gold standard, we asked three experienced radiologists (each of them having more than 10 years experience in radiology) to separately classify the 88 left/right pairs of CC-view and the 88 pairs of MLO-view mammograms, presented in random order on a laptop screen. The screen resolution was 3.6 pixels per millimeter, and brightness and contrast were adjusted before the reading session. The radiologists performed the classification individually, following the BI-RADS density scale definitions. Gold standard class was then defined for each of the 176 pairs of mammograms from the three radiologists' classifications (see later).

Classification Algorithms

The general purpose of pattern recognition is to determine to which category or class a given sample belongs (33). In this study, the samples are not directly the ROI: each ROI is characterized by an N-dimensional vector containing its computed statistical features ($N = 36$), and this observation vector serves as the input to a decision rule by which one of the given classes is attributed to the corresponding ROI. For the evaluation of the performance of the decision rule, the obtained classification is usually compared to a gold standard (also known as ground truth), which is assumed to represent the perfect classification of the samples.

All supervised classification algorithms require a set of training samples in order to establish the decision rule and a testing set to apply it. We used the leave-one-out method to avoid any bias introduced by testing on training samples. In this method, the tested ROI is always excluded from the learning process, while all other remaining ROIs are used to form the training set. Because the ROIs were strictly nonoverlapping, the 15 other ROIs selected from the same patient as the tested ROI were not excluded from the training set. This limitation allowed us to keep the number of training samples larger than N in all cases, which was a necessary condition for the computation of the features vectors covariance matrices.

We used three types of classification algorithms, namely a Bayesian classifier based on the measure of Mahalanobis distance, a naïve Bayesian classifier, and LDA. For all methods, the samples were the N-dimensional vectors characterizing the ROIs and the four density classes were used for both training and classification phases. Concretely, each ROI (represented by its projection onto the 36-dimensional features space) was successively considered as the test ROI. The decision rules for each classifier were computed from the training set consisting of the remaining 1407 ROIs, and a density class C_R attributed to the test ROI. The procedure was repeated until a class had been given to each ROI.

Classic Bayesian classifier based on Mahalanobis distance

For the Bayesian classifier, 50 ROIs per density class were chosen randomly from the actual training set and thus formed four subsets $\{S_k\}_{1\le k\le 4}$, each one containing 50 samples according to the gold standard (34). Assuming that the distribution of samples in each class could be approximated by an N-dimensional normal distribution, the probability of observing a given sample v in the class k is given by:

$$\psi_k(\mathbf{v}) = \frac{1}{\sqrt{(2\pi)^N \det \mathbf{K}_k}} \cdot \exp\left[-\frac{1}{2}(\mathbf{v}-\mu_k)^T \mathbf{K}_k^{-1}(\mathbf{v}-\mu_k)\right], \tag{1}$$

where μ_k represents the mean vector of class k and K_k is the covariance matrix of vectors in class k:

$$\mu_k = \frac{1}{n_k} \sum_{\mathbf{v_i}\in S_k} \mathbf{v_i} \tag{2}$$

$$\mathbf{K}_k = \frac{1}{n_k - 1} \sum_{\mathbf{v_i}\in S_k} (\mathbf{v_i}-\mu_k)^T(\mathbf{v_i}-\mu_k) \tag{3}$$

The product $(v - \mu_k)^T K_k^{-1}(v - \mu_k)$ in Equation 1 is known as the square of Mahalanobis distance and is a normalized measure of the distance between the sample vector v and the class center μ_k. K_k and μ_k were estimated from the sets $\{S_k\}_{1\le k\le 4}$ of 50 samples randomly chosen in the training set, to reduce computational cost and avoid unwanted rounding effect.

Under these assumptions, a Bayesian classifier could be defined. For a given sample v, the output of the classifier was a four-dimensional vector containing the four a posteriori probabilities $p(k|v)_{1\le k\le 4}$ for v to belong to class k as:

$$p(k|\mathbf{v}) = \frac{p(\mathbf{v}|k)p(k)}{p(\mathbf{v})} = \frac{\psi_k(\mathbf{v})p_a(k)}{\sum_k \psi_k(\mathbf{v})p_a(k)} \tag{4}$$

The attributed class was derived from the a posteriori probability vector components $p(k|v)$ as:

$$c_R = \textstyle\sum_{k=1}^{4} k \cdot p(k|\mathbf{v}), \tag{5}$$

with c_R being rounded to the nearest integer value to obtain the class attributed to the tested sample vector v.

In Equation 4, the a priori probability set $\{p_a(k)\}_{1\le k\le 4}$ was estimated as:

$$p_a(k) = \frac{1}{4}, \tag{6}$$

which represents the most conservative a priori assumption.

Naïve Bayesian classifier

For the second classifier, we implemented naïve Bayesian classification, which has been proven very powerful (35), even when the assumption of feature independence given the class, which is a sufficient condition for this method to be optimal, is violated (36). The proposed normalization forced the features to be independent and also greatly simplified the computation of p(*k*|v), since Equation 1 could be rewritten as:

$$\psi_k(\mathbf{v}_{n,k}) = \frac{1}{\sqrt{(2\pi)^N}} \exp\left[-\frac{1}{2}\mathbf{v}_{n,k}^T \mathbf{v}_{n,k}\right], \qquad (7)$$

where v has been normalized in the same way as training samples of class *k* to obtain the normalized vector $v_{n,k}$. The four a posteriori probabilities p(*k*|v) were then computed with Equation 4, and the attributed class with Equation 5.

We thus modified the Bayesian classifier procedure so that all feature distributions were within-class normalized. In order to force a distribution to be normal, its cumulative histogram was compared to the integral of the Gaussian density function: the normalized value $p_n^{j,k}$ of a given parameter $p^{j,k}$ is the solution of the equation:

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{p_n^{j,k}} e^{-t^2/2} dt = \frac{p^{j,k}}{p_{max}^{j,k}}, \qquad (8)$$

where $p_{max}^{j,k}$ is the highest value in the original distribution of feature *j* in class *k*.

Linear discriminant analysis

LDA implemented in Matlab Statistics Toolbox (37) is essentially similar to the first described algorithm, except that in Equation 3, only one pooled covariance matrix K_o is computed instead of the four K_k (homoscedasticity hypothesis), forcing the borders in the features space to be hyperplanes instead of quadrics. A multivariate normal density is then fitted to each class:

$$\psi_k(\mathbf{v}) = \frac{1}{\sqrt{(2\pi)^N \det \mathbf{K_o}}} \cdot \exp\left[-\frac{1}{2}(\mathbf{v}-\mu_k)^T \mathbf{K_o}^{-1}(\mathbf{v}-\mu_k)\right] \qquad (9)$$

Following, the decision rule used to attribute a class to a sample is in this case a simple linear combination of the features (37). The LDA classifier returns the class C_R corresponding to its position in the features space for each tested tample. This means that the a posteriori vector had only one non-zero component. As opposed to the classic Bayesian classifier described here, this variant made use of all ROIs present in the training set, without having to define one subset S_k per class.

Averaging the individual ROIs classifications

These three classifiers were used to individually classify all 1408 ROIs. However, since the BI-RADS density scale is based on an overall appreciation of the breast and since an overall dense breast may contain one or several ROIs that are essentially fatty, individual ROI classification may lead to results that differ from the radiologist's evaluations. Therefore, we also introduced two kinds of averaging to avoid decisions that were too localized. First, a posteriori probability vectors $[p(k|v)_{1 \le k \le 4}]$ were averaged for each mammogram over the four corresponding ROIs, and Equation 5 was used again to attribute a general class to each mammogram, instead of one per ROI. Second, we studied the effect of averaging on the 8 ROIs (four per mammogram) that had been defined for each left/right pair of CC or MLO views. This corresponds to the situation nearest to that of the three involved radiologists, who established the gold standard based on the display of a left/right pair of mammograms.

Reduction of the features' space size and number of classes

In order to reduce the original dimensionality of the features vector ($N = 36$) to a given $N' < N$ and to determine for that given N′ which parameters would lead to the best classification performance, we used standard features extraction techniques based on the maximization of the between-class scatter to the within-class scatter (Fisher linear discriminant) (38–40). Concretely, the Fisher linear discriminant gives a measure of the separability of the four density classes when only N′ features amongst the original N ones are considered for the classification. This process was conducted for $N' = 2$ and 5, and the separability measure was computed for every combination of N′ parameters (brute force testing). Once the best combination had been identified, all previously described algorithms were applied to the feature vectors orthogonally projected on the obtained subspaces, meaning that the classifiers only used the N′ best features for defining their classification rules.

We also examined the case of grouping BI-RADS 1 and 2 in the same density class, and BI-RADS 3 and 4 in another. We compared the performance obtained with this grouping being done before the training process, or after the classification (thus, respectively, two-class training – two-class classification and four-class training – two-class classification).

Evaluation of the performance

We used κ statistics with quadratic weights to evaluate the performance of the classification algorithms (41–45). This parameter represents the degree of chance-corrected agreement between two classifications (classification algorithm versus gold standard or radiologist versus radiologist) as:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{10}$$

where p_o is the observed agreement proportion and p_e the agreement expected by chance alone. Both are calculated from the confusion matrix and the quadratic weights matrix, and the values of κ stand between -1 and 1 (the minimum value actually depends on p_e but is always between -1 and 0). Benchmarks by Landis and Koch (46) (adjusted by Fleiss et al. [41] for taking the weighting process into account) are commonly used: κ values below 0.4 reflect poor agreement, between 0.4 and 0.6 moderate agreement, while it is substantial between 0.6 and 0.75 and excellent above 0.75. Weighted κ is particularly well adapted to multiclass tasks and when the classes are rather subjectively defined, which is the case for the BI-RADS density scale. The weighting process indeed differentiates between serious (more than one BI-RADS class difference) and slight disagreement (immediate neighbor class choice), and has been chosen as an evaluation parameter in numerous previous works on mammogram classification (16,17,20). Although much more sensitive to differences in class prevalence, the exact agreement proportion was also computed to be able to compare the performance with results from other studies (16,21,22).

RESULTS

The reference classifications by the three radiologists involved in this study are summarized in Table 2 and Figure 2. The exact agreement among the three classifications was 55%, while for the remaining 45% two of the radiologists chose a given BI-RADS density class and the last one chose an immediate neighboring class. When compared with each other, the three radiologists involved in our study obtained 67% to 79% exact agreement. The values of κ and exact agreement percentage, for each radiologist versus gold standard, are summarized in Table 2. Figure 2 presents the number of cases per radiologists' consensus level. The latter is defined as the number of radiologists having chosen the same BI-RADS category.

Table 2
Radiologist Classifications Compared to the Gold Standard Classification Defined in Text

	Radiologist # 1	Radiologist # 2	Radiologist # 3
Kappa	0.81 ± 0.07	0.88 ± 0.07	0.91 ± 0.08
Exact agreement	77%	89%	89%

Standard error for weighted κ was computed according to the formula given by Fleiss et al. (41).

Typical time periods to train and test the classifiers were 90 minutes for naïve Bayesian, 5 minutes for the Mahalanobis-Bayesian, and 1 minute for LDA classifier, on a Pentium 4 (3-GHz processor, 512 MB RAM). In the 36-dimensional feature space, Naïve Bayesian classification led to a κ value of 0.68 ± 0.07 and a percentage agreement with respect to the gold standard of 60%. This classifier was outperformed by the two others, since we obtained κ values of 0.78 ± 0.07 for Mahalanobis-Bayesian and 0.83 ± 0.08 for LDA. As one can expect from the overlap of standard errors, paired *t*-tests showed that none of these differences were significant at the 5% confidence level. The exact agreement proportions between these classifiers and the gold standard were, respectively, 76% and 83%. The confusion matrices given in Table 3 for the two best classifiers show that all but one mammogram pair were classified in the correct class or in one of its immediate neighbors. Moreover, this result was also valid when comparing breast density assessment of individual breasts before averaging the left/right pairs. The effect of the averaging process (individual ROI classification, averaging over the four ROIs defined for each mammogram, and averaging over the eight ROIs defined on a left/right pair of mammograms) is presented in Table 4.

The dimensionality reduction to N′ = 2 and 5 has as expected an effect on classification performance. As shown in Figure 3, κ decreases when the number of features is reduced, although both methods obtain already

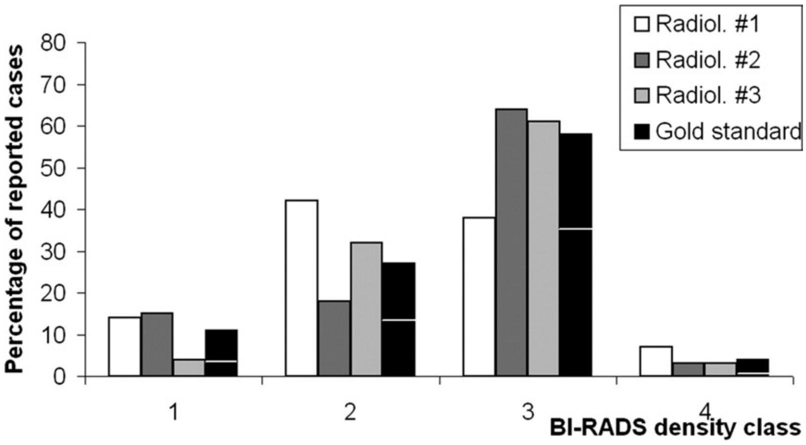


Figure 2. Repartition of the 176 breast pairs among BI-RADS density classes. The separation line in the gold standard column indicates the proportion of cases per consensus level: 3/3 (lower part of the column) or 2/3 (upper part).

Table 3
(a) Confusion Matrix Obtained for the Bayesian Classifier Based on Mahalanobis Distance. Results are Averaged Over Mammogram Pairs from the Same View. (b) Same for LDA Classifier

(a)	Gold Standard			
Bayesian Classifier	Density 1	Density 2	Density 3	Density 4
Density 1	14	3	0	0
Density 2	5	30	6	1
Density 3	0	14	86	3
Density 4	0	0	10	4

(b)	Gold Standard			
LDA Classifier	Density 1	Density 2	Density 3	Density 4
Density 1	16	3	0	0
Density 2	3	31	4	1
Density 3	0	13	95	3
Density 4	0	0	3	4

good results with five parameters only. The two optimal features for differentiating the four classes were homogeneity and coarseness, and the corresponding partition of the bi-dimensional subspace is given in Figure 4. For $N' = 5$, the optimal parameters were standard deviation, skewness, primitive length uniformity, fractal dimension, and coarseness, the latter parameter being computed from the block-averaged and the first four from the original ROI.

The reduction to a two-class problem led to the same results when the grouping of BI-RADS density classes was done before or after training. Naïve Bayesian classifier obtained κ values and percentage agreement of 0.68 ± 0.08 and 86%. Even if the difference is not significant at the 5% confidence level, it was once again outperformed by Mahalanobis Bayesian and LDA classifiers, for which the exact agreement were, respectively, 88% and 90% and weighted κ were 0.74 ± 0.08 and 0.78 ± 0.08. Thus, the performances of the last two classifiers for that particular two-class problem are excellent and nearly equivalent.

Finally, we observed no difference between the results obtained for CC and MLO views: performance of the automatic classifiers remained unchanged when the training phase was performed on one type of view and the classification on the other, or when training and classification processes were restricted to one view.

DISCUSSION

Because BI-RADS scale definitions allow for a certain freedom regarding interpretation, it was essential to carefully define the gold standard. The number of radiologists devoted to that task was between one and four among other published studies (16,17,20,21). The choice of three radiologists for this study was adequate, in the sense that there was no case where the three radiologists chose three different classes, or where one would have chosen a non-immediate neighbouring class respectively to the others. Thus, the odd number of radiologists permitted in all cases to unequivocally define the gold standard classification, as the class selected by at least two radiologists. The different case repartitions among the four BI-RADS classes are shown in Figure 2. The first radiologist tended to use the lowest categories more often than the other

Table 4
Weighted κ Values Obtained with the Different Averaging Processes and Classifiers. Exact Agreement is Given in Parenthesis

	Individual ROI Classification	Average per Mammogram (4 ROIs)	Average per View Type (8 ROIs)
Naïve Bayesian	0.50 ± 0.02 (39%)	0.65 ± 0.05 (55%)	0.68 ± 0.07 (60%)
Mahalanobis Bayesian	0.58 ± 0.03 (53%)	0.73 ± 0.05 (69%)	0.78 ± 0.07 (76%)
LDA	0.71 ± 0.03 (70%)	0.81 ± 0.05 (80%)	0.83 ± 0.08 (83%)

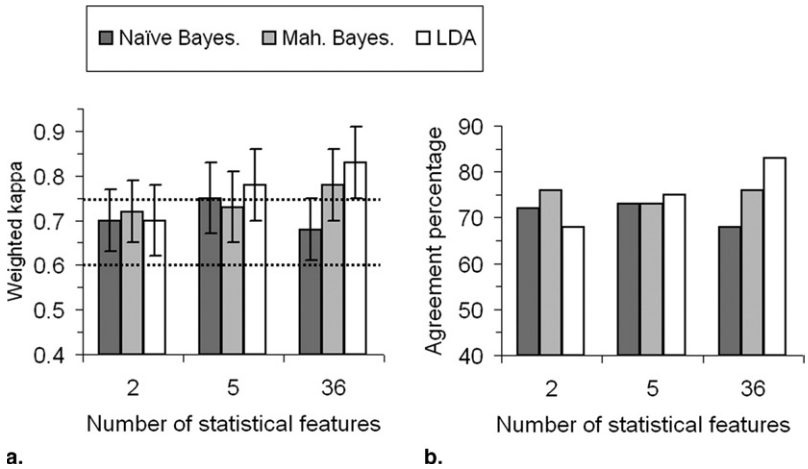


Figure 3. (**a**) Weighted κ value as a function of the features space dimensionality. Lines at 0.6 and 0.75 represent the limits for substantial and excellent agreement. (**b**) Corresponding percentage agreement.

two. The second observer classified the same proportion of mammograms between BI-RADS 1 and 2 categories, while reporting more than 60% in BI-RADS 3 category. The third observer barely used the extreme categories and concentrated most answers in BI-RADS 3 category as well.

The choice of presenting CC and MLO views separately to the radiologists allowed us to show that intraobserver reproducibility was excellent, even for different x-ray projections. The proportions of cases with one class difference between CC and MLO classifications were, respectively, 14%, 15%, and 9% for radiologists 1, 2, and 3. No difference greater than one BI-RADS density class was observed. Thus the corresponding confusion matrices (observer *i* CC classification versus observer *i* MLO classification) led to very high weighted κ values (0.90, 0.87, 0.87), showing that radiologists' classifications were nearly independent of the presented view. However, it was observed that the first observer attributed one class higher to MLO compared to CC for 10 of its 12 differences, while the second had the opposite trend (one class higher for CC view for 9 of the 13 differences), and the third observer had roughly equally distributed differences (5 of 8 cases with one class higher for CC).

The analysis of each within-class features distributions was in total agreement with the intuitive meaning of the statistical parameters and the two-scale analysis on normal and block-averaged ROIs provided very coherent results: the same trends were observed at millimeter and centimeter scales. Texture elements in low density breasts are small, fine, and well contrasted, with a high fractal dimension, while patterns in high density breasts are much coarser, due to the diffusive nature of glandular tissues.

The naïve Bayesian classifier obtained substantial agreement, but as some of the 36 features were strongly correlated, its performance was degraded as expected (35). The results of LDA and Bayesian classification based on the measure of Mahalanobis distance, in the 36-dimensional feature space, were remarkable, with cross-validated κ values of 0.78 and 0.83 respectively, and exact recognition proportions of 76% and 83%. LDA's slightly better performance is probably due to the fact that the whole 1407 ROIs training set was used for

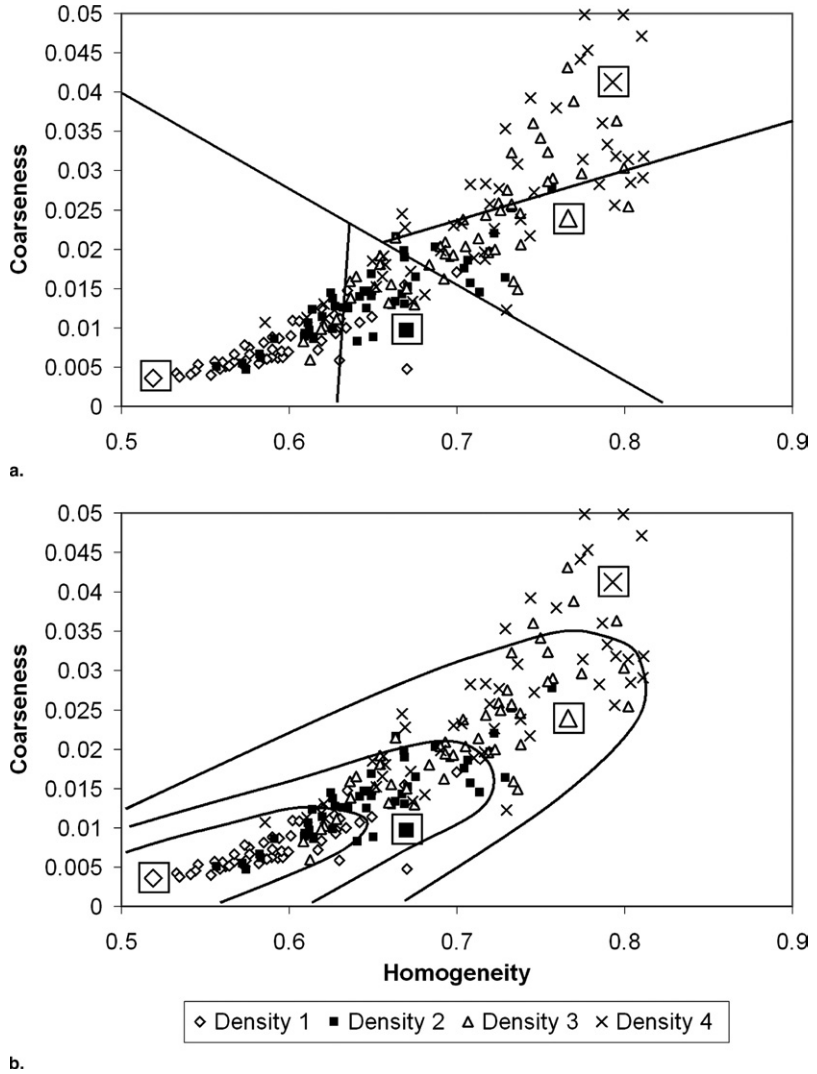


Figure 4. Partition of the optimal bidimensional feature subspace. (**a**) LDA leads to linear borders. (**b**) For Bayesian classifier based on Mahalanobis distance, the borders are conics. For visibility reasons, only 40 to 50 randomly chosen ROIs per density class are shown.

establishing the classification rules, whereas the same number of samples per density class, 50, was used for the Bayesian classifier, in order to avoid overtraining in the most represented classes. The confusion matrices given in Table 3 show an excellent differentiation of the four classes. However, half of the BI-RADS 4 cases were misclassified in density category 3 or 2. This may indicate that the sample size was too small for this category or that the gold standard assessment for this category was not accurate enough.

Compared to previous studies using a 4-grade scale (Wolfe scale [Caldwell et al. (16), Tahoces et al. (20)]), dense breast tissue proportion [Karrsemeijer (21)], BI-RADS [Petroudi et al. (22)], an improvement of κ and recognition rate was obtained: Caldwell et al. cite indeed κ values between 0.58 and 0.61, Tahoces et al., between 0.63 and 0.71, Karssemeijer, a global value of 0.73, and Petroudi et al., a recognition rate of 76%. This improvement can probably be explained by the homogeneity of the ROIs in size and localization in the breast, the absence of any background or pectoral muscle removal algorithm, the use of digital mammograms instead of digitized films, the wide range of complementary texture analysis techniques, and the averaging processes to take into account most of the dense breast region. Comparisons with these studies should proceed carefully; however, since little information is mentioned about preprocessing of the mammograms or case distributions. In addition, the comparison between the other scales with BI-RADS classification results is not a trivial point.

The reduction to a 2-grade scale (BI-RADS 1–2 versus BI-RADS 3–4, 88% recognition rate and higher) led to an excellent performance as well, comparable to the results given by Bovis and Singh (18).

When the algorithm had to make its decision based on two or five statistical parameters only, we found a substantial (weighted $\kappa > 0.7$) to excellent agreement with respect to the gold standard. The naïve Bayesian classifier proved that its performance is excellent for low-dimensional feature spaces, where the independence assumption can still be considered as valid. The performance of Mahalanobis Bayesian and LDA classifiers increased with the dimension of the optimal features subspace, with slightly better results for LDA for five features and above. It is interesting to note how the most optimal parameters were chosen in order to be complementary. For instance, with five parameters, two features related to gray level histogram (standard deviation and skewness), one from PM (plu), one from NGTDM (coarseness), and the fractal dimension were selected. This complementary nature between all texture analysis methods is one of the key points for obtaining a good classification even in a low-dimensional features space.

The improvement gained when averaging the results over four ROIs defined in the same breast and over the left/right breasts pair is clear for all classifiers. It shows that this process is the best way to take into account a significant part of both breasts, and thus avoid making a too local decision. Local classification, as shown in Table 4 for individual ROIs results, is not efficient for both Bayesian classifiers, although already substantially good for LDA.

Finally, according to Karssemeijer (21), the upper limit of the performance of an automatic classifier in terms of comparison with human observers remains an open question. It would be interesting to compare the gold standard defined in this study with other independent radiologists' classifications to have an idea of an empirical value of maximum κ and the exact agreement one could expect, the latter being evaluated by Karssemeijer (21) to be 80%. The exact agreement between the three radiologists involved in our study when compared with each other (67% to 79%) lies effectively in this range.

CONCLUSION

An excellent assessment of breast density according to BI-RADS was obtained with the semiautomated method presented in this study. A complete method was used combining complementary methods (moments, GLCM, PM, fractal dimension, and NGTDM) to describe ROIs manually chosen on digital mammograms, with widely used classification methods (LDA, Bayesian classification) and different averaging processes in order to take into account as much comprehensive information as possible. The results showed that the agreement between the radiologists and the automatic classifiers was notably higher than most previous published values, although extremely dense breasts (BI-RADS category 4) seemed somehow more difficult to classify accurately. Using $N = 36$ parameters led to high performance for the assessment of designing an automatic breast density classifier. The usefulness of mixing complementary methods was demonstrated by reducing the dimensionality of the feature space to five optimal parameters. The classifiers obtained excellent performances as well when tested in the two-class problem reduction. In a future phase, the validation procedure, currently limited to leave-one-ROI-out and justified by the fact that the ROIs do not overlap, could be extended to leave-one-patient-out on a larger patient database. The excellent results obtained with the most represented classes (BI-RADS categories 2 and 3) and with crossed-views training and testing suggest that the bias introduced by the leave-one-ROI-out method, if any, should not influence the overall performance of the classifiers, because in these cases training and testing on ROIs that had been taken from the same mammogram were less likely to happen.

The other key feature of the method resides in its simplicity. Apart from the fast computation of the 36 parameters, no additional algorithm is needed to remove the background, the pectoral muscle, and any potential imaging artefact, because a total control over the location of the ROIs is kept by manually selecting them. A fully automated classifier with a built-in location selection algorithm has not been investigated in this paper, but existing breast segmentation methods (12,47) could certainly be combined with the proposed classifiers to improve reproducibility and accuracy of the location choices. The automatization of ROI selection would help build a larger, more objective database, which is currently the main limitation of this study.

The proposed method represents a valuable tool for use in screening programs and could be inserted in a CAD device, in order to help radiologists in their density evaluation and diagnosis tasks. Intraobserver or interobserver variability in density assessment could indeed be avoided through the help of an automatic or semiautomatic classifier, and optimized data processing could be applied in order to display an optimal image to the radiologists for their diagnosis. An objective tool for determining breast density may find other potential applications in follow-up management for patients, with screening frequencies depending on breast density. Finally, training and education may benefit from such classifiers, in order to lower the variability of intraobserver and interobserver classifications inherent to the BI-RADS density class definitions.

ACKNOWLEDGMENTS

The authors are grateful to Elsabe Scott, MD, and Nigel Howarth, MD, for the classification of the mammograms and to Samuel De Laere for collecting the mammograms and creating the database.

REFERENCES

1. Fitzgibbons PL, Page DL, Weaver D, et al. Prognostic factors in breast cancer. College of American Pathologists Consensus Statement 1999. Arch Pathol Lab Med 2000; 124:966–978.
2. Ziv E, Smith-Bindman R, Kerlikowske K. Mammographic breast density and family history of breast cancer. J Natl Cancer Inst 2003; 95:556–558.
3. Colditz GA, Hankison SE, Hunter DJ, et al. The use of estrogens and progestins and the risk of breast cancer in postmenopausal women. N Engl J Med 1995; 332:1589–1593.
4. Kelsey JL, Gammon MD, John EM. Reproductive factors and breast cancer. Epidemiol Rev 1993; 15:36–47.
5. Boyd NF, Byng JW, Long RA, et al. Quantitative classification of mammographic densities and breast cancer risk: Results from the Canadian National Breast Screening study. J Natl Cancer Inst 1995; 87:670–675.
6. van Gils CH, Hendriks JH, Holland R, et al. Changes in mammographic breast density and concomitant changes in breast cancer risk. Eur J Cancer Prev 1999; 8:509–515.
7. Heine JJ, Malhotra P. Mammographic tissue, breast cancer risk, serial image analysis, and digital mammography: Part 1. Tissue and related risk factors. Acad Radiol 2001; 9:298–316.
8. Wolfe JN. Breast patterns as an index of risk for developing breast cancer. AJR Am J Roentgenol 1976; 126:1130–1137.
9. American College of Radiology (ACR) 2004. ACR Practice Guideline for the performance of screening mammography. Practice Guidelines and Technical Standards. 2004.
10. Harvey JA, Bovbjerg VE. Quantitative assessment of mammographic breast density: relationship with breast cancer risk. Radiology 2004; 230:29–41.
11. Brisson J, Diorio C, Mâsse B. Wolfe's parenchymal pattern and percentage of the breast with mammographic densities: Redundant or complementary classifications? Cancer Epidemiol Biomark Prevent 2003; 12:728–732.
12. Perconti P, Loew M. Analysis of parenchymal patterns using conspicuous spatial frequency features in mammograms applied to the BI-RADS density rating scheme. SPIE Medical Imaging: Image Processing 2006.
13. Kerlikowske K, Grady D, Barclay J, et al. Variability and accuracy in mammographic interpretation using the American College of Radiology Breast Imaging Reporting and Data System. J Natl Cancer Inst 1998; 90:1801–1809.
14. Berg WA, Campassi C, Langenberg P, Sexton MJ. Breast Imaging Reporting and Data System: Inter- and intraobserver variability in feature analysis and final assessment. AJR Am J Roentgenol 2000; 174:1769–1777.
15. Huo Z, Giger ML, Wolverton DE, Zhong W, Cumming S, Olopade OI. Computerized analysis of mammographic parenchymal patterns for breast cancer assessment: Feature selection. Med Phys 2000; 27:4–12.
16. Caldwell CB, Stappelton SJ, Holdsworth DW, et al. Characterisation of mammographic parenchymal pattern by fractal dimension. Phys Med Biol 1990; 35:235–247.
17. Byng JW, Boyd NF, Fishel E, Jong RA, Yaffe MJ. Automated analysis of mammographic densities. Phys Med Biol 1996; 41:909–923.
18. Bovis K, Singh S. Classification of mammographic breast density using a combined classifier paradigm. Proceedings of the 4th International Workshop on Digital Mammography, 2002, pp 177−180.
19. Zhou C, Chan HP, Petrick N, et al. Computerized image analysis: Estimation of breast density on mammograms. Med Phys 2001; 28:1056–1569.
20. Tahoces PG, Correa J, Souto M, Gómez L, Vidal JJ. Computer-aided diagnosis: The classification of mammographic breast parenchymal patterns. Phys Med Biol 1995; 40:103–117.
21. Karssemeijer N. Automated classification of parenchymal patterns in mammograms. Phys Med Biol 1998; 43:365–378.
22. Petroudi S, Kadir T, Brady M. Automatic classification of mammographic parenchymal patterns: A statistical approach. Proc IEEE Int Conf Eng Med Biol 2003; 798−801.
23. Vedantham S, Karellas A, Suryanarayanan S, et al. Full breast digital mammography with an amorphous silicon-based flat panel detector: Physical characteristics of a clinical prototype. Med Phys 2000; 27:558–567.
24. Burgess A. On the noise variance of a digital mammography system. Med Phys 2004; 31:1987–1995.
25. Muller S. Full-field digital mammography designed as a complete system. Eur J Radiol 1999; 31:25–34.
26. Hemdal B, Andersson I, Grahn A, et al. Can the average glandular dose in routine digital mammography screening be reduced? A pilot study using revised image quality criteria. Radiat Protect Dosimetry 2005; 114:383–388.
27. Li H, Giger ML, Olopade OI, Margolis A, Lan L, Chinander MR. Computerized texture analysis of mammographic parenchymal patterns of digitized mammograms. Acad Radiol 2005; 12:863–873.
28. Sonka M, Hlavak V, Boyle R. Image processing. In: Analysis and Machine Vision. 2nd ed. Pacific Grove, CA: Brooks/Cole, 1999.
29. Tuceryan M, Jain AK. Texture analysis. In: Chen CH, Pau LF, Wang P, editors. The Handbook of Pattern Recognition and Computer Vision. River Edge, NJ: World Scientific Publishing, 1998.
30. Haralick RM, Shanmugam K, Dinstein I. Textural features for image classification. IEEE Trans Syst Manage Cybern 1973; 3:610–662.
31. Lundhal T, Ohley WJ, Kunklinski WS, Williams DO, Gewirtz H, Most AS. Analysis and interpolation of angiographic images by use of fractals. Proceedings of IEEE Conference on Computers in Cardiology, 1985; 355−358.
32. Amadasun M, King R. Textural features corresponding to textural properties. IEEE Trans Syst Manage Cybern 1989; 19:1264–1274.
33. Fukunaga K. Introduction to Statistical Pattern Recognition. 2nd ed. San Diego, CA: Academic Press, 1990.
34. Chabat F, Guang-Zhong Y, Mansell DM. Obstructive lung diseases: Texture classification for differentiation at CT. Radiology 2003; 228: 871–877.
35. Friedman N, Geiger D, Goldszmidt M. Bayesian network classifier. Machine Learning 1997; 29:131–163.
36. Domingos P, Pazzani MJ. On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning 1997; 29:103–130.

37. Matlab Statistics Toolbox [computer program]. Natick, MA: MathWorks, 2004.
38. Nasri M, El Hitmy M. Algorithme génétique et Critère de la Trace pour l'Optimisation du vecteur Attribut: Application à la Classification Supervisée des Images de Textures. 15th International Conference on Vision Interface, 2002.
39. Yeom S, Javidi B. Three-dimensional distortion-tolerant object recognition using integral imaging. Optics Express 2004; 12:5795–5809.
40. Barrett HH, Myers KJ. Foundations of Image Science. Hoboken, NJ: Wiley, 2004.
41. Fleiss JL, Levin B, Paik MC. Statistical Rates and Proportions. 3rd ed. Hoboken, NJ: Wiley, 2003.
42. Ker M. Issues in the use of kappa. Invest Radiol 1991; 26:78–83.
43. Kundel HL. Measurement of Observer Agreement. Radiology 2003; 228:303–308.
44. Kraemer HC. Extension of the kappa coefficient. Biometrics 1980; 36: 207–216.
45. Kraemer HC, Periyakoil VS, Noda A. Kappa coefficients in medical research. In: D'Agostino RB, editor. Tutorials in Biostatistics, vol 1: Statistical Methods in Clinical Studies. Hoboken, NJ: Wiley, 2004.
46. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1973; 33:671–679.
47. Petroudi S, Brady M. Breast density segmentation using texture. Proceedings of the 8th International Workshop on Digital Mammography. Berlin/Heidelberg: Springer; 2006, pp 609–615.
48. Mandelbrot BB. The Fractal Geometry of Nature. San Francisco, CA: WH Freeman, 1982.

APPENDIX: DEFINITION OF THE STATISTICAL PARAMETERS

Parameters Computed From the Gray Level Histogram

The first four moments and balance parameter are computed from the individual pixel values x_i as follows:

$$\text{mean} \equiv \bar{x} = \frac{1}{N}\sum_i x_i \tag{11}$$

$$\text{standard dev.} \equiv \sigma = \frac{1}{\sqrt{N-1}}\Big(\sum_i (x_i - \bar{x})^2\Big)^{1/2} \tag{12}$$

$$\text{skewness} = \frac{1}{N\sigma^3}\sum_i (x_i - \bar{x})^3 \tag{13}$$

$$\text{kurtosis} = \frac{1}{N\sigma^4}\sum_i (x_i - \bar{x})^4 - 3 \tag{14}$$

$$\text{balance} = \frac{x_{70} - \bar{x}}{\bar{x} - x_{30}}, \tag{15}$$

where the summations are performed over the N pixels of the ROI, and x_p is the gray level yielding to pth percentile of the gray level distribution (15).

Gray Level Co-occurrence Matrices (GLCM)

The GLCM are computed as follows: first, the ROI is linearly rescaled to 16 gray levels only. Then for a given direction d and a given distance r, each element [i,j] of the co-occurrence matrix $(\mathbf{C})^{d,r}_{i,j}$ is given by the number of times that a couple of pixels separated by a distance r along a direction d have the values i and j, respectively. Each co-occurrence matrix is then normalized by the sum of its elements. The directions chosen for the GLCM are [1,0], [1,1], [0,1], and [−1,1], corresponding to angles of 0°, 45°, 90°, and 135°, respectively. The distances are 1, 3, 5, 7, and 9 pixels for each direction, which yields a set of 20 GLCM. Scalar parameters are then extracted from each matrix as follows:

$$\text{energy}(\mathbf{C}) = \sum_{i,j} \mathbf{C}^2_{i,j} \tag{16}$$

$$\text{entropy}(\mathbf{C}) = -\sum_{i,j} \mathbf{C}_{i,j} \log \mathbf{C}_{i,j} \tag{17}$$

$$\text{cmax}(\mathbf{C}) = \max_{i,j} \mathbf{C}_{i,j} \tag{18}$$

$$\text{contrast}(\mathbf{C}) = \sum_{i,j} |i-j|^2\, \mathbf{C}_{i,j} \tag{19}$$

$$\text{homogeneity}(\mathbf{C}) = \sum_{i,j} \frac{\mathbf{C}_{i,j}}{1 + |i-j|} \tag{20}$$

Primitives Matrix (PM)

Each element [a,r] of the primitives matrix $\mathbf{B}_{a,r}$ is the number of primitives of gray-level a and length r, a primitive being a contiguous set of pixels having the same value. As for GLCM, each ROI is rescaled to 4 bits before its primitives matrix is computed. Note that its dimensions are $(2^4 - 1, r_{max})$, because $0 \le a \le 2^4 - 1$ and $1 \le r \le r_{max}$, where r_{max} is the dimension of the ROI, corresponding to the maximal primitive length one could find in such an image. From this primitives matrix, four parameters are then extracted for each ROI: short primitive emphasis (spe), long primitive emphasis (lpe), gray level uniformity (glu), and primitive length uniformity (plu), defined by:

$$\text{spe} = \frac{1}{\mathbf{B}_{tot}}\sum_a \sum_r \frac{\mathbf{B}_{a,r}}{r^2} \tag{21}$$

$$\mathrm{lpe} = \frac{1}{\mathbf{B}_{tot}} \sum_{a} \sum_{r} \mathbf{B}_{a,r} r^2 \quad (22)$$

$$\mathrm{glu} = \frac{1}{\mathbf{B}_{tot}} \sum_{a} \left(\sum_{r} \mathbf{B}_{a,r} \right)^2 \quad (23)$$

$$\mathrm{plu} = \frac{1}{\mathbf{B}_{tot}} \sum_{r} \left(\sum_{a} \mathbf{B}_{a,r} \right)^2, \quad (24)$$

where $\mathbf{B}_{tot}$ is the sum of the elements of the primitives matrix $\mathbf{B}$: $\mathbf{B}_{tot} = \sum_a \sum_r \mathbf{B}_{a,r}$. Note that $\mathbf{B}$ could be defined for several directions, but we limited our investigations to one (34), corresponding to a scan of the image along direction [1,0].

Fractal Dimension

The fractal dimension of a two-dimensional (2D) image can be computed by a box-counting method as an extension to the one-dimensional (1D) case. Mandelbrot (48) first described the 1D problem of measuring a coastline on a map, with a ruler of a particular length ε. The smaller the ruler, the larger is the measured distance, because more and more details can be taken into account for the analysis. Mandelbrot gave the empirical relationship between the ruler size ε, and the measured length L, as:

$$L(\varepsilon) = \lambda \varepsilon^{1-D} \quad (25)$$

In Equation 25, λ is a scaling constant, and D is called the fractal dimension of the curve.

The generalization to a 2D image can be done as follows (16,17,31). First, the image to be analyzed is converted to a pseudo-3D surface, with the first two coordinates representing the spatial position of each pixel, the third one being the gray level. The total area A of the 3-D surface is then computed. For various values of the ruler size ε, the pixel values are then averaged over blocks of size $\varepsilon \times \varepsilon$, and the area $A(\varepsilon)$ is computed. For the 2D case, Equation 25 becomes:

$$A(\varepsilon) = \lambda \varepsilon^{2-D} \quad (26)$$

According to this equation, D can be estimated from a plot of $\log\{A(\varepsilon)\}$ versus $\log\{\varepsilon\}$.

Neighborhood Gray-Tone Difference Matrix (NGTDM)

NGTDM is a column matrix first defined by Amadasun and King (32) as follows: let $x_{k,l}$ be the gray level value of the pixel located at (k,l) on a two-dimensional image. The average neighbouring value is given by:

$$\bar{A}_{x_{k,l}} = \frac{1}{W-1} \left[\sum_{m=-d}^{d} \sum_{n=-d}^{d} x_{k+m,\, l+n} \right], \quad (m,n) \neq (0,0), \quad (27)$$

where d = 3 is the neighbouring size and W = $(2d+1)^2$. Denoting $\{X_i\}$ the set of all pixels with value i in the ROI, the ith entry of the NGTDM is given by:

$$s(i) = \sum_{x \in X_i} |i - \bar{A}_x| \quad (28)$$

Scalar parameters extracted from the NGTDM are:

$$\mathrm{coarseness} = \left[\varepsilon + \sum_{i=0}^{i_{\max}} p_i s(i) \right]^{-1} \quad (29)$$

$$\mathrm{contrast}' = \left[\frac{1}{N_g(N_g - 1)} \sum_{i=0}^{i_{\max}} \sum_{j=0}^{j_{\max}} p_i p_j (i-j)^2 \right] \cdot \left[\frac{1}{n^2} \sum_{i=0}^{i_{\max}} s(i) \right] \quad (30)$$

$$\mathrm{complexity} = \sum_{i=0}^{i_{\max}} \sum_{j=0}^{j_{\max}} \frac{|i-j|[p_i s(i) + p_j s(j)]}{n^2 (p_i + p_j)}, \quad p_i > 0,\ p_j > 0 \quad (31)$$

$$\mathrm{strength} = \frac{\sum_{i=0}^{i_{\max}} \sum_{j=0}^{j_{\max}} (p_i + p_j)(i-j)^2}{\varepsilon + \sum_{i=0}^{i_{\max}} s(i)}, \quad p_i > 0,\ p_j > 0, \quad (32)$$

where $p_i = |X_i| / \sum_{i=0}^{i_{\max}} |X_i|$ is the probability of occurrence of gray level i in the ROI, $i_{\max}$ the highest gray level and N_g the number of different gray levels effectively present in the ROI and ε a small number (10^{-12} in our case) to prevent coarseness and strength becoming infinite. The feature representing the contrast given by Equation 30 is called here *contrast'*, to make a distinction with the contrast derived from the primitives matrices (see Equation 19).

Mammographic texture synthesis: second-generation clustered lumpy backgrounds using a genetic algorithm

Cyril Castella[1,2], Karen Kinkel[3], François Descombes[4], Miguel P. Eckstein[5], Pierre-Edouard Sottas[1], Francis R. Verdun[1], and François O. Bochud[1*]

[1]Institut Universitaire de Radiophysique Appliquée,CHUV and University of Lausanne, Grand-Pré 1, CH-1007 Lausanne Switzerland
[2]LPHE,EPFL,CH-1015 Lausanne Switzerland
[3]Clinique des Grangettes, Chemin des Grangettes 7,CH–1224 Chêne-Bougeries Switzerland
[4]Haute Ecole Cantonale Vaudoise de la Santé, Bugnon 19,CH-1005 Lausanne,Switzerland
[5]Dept. of Psychology, University of California, Santa Barbara,California 93106-9660,USA
**Corresponding author: francois.bochud@chuv.ch*

Abstract: Synthetic yet realistic images are valuable for many applications in visual sciences and medical imaging. Typically, investigators develop algorithms and adjust their parameters to generate images that are visually similar to real images. In this study, we used a genetic algorithm and an objective, statistical similarity measure to optimize a particular texture generation algorithm, the clustered lumpy backgrounds (CLB) technique, and synthesize images mimicking real mammograms textures. We combined this approach with psychophysical experiments involving the judgment of radiologists, who were asked to qualify the visual realism of the images. Both objective and psychophysical approaches show that the optimized versions are significantly more realistic than the previous CLB model. Anatomical structures are well reproduced, and arbitrary large databases of mammographic texture with visual and statistical realism can be generated. Potential applications include detection experiments, where large amounts of statistically traceable yet realistic images are needed.

©2008 Optical Society of America

OCIS codes: (100.2960) Image analysis; (170.3830) Mammography; (330.5000) Vision – patterns and recognition

References and Links

1. P. F. Judy, R. G. Swensson, R. D. Nawfel, and K. H. Chan, "Contrast detail curves for liver CT," Med.Phys. **19**, 1167-1174 (1992).
2. S. E. Seltzer, P. F. Judy, R. G. Swensson, K. H. Chan and R. D. Nawfel, "Flattening of the contrast-detail curve for large lesions on liver CT images," Med. Phys. **21**, 1547-1555 (1994).
3. M.P. Eckstein and J.S. Whiting, "Visual signal detection in structured backgrounds. I. Effect of number of possible spatial locations and signal contrast," J. Opt. Soc. Am. A **13**, 1777-1787 (1996).
4. Y. Zhang, B. T. Pham, and M. P. Eckstein, "Evaluation of JPEG 2000 Encoder Options: Human and Model Observer Detection of Variable Signals in X-Ray Coronary Angiograms," IEEE Trans. Med. Imaging **23**, 613-632 (2004).
5. E. A. Krupinsky and H. Roehring, "Pulmonary nodule detection and visual search: P45 and P104 monochrome versus color monitor displays," Acad. Radiol. **9**, 638-645 (2002).
6. F.O. Bochud, J.-F. Valley, F.R. Verdun, C. Hessler and P. Schnyder, "Estimation of the noisy component of anatomical backgrounds," Med. Phys. **26**, 1365-1370 (1999).
7. D. S. Brettle, E. Berry, and M. A. Smith, "The effect of experience on detectability in local area anatomical noise," BJR **80**, 186-193 (2007).
8. R.F. Wagner and D.G. Brown., "Unified SNR analysis of medical imaging systems," Phys. Med. Biol. **30**, 489-518 (1985).
9. M. P. Eckstein, C. K. Abbey, and F. O. Bochud, "Practical guide to model observers in real and synthetic noisy backgrounds," in *Handbook of Medical Imaging, Physics & Psychophysics*, K. Beutel, H. Kundel, and K. Vanmetter, eds (SPIE Press, Bellingham, Washington, 2000).
10. H. H. Barrett and K. J. Myers, *Foundations of Image Science* (Wiley, Hoboken, NJ, 2004).

#91988 - $15.00 USD Received 23 Jan 2008; revised 10 Apr 2008; accepted 16 Apr 2008; published 12 May 2008

11. L. Chen and H. H. Barrett, "Task-based lens design with application to digital mammography," J. Opt. Soc. Am. A **22**, 148-167 (2005).
12. M. A. Kupinski, E. Clarkson, J. H. Hoppin, L. Chen, and H. H. Barrett, "Experimental determination of object statistics from noisy images," J. Opt. Soc. Am. A **20**, 421-429 (2003).
13. K. J. Myers, H. H. Barrett, M. C. Borgstrom, D. D. Patton, and G. W. Seeley, "Effect of noise correlation on detectability of disk signals in medical imaging," J. Opt. Soc. Am. A **2**, 1752-1759 (1985).
14. F. O. Bochud, C. K. Abbey, and M. P. Eckstein, "Visual signal detection in structured backgrounds. III. Calculation of figures of merit for model observers in statistically nonstationary backgrounds," J. Opt. Soc. Am. A **17**, 193-205 (2000).
15. F. O. Bochud, C. K. Abbey, and M. P. Eckstein, "Search for lesions in mammograms: Non-Gaussian observer response," Med. Phys. **31**, 24-36 (2004).
16. C. Castella, C. K. Abbey, M. P. Eckstein, F. R. Verdun, K. Kinkel, and F. O. Bochud, "Human linear template with mammographic backgrounds estimated with a genetic algorithm," J. Opt. Soc. Am. A **24**, B1-B12 (2007).
17. B. Bliznakova, Z. Bliznakov, V. Bravou, Z. Kolitsi, and N. Pallikarakis, "A three-dimensional breast software phantom for mammography simulation," Phys. Med. Biol. **48**, 3699-3719 (2003).
18. P. R. Bakic, M. Albert, D. Brzakovic, A. D. Maidment, "Mammogram synthesis using a 3D simulation. I. Breast tissue model and image acquisition simulation," Med. Phys. **29**, 2131-9 (2002).
19. P. R. Bakic, M. Albert, D. Brzakovic, A. D. Maidment, "Mammogram synthesis using a 3D simulation. II. Evaluation of synthetic mammogram texture," Med. Phys. **29**, 2140-51 (2002).
20. J. P. Rolland and H. H. Barrett, "Effect of random background inhomogeneity on observer detection performance," J. Opt. Soc. Am. A **9**, 649-658 (1992).
21. F. O. Bochud, C. K. Abbey, and M. P. Eckstein, "Statistical texture synthesis of mammographic images with clustered lumpy backgrounds," Opt. Express **4**, 33-43 (1999).
22. D. Whitley, "A Genetic Algorithm Tutorial," Stat. Comput. **4**, 65-85 (1994).
23. A. E. Eiden, R. Hinterding, and Z. Michalewicz, "Parameter Control in Evolutionary Algorithms," IEEE Trans. Evol. Comput. **3**, 124-141, 1999.
24. M. Sonka, V. Hlavak, and R. Boyle. *Image processing, Analysis and Machine Vision* (Brooks/Cole, Pacific Grove, Ca, 1999).
25. M. Tuceryan and A. K. Jain. "Texture Analysis," in *The Handbook of Pattern Recognition and Computer Vision*, C. H. Chen, L. F. Pau, and P. Wang, eds., (World Scientific Publishing Co, River Edge, NJ, 1998).
26. R. M. Haralick, K. Shanmugam, and I. Dinstein. "Textural Features for Image Classification," IEEE Trans. Syst. Man. Cybern. **3**, 610-62 (1973).
27. M. Amadasun and R. King, "Textural features corresponding to textural properties," IEEE Trans. Syst. Man, Cybern. **19**, 1264-1274 (1989).
28. C. B. Caldwell, S. J. Stappelton, D. W. Holdsworth, R. A. Jong, W. J. Weiser, G. Cooke, and M. J. Yaffe, "Characterisation of mammographic parenchymal pattern by fractal dimension," Phys. Med. Biol. **35**, 235-247 (1990).
29. C. Castella, K. Kinkel, M. P. Eckstein, P.-E. Sottas, F. R. Verdun, and F. O. Bochud, "Semiautomatic Mammographic Parenchymal Patterns Classification Using Multiple Statistical Features," Acad. Radiol. **14**, 1486-1499 (2007).
30. Z. Huo, M. L. Giger, D. E. Wolverton, W. Zhong, S. Cumming, O. I. Olopade, "Computerized analysis of mammographic parenchymal patterns for breast cancer assessment. Feature selection," Med. Phys. **27**, 4-12 (2000).
31. S. Vedantham, A Karellas, S. Suryanarayanan, D. Albagli, S. Han, E.J. Tkaczyk, C.E. Landberg, B. Opsahl-Ong, P.R. Granfors, I. Levis, C.J. D’Orsi, and R.E. Hendrick, "Full breast digital mammography with an amorphous silicon-based flat panel detector: Physical characteristics of a clinical prototype," Med. Phys. **27**, 558-567 (2000).
32. S. Muller, "Full-field digital mammography designed as a complete system," Eur. J. Radiol. **31**, 25-34 (1999).
33. T. Bäck and M. Schütz. "Intelligent mutation rate control in canonical genetic algorithms," in *Proceedings of the 9th International Symposium on Foundations of Intelligent Systems, number 1079 in Lectures notes in Artificial Intelligence*, Z. Ras and M. Michalewicz, eds., (Springer, London, UK, 1996), pp. 158-167.
34. American College of Radiology, *Breast Imaging Reporting and Data System Atlas* (American College of Radiology, Reston, Va, 2003).
35. A. Burgess and P. Judy, "Signal detection in power-law noise: effect of spectrum exponent," J. Opt. Soc. Am. A **24**, B52-B60 (2007).
36. J. R. Taylor, *An Introduction to Error Analysis*, (University Science Books, Mill Valley, Ca, 1982).

1. Introduction

The problem of human perception and performance in radiology detection tasks has been studied in numerous frameworks in the past: detection of a tumor on computer tomographic images of the liver [1,2], stenosis in a blood vessel on fluoroscopic images [3], filling defects in X-ray coronary angiograms [4], nodules on pulmonary radiographs [5], or microcalcifications on mammograms [6]. The aim of such studies is to determine the role on diagnostic detection of the inherent parameters of the images like resolution or contrast, the imaging unit acquisition parameters or the anatomy in the detection process. Many of such studies are psychophysical experiments involving radiologists or trained naïve [7] observers.

In particular, there has been a large interest in developing models that can predict human observer performance for detection tasks as a function of the image characteristics and the observer properties [8-10]. These models aim at avoiding subjective methods to evaluate image quality and/or objective yet time-consuming methods such as psychophysical studies [11,12]. Models for objects superimposed on various types of real backgrounds or computer generated noises patterns have been developed and applied to the detection of lesions in radiological images [13-16].

Both psychophysical and model observer approaches require a large number of images to obtain accurate results. Real images or regions of interest (ROIs) would be ideal, but in most cases the number of available clinical images is limited. In addition, the question arises about reproducibility of the results with sets of images obtained with other imaging systems, digitization methods, or image post-processing. An alternative to using real images is to use computer generated images. This would allow for generation of unlimited number of samples with known and well-controlled statistical properties. Such images might have adjustable properties that would not depend on imaging device characteristics or digitization processes.

Two major methods have been explored for producing synthetic images mimicking mammograms. First, complete three-dimensional simulation of the breast components and properties, in conjunction with imaging device simulation, which is expected to produce very realistic images [17-19]. However, the complexity and computational cost associated with such modeling and the difficulty of taking into account breast compression can often be a limitation in the quality of the resulting images. For that reason, 2D approaches have been investigated, using backgrounds constituted by the summation of elementary bright structures called blobs [11,20,21]. These lumpy backgrounds, as named originally by Rolland and Barrett [20], were designed to reproduce general lumpy textures. Bochud *et al.* [21] generalized the model to clustered lumpy backgrounds (CLB), matching the Wiener spectra of real mammograms and synthetic backgrounds and empirically optimizing the parameters to obtain images which were as visually realistic as possible. Lumpy backgrounds and CLB images have the advantage of having analytically computable statistical properties, and are stationary within their boundaries. Statistical descriptions of general lumpy and CLB objects have been further investigated by Kupinsky *et al.* [12]. However, for this model as for most of 3D or 2D methods, thorough and objective assessment of visual realism and similarity of statistical properties to real images has not been carried out. The main obstacle has been the difficulty of defining criteria for the assessment.

For synthetic images to be used by humans and model observers then necessary criteria are that the images look visually similar to the real images (visual realism) and that the statistical properties of the synthetic images match to larger degree those of the real images (statistical realism). Although these criteria are typically aimed at when creating synthetic backgrounds, the process is commonly approached through trial and error and comparison of a few synthetic and real images. The purpose of the current work was to systematically optimize the visual and statistical realism using a genetic algorithm as search optimization routine.

Specifically, our aim in this study was to extend and optimize the CLB model and to objectively assess the realism of the obtained images. For this purpose, we used a database of 1000 square ROIs selected from real mammograms, and defined a metric based on the Mahalanobis distance to compute the statistical distance between real images and synthetic CLB images. The CLB parameters were optimized using a genetic algorithm in order to minimize the Mahalanobis distance. Psychophysical experiments involving radiologists and radiographers were then designed in order to evaluate the visual realism of the synthetic images.

#91988 - $15.00 USD Received 23 Jan 2008; revised 10 Apr 2008; accepted 16 Apr 2008; published 12 May 2008

2. Material and methods

2.1 Clustered lumpy background (CLB) model

Lumpy backgrounds are synthetic, digital images generated by superposition of elementary bright blobs. The number of blobs is randomly sampled according to a Poisson process and the blob centers are placed at random locations uniformly distributed in the image. Lumpy backgrounds were originally designed by Rolland and Barrett [20] with circularly symmetric blobs b($\mathbf{r}$), so that the image **g** could be written as:

$$\mathbf{g}(\mathbf{r}) = \sum_{k=1}^{K} b(\mathbf{r} - \mathbf{r}_k), \tag{1}$$

where $\mathbf{r}_k$ is the center position of the kth blob, and K the total number of blobs in the image.

Later, Bochud, *et al.*, [21] generalized this model to clusters of exponential, not necessarily circular symmetric blobs. Clustered lumpy backgrounds (CLB) are produced by randomly choosing a number of clusters, K, following a Poisson process, and distributing them randomly on the image plane. For each cluster, a random number of blobs, N_k, are positioned randomly around the cluster center according to a probability density function (pdf) $\phi(\mathbf{r})$. Finally, all blobs belonging to the same kth cluster are rotated by an angle θ_k before being summed to obtain the final image g($\mathbf{r}$):

$$\mathbf{g}(\mathbf{r}) = \sum_{k=1}^{K} \sum_{n=1}^{N_k} b(\mathbf{r} - \mathbf{r}_k - \mathbf{r}_{kn}, \mathbf{R}_{\theta k}) \tag{2}$$

All parameters and their distributions are summarized in Table 1. The general functional expression of the blob has been chosen as:

$$b(\mathbf{r}, \mathbf{R}_\theta) = \exp\left(-\alpha \frac{\left\| \mathbf{R}_\theta \mathbf{r} \right\|^\beta}{L(\mathbf{R}_\theta \mathbf{r})} \right), \tag{3}$$

where α and β are real parameters, and L is the characteristic length of an ellipse with half axes equal to L_x and L_y [21]. One of the major advantages of CLB technique is that some statistical properties of g($\mathbf{r}$) like its power spectrum can be analytically computed from the model parameters.

Table 1. Definitions and distributions of the CLB model parameters.

Variable	Definition	Distribution
K	number of clusters	poisson with mean value K_0
$\mathbf{r}_k$	position of the kth cluster	uniform across image
N_k	number of blobs within the kth cluster	poisson with mean value N_0
θ_k	rotation angle of the blobs in the kth cluster	uniform between 0 and 2π
$\mathbf{R}_\theta$	rotation matrix of angle θ	N/A
$\mathbf{r}_{kn}$	position of the nth blob within the kth cluster	Gaussian pdf $\phi(\mathbf{r})$
b($\mathbf{r}$,$\mathbf{R}_\theta$)	blob profile rotated at angle θ	N/A

#91988 - $15.00 USD Received 23 Jan 2008; revised 10 Apr 2008; accepted 16 Apr 2008; published 12 May 2008

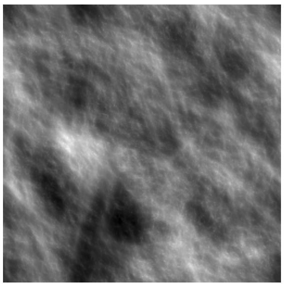


Fig. 1. (1.55 MB) Movie showing the construction of a CLB image. This example has two CLB layers with isotropic orientation of the blobs.

The free parameters of the CLB model are thus $\{\alpha,\beta,L_x,L_y,\sigma_x,\sigma_y,K_0,N_0\}$, where σ_x and σ_y are the standard deviations of the Gaussian pdf, $\phi(\mathbf{r})$, in x and y directions respectively. These 8 parameters had been empirically optimized in the original study [21], on the basis of visual inspections of the images and comparison of Wiener spectrum with that of real mammograms. These values were used as a starting point for our study.

In order to improve the realism of CLB images, we introduced two variations into the model. First, we superimposed another CLB onto the image computed from Eq. (2), with fixed parameters $\alpha = 2.0$, $\beta = 0.9$, $L_x = 50$ pixels, $L_y = 5$ pixels, $\sigma_x = 10$ pixels, $\sigma_y = 10$ pixels, and free parameters $K_0' << K_0$ and N_0'. The inclusion of a small amount of long and narrow blobs aims to better reproduce the fibrous structures of real mammograms.

The second variation was included to favor oriented structures similar to those visible on real mammograms. At the whole breast scale, these structures arise from the projection of the ducts converging towards the nipple, or from suspensory ligaments. For this purpose, the pdf of the rotation angle was changed from uniform to Gaussian with a mean equal to θ_0 and a standard deviation of $\pi/6$. With this change, the large scale oriented structures were constructed by the summation of clusters with similar orientation. The mean parameter, θ_0, was changed randomly with uniform pdf between 0 and 2π for each realization. If two superimposed layers were used for the image, both used the same θ_0.

An example of a synthetic image generation is shown on Fig. 1.

2.2 Optimization of the CLB parameters with a genetic algorithm

Genetic algorithms are a family of computational models inspired by evolution [22]. The free parameters of a given optimization problem are encoded on a chromosome-like data structure, and selection and recombination operators are applied in order to allow a population of potential solutions to evolve towards the optimal solution of the problem. The initial population is usually chosen randomly in the search space, and the corresponding chromosomes are evaluated through a fitness function. The best chromosomes are given better reproduction and survival opportunities. Following, crossover and mutation operators are applied in order to generate a new population of equal cardinality. These processes of evaluation, crossover and mutation are repeated until a user-defined (sub-)optimal value of the fitness function is reached, or when the best chromosome of the population has not been improved for a given number of generations.

Genetic algorithms have a great potential for non-linear function optimization in multi-dimensional spaces, since the intrinsic parallel structure of the optimization process is highly efficient for exploring multiple locations in the search space simultaneously, and avoiding local extrema. They can be used for binary or real coded problems, and many specific reproduction/mutation operators and techniques have been designed [23] in order to create specific algorithms for handling a wide range of optimization problems.

According to Eqs. (2) and (3), a classical CLB implementation requires a set of eight real parameters $\{\alpha,\beta,L_x,L_y,\sigma_x,\sigma_y,K_0,N_0\}$. For the 2-layer CLB, the addition of $\{K_0',N_0'\}$ increases the number of parameters to ten. The statistical properties of CLB images depend in a non-analytical way on the parameters. Their optimization is furthermore complicated by the stochastic nature of the realizations for a same set of parameters. The optimization of the eight parameters of the previously published CLB model [21] were limited to maximize the similarity of basic gray level (GL) histogram properties and Wiener spectrum of the synthetic and real mammographic textures, and to produce images qualitatively similar to real mammograms ROIs. No other consideration was taken into account in order to evaluate the mathematical realism of the obtained synthetic images. One key aspect of the present study was to introduce a metric based on Mahalanobis distance for quantifying similarity between synthetic and real images.

For this purpose, 36 statistical features based on complementary textural patterns analysis methods were computed. We used the GL histogram properties, the gray-level co-occurrence matrices (GLCM) [24-26], the primitives matrices [25], the neighborhood gray tone difference matrix (NGTDM) [27], and the fractal dimension [28], and computed the features for 1000 square ROIs within digital mammograms [29]. These 256 by 256 pixels square regions were selected from the central breast areas of digital mammograms. We used a database of 88 patients who underwent screening exams on a GE Senograph 2000D full-field digital detector [31,32] , with one craniocaudal (CC) and one mediolateral oblique (MLO) view per breast per patient.

Features derived from the GL histogram were standard deviation, skewness, kurtosis, and balance [30]. They describe the general properties of the overall gray level distribution, including the histogram shape and symmetry. GLCM features were energy, entropy, maximum, contrast, and homogeneity. GLCM give information about the spatial relationships of GL in structural patterns. Primitives matrices (also known as run-length matrices) characterize the size and shape of textural patterns in an image. Short primitive emphasis, long primitive emphasis, gray level uniformity, and primitive length uniformity provided four more features. Additionally, four statistical parameters were computed from NGTDM: coarseness, contrast, complexity, and strength. These features were designed by Amadasun and King in order to give mathematical descriptions of the subjective aspect of images with such textural properties [27]. Finally, the fractal dimension was computed. This feature is related to the complexity of textural patterns, a low fractal dimension denoting a rather homogeneous image structure. These 18 statistical quantities were computed for each of the 1000 mammograms ROIs, providing information about the structural patterns from the mm to the cm scale. As structures in mammograms typically range from about 1 mm to a few cm, this statistical analysis was also performed at another scale on the same ROIs in order to characterize the larger scale textural properties. For this purpose, each ROI was averaged on square 8 x 8 pixels blocks, and the same 18 parameters were computed again, making a total of 36 features. An exhaustive description of the mathematical definitions of the statistical features used in this work have been published in a previous study [29].

Once all 36 features of a given synthetic or real image were measured and grouped into a single vector **v**, the Mahalanobis distance *d* was given by:

$$d = \left[(\mathbf{v} - \boldsymbol{\mu})^T \mathbf{K}^{-1} (\mathbf{v} - \boldsymbol{\mu}) \right]^{1/2} , \tag{4}$$

where $\boldsymbol{\mu}$ represents the mean vector over the real images and **K** is the covariance matrix:

$$\boldsymbol{\mu} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{v_i} \tag{5}$$

$$\mathbf{K} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{v_i} - \boldsymbol{\mu})^T (\mathbf{v_i} - \boldsymbol{\mu}) , \tag{6}$$

with n = 1000 being the size of the reference database.

The chromosomes in our genetic algorithm implementation were sets of 8- or 10-dimensional real vectors representing CLB parameters values. The genetic algorithm used the average Mahalanobis distance *d* computed over m = 10 successive CLB realizations as the fitness function for evaluating the chromosomes, and was designed to minimize it. This

#91988 - $15.00 USD Received 23 Jan 2008; revised 10 Apr 2008; accepted 16 Apr 2008; published 12 May 2008

(C) 2008 OSA

averaging was done in order to avoid erroneous evaluation caused by the random nature of the CLB algorithm. Preliminary trials with smaller values of m had indeed been unsuccessful, because the fitness function was too unstable for accurately evaluating the chromosomes. Since the feature distributions are rather compact around their average values for a given set of CLB parameters, the choice of m=10 was a good trade-off between computational cost and fitness function stability. Rank-weighted selection of the parents, and elitist strategy were employed for the reproduction operators.

Crossover of two chromosomes $\mathbf{c}^1$ and $\mathbf{c}^2$ consisted in averaging half of the genes, keeping the others unchanged. The genes to be averaged were chosen randomly with equal probabilities. The crossover between $\mathbf{c}^1$ and $\mathbf{c}^2$ occurred with probability p_c, leaving both genes unchanged otherwise. The best chromosome remained unchanged from one generation to the next, which is the definition of elitist strategy. After the crossover processes, all but the elite chromosome underwent individual gene mutation with probability p_m, monotonically decreasing during the evolution [33].

For each gene G, evolution was restricted to an interval $[G_{min}, G_{max}]$, starting from random values between these bounds. The latter were deduced from the original CLB model as: $\{G_{\min}, G_{\max}\} = \{.8G_{Opex99}, 1.2G_{Opex99}\}$. These figures come from the assumption that the original model [21], referred as *Opex99* in this text, could be used as a starting point for the optimization process.

Preliminary optimizations had indeed shown that this restriction of the search space ensured that the Wiener spectrum of the synthetic images remained close to the one of the original CLB model, which had been designed in order to match the spectrum of real mammograms. All parameters of the genetic algorithm and their meaning are given in Table 2. Four variations of the CLB model were successively optimized: 1-layer classical CLB with isotropic orientation of the clusters (referred further in text as *simpiso* type), 2-layer CLB with isotropic orientation of the clusters (*doubiso*), 2-layer CLB with favored orientation of the clusters (*doubori*), and 1-layer CLB with favored orientation of the clusters (*simpori*).

Table 2. Genetic algorithm parameters used for optimizing CLB variables.

Parameter	Meaning	Value
L	Number of genes in a chromosome	8 or 10
S	Size of the chromosomes population	51
m	Number of realizations of a chromosome for evaluating its fitness function	10
p_c	Crossover probability	0.8

2.3 Evaluation of the visual realism of the synthetic images

The role of the genetic algorithm was to ensure that the synthetic CLB images would have statistical properties similar to real images. Although this point was necessary for future model observer experiments for example, it was certainly not a sufficient condition for using them in psychophysical detection experiments. Human perception is highly dependent on properties of the background as well as those of the neural processing and coding of visual information. Thus, similar statistical properties for a pair of images does not necessarily imply their visual resemblance to human observers. To evaluate the visual realism of the four optimized CLB types and compare it to the original CLB, a study was conducted with three radiologists and two radiographers.

The three main structures types that are likely to be found in real mammograms were evaluated: glandular areas, fatty areas, and fibers [34]. The observers were first presented a series of 20 real images representative of each structure type. The selection of these reference images was based on the choices of one of the radiographers, and then confirmed by the opinion of a radiologist. The presentation of the reference images also allowed the radiologists

#91988 - $15.00 USD Received 23 Jan 2008; revised 10 Apr 2008; accepted 16 Apr 2008; published 12 May 2008

to get acquainted to the display screen, light conditions, and definitions used for the three structure types. After this training phase, 50 realizations of each CLB model variation were presented in random order. The four variations developed with the GA, and the original CLB [21] were displayed in 10 blocks of 25 images. The order of presentation for each CLB type was randomized within each block.

For each image, the observers were asked to tell whether or not they observed a given structure (glandular areas, fatty areas, fibers). For each affirmative answer, they were asked to grade the realism of the structure, based on a 10-grade scale evaluation. In order to ensure a consistent inter-observer use of the scale, the observers were clearly informed before the rating experiments that they should use grades 7 to 10 for images that could be expected to be observed on real mammograms, and grades 1 to 6 for insufficiently realistic images. In the latter case, the observers were given the possibility to further evaluate which features looked unrealistic by using one or more checkboxes representing possible defaults: too disorganized, too rectilinear, too much contrast, too fuzzy, or appearance of 3D-like artifacts. Additionally, the radiologists were asked to mention if some structure resembled a tumor (mass). This latter question was aimed at determining whether unwanted pathological (tumor-like) patterns arose from the CLB superimposition algorithm.

The 12-bits CLB images were converted to 256 gray levels before being displayed on a laptop screen. Their mean gray level value and standard deviation were adjusted to 110 and 35 respectively, in order to obtain images lying in the central dynamic range region of the display. The observers had the possibility to adjust the display brightness and contrast by observing a mammography test pattern at the beginning of the experiment. The laptop display was a practical choice, since the visualization experiments were to be conducted in several dark rooms. For the proposed task, all radiologists and radiographers unanimously reported adequate conditions to confidently assess the realism of the three structure types, since they were to be compared to real digital mammograms ROIs displayed on the same screen at the beginning of the test, and since no detection and/or classification tasks had to be conducted for this study. The 256 by 256 pixels synthetic images display size was 9 by 9 cm. According to preliminary discussions with the radiologists, the size of the image structures at this scale corresponded to the typical scale obtained when zooming on a digital mammography display unit.

3. Results

3.1 CLB parameters optimizations

Although genetic algorithms with elitist strategy usually have the property to be monotonically converging towards extrema of the fitness function, the example fitness function history on Fig. 2 shows that it decreased relatively regularly during 20-30 generations, and then had a more chaotic behavior. This was observed for all model variations, and can be explained by the random nature of the m realizations per chromosome that were computed for evaluating its fitness function. The same CLB parameters lead to images with similar overall statistical properties, but the 36 features we used in this study allowed for evaluating their variations much more precisely. The fitness function of a given chromosome could thus vary from a generation to another, and the best chromosome of generation T’ could be rejected to a higher rank at T’+1, even by chromosomes that had worse performance at generation T’. The upper series in Fig. 2 shows that the median fitness function of the population was less sensitive to this phenomenon. The evolution process was conducted during 100 generations for each of the variation of the CLB model, and the best chromosome of the evolution history was selected for computing the fitness function averages presented on Fig. 3, on the basis of 200 realizations per model.

#91988 - $15.00 USD Received 23 Jan 2008; revised 10 Apr 2008; accepted 16 Apr 2008; published 12 May 2008

(C) 2008 OSA

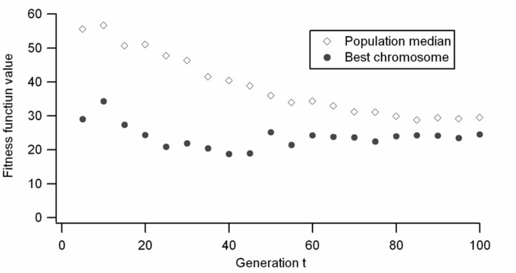


Fig. 2. Example of fitness function history. The upper series represents the median value of the fitness function evaluated on the population at generation t, and the lower series indicates the value for the best chromosome.

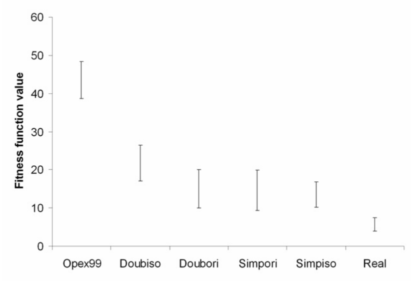


Fig. 3. Fitness function computed from 200 realizations with the optimized set of CLB parameters for all model variations. The error bars represent the standard deviation of the realizations' fitness function. The fitness function averaged over 200 real images is shown for comparison.

Figure 3 shows that the gain obtained by tuning the CLB parameters with the genetic algorithm is at least a factor of 2 for average Mahalanobis distance, compared to the original values (*Opex99* [21] series), depending on the model used. ANOVA analysis and Tukey HSD Test were performed in order to compare the fitness function values for all series. The results (F = 615.7, $p < 0.001$, HSD[.01] = 2.32) indicate that the difference in Mahalanobis distance between *Opex99* and all others series is statistically significant ($p < .01$). The difference between *doubiso* series and the three other optimized models is also significant ($p < .01$). Finally, even after the optimization, a significant difference between real images and each synthetic series remained ($p < .01$).

Figure 4 presents typical examples of images created with the different CLB parameters. The real mammogram ROI was selected from a medium-density breast. The optimized CLB parameters for generating these 256 by 256 pixels images are detailed in Appendix A. Typical computation time needed for computing the 200 realizations and their associated Mahalanobis distance was 40 minutes, which represents 12 seconds/realization.

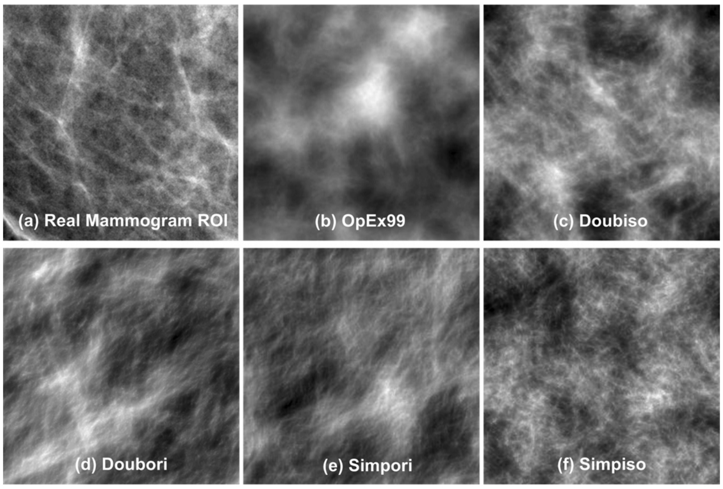


Fig. 4. Examples of realizations for the different types of CLB variations. (a) ROI selected from a real mammograms; (b) 1-layer CLB, Opex99 [21] parameters (referred in text as *Opex99*); (c) 2-layer CLB, isotropic orientation of the clusters (*doubiso*); (d) 2-layer CLB, favored orientation of the clusters (*doubori*); (e) 1-layer CLB, favored orientation of the clusters (*simpori*); (f) 1-layer CLB, optimized version of (b) (*simpiso*).

#91988 - $15.00 USD Received 23 Jan 2008; revised 10 Apr 2008; accepted 16 Apr 2008; published 12 May 2008

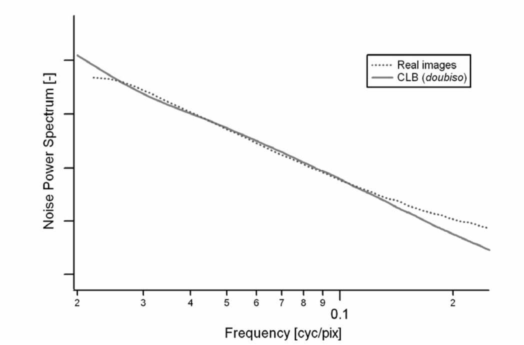


Fig. 5. Comparison of the real images and optimized CLB Wiener spectra. Pixel size is 0.1 mm. Only one series of synthetic images (*doubiso*) is shown. Other series have very similar spectra.

The Wiener spectra of real and synthetic mammograms is shown on Fig. 5. The spectra have a power-law form $W(f) = K / f^b$, where *f* is the radial frequency [6,35]. The exponent values are b=3.02 ± 0.02 for the real images, and b=2.92 ± 0.01 for the CLB (mean ± standard error).

3.2 Evaluating the realism of synthetic textures

Figure 6 summarizes the results for visual realism evaluation experiments performed by the radiologists (KK, ES, NH) and the radiographers (FD, PS). About 2% of the grades were classified as outliers according to Chauvenet criterion [36]. The corresponding data were removed before the statistical analysis presented in Table 3. The rejected outliers did not change any of the values of the 10, 25, 50, 75, and 90th percentiles shown on Fig. 6, where the box plots summarize all marks given by the five observers to each series of images.

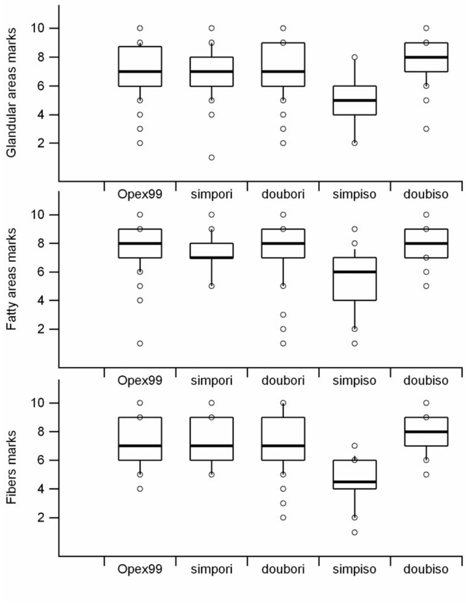


Fig. 6. Realism marks given by the observers (radiologists and radiographers) for the glandular areas, fatty areas, and fibers. The boxes represent the 25th, 50th, and 75th percentiles, and the whiskers the limits for the 10th and 90th percentiles.

#91988 - $15.00 USD Received 23 Jan 2008; revised 10 Apr 2008; accepted 16 Apr 2008; published 12 May 2008

Table 3. Visual realism evaluation by the five observers for glandular (GL) and fatty (FA) areas, and fibers (FI). Bold values indicate statistically significantly realistic evaluations (one sided Student t-test, α=5%, β=0.8, H_0: μ=6.5). Italic values correspond to the mean and standard error.

Obs.	Opex99			Doubiso			Doubori			Simpori			Simpiso		
	GL	**FA**	**FI**	**GL**	**FA**	**FI**	**GL**	**FA**	**FI**	**GL**	**FA**	**FI**	**GL**	**FA**	**FI**
KK	6.80	**7.18**	**7.85**	**7.18**	**7.96**	**7.94**	**7.40**	**7.96**	**8.78**	**7.22**	**7.76**	**7.92**	5.00	6.47	4.93
ES	6.65	**6.87**	**6.87**	**7.68**	**7.70**	**7.73**	6.68	**6.81**	**6.81**	**6.98**	**7.05**	**7.11**			
NH	**7.68**	**7.68**	**7.68**	**9.00**	**9.00**	**9.00**	**7.68**	**7.68**	**7.68**	**6.96**	**6.96**	**6.96**		(b)	
FD	**7.47**	**8.16**	**7.00**	**7.68**	**8.04**	**6.87**	**7.54**	**7.93**	**7.03**	**7.44**	**7.71**	**7.08**			
PS	**7.65**	**7.77**	(a)	**8.40**	**8.14**	**7.58**	**7.23**	**7.15**	6.18	**7.55**	**7.46**	7.67	4.67	4.95	4.09
Mean	*7.3*	*7.5*	*7.4*	*8*	*8.2*	*7.8*	*7.3*	*7.5*	*7.3*	*7.2*	*7.4*	*7.3*	*4.8*	*5.7*	*4.5*
	± 0.2	*± 0.2*	*± 0.3*	*± 0.3*	*± 0.2*	*± 0.4*	*±0.2*	*± 0.2*	*± 0.4*	*± 0.1*	*± 0.2*	*± 0.2*	*± 0.2*	*± 0.8*	*± 0.4*

For each synthetic image, the radiologists evaluated the realism of the three structures types (glandular areas, fatty areas, and fibers), whereas the two radiographers chose not to give their opinion in some cases, when they judged that a given structure covered a too small part of the ROI to be evaluated. This mainly happened for the evaluation of fibers, which are less visible in *Opex99* and *simpiso* series. This latter series (*simpiso*) was only evaluated by the first two observers which took part in the study. T-tests of the first two observers' data (α=5% and β=.8) showed that the *simpiso* series was significantly lower from the apriori selected threshold for realism (6.5) and thus statistically significantly visually unrealistic. It may be useful to repeat here that the observers were asked before the experiments to use a 10-grade scale with a threshold separating sufficiently realistic (grades 7 to 10) from insufficiently realistic (grades 1-6) images.

Bold values in Table 3 indicate that nearly all structures were considered significantly realistic at 5% confidence level (above the a priori selected threshold of grade 6.5) for the first four CLB models. After the evaluation of the first two observers, it was decided to further discard the *simpiso* series, since the grades given by these observers indicated that these images lack visual realism, compared to the other series.

When compared to each other, *Opex99*, *doubori* and *simpori* series obtain comparable overall performance for all structures types, while *doubiso* series outperforms them for all structures types when the results are averaged over the 5 observer. 1-way ANOVA analysis and Tukey HSD Test performed among these four series indicated that the average grade for *doubiso* images is significantly higher than for all other series for the glandulary areas ($p < .01$ in each case), fatty areas ($p < .01$), and fibers ($p < .05$).

Additionally, a two-way ANOVA was conducted in order to estimate separately the influence of the two model variations: the addition of the second layer with large blobs on one hand, and the preferred orientation of the blobs on the other. The grades given by the five observers were pooled together, and separate analyses were carried out for glandular areas, fatty areas, and fibers. The visual realism was significantly better for the images containing two layers (*doubiso, doubori*) than for images created with one layer only (*simpiso, simpori*), for all structure types ($p < .01$ in all cases). Images with an isotropic distribution of blob orientation (*simpiso, doubiso*) were judged significantly ($p<.01$) more realistic than images with preferred blob orientations (*simpori, doubori*) for the glandular areas. However, this effect was not statistically significant for fatty areas ($p = .08$) and fibers ($p = .27$).

4. Discussion

Although the implementation of the genetic algorithm became complex due to the inherent random nature of CLB model, the optimization produced images which statistical properties were significantly closer to real mammographic images than the original CLB. As the chromosomes' evolution continued for more than 50 generations after the optimal parameters presented in Table 4 had been found, these values can thus be confidently considered as optimal for the developed model variations. It is difficult to intuitively interpret the absolute Mahalanobis distances in Fig. 3, since several statistical parameters, among the 36 used for defining that metric, are correlated in a complex way. However, a benchmark can be given by the distance computed for the 200 real mammograms ROIs, which is equal to 5.7 ± 1.8 (mean ± SD, see Fig. 3). This indicates that from the statistical point of view, the synthetic images obtained by the models tuned by the genetic algorithm are much closer to real images than the original *Opex99* series, but also that they cannot be considered indistinguishable from real images. Allowing enlarged bounds for G_{min} and G_{max} would lead to optimized chromosomes with better fitness function, but preliminary tests had shown that when given more freedom, the blobs dimensions evolved to points as small as $G_{min,Lx}$ by $G_{min,Ly}$, lowering the average Mahalanobis distance down to about 10-15 depending on the model, but losing all visual realism. This emphasized the need for a realism evaluation conducted not only for the objective, statistical point of view through the Mahalanobis distance metric, but also for the subjective, visual aspect of the optimized model.

Concerning the model variations and their effect on the visual evaluation by the radiologists and radiographers, the favored orientation of the structures in *simpori* and *doubori* series was generally recognized as such by the observers, and their main drawback was that in some cases this orientation was too obvious and artificial, giving them the feeling of seeing three-dimensional structures instead of flat projections. This defect was particularly mentioned in *simpori* series, while the few large scale structures of *doubori* seemed to hide or mask the main layer composed of the smaller blobs. On the other hand, the observers found that some of the isotropic images were too disorganized to correctly represent real mammograms. This was the main reason for discarding the display of *simpiso* series for the last three observers in the psychophysical study. The presence of the second layer CLB in *doubori* did not improve or deteriorate significantly the visual aspect of *simpori* series, but the difference was clearly shown by the observers' evaluations for the isotropic series: they reported unorganized images with too much contrast for the 1-layer series, and selected the 2-layer *doubiso* images as best overall series. The only limitation mentioned by the radiologists for that series was that for some images (about 10% of the set), bright points caused by blobs superimposition might be interpreted as clusters of microcalcifications. However, for visual experiments of mass detection, they confirmed that this downside would not be critical, since they are not affected by the presence of mm-scale microcalcifications when looking for cm-scale structures like masses.

A remaining question is the visual variability of the synthetic textures. As for most of other models, it is much smaller than that of real images. This has been partly solved by converting the CLB output float images to 12-bit images using randomly chosen values of mean gray level and standard deviation following real image corresponding distributions. One could have imagined using "floating" values for CLB parameters as well, but this possibility was not applied in our study. Another possibility is to separately use genetic optimization to fit the CLB parameters to ROIs from mammograms corresponding to individual patients or groups of patients. Since the real images have intrinsic variability sources like breast dimension and composition or quantum noise that are difficult to fully reproduce with CLB model, experienced radiologists would probably be able to distinguish between real and synthetic backgrounds. In order to focus the observers' rating task on the different optimized CLB models, real mammograms ROIs were used as reference images only in the experiments, and the observers were not asked to rate other real images. However, we are confident that the optimized CLB provide excellent candidates for designing and conducting realistic mass detection tasks and that the results can be generalized to clinically relevant tasks, since recent findings [16] with these backgrounds suggest that human observers use similar strategies with both background types.

5. Conclusion

Using a genetic algorithm and variations of the original CLB model, we were able to synthesize images which resulted in significantly closer visual and statistical properties to real images than those arising from the original CLB model. These models and parameters allow for generating an arbitrary number of such images while improving their realism. The synthetic images may find direct applications in detection experiments involving human or model observers since the visual and statistical characteristics have both been deemed by the current study to be similar to that of real images. In particular, the *doubiso* series were deemed to have visual characteristics very close to real images, even if their statistical properties are more distant from real images than for the other model variations *simpori* and *doubori*.

Compared to other image synthesis techniques, our technique is limited to the generation of square ROIs. However, it has the advantage of being able to quickly generate a large number of images, with traceable statistical properties, and visually representing all major structures types (glandular areas, fatty areas, fibers) that are visible on real mammograms. An interesting application of the technique described in this work would be to generate separate optimizations of the CLB model for different breast density classes. However, the methodology presented in this study is not limited to mammography and may be easily generalized to other medical or non-medical images. The only need is a sufficiently large database of reference textures for defining the Mahalanobis distance used as fitness function by the genetic algorithm for tuning the CLB parameters. Further work may also focus on other blob functional forms than the exponential blobs used in this study, and their influence on visual and statistical properties of the synthetic images.

Appendix A: Optimal CLB parameters for each model variation

The CLB parameters mentioned used for generating the ROIs of Fig. 4 are given in Table 4.

Table 4. Optimized CLB parameters for the various CLB models.

Series	α	β	L_x	L_y	σ_x	σ_y	K_0	N_0	K_0'	N_0'
Opex99	2.1	0.5	5	2	12	12	600	20	N/A	N/A
Doubiso	2.31	0.57	4.09	1.76	13.27	13.92	643.81	20.21	61.47	5.60
Doubori	2.49	0.53	4.14	1.61	11.17	14.29	709.96	20.37	78.00	5.01
Simpori	2.51	0.54	4.53	1.66	10.67	13.33	714.44	23.14	N/A	N/A
Simpiso	2.47	0.59	4.19	1.63	11.02	12.27	674.97	16.53	N/A	N/A

Acknowledgments

The authors are grateful to Elsabe Scott, MD, Nigel Howart, MD, Christel Elandoy, and Philippe Spring for their participation in the psychophysical experiments. This work was supported by Swiss National Science Foundation under Grant No.320000-113863/1.

Human linear template with mammographic backgrounds estimated with a genetic algorithm

Cyril Castella,[1] Craig K. Abbey,[2] Miguel P. Eckstein,[2] Francis R. Verdun,[1] Karen Kinkel,[3] and François O. Bochud[1,*]

[1]*University Institute for Radiation Physics, Centre Hospitalier Universitaire Vaudois (CHUV) and University of Lausanne, Grand-Pré 1, CH–1007 Lausanne, Switzerland*
[2]*Department of Psychology, University of California, Santa Barbara, California 93106-9660, USA*
[3]*Clinique des Grangettes, Chemin des Grangettes 7, CH–1224 Chêne-Bougeries, Switzerland*
**Corresponding author: francois.bochud@chuv.ch*

Received May 1, 2007; accepted June 29, 2007;
posted July 17, 2007 (Doc. ID 82634); published September 21, 2007

We estimated human observer linear templates underlying the detection of a realistic, spherical mass signal with mammographic backgrounds. Five trained naïve observers participated in two-alternative forced-choice (2-AFC) detection experiments with the signal superimposed on synthetic, clustered lumpy backgrounds (CLBs) in one condition and on nonstationary real mammographic backgrounds in another. Human observer linear templates were estimated using a genetic algorithm. A variety of common model observer templates were computed, and their shapes and associated performances were compared with those of the human observer. The estimated linear templates are not significantly different for stationary CLBs and real mammographic backgrounds. The estimated performance of the linear template compared with that of the human observers is within 5% in terms of percent correct (Pc) for the 2-AFC task. Channelized Hotelling models can fit human performance, but the templates differ considerably from the human linear template. Due to different local statistics, detection efficiency is significantly higher on nonstationary real backgrounds than on globally stationary synthetic CLBs. This finding emphasizes that nonstationary backgrounds need to be described by their local statistics. © 2007 Optical Society of America

OCIS codes: 100.2000, 110.3000, 170.3830, 330.1880, 330.4060, 330.5510.

1. INTRODUCTION

As long as medical diagnosis decisions are based on visual inspection of medical images, it will be necessary to better understand human decision-making strategies. If we want to extract as much information as possible from an image, it has to be processed and displayed in such a way that the human observer can most efficiently read it. In other words, imaging systems need to be adjusted and the image quality assessed and optimized for human visual and decision capabilities [1]. There are basically two ways of conducting a simple detection task: either by hiring human observers or by using mathematical models that attempt to mimic human performance. This latter method has the advantage of reducing the time and the cost of the optimization process but requires knowing how the human observer actually processes the image and how the image backgrounds influence the detection task.

Human performance in radiology detection tasks has been studied within numerous model frameworks in previous decades. In most of the studies, human observer characteristics were assessed through psychophysical experiments reproducing the clinical task. Such experiments were conducted in the context of tumor detection on computer tomographic images of the liver [2,3], stenosis in a blood vessel on fluoroscopic images [4], nodules on pulmonary radiographs [5], or microcalcifications and tumors on mammograms [6,7]. Human performance was then compared with that of hypothetical linear models derived from the theory of signal detectability [8]. Each of the models has an underlying template that is assumed to mediate human visual detection.

An alternative approach to studying human observer strategy is to attempt to directly estimate the linear-template model from the observers' trial-to-trial decisions and the images presented. This method, known as classification images, was originally developed by Ahumada for audition and generalized by Abbey *et al.* to a variety of visual tasks, including two-alternative forced-choice (2-AFC), correlated Gaussian noise processes and real medical image backgrounds [9–12]. The method gives an estimate of the linear template used by a human observer by analyzing the results of 2-AFC experiments. Aside from assuming the linearity of the system, it does not make any *a priori* assumption on the way the observer processes the background and is influenced by its statistical characteristics. The direct estimation of the underlying observer template might lead to better prediction of human performance than the use of a hypothetical model observer's template. Thus, the first goal of the present paper is to estimate human observer templates for the detection of simulated masses in real x-ray mammogram backgrounds. Use of real mammographic images rather than images with Gaussian noise processes precludes the use of the standard linear-weighted sum of images to estimate the templates [13–15] and thus demands developing an iterative method to find the best estimate of the human template. Here we propose a method using genetic algorithms (GAs) to find the template that maximizes the

1084-7529/07/1200B1-12/$15.00 © 2007 Optical Society of America

likelihood of observing the trial-to-trial human decisions. In addition, we assess the ability of the estimated template for each individual observer in predicting human performance.

One challenge when studying human detection performance and/or evaluating and optimizing image quality is that it require a few hundreds if not thousands of images. Because access to a sufficient number of real clinical images might become difficult, there has been a strong motivation to use synthetic backgrounds that are realistic looking. For that reason, and also because it is useful for controlling image statistics, a large part of theoretical and experimental studies have been conducted using synthetic backgrounds. The first studies consisted in simple objects superimposed on white noise [16,17]. Although being particularly convenient for theoretical considerations, white noise is actually a too simplified view of noise encountered in medical imaging, and this framework was then extended to more realistic noise models such as correlated noise applied to radiology [18] or nuclear medicine [19]. Researchers have since been focused on images containing anatomical (or pseudoanatomical in the case of computer-synthesized images) variations. Burgess and colleagues [20,21] showed that real or realistic backgrounds (in the sense of visual and statistical realism) are necessary for evaluating observers' performance for detection tasks and that noise-limited experiments with simple phantoms are too limited for the mass detection task applied to mammography. Lumpy backgrounds [22] and clustered lumpy backgrounds [23] (CLBs) were developed in order to better reproduce anatomical variations. A new version of CLBs [24] was specifically developed for mammographic textures. Yet aside from the visual similarity and commonality of global image statistics between the CLBs and the real mammographic backgrounds, it is still unknown whether human observer strategies are similar across both types of backgrounds. One important difference is that CLBs are statistically stationary, while real mammographic backgrounds are not. A recent study by Zhang *et al.* [25] has shown that human observers can adapt their detection mechanisms to the local statistical properties of spatially oriented backgrounds. Thus, the second goal of this paper is to investigate whether human visual detection strategies are similar across CLBs and real mammographic backgrounds by estimating the underlying templates of human observers for both backgrounds.

2. MATERIALS AND METHODS

A. Human Linear Observer-Template Model

The linear observer-template model, further referred in this text as human linear template (HLT), has already been described previously [1,9–11], and it will only briefly be reviewed here. In this model, human observers are assumed to perform 2-AFC tasks by formulating a linear internal response λ to each image as

$$\lambda = \mathbf{w}^t \mathbf{g} + \epsilon, \tag{1}$$

where $\mathbf{g}$ is the image shown to the observer, $\mathbf{w}$ is the observer template (both described as vectors), and ϵ is the observer's internal noise. A binary variable o_i can be defined, which represents the outcome of the ith trial:

$$o_i = step(\lambda^+ - \lambda^-) = step(\mathbf{w}^t \boldsymbol{\Delta}\mathbf{g}_i + \Delta\epsilon_i). \tag{2}$$

In this equation, the variables related to the signal-present image are denoted by the $^+$ superscript, the signal-absent by a $^-$, and $\boldsymbol{\Delta}\mathbf{g}_i$ is the difference between $\mathbf{g}_i^+$ and $\mathbf{g}_i^-$. The trial outcome o_i is equal to 1 if the observer chose the signal-present image during the ith trial, and 0 otherwise. If $\Delta\epsilon$ is assumed to follow a Gaussian distribution with zero mean and a variance of $2\sigma^2$, then the probability that $o_i=1$ is given by

$$p(o_i = 1) \equiv p_i = \Phi\left(\frac{\mathbf{w}^t \boldsymbol{\Delta}\mathbf{g}_i}{\sqrt{2}\sigma}\right), \tag{3}$$

where Φ is the Gaussian cumulative density function. Since this probability is invariant to a common scaling of $\mathbf{w}$ and ϵ, they can be scaled so that the magnitude of the internal noise is fixed to a value of $\sigma=1$, which yields

$$p_i = \Phi\left(\frac{\mathbf{w}^t \boldsymbol{\Delta}\mathbf{g}_i}{\sqrt{2}}\right). \tag{4}$$

The definitions above lead to a conditional Bernouilli probability distribution for o_i given by

$$\Pr(o_i|\boldsymbol{\Delta}\mathbf{g}_i, \mathbf{w}) = p_i^{o_i}(1 - p_i)^{1-o_i}, \tag{5}$$

where it is understood that p_i is a function of $\mathbf{w}$ and $\boldsymbol{\Delta}\mathbf{g}_i$ as shown in Eq. (4).

Under these assumptions, it is possible to analyze the likelihood of the human observer response of the ith trial as a function of the template $\mathbf{w}$ and the differences between the images $\boldsymbol{\Delta}\mathbf{g}_i$ for that trial. Since the images are independent of the observer template, the ith trial likelihood can be written as

$$L_i(\mathbf{w}) = f_{\text{joint}}(o_i, \boldsymbol{\Delta}\mathbf{g}_i; \mathbf{w}) = \Pr(o_i|\boldsymbol{\Delta}\mathbf{g}_i; \mathbf{w}) f_{\text{marg}}(\boldsymbol{\Delta}\mathbf{g}_i), \tag{6}$$

where $f_{\text{joint}}(o_i, \boldsymbol{\Delta}\mathbf{g}_i; \mathbf{w})$ is the joint distribution of the difference image and the observer response, while $f_{\text{marg}}(\boldsymbol{\Delta}\mathbf{g}_i)$ is the marginal distribution of the difference image alone. Assuming that the images used in the N_T different trials of the entire experiment and the trial scores are independent from trial to trial, the likelihood of the template $\mathbf{w}$ for the entire experiment is given by the product of the individual likelihoods. The log likelihood is then given by

$$l(\mathbf{w}) = \ln\left(\prod_{i=1}^{N_T} L_i(\mathbf{w})\right) = \sum_{i=1}^{N_T} [o_i \ln(p_i(\mathbf{w})) + (1 - o_i)\ln(1 - p_i(\mathbf{w})) + \ln(f_{\text{marg}}(\boldsymbol{\Delta}\mathbf{g}_i))]. \tag{7}$$

The last term of Eq. (7) is independent of $\mathbf{w}$ and is irrelevant for finding extremal values of the log likelihood. Therefore, in order to find the maximum-likelihood (ML) estimate of $\mathbf{w}$, one has to maximize the function [11]

$$Q(\mathbf{w}) = \sum_{i=1}^{N_T} [o_i \ln(p_i(\mathbf{w})) + (1 - o_i)\ln(1 - p_i(\mathbf{w}))]. \tag{8}$$

B. Template Estimation

The optimization of Eq. (8) is not trivial, since the number of unknowns is equal to N^2 for $N\times N$ pixel images. We therefore used the methodology described by Abbey *et al.* [11] in order to reduce the number of parameters: We assumed that the observer template could be represented by a limited set of known linear feature vectors. The hypothesis is that the Fourier decomposition of the observer template can be expressed as a weighted sum of a limited set of frequency band channels. In the spatial domain, the inverse Fourier transform of these channels can also be used as a base, which yields

$$\mathbf{w} = \sum_{j=1}^{N_C} \beta_j \mathbf{t}_j, \tag{9}$$

where N_c is the total number of frequency channels; $\mathbf{t}_i$ is the ith base template, equal to the inverse Fourier transform of the frequency band channel i; and $\boldsymbol{\beta}$ is the weights vector. The $\mathbf{t}_i$'s were computed by performing the fast Fourier transform (FFT) of circularly symmetric images representing the frequency channels and then windowed in the spatial domain with a fourth-order Butterworth filter to reduce ringing. The channel width was 0.0078 cycles/pixel (0.22 cycles/deg), and the radial coordinates of the channels' centers were equally spaced between 0.01 and 0.25 cycles/pixel (0.28 and 6.98 cycles/deg). We used a total of $N_c = 50$ overlapping channels, which reduced the dimensionality of the optimization problem from N^2 to N_c. Using Eq. (9), Eq. (4) can be rewritten as

$$p_i = \Phi\left(\frac{\left(\sum_{j=1}^{N_c} \beta_j \mathbf{t}_j^T\right)\Delta \mathbf{g}_i}{\sqrt{2}}\right) \equiv \Phi\left(\frac{\sum_{j=1}^{N_c} \beta_j X_{ij}}{\sqrt{2}}\right), \tag{10}$$

where X_{ij} is the dot product between the jth base template and the difference image of the ith trial; X_{ij} does not depend on $\mathbf{w}$ and can be used throughout the whole optimization process once computed.

We used a standard GA for the derivation of the ML estimate of the observer template $\mathbf{w}$. GAs are a family of computational models inspired by evolution [26]. They are well suited for complex optimization problems where the search space has a high dimension. The free parameters of a given optimization problem are encoded on a chromosomelike data structure, and selection and recombination operators are applied in order to allow a population of potential solutions to evolve toward the optimal solution of the problem. The initial population is usually chosen randomly in the search space, and the corresponding chromosomes are evaluated through a fitness function. The best chromosomes are given better reproduction and survival opportunities. Then, crossover and mutation operators are applied in order to generate a new population of equal cardinality. These processes of evaluation, crossover, and mutation are repeated until a user-defined (sub)optimal value of the fitness function or number of generations is reached, or when the best chromosome of the population has not been improved for a given number of generations.

In this study, the fitness function was Eq. (8), with p_i computed with Eq. (10). The chromosomes were the channel weight vectors $\boldsymbol{\beta}$, and the genes the N_c individual weights. We constrained the chromosomes to have a Euclidian norm of 1, which is not a restrictive constraint since two weight vectors that differ only by a scaling factor will represent observer templates that yield the same results in a 2-AFC task. The mutation operator was a two-gene swap with a probability of $p_m = 0.3$ per chromosome per generation, while the crossover operator randomly exchanged half of the genes of two chromosomes of generation G with a probability $p_c = 0.8$ to generate two new chromosomes for generation $G+1$. We used elitist strategy, which consists in keeping the chromosome with the best fitness function value at generation G unaffected by mutations and crossovers when defining the population of generation $G+1$. This guarantees that the fitness function is monotonically optimized. We also used rank selection, which ensures that the best chromosomes are given better survival opportunities: At generation G, the chromosomes were ranked according to their fitness function value, and the probability of being chosen for the crossover operation was linearly dependent on the rank. Population size was 21 normalized chromosomes that were initialized with random real numbers following a Gaussian distribution at generation 0. We let the GA evolve during 10,000 generations.

C. ROI, NPW, and NPWE Models

The first group of linear models observers used in this work include the region of interest (ROI), the nonprewhitening matched filter (NPW), and the NPW with an eye filter (NPWE). These models incorporate different degrees of knowledge about the signal but do not make any assumption about the backgrounds. Complete descriptions can be found in [1,27], and they will be only briefly reviewed here.

The ROI model is rather primitive and uses only information about the spatial extent of the signal. The profile of the signal is not taken into account, and the template consists in a uniform activation area integrating the pixel values of a ROI. This model is very limited in the presence of real or realistic backgrounds and signal.

The NPW uses full knowledge of the signal shape: Its template $\mathbf{w}_{\mathrm{NPW}}$ matches exactly the signal profile $\mathbf{s}$. This approach has been shown to be optimal when the backgrounds consist in pure white noise but is suboptimal for correlated noise.

The NPWE is an extension of the NPW model, in which the signal is filtered in the Fourier domain by the human contrast sensitivity function, also known as the eye filter. The eye filter takes into account the different sensitivities of the human visual system at different scales. Its functional form in the frequency domain can be modeled by [28]

$$\mathbf{E}(\rho) = \rho^n \exp(-c\rho^2), \tag{11}$$

where ρ is the radial frequency in cycles/deg, $n = 1.3$, and $c = 0.041$ (values from Burgess [29]). The NPWE template is given by

$$\mathbf{w}_{NPW} = \mathbf{E}^t \mathbf{E} \mathbf{s}, \tag{12}$$

with the matrix $\mathbf{E}$ implementing the effect of the eye filter and defined by Eq. (11).

As with the HLT, a Gaussian-distributed random variable can be added in any of these models to the observer response, as in Eq. (1), ϵ, to account for the observer internal noise. The value of ϵ is usually chosen in order to degrade the model's performance to match human observer performance.

D. Channelized Hotelling Observer

The Hotelling observer [18,22,30,31] uses information about both signal profile and background statistics. It is the ideal observer for images with a multivariate Gaussian distribution and the same correlation structure in signal-absent and signal-present images [32]. For a given signal $\mathbf{s}$, its template $\mathbf{w}_{\mathrm{HOT}}$ is given by

$$\mathbf{w}_{HOT} = \mathbf{K}_b^{-1}\mathbf{s}, \tag{13}$$

where $\mathbf{K}_b$ is the covariance matrix of the background, which describes the variance of each image pixel and the covariance between pairs of pixels; $\mathbf{K}_b$ can be derived directly from the noise power spectrum (NPS) for spatially stationary backgrounds (like some computer-generated textures [33]), but in the general case, for $N \times N$ images, $\mathbf{K}_b$ is an $N^2 \times N^2$ matrix that has to be estimated from the variance–covariance computation. Due to the large size of the resulting covariance matrix, the practical implementation of the Hotelling observer is a problem for real backgrounds. One way to avoid this is to reduce the image to a smaller set of channel response variables [27,34]. The channelized Hotelling template is then given by [35]

$$\mathbf{w}_{CH} = \mathbf{T}(\mathbf{T}^t\mathbf{K}_b\mathbf{T} + \mathbf{K}_\epsilon)^{-1}\mathbf{T}^t\mathbf{s}, \tag{14}$$

where the column vectors of the matrix $\mathbf{T}$ each represent the spatial profile of a channel and $\mathbf{K}_\epsilon$ is the covariance matrix of the internal noise. Internal noise is assumed to be zero mean, independent in each channel, with variance proportional to the variance of the background noise in each channel.

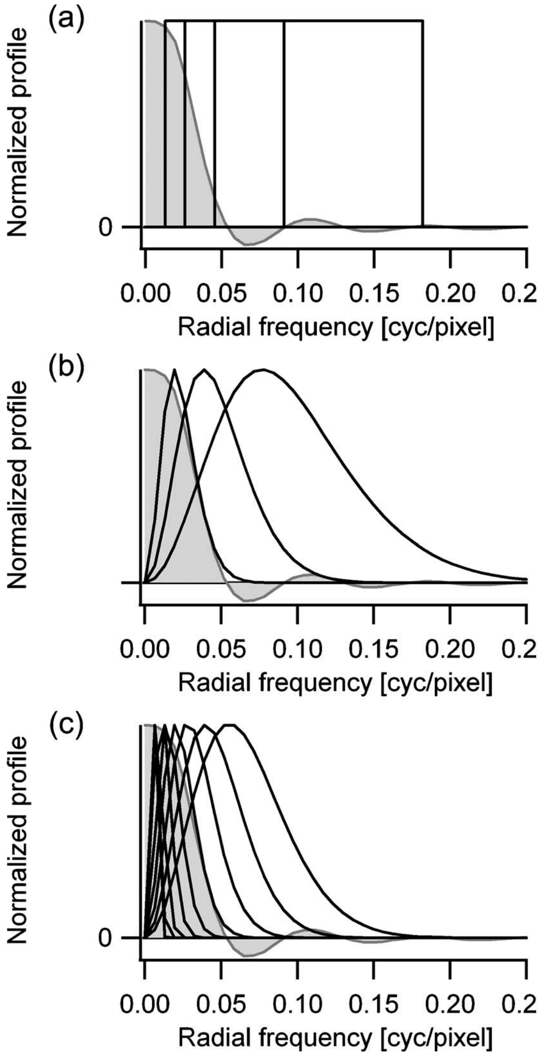


Fig. 1. Channels used for (a) the SQR model, (b) the S-DOG model, and (c) the D-DOG model. The gray area represents the signal frequency profile.

In this work, we used two types of radially symmetric channels: channels with a square bandpass profile in radial frequency [33] [SQR channels, Fig. 1(a)], and overlapping difference-of-Gaussians (DOG) channels [36,37]. The two DOG-channel models, namely sparse [S-DOG channels, Fig. 1(b)] and dense [D-DOG channels, Fig. 1(c)], as well as the SQR-channel model, are described intensively in the work by Abbey and Barrett [35].

We estimated $\mathbf{K}_b$ by assuming stationarity of the backgrounds and using the Wiener–Khinchin theorem to estimate the covariance matrix from the NPS [38]. This assumption is valid for the CLBs, which are stationary by construction, but is only an approximation for real images [13]. The NPS for the CLBs and the real images were computed separately and averaged over 4000 (CLBs) and 1400 (real) images.

E. Images

Two types of backgrounds were used in the 2-AFC experiments: real and synthetic. The real backgrounds were extracted from digital mammograms. We used a database [24] of 88 patients who underwent screening exams on a GE Senograph 2000D full-field digital detector [39,40] with one craniocaudal (CC) and one mediolateral oblique (MLO) processed mammogram per breast to collect square 256×256 pixel ROIs. All ROIs were selected manually in the central breast region in order to avoid any imaging artifact or pathological abnormality. Fatty and dense breasts were represented in the database. We gathered a total of 1400 nonoverlapping 16-bit ROIs having a pixel size of 0.1×0.1 mm.

The synthetic backgrounds were second-generation CLBs [23,24]. These backgrounds consist of the superposition of clusters of elliptical blobs. In order to reproduce both visual and statistical properties of real mammograms, we used CLB images $\mathbf{g}(\mathbf{r})$ that were generated by

$$\begin{aligned}\mathbf{g}(\mathbf{r}) = {} & \sum_{k=1}^{K,\mathrm{small}} \sum_{n=1}^{N_{k,\mathrm{small}}} b_{\mathrm{small}}(\mathbf{r} - \mathbf{r}_{k,\mathrm{small}} \\ & - r_{kn,\mathrm{small}}, \mathbf{R}_{\theta k,\mathrm{small}}) \\ & + \sum_{k=1}^{K,\mathrm{large}} \sum_{n=1}^{N_{k,\mathrm{large}}} b_{\mathrm{large}}(\mathbf{r} - \mathbf{r}_{k,\mathrm{large}} \\ & - \mathbf{r}_{kn,\mathrm{large}}, \mathbf{R}_{\theta k,\mathrm{large}}),\end{aligned} \tag{15}$$

where b_{small} and b_{large} were exponential blob functions of the form

$$b(\mathbf{r}, \mathbf{R}_\theta) = \exp\left(-\alpha \frac{\|\mathbf{R}_\theta \mathbf{r}\|^\beta}{L(\mathbf{R}_\theta \mathbf{r})}\right). \tag{16}$$

Table 1. Definitions and Values of the Variables Used in the CLB Model

Variable	Definition	Distribution/Value
K	Number of clusters	Poisson with mean values $K_{0,\text{small}}=643.81$ and $K_{0,\text{large}}=61.47$
$\mathbf{r}_k$	Position of the kth cluster	Uniform across image
N_k	Number of blobs within the kth cluster	Poisson with mean value $N_{0,\text{small}}=20.21$ and $N_{0,\text{large}}=5.60$
θ_k	Rotation angle of the blobs in the kth cluster	Uniform between 0 and 2π
$\mathbf{R}_\theta$	Rotation matrix of angle θ	N/A
$\mathbf{r}_{kn}$	Position of the nth blob within the kth cluster	Gaussian pdf $\phi(\mathbf{r})$, with standard deviations $\sigma_x=13.27$, $\sigma_y=13.92$ pixels
$b(\mathbf{r},\mathbf{R}_\theta)$	Blob profile rotated at angle θ	N/A
α	Blob shape factor	$\alpha=2.31$
β	Blob shape factor	$\beta=0.57$
L_x,L_y	Characteristic dimensions of the elliptical blobs	$L_x=4.09$, $L_y=1.76$

The parameters in Eqs. (15) and (16) are given in Table 1. Values for α, β, K, N, and L are from [24]. Since the distribution of cluster orientation is uniform between 0 and 2π, CLB images are isotropic. By construction, they are also stationary within their boundaries. The two blob scales (small and large) mimic textures that are found on real mammograms: fatty areas, glandular areas, and fibers.

F. Signal

For these experiments, we wanted a realistic signal that would mimic a medium-sized mass. We chose to extract a 4 mm diameter mass from a high-dose image of the Kodak ITO mammography phantom. The synthetic masses contained in this breast-tissue equivalent phantom are acetate spherical beads having a density of 1.15 g/cm^3. The image of the phantom was acquired on a GE Senograph 2000D (Mo/Rh, 200 mAs), the same as used for the real mammograms in this study.

The mass and the surrounding background were extracted from the phantom image, and the mean pixel value of the background was subtracted. The resulting signal image was then convolved with the radial attenuation profile shown in Fig. 2. The amplitude A of the signal was defined as the mean value of the central 3×3 pixel area of the signal. The signal was added onto existing backgrounds after having been scaled to the desired amplitude. Figure 3 presents the spatial and Fourier space profiles of the mass in the same conditions as they were displayed to the observers (see Subsection 2.G for magnification and resampling specifications). The idea behind that particular signal choice was to use a synthetic mass that would be as realistic as possible, yet nearly circularly symmetric. The advantage of using a phantom image is that it was possible to incorporate in a realistic way the transition between the signal and the surrounding uniform background, therefore avoiding the too-synthetic aspect of an isolated bright spot.

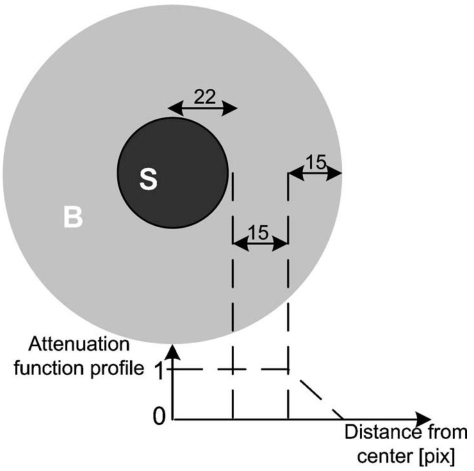


Fig. 2. Dimensions of the original signal (S) and attenuation function used for the surrounding background. Display screen pixel size is 0.25 mm.

G. 2-AFC Psychophysical Experiments Setup

Four naïve observers and one co-author (named Cy in the results) took part in a classical signal-known-exactly (SKE), location-known-exactly (LKE) 2-AFC detection experiment in which the observers were presented two images simultaneously and were asked to indicate the one that contained the signal. For each such trial, the image containing the signal was chosen randomly and the mass was added digitally to the corresponding background before display.

The image pairs were displayed on a Siemens SMM 21140 P high-contrast gray-scale monitor (Siemens, Karlsruhe, Germany), which has a pixel size of 0.25 mm. All backgrounds were scaled so that their mean pixel value was equal to 128 and their standard deviation to 30 gray levels (GL). This way, all images were presented in the middle of the screen dynamic range. In order to reproduce typical clinical settings, the backgrounds were magnified (1.5 magnification factor) before being displayed. The displayed backgrounds were nearest-neighbor resampled 154×154 pixel versions of the original ones. Fiduciary cues were added to both signal-present and signal-absent images in order to assist the observers focusing on the possible signal locations and thus minimize location uncertainty. Examples of displayed images are presented in Fig. 4.

The target contrast resulted in a percentage of correct answers (Pc) ranging from 70% to 85%. After preliminary experiments by one of the authors, the amplitude of the

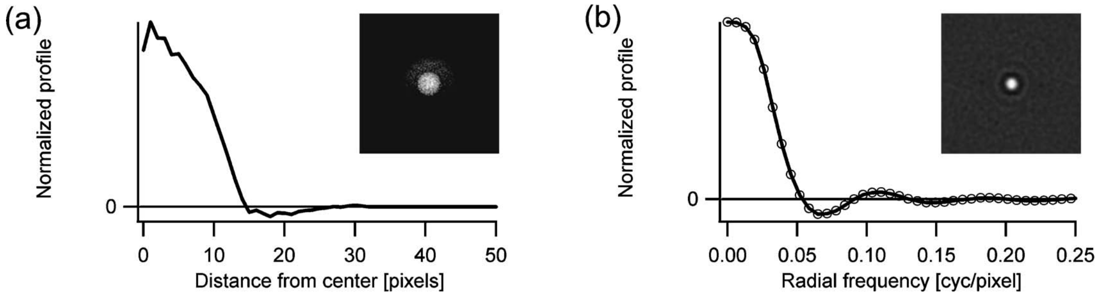


Fig. 3. (a) Normalized profile of the signal used in the psychophysical experiments. (b) Fourier space representation. Display screen pixel size is 0.25 mm.

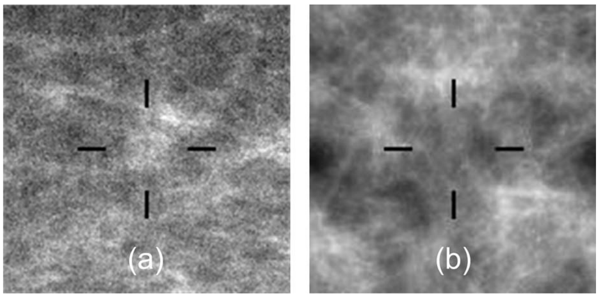


Fig. 4. Examples of (a) signal-present real background and (b) signal-absent CLB. The amplitude of the signal has been increased for printing purposes.

signal was set to $A=15$ GL. The observers were trained before the beginning of each viewing session: They were presented a series of decreasing contrast trials from $A=35$ down to $A=15$, until they had stabilized their Pc between 70% and 85%. The training was performed with images taken from the same database as for the detection experiments, but the pairs that were presented during training sessions were different from those of the detection sessions. All observers were given feedback—correct/incorrect—after each trial, as well as their Pc after each series of 25 trials. Each observer was presented with a different series of randomized image pairs. They were not given any time limit within a trial to reach a decision. Viewing distance was about 40 cm.

The total number of trials was 1400 for the real backgrounds and 4000 for the CLBs. Thus, each observer performed a series of about 15–20 one-hour sessions. The real images and CLB trials were not interleaved for a given observer in order to keep the analysis and the estimation of the corresponding human linear templates as independent as possible. Two observers (Ce and Ga) began with the real image trials, while the other three (Cy, Ro, and Va) began with the CLBs. The image pairs displayed at each trial were randomly chosen and were different for each observer. Two separate image databases (4000 signal present and 4000 signal absent) were used for CLB trials, while each image of the real background database was displayed once as a signal present and once as a signal absent.

In order to estimate the performance of the templates obtained at the end of the optimization process by the GA, we generated a new series of 2-AFC experiments by randomly choosing 1400 real image pairs and 4000 CLB pairs. We estimated the observer templates' Pc on the basis of these new trials by directly computing the dot product with each image of the pair and choosing the highest noiseless response. Note that each individual observer was presented different trials during the detection experiment but that all templates were tested on the same trials. The estimation of the performance of the other models (ROI, NPW, NPWE, SQR, S-DOG, and D-DOG) was done using the same approach, by testing the templates on the simulated 2-AFC.

Finally, the performances of the templates were estimated by a radius of curvature (ROC) curve based on the linear response λ of 1400 real images and 4000 CLBs. We used a Java code from the Johns Hopkins University and translated by J. Eng from the original Fortran program LABROC4 by Charles Metz and colleagues, which is available on the Internet [41].

3. RESULTS AND DISCUSSION

A. Robustness of the Experiments

None of the human observers showed significant bias at the 5% confidence level toward one of the two alternatives (left versus right image) in any of the experiments. The results obtained after the training sessions showed no increasing trend in the performances expressed in term of Pc for each series of 25 trials. The fastest two observers did not improve their decision time, while the other three reduced it by about 50% between the first and the last series.

B. Template Estimation

Figure 5 shows the obtained human observer templates presented as 2D images as well as radially averaged in the spatial and Fourier domains. Only the average of the five different observers is presented, but the individual templates are very similar.

Visual inspection indicates that almost all templates present the same transition at the signal edges: A bright circular zone surrounded by a dark inhibition region. This suggests that human observers concentrate on both the center and the edges of the signal location, rather than just on a homogeneously brighter area. The radial profiles show that the templates are positive within the signal area (below about 15 pixels), then become negative in a surrounding ring of about 5 pixels around the signal location, and terminate with oscillations decreasing rapidly in amplitude. In the frequency domain, the template profiles oscillate around zero, with nodes very close to those of the signal profile. The main differences between template and

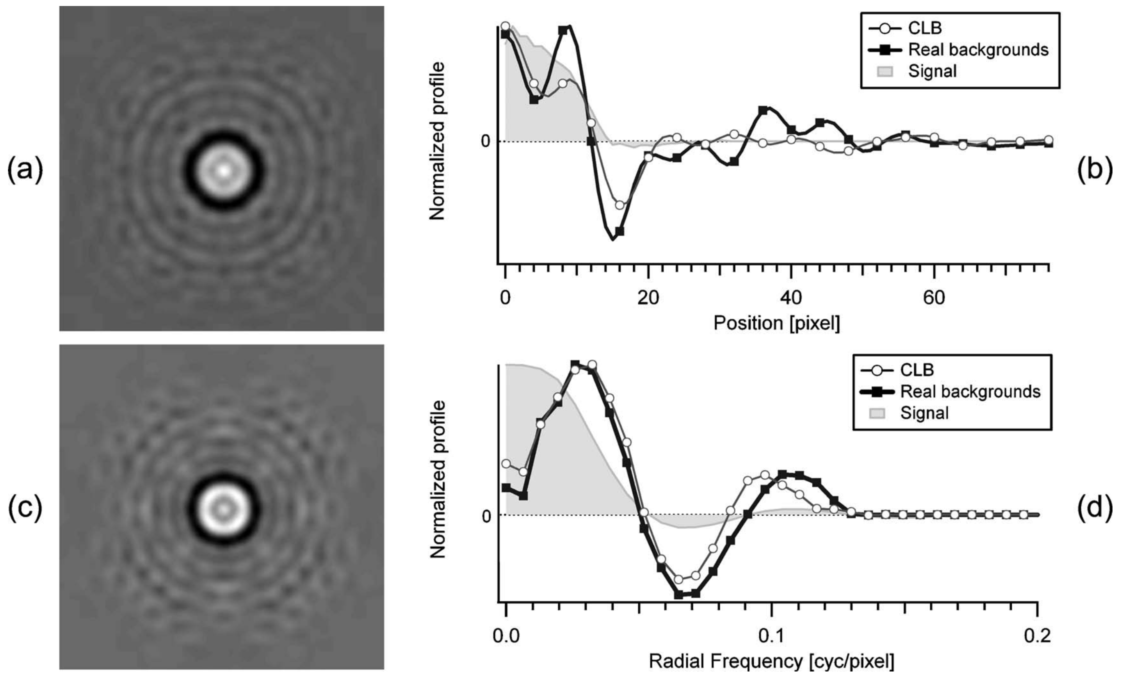


Fig. 5. Human observer template obtained when pooling the five observers' data. (a) CLBs. (b) Real backgrounds. Profiles of both templates in (c) the spatial domain and in (d) the Fourier domain. Display screen pixel size is 0.25 mm.

signal are in the low frequency domain where the template hits a maximum at around 0.03 cycles/pixel (0.84 cycles/deg) and the following oscillations are much more pronounced in the template than the signal, showing an inhibition process.

The metric we used for comparing the radial frequency profiles of the templates estimated for CLB and real images is the root-mean-square difference, weighted by the inverse of the variability of each point (wRMSD [1]). The variability was computed as the square root of the mean interobserver variance. Assuming a standard normal distribution for the wRMSD, the p-value for a t-test under the null hypothesis that the wRMSD is equal to 0 is greater than 0.05 if wRMSD is smaller than 2. The wRMSD computed for frequencies between 0 and 0.25 cycles/pixel (6.98 cycles/deg) is equal to 0.81: Such a low value thus indicates that the profiles are not significantly different, which suggests that human observers apply the same strategy for both types of backgrounds.

C. Observers and Linear Template Performances

Figure 6(a) shows the individual performances of each observer for the two types of backgrounds together with the performance predicted by the corresponding estimated linear observer templates. Individual and average observer performance (Pc) were significantly superior by 6% to 15% for visual detection with the real images than for CLBs ($p < 0.0001$ for all cases). As for Pc, the area under the ROC curve [AUC; Fig. 6(b)] for the task with real backgrounds was significantly larger than for the CLBs (0.84 ± 0.02 versus 0.73 ± 0.02). Mean decision times, averaged over the five human observers, were 2.2 s for the real images and 4.8 s for CLBs, confirming that the task on CLBs is perceived as more complex for the observers.

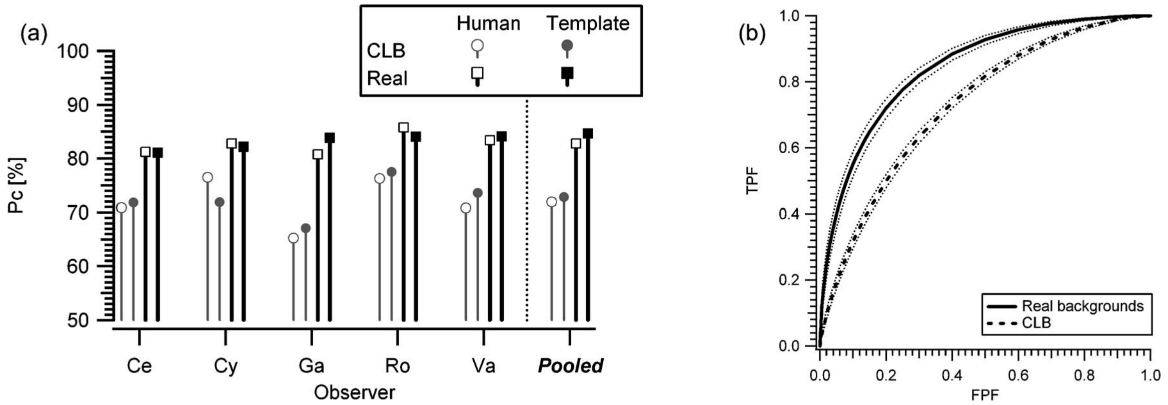


Fig. 6. Results of the 2-AFC experiments performed on the CLBs and real backgrounds. (a) Performance of each human observer and the corresponding linear templates presented in terms of Pc. (b) ROC curve computed with the templates obtained from the pooled data (the dim bands around the curves show the 95% confidence intervals).

It is perhaps not accidental that the smallest performance difference is obtained by observer Cy who, as the only co-author of the five observers, had been seeing and working with CLBs for months. The more extensive experience of Cy with the CLBs could have induced a better knowledge of the structures that were naturally arising from the superposition of blobs and those that were due to the signal.

Although interobserver performances can vary significantly for a given background, performance for a given human observer and the predicted performance from his/her corresponding estimated template are very similar. Differences between the human and the estimated performance range between −4.6% and 3.07% (in units of Pc).

The difference, however, is neither systematic nor random. For the real images, three human observers out of five are outperformed by their linear templates. For the CLB images, only one observer performs better than his/her estimated linear template. This could probably be explained by the fact that the templates are applied without adding internal noise, while this is not the case for human observers.

Given that the CLB images are by construction globally stationary, while the real backgrounds are not, the superior human performance for the nonstationary backgrounds agrees with the findings of Zhang *et al.* [25].

Performance of the estimated template (pooled across all observers) with CLBs has also been applied on real backgrounds (AUC=0.84±0.02) and vice versa (AUC=0.74±0.01). This shows that although the strategy is the same for both types of images, it is significantly more efficient on real backgrounds.

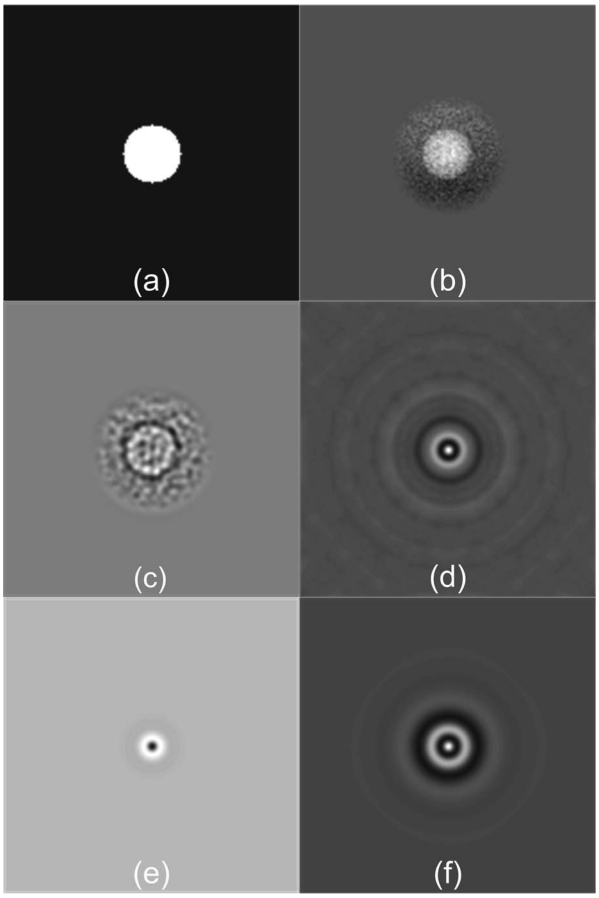


Fig. 7. Model observer templates presented as 2D images. (a) ROI, (b) NPW, (c) NPWE, (d) SQR, (e) S-DOG, (f) D-DOG.

D. Comparison with Other Model Observers

Figure 7 shows the templates computed for the other model observers, displayed as 2D images. Only the templates computed for the real background detection task are shown, since CLB-model observer templates are either identical by definition (ROI, NPW, NPWE) or were observed to be nearly undistinguishable (SQR, S-DOG, D-DOG). As for the HLT, the NPWE template reveals an activation/inhibition transition at around 15 pixels. This transition is also visible on the SQR and D-DOG templates, even though the latter two have another inhibition circular area very close to the center of the signal location. The S-DOG template is the only template to be negative at the very center of the signal location. Its activation area is a crown between 3 and 6 pixels (0.11 to 0.22 deg) around it.

The radial profiles of all templates in the frequency domain are shown on Fig. 8. Only those corresponding to the real image detection task have been displayed, since the ones corresponding to the task on CLBs are very similar. This similarity was expected, since NPWE does not take any property of the background into account and the three channelized observers depend on the noise power spectra, which had been matched between the two image types while generating the optimized CLBs [24]. As for HLT, all templates peak at around 0.04 cycles/pixel (1.12 cycles/deg) before oscillating as the frequency increases. The eye filter in the NPWE model enhances the signal oscillations at higher frequencies and therefore spans the area of interest up to 0.2 cycles/pixel (5.59 cycles/deg). The values of the wRMSD between the HLT and the NPW, NPWE, and channelized Hotelling templates are given in Table 2. If the whole signal frequency range (up to 0.15 cycles/pixel, 4.19 cycles/deg) is considered, the extremely high values of the wRMSD show that all model templates are substantially different from the HLT. However, for frequencies up to 0.1 cycles/pixel (2.79 cycles/deg) only, the template profiles depart less strongly from the HLT. The wRMSD values of all models lie in the same range (3.9–6.4), with the NPWE being the closest to the HLT. The reduced ability of the channelized models to be tuned to the specific signal is probably responsible for the poor agreement with the HLT.

The performance of the model observers in the 2-AFC experiment, defined by the AUC, is given in Fig. 9. As expected, the ROI and the NPW models perform very poorly on both CLB and they real images, and they are below the human observers. The addition of the eye filter in the NPWE, however, increases its performance significantly above the human observers. Noise was introduced in the NPWE according to the scheme presented in Subsection 2.C, but with the same noise level, human performance could be matched only for one background type at a time (see NPWE* in Fig. 9).

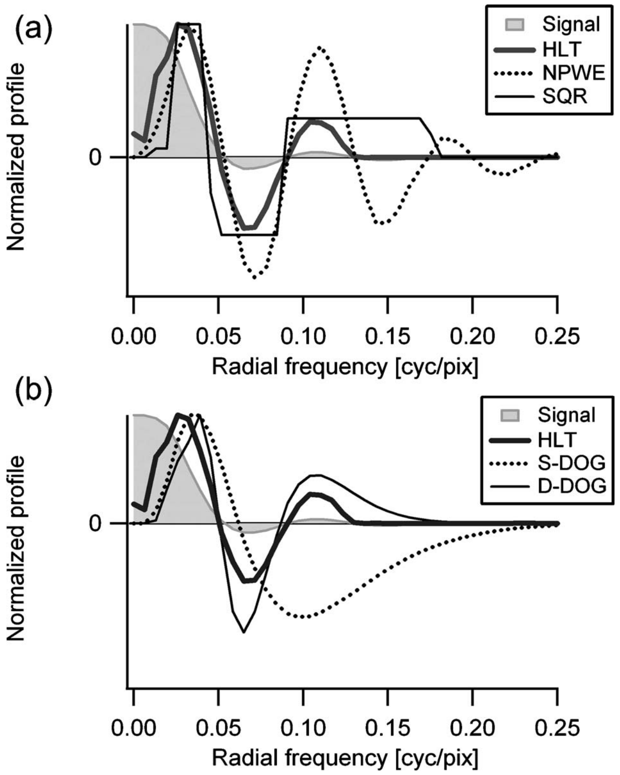


Fig. 8. Radial frequency profiles of the model observer templates estimated for the detection task with real backgrounds. The frequency profiles of the signal and the human linear template are shown for comparison. (a) NPWE and SQR models. (b) S-DOG and D-DOG models. Similar results are obtained with CLB images.

Table 2. Weighted wRMSD between the Frequency Profiles of the HLT Estimated from the Pooled Observer Data for Real Images and the Other Model Observer Templates[a]

f_{max} [cyc/pix]	$\mathrm{HLT_{Real}}$ vs. $\mathrm{HLT_{CLB}}$	$\mathrm{HLT_{Real}}$ vs. NPW	$\mathrm{HLT_{Real}}$ vs. NPWE	$\mathrm{HLT_{Real}}$ vs. SQR	$\mathrm{HLT_{Real}}$ vs. S-DOG	$\mathrm{HLT_{Real}}$ vs. D-DOG
0.05	0.91	7.7	5.0	6.0	6.4	8.8
0.10	0.88	5.6	3.9	4.9	6.2	6.4
0.15	0.84	22.4	536.5	317.2	471.0	126.8

[a] Each line corresponds to the wRMSD computed for frequencies from 0 up to f_{max} only.

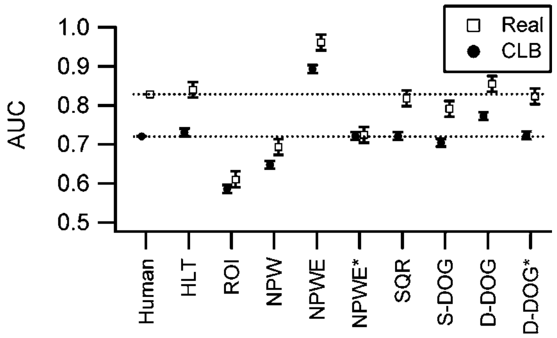


Fig. 9. Performance of the various model observers in the 2-AFC experiment and comparison with human data (dotted lines). The performance is given by the AUC for all model observers and for the percentage of correct responses for the pooled human observers. NPWE* and D-DOG* values correspond to the NPWE and D-DOG with noise added in the model, respectively, in order to match human observer performance on CLBs.

The three channelized Hotelling models have very close performances, with the D-DOG model outperforming the other two by 0.05–0.07 depending on the background type. The D-DOG model is the only one to outperform the human observer. However, by adding internal noise to this model (D-DOG* in Fig. 9; see Subsection 2.D), the performance obtained by all channelized models are very close to those of the pooled human observer on both backgrounds. The fact that the SQR and S-DOG models do not outperform human results could either mean that the assumption of image stationarity for using the Wiener–Kintchin theorem when estimating the covariance matrix was too strong or that there is excessive information loss that occurs when restricting the Hotelling observer to the limited number of channels. However, given that the SQR and S-DOG models are also slightly outperformed by humans with the stationary CLBs, the limited number of channels might be the likely explanation for the inferior performance of the channelized models. Increasing the number of channels and/or testing other channel types (Gabor, Laguerre–Gauss) and identifying the best set of channels for this task may lead to an increased overall performance, but not necessarily to a better matching with human results.

Similarly to human observers, all models perform better on the real backgrounds than on the CLBs, although using the same or very similar templates for both tasks. Again, this suggests that detecting a signal from a nonstationary background is easier than on a CLB, where local properties cannot be used. Minor differences were expected for the channelized Hotelling models using background-specific information contained in the NPS [although having been matched as closely as possible, the NPS are not completely identical; see Fig. 10(a)], but they are more surprising for the ROI, NPW, and NPWE models because these models do not depend explicitly on background properties.

E. Dissociation in Human Performance between Real Mammograms and CLBs

Zhang *et al.* [25] found a dissociation of performance between stationary and nonstationary noise for human observers and for model observers that adjusted their local strategy (template) to the nonstationary noise. In contrast, nonadaptive model observers that used a fixed template for all areas of the nonstationary images (i.e., global prewhitening or NPWE) did not show such performance dissociation. Thus, Zhang *et al.* [25] concluded that the improved human performance with nonstationary images was due to the ability of human observers to adapt their strategy to the local statistics of the nonstationary backgrounds.

The present study also finds a dissociation of human performance across nonstationary (real mammograms) and stationary (CLB) backgrounds but, in contrast to the Zhang *et al.* study, this difference cannot be attributed to an adaptive strategy of human observers. First, the tem-

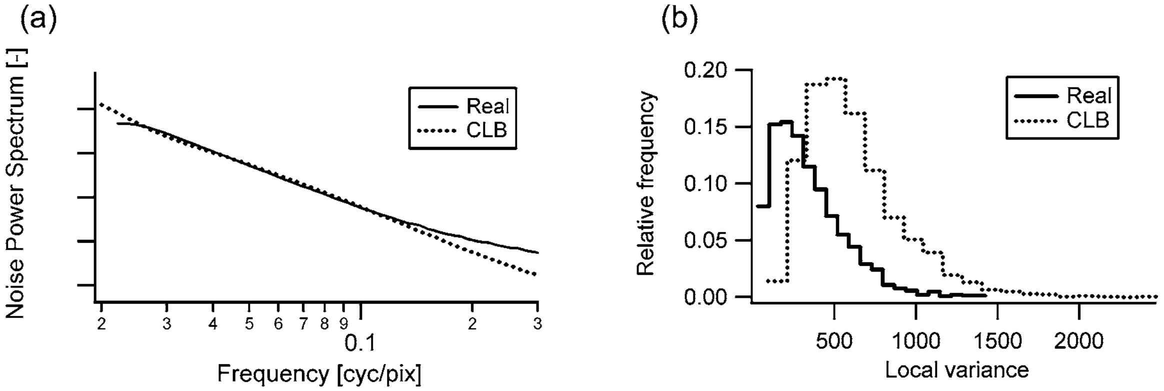


Fig. 10. (a) Noise power spectrum of the real backgrounds and the CLBs. Slope in the linear part is −3.23 for the real images and −3.17 for the CLBs. (b) Distribution of the local variance across real backgrounds and CLB images. The local variance is defined as the variance computed for a 40×40 pixel area around the center of the image.

plate estimation methods suggest that human observers apply the same strategy in the two background categories. Second, the fact that nonadaptive models (NPWE, NPW, and ROI) also result in better performance with the real mammogram backgrounds suggests that the dissociation of performance in the present study is related to image properties.

Figure 10 suggests an explanation for the better performance on real backgrounds. As mentioned in Subsection 2.G, the global variance—computed across the whole background—is the same for each image presented to the observers, and the noise power spectra have been matched between the two background types [Fig. 10(a)]. However, the mean local variance, defined as the variance computed for a 40×40 pixel area around the center of the signal location, changes from one background type to the other. It is well known [16,17,42] that the local variance has an effect on the detection performance for both human and model observers. According to Fig. 10(b), the mean local variance is lower for real backgrounds ($\langle\sigma^2\rangle=344$) than for CLBs ($\langle\sigma^2\rangle=617$). This reflects the fact that, although histogram equalization implemented in the 2-AFC experiment forces both background types to have the same global variance, the real images are locally less noisy, which makes the signal easier to detect for human and model observers.

Thus, for the Zhang *et al.* [25] images, which by construction had strong nonstationarities that were more exaggerated than those in real mammograms, human observers could adapt their strategy. In the present study, however, the dissociation in performance across stationary and nonstationary backgrounds can be attributed to the lower variance at the possible signal locations and not necessarily to a different human strategy across background types.

4. SUMMARY AND CONCLUSIONS

In the framework of objective assessment of image quality for a realistic mammographic detection task, we developed a method to estimate human linear templates for real backgrounds using a GA. We showed that for the simple task of spherical signal detection on mammographic backgrounds, the HLT can fit human observer results accurately. For images and conditions used in this study, human observers seem to use the same strategy whether the background is nonstationary (real images) or stationary (CLBs). However, this does not imply that the performances are the same in both types of backgrounds. Our results show that even though both types of backgrounds have very similar global statistical features, human observers (and their derived linear templates) perform significantly better on nonstationary backgrounds.

Local properties also influence the performance of other model observers, since whether they use the same (ROI, NPW, NPWE) or similar (SQR, S-DOG, D-DOG) templates for the two backgrounds types, the matching of global properties such as global variance or noise power spectrum does not prevent a significantly better performance on real backgrounds than on CLBs.

However, although human results differ from one background type to another, they can still be accurately reproduced by the HLT and by the models incorporating knowledge of the background, such as channelized Hotelling observers. In our experiments, the HLT does not need the addition of internal noise in the response in order to match results on both CLBs and real backgrounds. This is also the case for the SQR and S-DOG models, whereas the noiseless D-DOG model outperforms human observers and needs the addition of internal noise in its channels in order to match human data. The NPWE model cannot be fitted to human observer results for both backgrounds with the same amount of noise.

This study thus confirms that the HLT provides the possibility to estimate templates that fit human observer performance from experiments performed on easily available synthetic backgrounds and to extrapolate them to real backgrounds, since the templates for both tasks are nearly identical and lead to the same performance when tested on the same background type. This method avoids complications related to the computation of the covariance matrix in the case of Hotelling models and, for the present conditions, *a posteriori* addition of noise.

To further test the adaptive or nonadaptive strategy of human observers in real mammograms, further work may concentrate on calculating the HLT for different areas (with different statistics) of the mammogram. Such work would require classification of background areas based on statistical properties and sufficient data to accurately estimate local human linear templates.

ACKNOWLEDGMENTS

The authors are grateful to Cécile Althaus, Gaëlle Bocksberger, Valentine Clémence, and Roman Naegeli for their participation in the experiments. This work was supported by the Swiss National Science Foundation FN 3252BO-104273.

REFERENCES

1. H. H. Barrett and K. Myers, *Foundations of Image Science* (Wiley, 2004).
2. P. F. Judy, R. G. Swensson, R. D. Nawfel, and K. H. Chan, "Contrast detail curves for liver CT," Med. Phys. **19**, 1167–1174 (1992).
3. S. E. Seltzer, P. F. Judy, R. G. Swensson, K. H. Chan, and R. D. Nawfel, "Flattening of the contrast-detail curve for large lesions on liver CT images," Med. Phys. **21**, 1547–1555 (1994).
4. M. P. Eckstein and J. S. Whiting, "Visual signal detection in structured backgrounds. I. Effect of number of possible spatial locations and signal contrast," J. Opt. Soc. Am. A **13**, 1777–1787 (1996).
5. C. Herrmann, E. Buhr, and D. Hoeschen, "Bildrauschen und Diagnose von Rundherden in Thoraxaufnahme," Z. Med. Phys. **6**, 80–86 (1996).
6. F. O. Bochud, J.-F. Valley, F. R. Verdun, C. Hessler, and P. Schnyder, "Estimation of the noisy component of anatomical backgrounds," Med. Phys. **26**, 1365–1370 (1999).
7. F. O. Bochud, C. K. Abbey, and M. P. Eckstein, "Search for lesions in mammograms: non-Gaussian observer response," Med. Phys. **31**, 24–36 (2004).
8. W. W. Peterson, T. G. Birdsall, and W. C. Fox, "The theory of signal detectability," IEEE Trans. Inf. Theory **4**, 171–212 (1954).
9. C. K. Abbey and M. P. Eckstein, "Classification image analysis: estimation and statistical inference for two-alternative forced-choice experiments," J. Vision **2**, 66–78 (2002).
10. C. K. Abbey and M. P. Eckstein, "Optimal shifted estimates of human-observer templates in two-alternative forced choice experiments," IEEE Trans. Med. Imaging **21**, 429–440 (2002).
11. C. K. Abbey, M. P. Eckstein, S. S. Shimozaki, A. H. Baydush, D. M. Catarious, and C. E. Floyd, "Human-observer templates for detection of a simulated lesion in mammographic images," Proc. SPIE **4686**, 25–36 (2002).
12. C. Castella, K. Kinkel, F. Descombes, M. P. Eckstein, P. Sottas, F. R. Verdun, and F. O. Bochud, "Mass detection on real and synthetic mammograms: human observer templates and local statistics," Proc. SPIE **6515**, 65150U (2007).
13. A. J. Ahumada, Jr., "Classification image weights and internal noise level estimation," J. Vision **2**, 121–131 (2002).
14. J. A. Solomon, "Noise reveals visual mechanisms of detection and discrimination," J. Vision **2**, 105–120 (2002).
15. R. F. Murray, P. J. Bennett, and A. B. Sekuler, "Optimal methods for calculating classification images: weighted sums," J. Vision **2**, 79–104 (2002).
16. A. E. Burgess, R. F. Wagner, R. J. Jennings, and H. B. Barlow, "Efficiency of human visual signal discrimination," Science **214**, 93–94 (1981).
17. A. E. Burgess and H. Ghandeharian, "Visual signal detection. I. Ability to use phase information," J. Opt. Soc. Am. A **1**, 900–905 (1984).
18. K. J. Myers, H. H. Barrett, M. C. Borgstrom, D. D. Patton, and G. W. Seeley, "Effect of noise correlation on detectability of disk signals in medical imaging," J. Opt. Soc. Am. A **2**, 1752–1759 (1985).
19. K. J. Myers, J. P. Rolland, H. H. Barrett, and R. F. Wagner, "Aperture optimization for emission imaging: effect of spatially varying background," J. Opt. Soc. Am. A **7**, 1279–1293 (1990).
20. A. E. Burgess, F. L. Jacobson, and P. F. Judy, "Lesion detection in digital mammograms," Proc. SPIE **4320**, 555–560 (2001).
21. A. E. Burgess and P. F. Judy, "Detection in power-law noise: spectrum exponents and CD diagram slopes," Proc. SPIE **5034**, 57–62 (2003).
22. J. P. Rolland and H. H. Barrett, "Effect of random background inhomogeneity on observer detection performance," J. Opt. Soc. Am. A **9**, 649–658 (1992).
23. F. O. Bochud, C. K. Abbey, and M. P. Eckstein, "Statistical texture synthesis of mammographic images with clustered lumpy backgrounds," Opt. Express **4**, 33–43 (1999).
24. C. Castella, K. Kinkel, F. Descombes, M. P. Eckstein, P.-E. Sottas, F. R. Verdun, and F. O. Bochud, "Mammographic texture synthesis using genetic programming and clustered lumpy background," Proc. SPIE **6146**, 238–249 (2006).
25. Y. Zhang, C. K. Abbey, and M. P. Eckstein, "Adaptative detection mechanisms in globally statistically nonstationary-oriented noise," J. Opt. Soc. Am. A **7**, 1549–1558 (2006).
26. D. Whitley, "A genetic algorithm tutorial," Stat. Comput. **4**, 65–85 (1994).
27. J. Beutel, H. L. Kundel, and R. L. Van Metter, *Handbook of Medical Imaging Volume 1. Physics and Psychophysics* (SPIE, 2000).
28. P. G. J. Barten, "The SQRI method: a new method for the evaluation of visible resolution on a display," Proc. Soc. Inf. Display **28**, 253–262 (1987).
29. A. E. Burgess, "Statistically defined backgrounds: performance of a modified nonprewhitening observer," J. Opt. Soc. Am. A **11**, 1237–1242 (1994).
30. R. D. Fiete, H. H. Barrett, W. E. Smith, and K. J. Myers, "The Hotelling trace criterion and its correlation with human observer performance," J. Opt. Soc. Am. A **4**, 945–953 (1987).
31. H. H. Barrett, J. Yao, J. P. Rolland, and K. J. Myers, "Model observers for assessment of image quality," Proc. Natl. Acad. Sci. U.S.A. **90**, 9758–9765 (1993).
32. H. H. Barrett, C. K. Abbey, and E. Clarkson, "Objective assessment of image quality. III. ROC metrics, ideal observers, and likelihood generating functions," J. Opt. Soc. Am. A **15**, 1520–1535 (1998).
33. F. O. Bochud, C. K. Abbey, and M. P. Eckstein, "Visual signal detection in structured backgrounds. III. Calculation of figures of merit for model observers in statistically nonstationary backgrounds," J. Opt. Soc. Am. A **17**, 193–205 (2000).
34. K. J. Myers and H. H. Barrett, "The addition of a channel mechanism to the ideal-observer model," J. Opt. Soc. Am. A **4**, 2447–2457 (1987).
35. C. K. Abbey and H. H. Barrett, "Human- and model-observer performance in ramp-spectrum noise: effects of regularization and object variability," J. Opt. Soc. Am. A **18**, 473–487 (2001).
36. C. K. Abbey and H. H. Barrett, "Linear iterative reconstruction algorithms: study of observers performance," in *Proceedings of the 14th International Conference on Information Processing in Medical Imaging*, Y. Bizais, C. Barillot, and R. Di Paola, eds. (Kluwer Academic, 1995), pp. 65–76.
37. C. K. Abbey, H. H. Barrett, and D. W. Wilson, "Observer signal-to-noise ratios for the ML–EM algorithm," Proc. SPIE **2712**, 47–58 (1996).
38. A. Papoulis, *Probability, Random Variables, and Stochastic Processes* (McGraw-Hill, 1991).
39. S. Vedantham, A. Karellas, S. Suryanarayanan, D. Albagli, S. Han, E. J. Tkaczyk, C. E. Landberg, B. Opsahl-Ong, P. R. Granfors, I. Levis, C. J. D'Orsi, and R. E. Hendrick, "Full breast digital mammography with an amorphous silicon-based flat panel detector: physical

characteristics of a clinical prototype," Med. Phys. **27**, 558–567 (2000).

40. S. Muller, "Full-field digital mammography designed as a complete system," Eur. J. Radiol. **31**, 25–34 (1999).
41. J. Eng, "JLABROC4: Maximum likelihood estimation of a binormal ROC curve from continuously distributed test results," Version 1.0.1 (The Russell H. Morgan Department of Radiology and Radiological Science, Johns Hopkins University), http://www.rad.jhmi.edu/jeng/javarad/roc/main.html.
42. A. E. Burgess and B. Colborne, "Visual signal detection. IV. Observer inconsistency," J. Opt. Soc. Am. A **5**, 617–627 (1988).

Mass detection on mammograms: influence of signal shape uncertainty on human and model observers

C. Castella,[1,2] M. P. Eckstein,[3] C. K. Abbey,[3] K. Kinkel,[4] F. R. Verdun,[1] R. S. Saunders,[5] E. Samei,[5] and F. O. Bochud[1,*]

[1]*University Institute for Radiation Physics, University Hospital Center, and University of Lausanne, CH-1007 Lausanne, Switzerland*
[2]*Laboratory for High Energy Physics, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland*
[3]*Department of Psychology, University of California, Santa Barbara, California 93106-9660, USA*
[4]*Clinique des Grangettes, CH-1224 Chêne-Bougeries, Switzerland*
[5]*Duke Advanced Imaging Laboratories, Duke University, Durham, North Carolina 27705, USA*
**Corresponding author: francois.bochud@chuv.ch*

Received September 9, 2008; revised December 11, 2008; accepted December 14, 2008;
posted December 15, 2008 (Doc. ID 101392); published January 30, 2009

We studied the influence of signal variability on human and model observers for detection tasks with realistic simulated masses superimposed on real patient mammographic backgrounds and synthesized mammographic backgrounds (clustered lumpy backgrounds, CLB). Results under the signal-known-exactly (SKE) paradigm were compared with signal-known-statistically (SKS) tasks for which the observers did not have prior knowledge of the shape or size of the signal. Human observers' performance did not vary significantly when benign masses were superimposed on real images or on CLB. Uncertainty and variability in signal shape did not degrade human performance significantly compared with the SKE task, while variability in signal size did. Implementation of appropriate internal noise components allowed the fit of model observers to human performance. © 2009 Optical Society of America

OCIS codes: 330.1880, 170.3830, 330.4060, 330.5020, 330.5510.

1. INTRODUCTION

Detection and classification tasks are fundamental to many medical imaging applications. In radiology, these tasks involve first determining whether candidate signals are present in the image, then evaluating each candidate and rating its likelihood of being an actual lesion. Likely lesions are then classified based on their characteristics such as size, shape, and malignancy. While modeling the full clinical detection and classification process is still out of the scope of current psychophysical studies, numerous authors have reported results and models with the aim of improving the understanding of the processes behind various detection tasks. In most instances, these experiments were simplifications of real clinical tasks, using statistically or exactly known backgrounds, signals, and/or signal locations. The simpler tasks facilitate data collection and the robustness of the analysis.

When studying mass detection in a typical radiological task, use of real backgrounds and masses is desirable to achieve realism, but the collection of hundreds or thousands of similar images can be difficult and time-consuming. Therefore, mammographic nonstationary backgrounds are typically replaced by synthesized white noise [1,2], power-law filtered white noise [3–5], or lumpy backgrounds [6–9]. Similarly, masses are generally approximated by disks [2,3,10], phantom elements [9], and Gaussian or Gabor functions [11–13]. Signal location uncertainty in the clinical task can be simplified by controlling the number and the location of the signals under the M-alternative forced-choice (M-AFC) paradigm, or receiver operating characteristic (ROC) studies [14,15]. For such controlled signal conditions, Brettle *et al.* [16] recently showed that trained naïve (nonphysician) observers' performance was very close to that of radiologists.

To reflect the image properties of clinical tasks, recent developments have focused on producing synthetic yet realistic backgrounds and signals that mimic medical images while preserving data collection and computational efficiency. Lumpy backgrounds, initially developed by Rolland and Barrett [6], have been extended to clustered lumpy backgrounds (CLB) [7]. Later, they have been further optimized [8] (second-generation CLB) to reproduce visual and statistical properties of mammograms. On the signal front, Saunders *et al.* [17,18] recently developed an algorithm capable of generating benign or malignant breast mass signals based on the analysis of real masses' characteristics.

An important aspect of clinical relevance that is introduced in our study is signal uncertainty. While most studies have concentrated on signal-known-exactly (SKE) tasks, where the signal presented to the observers is known and does not vary throughout the entire experiment, less is known about more realistic conditions in which each image presents a different realization of the signal and the exact physical characteristics of the signal are not known to the observer. To model experiments involving various signals, signal-known-exactly but variable (SKEV) and signal-known-statistically (SKS) para-

1084-7529/09/020425-12/$15.00 © 2009 Optical Society of America

digms have been introduced [19–23] in M-AFC tasks. In the SKEV task, a pool of different signals is used throughout the detection experiment. Although the signal is randomly selected from one trial to the next, a high-contrast replica of the actual signal is displayed in addition to the M-AFC images. The observer thus always knows which signal is present. The SKEV task allows for generalization beyond a specific signal, while retaining the simplicity in analysis and modeling of the SKE task. In the SKS task, the signal is also randomly chosen for each trial out of a pool of different candidates, but the observer does not know which signal was selected. This scenario closely reflects a real radiological task, but its results are more complex to analyze and model than SKE or SKEV tasks.

To reproduce or predict human performance in detection tasks, model observers have been developed and successfully adapted to SKE experiments [14,16,24]. Later, models have been adapted to SKS tasks [19], and Eckstein *et al.* [20,21] and Zhang *et al.* [22] showed a good correlation with human results for SKS experiments with x-ray coronary angiograms. However, very little is known about the ability of SKS-adapted model observers to accurately predict human performance in mammography for SKS conditions.

The first purpose of this study is to evaluate the influence of background, signal shape and signal size, on the detection performance of human observers by conducting psychophysical tasks with mammographic backgrounds and second-generation CLB, combined with synthetic benign and malignant breast masses to produce fully realistic yet controlled images.

The second purpose is to use the human observers' data to evaluate linear model observers in their ability to predict human observer performance across the various psychophysical conditions, and to evaluate alternative methods for introducing internal noise in the models to best predict human observers' performance in SKS tasks.

2. MATERIAL AND METHODS

A. 2-AFC Setup

Four nonphysician observers participated in this study. All had experience with 2-AFC experiments, since they had participated in a previous SKE study with mammographic backgrounds and CLB [9]. For each of the 13 background and signal combinations described in Fig. 1, the observers were presented 1,400 image pairs, or trials. The signal was randomly embedded in one of the two images for each trial.

The observers had to determine which image was the most likely to contain the signal. Fiduciary highly visible cues were provided to precisely locate the two possible signal locations, one per image. There was no time limit, and feedback was provided after each trial (correct or incorrect answer) and as a summary performance measure after every 25 trials (percent correct, P_c). The order of the 13 different tasks was the same for all observers. The observers performed each task during sessions distributed over one day or two consecutive days.

The images were displayed in a dark room on a Siemens SMM 21140 P high-contrast gray-scale monitor (Siemens, Karlsruhe, Germany) calibrated to the DICOM Grayscale Display Function and TG18 standards [25]. Pixel size of the display screen was 0.25 mm. The observers were free to select the viewing distance, which was typically about ~40 cm, but they were not allowed to modify any display settings.

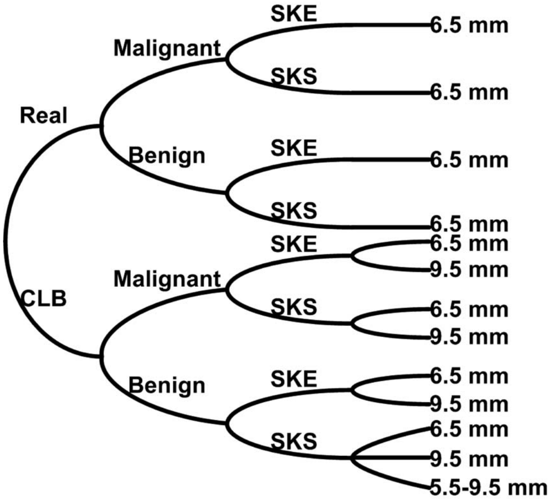


Fig. 1. Backgrounds, mass type, and signal conditions for the 13 2-AFC detection task experiments. Benign and Malignant characterize simulated masses. SKE stands for an experiment with a single signal of given shape and size and SKS an experiment with a signal with variable shape but a given size, except for the last experiment where both shape and size were variable.

B. Backgrounds

We used two kinds of backgrounds: real regions of interest (ROIs) extracted from digital mammograms, and synthetic second-generation CLB. For the real images, we used a database of 88 disease-free patients who underwent screening exams on a GE Senograph 2000D full-field digital detector (pixel size: 0.1 by 0.1 mm) [26,27]. A total of 2,800 square ROIs (256 by 256 pixels) were manually selected from the processed mammograms and resampled to 154 by 154 pixels to emulate a magnification factor of 1.5 on the display screen, reproducing typical clinical settings [9]. The CLB had been designed to mimic digital mammogram ROIs and their statistical and visual properties assessed by radiologists in a previous study [8].

To obtain comparable conditions between the real and the synthetic images, and because of the prominent importance of local statistics [9], we matched the first two moments of the gray-level distributions of real and CLB images over the 40×40 pixel central area of the displayed images. This corresponds to the area covered by the fiduciary cues over which the human observers were hypothetically processing to perform the task. Over this area, the mean gray level was set to 128 and the standard deviation to 20, ensuring that the rescaled images would be in the middle of the display screen dynamic range. This change of the first-order statistics implied a shift of the ROI power spectrum without significantly altering its slope.

C. Signals

The signals were synthetic breast masses developed by Saunders *et al.* [18] and based on the analysis of real breast lesions from the Digital Database for Mammography Screening (DDSM) [28]. They provided signals that closely resemble real masses and represent perfect knowledge of ground truth. We chose to use two kinds of simulated breast masses: oval-shaped benign circumscribed masses, and irregularly shaped malignant masses with ill-defined borders [29]. In this work, these two kinds of simulated lesions will be referred to as benign masses and malignant masses, respectively.

Both mass types were constructed using concentric elliptical rings as a basis, and their edges characteristics were matched to those of corresponding real benign or malignant breast lesions from the DDSM. Their size was defined as the major axis of the ellipse corresponding to the half-maximum of the signal intensity [18]. During the 2-AFC experiments, these masses were embedded in real time in the different backgrounds, real or synthetic, resulting in controlled signal-present images. Detection tasks with 6.5 and 9.5 mm masses were conducted with CLB, while 6.5 mm masses only were used with the real backgrounds to limit the number of psychophysical experiments.

The masses were added linearly to the backgrounds. The reason for the linear addition is that the processed mammograms of the GE mammography unit are obtained by windowing the logarithm of the exposure data [30]. The signals, multiplicative in the exposure domain due to exponential attenuation through the lesion, were thus additive in the log(exposure) images. The amplitude of the signal, defined as the maximum intensity of the mass, was set after preliminary experiments by one of the authors to obtain a P_c between 0.70 and 0.85 for each condition. For benign masses, this resulted in an amplitude of 10 gray levels (GL), whereas a higher amplitude of 15 GL had to be used for malignant masses, which tend to be less conspicuous due to smoother borders. At the beginning of each of the 13 different experimental conditions presented in Fig. 1, the observers were trained with sets of 25 trials with decreasing signal amplitudes until they had reached a P_c of at least 0.70 for the actual experimental contrast conditions. Depending on the observers and series conditions, this training phase lasted from 100 to ~500 trials. The target P_c range 0.70–0.85 is low compared with traditional 2-AFC studies [31]. We chose this range of P_c for efficient estimation from samples of one model observer whose template (Human Linear Template; see Subsection 2.D), is derived from the observers' correct and incorrect answers and the backgrounds. Template estimation would have been much less efficient with a higher value of P_c.

For SKE experiments, the signal was identical during the whole experiment, including high-contrast and low-contrast training phases. The observers were aware that they were being trained with the same signal that would be presented during the actual experiment. For the SKS task, the signal was chosen randomly for each trial from a pool of 50 similar candidates of the same mass type (benign or malignant) and with the same size: the actual signal to be detected thus changed from trial to trial, and was not known by the observer. In a similar manner to the SKE task, for the SKS conditions, the 50 similar signals were randomly chosen from the same set during the training phase and the following experiment. Finally, a last experiment was conducted with CLB and benign masses having sizes of 5.5, 6.5, 7.5, 8.5, and 9.5 mm. Ten masses per size constituted the pool of SKS signals. The aim was to compare SKS results when the mass size was kept constant and only its shape and orientation changed, and when the sizes covered the range of interest in screening mammography. Examples of displayed images, including fiduciary cues, are shown in Fig. 2.

D. Model Observers

Linear model observers were implemented and compared with human observers' results. We used the nonprewhitening matched filter (NPW) [14], the NPW with an eye-filter (NPWE) [24,32], channelized Hotelling observers (CH) [14,15,33] with square (SQR) channels, sparse (SDOG) or dense (DDOG) difference-of-Gaussians channels, and Gabor function channels [9,34], and estimated Human Linear Template (HLT) models [9,35–40]. These models have been described extensively in the literature, and they will only be briefly reviewed here.

The decision variable of a general linear observer of an image $\mathbf{g}_i$ is given by the product of a template $\mathbf{w}$ and the image:

$$\lambda_i = \mathbf{w}^T\mathbf{g}_i + \epsilon. \tag{1}$$

In Eq. (1), both $\mathbf{w}$ and $\mathbf{g}$ are expressed as 1-D vectors. ε is an optional noise term that reflects the observer's internal noise. The templates for the NPW and NPWE observers in SKE tasks were defined as

$$\mathbf{w}_{\mathrm{NPW}} = \mathbf{s}, \tag{2}$$

$$\mathbf{w}_{\mathrm{NPWE}} = \mathbf{E}^T\mathbf{E}\mathbf{s}, \tag{3}$$

where $\mathbf{s}$ is the signal, and $\mathbf{E}(\rho)=\rho^n \exp(-c\rho^2)$ is an eye filter that accounts for human eye different sensitivity to radial frequency ρ. The parameters used for the eye filter ($n=1.3$, $c=0.0041$) are from Burgess [24].

The general Hotelling observer template is derived from the covariance matrix of the backgrounds $\mathbf{K}_b$ as

$$\mathbf{w}_{\mathrm{Hot}} = \mathbf{K}_b^{-1}[\langle \mathbf{g}_s\rangle - \langle \mathbf{g}_n\rangle], \tag{4}$$

where $\langle \mathbf{g}_s\rangle$ and $\langle \mathbf{g}_n\rangle$ are respectively, the means of the images containing the signal and the background and containing the background only. If the signal is identical for all signal-present images, then $\langle \mathbf{g}_s\rangle-\langle \mathbf{g}_n\rangle$ is equal to $\mathbf{s}$.

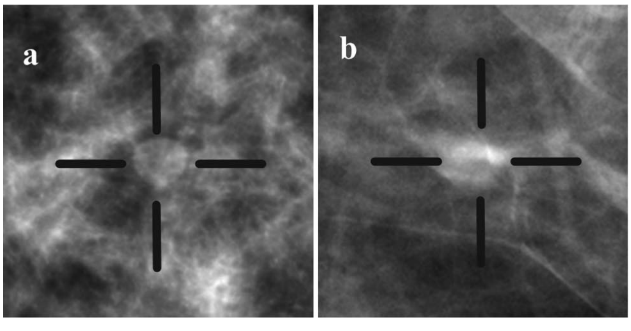


Fig. 2. (a) Example of CLB with a digitally embedded benign simulated mass. (b) Real mammogram ROI with a digitally embedded malignant simulated mass. Signal contrast has been strongly increased for illustration purposes.

The covariance matrix inversion in Eq. (4) is often impractical to implement, since for $N \times N$ pixel images, the size of this matrix is $\mathrm{N}^2 \times \mathrm{N}^2$. Moreover, the large number of independent images needed for getting a nonsingular estimate of the covariance matrix is rarely reached in a typical experimental study. To overcome these computational issues, the Hotelling observer may be approximated by reducing the images to a small set of variable response channels [5,14,15,33,34]. The CH observer template is then given by

$$\mathbf{w}_{\mathrm{CH}} = (\mathbf{K}_{b,c} + \mathbf{K}_{\epsilon})^{-1}\mathbf{s}_c. \tag{5}$$

In Eq. (4), $\mathbf{K}_{b,c}$ is the channelized covariance matrix which represents the external noise source. It is computed from the background images as $\mathbf{K}_{b,c} = \langle (\mathbf{T}^T\mathbf{g}_n - \langle \mathbf{T}^T\mathbf{g}_n \rangle)(\mathbf{T}^T\mathbf{g}_n - \langle \mathbf{T}^T\mathbf{g}_n \rangle)^T \rangle$, where the column vectors of the matrix $\mathbf{T}$ each represent the spatial profile of a channel. The noiseless covariance matrices $\mathbf{K}_{b,c}$ were estimated by sampling using the 1,400 signal-absent images of the 2-AFC experiments. $\mathbf{s}_c$ is the expectation of the signal seen through the channels: $\mathbf{s}_c = \mathbf{T}^T[\langle \mathbf{g}_s \rangle - \langle \mathbf{g}_n \rangle]$. $\mathbf{K}_\varepsilon$ is the covariance matrix of the internal noise expressed in the channels' basis (see Subsection 2.F). The four kinds of channels used for this study are the SQR, SDOG, and DDOG channels as described by Castella *et al.* [9] and Abbey and Barrett [34], and channels defined by Gabor functions. The Gabor channels were constructed as

$$G(x,y,\Lambda,\theta,\varphi) = \exp\left[-\frac{(x'^2 + y'^2)}{2\sigma^2}\right]\cos(2\pi x'/\Lambda + \varphi). \tag{6}$$

In Eq. (6), Λ is the wavelength in pixels, φ is equal to 0 for odd-phase channels and $\pi/2$ for even-phase channels, $\sigma = 0.56\Lambda$ for a bandwidth of one octave, $x' = x\cos\theta + y\sin\theta$, and $y' = -x\cos\theta + y\cos\theta$. We used a total of five orientations, eight wavelengths and two phases (odd and even), making a total of 80 channels. The wavelengths were chosen according to the DDOG channels' peak frequencies, with values ranging from $\Lambda_{\min} = 18$ pixels (0.64 deg visual angle) to $\Lambda_{\max} = 192$ pixels (6.8 deg visual angle) in discrete steps spaced by a multiplicative factor of 1.4.

The HLTs [9,35–40] were estimated using a genetic algorithm (GA). The details of the procedure are described in a paper by Castella *et al.* [9]. With this method, the template itself $\mathbf{w}_{\mathrm{HLT}}$ is derived *a posteriori* from the individual 2-AFC decisions made by the human observers. The GA finds the linear template that maximizes the likelihood function of observing the individual trials' choices of human observers [38].

For each experimental condition given in Fig. 1, ten estimates of $\mathbf{w}_{\mathrm{HLT}}$ were obtained by running ten times the GA with different seeds. The results and standard errors presented in the next sections correspond to the average performance of the model for the ten estimates.

E. Performance Evaluation

Human observers' performance on a given task was measured using the proportion of correct answers P_c. Individual observers' standard errors for P_c were derived by computing P_c for subsets of 50 consecutive trials. Additionally, all human observers' results were pooled to determine the generic observer performance and associated variance estimates. We used the Gallas *et al.* [41] multireader multicase variance analysis method for binary data and generic study designs. This method provides an unbiased estimate of the generic observer's performance and variance (in terms of P_c) when different observers with possibly different skills perform a binary task with possibly different cases and case numbers among observers. For our study, a case consisted of a randomly chosen pair of one signal-present and one signal-absent image. As the database of possible signal-present and signal-absent images contained as much as N_T different backgrounds for each condition, we assumed that the cases were independent for the statistical analysis.

P_c was then converted to an empirically obtained index of detectability d' by generating a lookup table for P_c versus d' from the usual cumulative Gaussian relationship under the assumption that the variances of the responses to the signal-present and signal-absent locations be identical [15,42]. The use of backgrounds that are not contiguous in the patient images justifies the transformation from P_c to d' with the assumption of statistically independent internal responses.

The model observers' performances were assessed using Monte Carlo experiments. For each model and experimental condition, the decision-variable distributions were estimated by directly computing the dot product between the corresponding templates and 1,400 random signal-present and signal-absent image pairs, using the same backgrounds and signals as for the human observers. For SKE tasks, the area under the ROC curve was computed from the decision-variable distributions for signal-present and signal-absent images using JROCFIT [43], and then transformed into a detectability index [42]. For SKS tasks, individual trial decisions were computed using the sum of likelihood rule leading to an estimate of P_c, which was then also transformed into a detectability index. The use of the same relationship between P_c and d' for the SKE and SKS tasks is a simplified approach, since one should also include the signal uncertainty in the conversion for SKS tasks. But here we were using the conversion to d' only as a transform to a performance measure that was comparable to the d' from the SKE.

F. Selecting the Internal Noise Level

Human observers are known to be subject to internal noise, and there are various ways to implement it in model observers [14,15,44]. For models using channels mechanisms (all four CH observers in this study), internal noise was assumed to be zero mean, independent in each channel, with variance proportional to the variance of the external noise in each channel, and with a proportionality factor p_n. The noisy channelized background covariance matrix $\mathbf{K}_{b,c,n}$ could thus be defined as

$$\begin{cases} (\mathbf{K}_{b,c,n})_{i,j} = (\mathbf{K}_{b,c})_{i,j} & \text{if } i \neq \mathrm{j} \\ (\mathbf{K}_{b,c,n})_{i,j} = (1 + p_n)(\mathbf{K}_{b,c})_{i,j} & \text{if } i = j \end{cases}. \tag{7}$$

Under these assumptions, the decision variable in Eq. (1) becomes

$$\lambda_i = \mathbf{w}_{CH}^T \mathbf{g}_i + \sum_k w_k \epsilon_k, \tag{8}$$

where the w_k are the components of $\mathbf{w}$ in the channel's basis. Each ε_k is Gaussian distributed with zero mean and has a variance of $p_n \sigma_k^2$.

For the other model observers (NPW, NPWE, and HLT), noise was added in the decision variable as

$$\begin{cases} \lambda_i = \mathbf{w}^T \mathbf{g}_i + \epsilon \\ \epsilon \sim N(0, p_n \sigma_{\text{ext}}^2) \end{cases}, \tag{9}$$

where ε is Gaussian distributed with zero mean and variance equal to $p_n \sigma_{\text{ext}}^2$. The variability σ_{ext}^2 was estimated by computing the variance of λ without internal noise from 1,400 signal-absent images.

On the basis of human observer results in the psychophysical tasks, we then defined the optimal value of p_n in Eqs. (7) or (9) for a given model observer as the one that best matched the performance of this model to that of the human observers. To determine the optimal internal noise level for each model, we used Monte Carlo trials with the same backgrounds and signals as for human observers. Optimal p_n for benign and malignant masses were separately assessed. We iteratively changed p_n until the root-mean-square error (RMSE) between the model and the generic human observer performance was minimized for the SKE tasks:

$$RMSE = \sqrt{\frac{1}{3}[(d'_h - d'_m)^2_{\text{CLB},6.5\ mm} + (d'_h - d'_m)^2_{\text{CLB},9.5\ mm} + (d'_h - d'_m)^2_{\text{real},6.5\ mm}]}, \tag{10}$$

where the subscripts h and m, respectively, stand for generic human and model observer performances.

G. Model Observers and Internal Noise in SKS Tasks

All models were also adapted to the SKS tasks using the sum-of-likelihood rule described by Zhang *et al.* [45]. In this approach, the response of a model observer is obtained by combining the individual responses of the templates corresponding to the J different possible signals and assuming Gaussian internal responses:

$$\begin{cases} l_+ = \displaystyle\sum_{j=1}^{J} \left(\frac{1}{2\pi\sigma_j^2}\right) \exp\left[-\frac{(\lambda_{+,j} - \mu_{+,j})^2}{2\sigma_j^2}\right] \exp\left[-\frac{(\lambda_{-,j} - \mu_{-,j})^2}{2\sigma_j^2}\right] \\ l_- = \displaystyle\sum_{j=1}^{J} \left(\frac{1}{2\pi\sigma_j^2}\right) \exp\left[-\frac{(\lambda_{-,j} - \mu_{+,j})^2}{2\sigma_j^2}\right] \exp\left[-\frac{(\lambda_{+,j} - \mu_{-,j})^2}{2\sigma_j^2}\right] \end{cases}, \tag{11}$$

where the decision variables l_+ and l_- are the sums of likelihoods for the signal-present and signal-absent image, respectively; $\lambda_{\pm,j}$ is given by Eq. (1); and $\mu_{+,j}$ is the expected response of the jth template to the images containing the jth signal and $\mu_{-,\text{j}}$ the response to the signal-absent images; assuming an equal variance σ_j^2. The expectations and variances of the responses of each template were estimated from the $\lambda_{\pm,j}$ distributions with the 1,400 signal-present and 1,400 signal-absent images used in the experiments.

Adding internal noise to this process is not trivial. Three alternatives were tested in this paper:

(i) *Internal noise in the individual template responses*. In this scheme, the optimal internal noise level p_n that had been found for the SKE tasks was used in Eq. (8) and (9) to alter the distributions of $\lambda_{\pm,j}$ in Eq. (11).

(ii) *Internal noise added to the maximum of the logarithm of the likelihoods*. The template with the maximum likelihood only was used instead of the sum in Eq. (11). Internal noise was then added to the logarithms of l_+ and l_- as a Gaussian random variable with zero mean and $\text{p}_n \sigma_{\text{ext}}^2$ variance, the latter being estimated as in Eq. (9).

(iii) *Internal noise assuming that a single template is used*. This alternative assumes that the model performs SKS and SKE tasks the same way, using a single template. For each task, the model observers' templates derived for SKE tasks were used for the corresponding SKS tasks, and internal noise was added as in Eq. (8) and (9)

Using the optimal internal noise levels p_n that had been found for the SKE tasks, Monte Carlo trials were conducted for the SKS tasks to test each of these internal noise schemes. An overall measure of agreement between a given model and the generic human observer $RMSE_{\text{overall}}$ could then be computed with all seven tasks with benign masses and all six tasks with malignant masses (see Fig. 1). For statistical reasons, only one HLT per SKS detection task given in Fig. 1 was estimated, corresponding to the generic observer data.

The statistical significance in the differences between generic human and model observers' performance were assessed with a F-test with (number of tasks-1) degrees of freedom (*df*) for the numerator and (number of observers-1) *df* for the denominator. This test uses the mean-square error between the generic human and the model observer, and compares it to the mean variability across individual human observers [44]:

$$F = \frac{\left(\sum\limits_{i \text{ tasks}} (d'_{i,\text{model}} - d'_{i,\text{human}})^2\right) / df_{\text{numerator}}}{\left(\sum\limits_{i \text{ tasks}} \text{var}_{i,\text{human}}\right) / df_{\text{denominator}}}. \tag{12}$$

3. RESULTS

A. Robustness of the Results

Potential sources of bias in the human observer results were statistically tested for each of the 13 experiments. At a 5% confidence level, there was no significant deviation from an equal proportion of left versus right image choice for any observer.

Possible learning effects were tested by comparing the proportion of correct answers of the first and last 200 trials to P_c for the whole 1,400 trials for each experiment and observer. We found no significant deviation from random differences in performance across the beginning, the end, and the experiment as a whole for all observers. These results suggest that the observers had effectively stabilized their performance after the training phases, and were performing consistently during the actual detection experiments.

Finally, potential correlation between decision time (time used to give an answer for a given trial) and P_c was also tested. For each condition, the 1,400 trials were divided into 28 subsets of 50 consecutive trials. The mean decision time and P_c were then computed for each subset, and the correlation coefficient between these quantities assessed using a 2-sided T-test [46]. The correlation coefficient was not significantly different from 0 for any observer. This suggests that for a given observer and conditions, no improvement or degradation of the performance resulting from an increased decision time t could be statistically demonstrated.

However, it is of interest to note that while P_c did not change significantly during a given experiment, t generally decreased by 30 to 50% for each observer between the first trials and the last ones. Absolute mean values for t ranged from 1 to 4 seconds, depending on the observer and the conditions.

Table 1. Generic Observer Performance (d') and 95% Confidence Interval for the 13 Different Tasks Involving Real or CLB Backgrounds, SKE or SKS Detection, and Simulated Mass Size of 6.5, 9.5 or 5.5–9.5 mm

Conditions	Benign Masses	Malignant Masses
Real,SKE,6.5	1.07 [1.01–1.14]	1.01 [0.96–1.07]
CLB,SKE,6.5	1.11 [1.01–1.21]	0.81 [0.73–0.89]
Real,SKS,6.5	1.14 [1.09–1.19]	1.06 [0.99–1.14]
CLB,SKS,6.5	1.06 [0.97–1.16]	0.81 [0.77–0.87]
CLB,SKE,9.5	0.89 [0.82–0.96]	0.92 [0.82–1.03]
CLB,SKS,9.5	0.98 [0.89–1.07]	1.02 [0.87–1.17]
CLB,SKS,5.5–9.5	0.90 [0.86–0.93]	–

Concerning the robustness of the HLT estimation, the performances of the ten estimates of $\mathbf{w}_{\mathrm{HLT}}$ per experimental condition were first assessed separately using Monte Carlo trials, in a way similar to the other models. The ten P_c were then averaged to determine the overall HLT model performance. Statistical analysis showed no significant difference between the performance averaged across $\mathbf{w}_{\mathrm{HLT}}$ estimations, and the performance of the template obtained by spatially averaging the ten estimates.

B. Human Observer Results

1. Benign Masses

The index of detectability d' and associated confidence intervals averaged over the four human observers for each of the 13 experimental conditions described in Fig. 1 are given in Table 1. As an example of typical individual experimental results, human observers' d' for the 6.5 mm benign masses are given in Figs. 3(a) for SKE tasks and 3(b) for SKS. For SKE experiments, d' averaged over the four observers (generic human observer) was 1.07 for the real backgrounds, and 1.11 for the CLB ($p=0.57$). For the SKS tasks, the difference between real images ($d'=1.14$) and CLB (1.06) was not statistically significant either ($p=0.14$).

When comparing SKE and SKS tasks for the 6.5 mm masses, there was no significant difference for either real backgrounds ($p=0.10$) or CLB ($p=0.53$).

In the experiments with fixed signal size, the 6.5 mm masses were better detected than the 9.5 mm masses. The difference is especially visible in the SKE experiment (difference of 0.22 in d' units, $p=0.0003$), while smaller in the SKS task (0.08 in d' units, $p=0.18$); see Table 1. This trend is also visible in the size uncertainty experiment with signal size ranging from 5.5 to 9.5 mm (Fig. 4). A 2-way ANOVA performed across signal sizes and observers showed that neither observer ($df=3$, $F=0.49$, $p=0.69$) nor mass size ($df=4$, $F=1.71$, $p=0.21$) dependency were significant. Finally, we compared the performance of the generic observer for SKS with fixed size (shape uncertainty) versus size uncertainty. There is no significant difference in the performance for the 6.5 mm masses (0.08 in d' units, $p=0.51$), but the 9.5 mm are clearly better detected when the observer knows the signal size than in the size uncertainty task (0.19 in d' units, $p=0.01$).

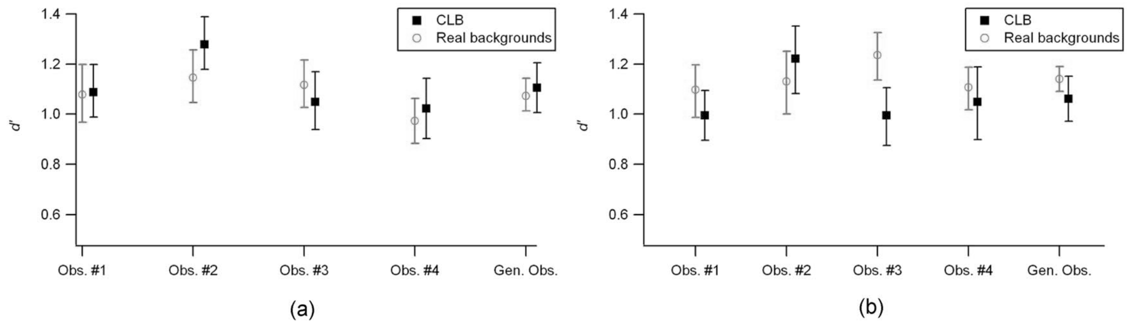


Fig. 3. (Color online) (a) Human observers' performance (d') for the SKE tasks with the 6.5 mm benign simulated masses. The rightmost values for each figure (generic observer, Gen. Obs.) were obtained by pooling all observer data. The error bars represent the 95% confidence interval. (b) Same for the SKS tasks.

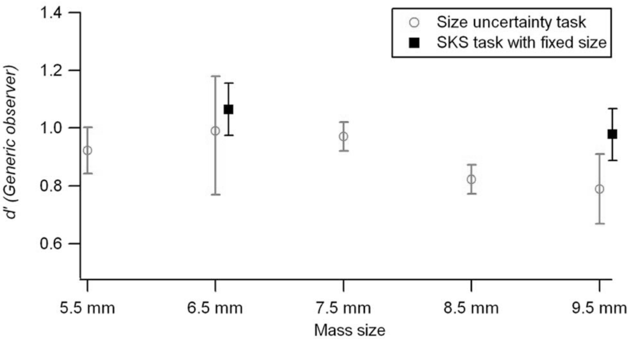


Fig. 4. (Color online) Generic human observer results for the size uncertainty task with benign simulated masses (open circles). For comparison, the performance in the SKS experiments with fixed size signals are shown (black squares). Error bars represent the 95% confidence interval.

2. Malignant Masses

Human observers' performance for the 6.5 mm malignant masses is shown in Figs. 5(a) (SKE) and 5(b) (SKS). For these masses, generic observer's d' was significantly higher with the real images than with the CLB for both SKE (0.20 in d' units, $p < 10^{-4}$) and SKS tasks (0.25 in d' units, $p < 10^{-4}$).

As for benign masses, there was no significant difference between SKE and SKS tasks for both real backgrounds (0.05 in d' units, $p = 0.31$) and CLB (<0.01 in d' units, $p = 0.99$).

The mass size effect was different than for benign masses. The 9.5 mm malignant masses were better detected than 6.5 mm masses: the difference is visible in Table 1 for SKS task (0.21 in d' units, $p = 0.01$) and for SKE, although not statistically significant (0.11 in d' units, $p = 0.10$).

C. Model Observers

Figures 6(a) and 6(b) present the RMSE of the different models after adjustment of the internal noise level. As mentioned previously, the internal noise parameters were varied to match human performance on the SKE tasks. The same values for the internal noise parameters were then used for the other conditions. Some models (DDOG with benign masses, and SDOG, DDOG for both mass types) already had a performance level that was below that of human observers before any internal noise addition. For this reason, there is no nonzero optimal value of the noise level parameters p_n for these models, since any amount of internal noise would further degrade their performance. Table 2 shows the $RMSE_{\text{overall}}$ values, which represent the difference in performance between the generic human observer and the models with optimal value of p_n. The $RMSE_{\text{overall}}$ includes SKE tasks, and SKS tasks with the three alternative ways of adding internal noise to the models presented in Subsection 2.G.

Figure 7 shows representative examples of model observer templates (SKE tasks, CLB) and Fig. 8, typical in-

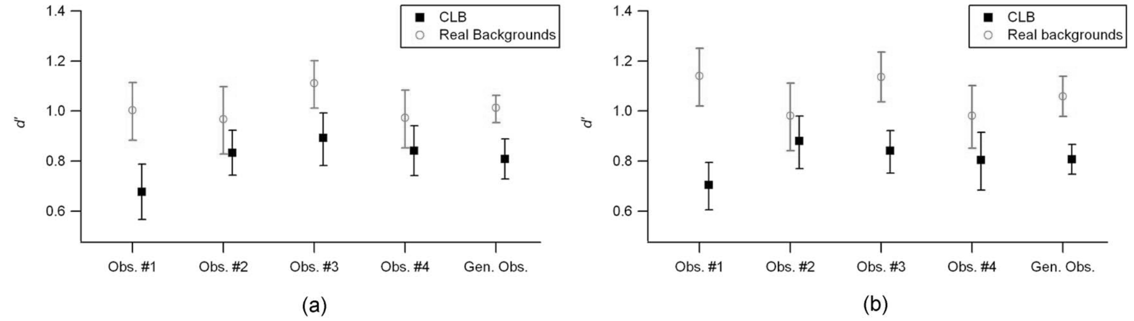


Fig. 5. (Color online) (a) Human observers' performance (d') for the SKE tasks with the 6.5 mm malignant simulated masses. The rightmost values for each figure (generic observer, Gen. Obs.) were obtained by pooling all observers data. The error bars represent the 95% confidence interval. (b) Same for the SKS tasks.

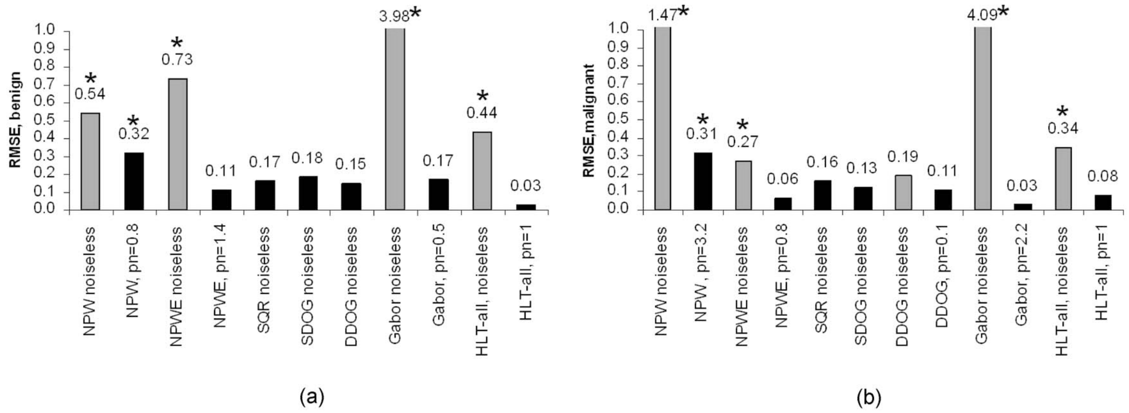


Fig. 6. (Color online) RMSE in d' units between the generic human observer and the different model observers for SKE tasks for noiseless (shaded) and noise-level-optimized (black) models with (a) benign and (b) malignant simulated masses. Stars indicate performance levels that are significantly different from humans (F-test, $p < 0.05$) [44].

Table 2. $RMSE_{\text{overall}}$ in d' Units Between the Generic Human Observer and the Different Model Observers for the Different Noise Addition Schemes Described in Subsection 2.G[a]

$RMSE_{\text{overall}}$	Internal Noise	NPW	NPWE	SQR	SDOG	DDOG	Gabor	HLT
Benign	Indiv. responses	*0.44*	**0.14**	**0.23**	**0.34**	**0.13**	*0.43*	N/A
	Max. likelihood	*0.36*	0.29	0.31	*0.38*	0.15	*0.33*	N/A
	Single template	**0.29**	0.31	0.25	**0.34**	**0.13**	**0.17**	**0.11**
Malignant	Indiv. responses	*0.79*	**0.13**	**0.21**	**0.17**	**0.14**	0.29	N/A
	Max. likelihood	0.56	0.25	0.22	**0.17**	0.17	0.30	N/A
	Single template	**0.54**	0.19	**0.21**	**0.17**	**0.14**	**0.11**	**0.07**

[a] The $RMSE_{\text{overall}}$ is computed for all seven tasks with benign simulated masses and all six tasks with malignant simulated masses. The best internal noise scheme for each model and mass type is indicated in bold. Italic values indicate performance levels that are significantly different from those of humans (F-test, $p < 0.05$) [44].

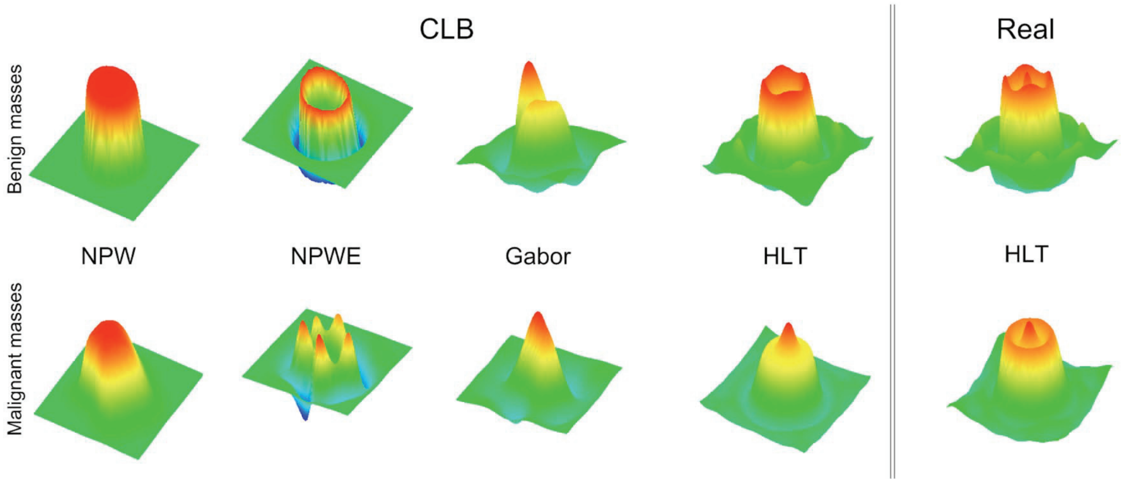


Fig. 7. (Color online) Templates derived for the NPW, NPWE, CH with Gabor channels, and HLT models for the SKE tasks with CLB and benign (upper row) and malignant (lower row) simulated masses. HLT estimated for the tasks with real backgrounds are shown in the last column.

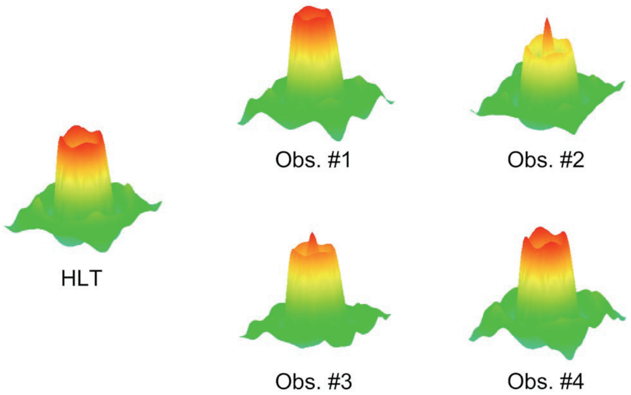


Fig. 8. (Color online) Examples of individual HLT (SKE task, 6.5 mm benign simulated masses, CLB). The leftmost template is the one corresponding to the generic observer.

dividual HLTs compared with the one corresponding to the generic observer (SKE task with 6.5 mm benign masses and CLB).

4. DISCUSSION

A. Influence of the Background and the Local Statistics

In a previous study [9], we showed that human strategy was similar between real mammographic and second-generation CLB for a SKE detection experiment with a mass signal extracted from a mammographic phantom. However, we also observed dissociations in performance for human and model observers between the two background types, and argued that matching the first two orders' statistics over the backgrounds as a whole was not sufficient to ensure comparable conditions. For this reason, we tried to follow a more local approach in the current study, and matched these statistics specifically in the central part of the images, where the observers supposedly performed the task.

For benign masses, human observers achieved very close performance for both backgrounds. Individual observer's differences in P_c did not exceed ±2.6%, except for one observer in the SKS task (Obs. No. 3, 5%). For these signals with sharp edges, comparable to that used in the previous study [9], matching local statistics resulted in backgrounds that are comparable in terms of detection performance. For malignant masses, however, the significant performance difference between real images and CLB could indicate that a different strategy involving more complex properties than first two orders' statistics is used by the human observers: we hypothesize that the systematic lower human performance with CLB suggests that their random, stationary nature containing blobs with smooth edges by construction is more likely to hide signals with similarly smooth edges than nonstationary images, in a way similar to the findings of Zhang *et al.* [47] with highly nonstationary backgrounds.

B. Influence of the Signal Shape Uncertainty

Quite surprisingly, when the signal size was constant over the experiment and the only uncertainty was its shape, human observers performed as well for the SKS tasks as for the corresponding SKE tasks. This shows that, although they had been trained with high-contrast versions of the same signals as in the actual experiments, human observers were not able to develop a better strategy for the SKE task than for the SKS. Zhang *et al.* [22] had reached somewhat different conclusions when comparing SKS to the *a priori* easier SKEV task: in their 4-AFC experiments with x-ray coronary angiograms, a high contrast copy of the actual signal used for the given trial was shown to the observer. Both signal shape and size varied from one trial to the next, and human observers performed better in the SKEV than in SKS tasks. This difference may arise from the fact that in all but the last experiment in our study, SKS experiments were performed with a constant signal size, whereas Zhang *et al.* used projected ellipsoid signals ranging from 1 mm $\times 1$ mm to 7.5 mm$\times$3 mm, introducing much more uncertainty about the actual signal size. Furthermore, our last experiment with signal size ranging from 5.5 to 9.5 mm confirmed that introducing signal size uncertainty lowered the detection performance of the observers compared with the SKS experiments with fixed size, especially for the largest masses (comparison points in Fig. 4, $p=0.51$ for 6.5 mm masses, $p=0.01$ for 9.5 mm). When mass sizes are mixed, the lower-bending performance curve for the largest masses is similar to the results of Judy *et al.* [10] with disk signals on correlated noise, or to more general findings with contrast-detail experiments by Burgess *et al.* with mammograms or power-law noise [4,5] for SKE tasks and search within a defined area. Judy *et al.* [10] also compared SKE with SKS experiments and showed that size uncertainty degraded human observer performance mainly for the largest disks with diameters larger than 1 cm, which seems consistent with our findings in Fig. 4, as far as the two studies can be compared. Judy *et al.* indeed used different contrasts for the different signal sizes to maintain a constant nonprewhitening matched filter observer performance, whereas we used a fixed signal contrast for all mass sizes.

In our study, the effect of signal shape uncertainty was investigated with CLB and real images, while the size uncertainty experiment was performed only with the CLB. However, the other results with benign masses (Subsection 4.A, last paragraph) suggest that size uncertainty experiments with real backgrounds should lead to the same conclusions, since the detection performance for SKE and SKS tasks with benign masses was very similar for both background types.

C. Model Observers versus Human Observers

The RMSE presented in Table 2, which take into account SKE and SKS tasks, show that most models can fit human results with appropriate internal noise level adjustments. The only exception is the NPW model: without internal noise, it performs better than human observers. This result may be surprising at first when compared with previous reports in literature, which typically find that the NPW performs worse than human observers in anatomical backgrounds. The apparent discrepancy is the consequence of the fact that unlike previous studies [4,9,12], in the current experiments the local means for all signal-absent images were matched over an area close to the size of the largest signals to ensure comparable conditions between the real and the synthetic backgrounds (see Subsection 2.B). The NPW model is highly degraded by variations of the mean local luminance, and thus the preprocessing that matches the means of local background areas highly improves the NPW model's performance. However, even with the optimal internal noise level, the difference between the performance of NPW model and human results is much larger than for the other models. As in our previous study with phantom masses on real and synthetic backgrounds [9], this suggests that human observers' detection strategy is more complex than that of the basic NPW model.

The addition of the eye filter (NPWE model) has a major effect on the ability of the model to predict human results. As illustrated in Fig. 7, the NPWE model acts mostly as an edge enhancement filter. The internal noise level that has to be added to this model to match human results is greater for benign ($p_n=1.4$) than for malignant ($p_n=0.8$) masses. This is probably due to the edge enhancement being more efficient with benign masses, which intrinsically possess much sharper edges than malignant masses. With the optimal values for the internal noise level, the NPWE model is a very good predictor of human results for all tasks. The best way to incorporate internal noise into this model for the SKS tasks is to add the noise to the individual template responses. This may be because the J templates corresponding to the J possible signals are quite different one from each other since they are essentially enhancing the signal edges: combining the J likelihoods brings useful information to the model even if the signals are similar in size, and adding noise in the individual responses appeared to be the best solution to lower the performance of the NPWE observer to match that of the generic human observer. The other methods for noise addition in the SKS tasks (maximum of the logarithm of the likelihood and single SKE template) lead to a performance that is below that of humans: if the NPWE is no longer able to combine the individual responses, it seems to be much less efficient in SKS tasks.

Results for the CH model observers can be divided into two classes. For the SQR, SDOG, and DDOG channels, the performance without any internal noise was close to that of human observers, or slightly below. As human observers are known to be subject to internal noise [1], these models cannot fully account for human observers' decision processes for the current tasks. The low performance of these models is likely related to the information reduction through the preprocessing of the image by a small number of channels that might not fully capture the important signal features. For these channelized models (without internal noise), the use of the sum-of-likelihoods rule, the maximum likelihood rule, or the SKE template led to very similar results in SKS tasks (see Table 2).

In contrast to the SQR, SDOG, and DDOG channel models, the CH observer defined with Gabor function channels performed much better than human observers.

Performance of the CH-Gabor model was close to perfect ($P_c = 100\%$) for all signal and background combinations studied in this paper. Adding internal noise lowered the performance of the CH-Gabor model down to the level of human observers for SKE tasks [Figs. 6(a) and 6(b)]. The internal noise level that had to be added to the CH-Gabor model was higher for malignant ($p_n = 2.2$) than for benign masses ($p_n = 0.5$). This result was opposite from what we found for the NPWE model ($p_n = 0.8$ for malignant masses, 1.4 for benign masses). The dissociation in results might relate to the fact that, unlike the NPWE, the CH with Gabor function templates do not specifically emphasize the signal edges, but rather extend over the whole area covered by the signal (see, for example, Fig. 7). The malignant masses have higher luminance contrast than the benign ones. Thus, templates with spatial integration areas that extend over the whole signal area might be particularly efficient for malignant mass detection. This is the case of the CH observer with Gabor functions model, which is particularly efficient with malignant masses and requires more internal noise than other models to match human observers' performance. As for the NPW model, it is likely that our results might change if we did not normalize the images to match the local mean luminance of the backgrounds, since the process of matching the local means removes some of the low-frequency noise.

In relation to the SKS tasks, if the SKE template only is used even though the signals are randomly chosen (third internal noise scheme in section 2.G), the CH observer with Gabor function channels is one of the best models in matching human performance level (see Table 2). For the other internal noise addition methods (sum of likelihoods with noise in the individual responses, and maximum likelihood rule), a problem appears with SKS tasks. The internal noise is integrated over the J templates when combining the individual responses, which leads to a better performance than in SKE tasks. This difference is especially large for the 9.5 mm masses, and may be explained by the J templates for this model being very similar, since they essentially use information contained at the center of the potential location. For this reason, the $RMSE_{\mathrm{overall}}$, which takes into account SKE and SKS tasks, is much higher than the RMSE computed for SKE tasks only with these two internal noise schemes.

Finally, we evaluated a model (HLT) that uses a template estimated directly from observer choices for the set of test images. The estimation of the HLT assumes that human observers use only a single template per task for both the SKE and SKS conditions. This is likely a simplification of the strategy used by human observers but might still capture some important aspects of human performance. Estimation of specific templates for each of the signals presented in the SKS task is not plausible for our study because statistical considerations would require many more signal specific trials to obtain stable estimates of the individual HLTs. We thus computed only one overall template per task, acknowledging that this constituted a simplified analysis. The hypothesis of a single human template per task was further supported *a posteriori* by the results, which show that the HLT model predicts human results remarkably well for all experimental conditions. The $RMSE_{\mathrm{overall}}$ (0.11 for benign masses, 0.07 for malignant masses, with the same p_n value equal to 1) is the lowest of all models.

The spatial profiles of the HLT for the 6.5 mm masses (see Fig. 7) show that human observers' strategy was essentially concentrated on the signal edges for the benign masses with real and synthetic backgrounds and malignant masses with real backgrounds. This is perfectly consistent with our previous study with a phantom mass with relatively sharp borders embedded in real and synthetic backgrounds [9]. However, the template for malignant masses with CLB suggests that the observers concentrated more on the central part of the signal location. This may explain why their performance in detecting malignant masses was not as good with the CLB as with the real backgrounds. This assumption should be taken with care, since individual detection strategy may vary across human observers: the individual templates shown in Fig. 8 suggest that for the corresponding task all observers focused on signal edges as in our previous study [9], but also that two of them used information contained at the center of the potential signal location. Further analysis in the spatial or frequency domains was not carried out in this study for two reasons: first, the wide range of tasks and the number of observers would have led to computing time issues, since every HLT estimation had to be repeated ten times. Second, using radial-averaged parameters to reduce the analysis to one dimension [9] would have been problematic with signals that are not circularly symmetric. Objective comparison of the templates across models, tasks, and individual observers, should be explored in further work.

5. CONCLUSIONS

By conducting detection experiments with realistic benign and malignant breast masses superimposed on real mammographic backgrounds and realistic, second-generation CLB, we were able to study the influence of signal variations and uncertainty on human observers' detection performance. We showed that human observers' performance did not differ significantly between SKE and SKS tasks when the signal size was kept constant. However, human observers were sensitive to signal size uncertainty, and their performance diminished between fixed-size and size-uncertainty experiments, especially for the largest masses. Following this idea, assessing human observer detection performance for such nontrivial signals as benign or malignant masses would already be possible with a limited set of signals covering the size range of interest: there would be no need to use sets with large numbers of signals covering the possible orientations and shapes in psychophysical studies.

Excellent agreement with human observers was obtained for NPWE, CH observer with Gabor function channels, and HLT models with adapted internal noise levels. The NPW observer appears to be too simplistic to correctly model human results. The performance of the other CH observers (SQR, SDOG, DDOG channels) seems to be too low to correctly model human decision processes for SKE and SKS tasks. Even if the detection performance level between humans and models has been matched with success in this study, further work is still needed for objectively comparing the different model observer templates and the HLT to study not only the performance level, but also the similarity or differences in detection strategies between human observers and models.

Finally, one has to keep in mind that the SKS approach (or SKEV, for which results have been shown to be highly correlated with SKS [19,21,22]) is still far from the actual clinical situation. Many other factors influence the radiologist's ability to correctly detect masses on mammograms: much wider search space, signal location uncertainty, and extremely low prevalence of the order of 7 per 1000 cases [48,49], for example. Moreover, real clinical strategies also include comparison with the contra-lateral breast and global breast architecture. However, these elements are currently beyond the scope of most psychophysical experiments and would lead to overly hard to interpret results. For this reason, the limitations of the current study (square regions of interest instead of whole breast, 2-AFC, controlled signal location) are still necessary to investigate human observer detection strategies and performance.

ACKNOWLEDGMENTS

The authors thank Cécile Althaus, Gaëlle Bocksberger, and Valentine Clémence for their participation in the detection experiments. This work was supported by Swiss National Science Foundation (SNSF) grant 320000-113863/1 and National Institutes of Health (NIH) 53455.

REFERENCES

1. A. E. Burgess, R. F. Wagner, R. J. Jennings, and H. B. Barlow, "Efficiency of human visual signal discrimination," Science **214**, 93–94 (1981).
2. A. E. Burgess and H. Ghandeharian, "Visual signal detection. I. Ability to use phase information," J. Opt. Soc. Am. A **1**, 900–905 (1984).
3. K. J. Myers, H. H. Barrett, M. C. Borgstrom, D. D. Patton, and G. W. Seeley, "Effect of noise correlation on detectability of disk signals in medical imaging," J. Opt. Soc. Am. A **2**, 1752–1759 (1985).
4. A. E. Burgess, F. L. Jacobson, and P. F. Judy, "Human observer detection experiments with mammograms and power-law noise," Med. Phys. **28**, 419–437 (2001).
5. A. E. Burgess and P. F. Judy, "Signal detection in power-law noise: effect of spectrum exponents," J. Opt. Soc. Am. A **24**, B52–B60 (2007).
6. J. P. Rolland and H. H. Barrett, "Effect of random background inhomogeneity on observer detection performance," J. Opt. Soc. Am. A **9**, 649–658 (1992).
7. F. O. Bochud, C. K. Abbey, and M. P. Eckstein, "Statistical texture synthesis of mammographic images with clustered lumpy backgrounds," Opt. Express **4**, 33–43 (1999).
8. C. Castella, K. Kinkel, F. Descombes, M. P. Eckstein, P.-E. Sottas, F. R. Verdun, and F. O. Bochud, "Mammographic texture synthesis: second-generation clustered lumpy backgrounds using a genetic algorithm," Opt. Express **16**, 7595–7607 (2008).
9. C. Castella, C. K. Abbey, M. P. Eckstein, F. R. Verdun, K. Kinkel, and F. O. Bochud, "Human linear template with mammographic backgrounds estimated with a genetic algorithm," J. Opt. Soc. Am. A **24**, B1–B12 (2007).
10. P. F. Judy, M. F. Kijewski, and R. G. Svensson, "Observer detection performance loss: target-size uncertainty," Proc. SPIE **3036**, 39–47 (1997).
11. C. K. Abbey and M. P. Eckstein, "Maximum-likelihood and maximum-a-posteriori estimates of human-observer templates," Proc. SPIE **4324**, 114–122 (2001).
12. C. K. Abbey, M. P. Eckstein, S. S. Shimozaki, A. H. Baydush, D. M. Catarious, and C. E. Floyd, "Human observer templates for detection of a simulated lesion in mammographic images," Proc. SPIE **4686**, 25–35 (2002).
13. M. P. Eckstein, A. J. Ahumada, and A. B. Watson, "Image discrimination models predict signal detection in natural medical image backgrounds," Proc. SPIE **3016**, 44–56 (1997).
14. H. H. Barrett and K. J. Myers, *Foundations of Image Science* (Wiley, 2004).
15. M. P. Eckstein, C. K. Abbey, and F. O. Bochud, "A practical guide to model observers for visual detection in synthetic and natural noisy images," in *Handbook of Medical Imaging*, Vol. 1, Physics and psychophysics, J. Beutel, H. L. Kundel, R. L. Van Metter, eds. (SPIE Press, 2000), pp. 593–628.
16. D. S. Brettle, E. Berry, and M. A. Smith, "The effect of experience on detectability in local area anatomical noise," Br. J. Radiol. **80**, 186–193 (2007).
17. R. Saunders and E. Samei, "Characterization of breast masses for simulation purposes," Proc. SPIE **5372**, 242–250 (2004).
18. R. Saunders, E. Samei, J. Baker, and D. Delong, "Simulation of mammographic lesions," Acad. Radiol. **13**, 860–870 (2006).
19. M. P. Eckstein and C. K. Abbey, "Model observers for signal-known-statistically tasks (SKS)," Proc. SPIE **4324**, 91–102 (2001).
20. M. P. Eckstein, B. Pham, and C. K. Abbey, "Effect of image compression for model and human observers in signal-known-statistically tasks," Proc. SPIE **4686**, 13–24 (2002).
21. M. P. Eckstein, Y. Zhang, B. Pham, and C. K. Abbey, "Optimization of model observer performance for signal known exactly but variable tasks leads to optimized performance in signal known statistically tasks," Proc. SPIE **5034**, 123–134 (2003).
22. Y. Zhang, B. P. Pham, and M. P. Eckstein, "Task-based model/human observer evaluation of SPIHT wavelet compression with human visual system-based quantization," Acad. Radiol. **12**, 324–336 (2005).
23. C. Castella, K. Kinkel, M. P. Eckstein, C. K. Abbey, F. R. Verdun, R. S. Saunders, E. Samei, and F. O. Bochud, "Mass detection on mammograms: signal variations and performance changes for human and model observers," Proc. SPIE **6917**, 69170K (2008).
24. A. E. Burgess, "Statistically defined backgrounds: Performance of a modified nonprewhitening observer," J. Opt. Soc. Am. A **11**, 1237–1242 (1994).
25. National Electrical Manufacturers Association, *Digital Imaging and Communications in Medicine (DICOM) Part 14: Grayscale Display Standard Function* (NEMA, 2000).
26. S. Muller, "Full-field digital mammography designed as a complete system," Eur. J. Radiol. **39**, 25–34 (1999).
27. S. Vedantham, A. Karellas, S. Suryanarayanan, D. Albagli, S. Han, E. J. Tkaczyk, C. E. Landberg, B. Opsahl-Ong, P. R. Granfors, I. Levis, C. J. D'Orsi, and R. E. Hendrick, "Full breast digital mammography with an amorphous silicon-based flat panel detector: Physical characteristics of a clinical prototype," Med. Phys. **27**, 558–567 (2000).
28. M. Heath, K. Bowyer, D. Kopans, R. Moore, and W. P. Kegelmeyer, "The Digital Database For Screening Mammography," in *Proceedings of the Fifth International Workshop on Digital Mammography*, M. J. Yaffe, ed (Medical Physics Publishing, 2001), pp. 212–218.
29. C. J. D'Orsi, *Illustrated Breast Imaging Reporting and DATA System (BIRADS)* (American College of Radiology, 1998).
30. A. E. Burgess, "On the noise variance of a digital mammography system," Med. Phys. **31**, 1987–1995 (2004).
31. A. E. Burgess, "Comparison of receiver operating characteristics and forced choice observer performance measurement methods," Med. Phys. **22**, 643–655 (1995).
32. P. G. J. Barten, "The SQRI method: a new method for the evaluation of visible resolution on a display," Proc. S.I.D.

28, 253–262 (1987).

33. K. J. Myers and H. H. Barrett, “Addition of a channel mechanism to the ideal-observer model,” J. Opt. Soc. Am. A **4**, 2447–2457 (1987).
34. C. K. Abbey and H. H. Barrett, “Human- and model-observer performance in ramp-spectrum noise: effects of regularization and object variability,” J. Opt. Soc. Am. A **18**, 473–487 (2001).
35. A. J. Ahumada, Jr., “Classification image weights and internal noise level estimation,” J. Vision **2**, 121–131 (2002).
36. J. A. Solomon, “Noise reveals visual mechanisms of detection and discrimination,” J. Vision **2**, 105–120 (2002).
37. R. F. Murray, P. J. Bennett, and A. B. Sekuler, “Optimal methods for calculating classification images: Weighted sums,” J. Vision **2**, 79–104 (2002).
38. C. K. Abbey, M. P. Eckstein, S. S. Shimozaki, A. H. Baydush, D. M. Catarious, and C. E. Floyd, “Human-observer templates for detection of a simulated lesion in mammographic images,” Proc. SPIE **4686**, 25–36 (2002).
39. C. K. Abbey and M. P. Eckstein, “Classification image analysis: Estimation and statistical inference for two-alternative forced-choice experiments,” J. Vision **2**, 66–78 (2002).
40. C. K. Abbey and M. P. Eckstein, “Optimal shifted estimates of human-observer templates in two-alternative forced choice experiments,” IEEE Trans. Med. Imaging **21**, 429–440 (2002).
41. B. D. Gallas, G. A. Pennello, and K. J. Myers, “Multireader multicase variance analysis for binary data,” J. Opt. Soc. Am. A **24**, B70–B80 (2007).
42. F. O. Bochud, C. K. Abbey, and M. P. Eckstein, “Visual signal detection in structured backgrounds. III. Calculation of figures of merit for model observers in statistically nonstationary backgrounds,” J. Opt. Soc. Am. A **17**, 193–205 (2000).
43. J. Eng, “ROC analysis: web-based calculator for ROC curves,” http://www.jrocfit.org.
44. Y. Zhang, B. T. Pham, and M. P. Eckstein, “Evaluation of internal noise methods for Hotelling observer models,” Med. Phys. **34**, 3312–3322 (2007).
45. Y. Zhang, B. T. Pham, and M. P. Eckstein, “Automated optimization of JPEG 2000 encoder options based on model observer performance for detecting variable signals in X-ray coronary angiograms,” IEEE Trans. Med. Imaging **23**, 459–474 (2004).
46. R. Lowry, “VassarStats: Web Site for Statistical Computation,” http://faculty.vassar.edu/lowry/VassarStats.html.
47. Y. Zhang, C. K. Abbey, and M. P. Eckstein, “Adaptive detection mechanisms in globally statistically nonstationary-oriented noise,” J. Opt. Soc. Am. A **23**, 1549–1558 (2006).
48. N. Perry, M. Broeders, C. de Wolf, S. Törnberg, R. Holland, and L. von Karsa, *European Guidelines for Quality Assurance in Breast Cancer Screening and Diagnosis*, 4th ed. (Office for Official Publications of the European Communities, 2006).
49. Y. Jiang, D. L. Miglioretti, C. E. Metz, and R. A. Schmidt, “Breast cancer detection rate: designing imaging trials to demonstrate improvements,” Radiology **243**, 360–367 (2007).

Masses detection in breast tomosynthesis and digital mammography: a model observer study

C. Castella[a,b], M. Ruschin[c], M. P. Eckstein[d], C. K. Abbey[d], K. Kinkel[e], F. R. Verdun[a], A. Tingberg[f], F. O. Bochud[a*]
[a]Univ. Institute for Radiation Physics, CHUV and Univ. Lausanne, CH-1007 Lausanne, Switzerland;
[b]LPHE, EPFL, CH-1015 Lausanne Switzerland;
[c]Dept. of Radiation Physics, Princess Margaret Hospital, Toronto On, M5G 2M9, Canada;
[d]Dept. of Psychology and Inst. of Collaborative Biotechnology, University of California, Santa Barbara, CA 93106-9660, USA;
[e]Clinique des Grangettes, Chemin des Grangettes 7, CH–1224 Chêne-Bougeries, Switzerland;
[f]Dept. of Medical Radiation Physics, Lund University, Malmö University Hospital, SE-205 02, Malmö, Sweden

ABSTRACT

In this study, we adapt and apply model observers within the framework of realistic detection tasks in breast tomosynthesis (BT). We use images consisting of realistic masses digitally embedded in real patient anatomical backgrounds, and we adapt specific model observers that have been previously applied to digital mammography (DM). We design alternative forced-choice experiments (AFC) studies for DM and BT tasks in the signal known exactly but variable (SKEV) framework. We compare performance of various linear model observers (non-prewhitening matched filter with an eye filter, and several channelized Hotelling observers (CHO) against human.

A good agreement in performance between human and model observers can be obtained when an appropriate internal noise level is adopted. Models achieve the same detection performance across BT and DM with about three times less projected signal intensity in BT than in DM (humans: 3.8), due to the anatomical noise reduction in BT. We suggest that, in the future, model observers can potentially be used as an objective tool for automating the optimization of BT acquisition parameters or reconstruction algorithms, or narrowing a wide span of possible parameter combinations, without requiring human observers studies.

Keywords: Model observers, observer performance evaluation, image perception, breast tomosynthesis, digital mammography

1. INTRODUCTION

Several studies have demonstrated that detection tasks in mammography are mainly limited by the superposition of anatomical structures on the projected images [1,2]. This masking effect, known as anatomical noise, lowers both sensitivity and specificity of mammography, by hiding abnormalities or creating suspicious structures in the projected image. On the other hand, emerging breast tomosynthesis (BT) technique has been reported to lead to excellent results in detection experiments involving human observers [3-6]. The good performance of the observers has been typically explained by a reduction of the anatomical noise in the three-dimensional breast reconstruction, allowing the human observers to isolate the lesions easier than with mammograms.

However, many aspects of BT still need to be optimized: acquisition techniques (tube load, mean glandular dose, number of projections, angular scanning span), reconstruction and filtering algorithms, or image display [7]. Conducting comparison studies involving radiologists and/or medical physicists is time consuming and hardly practical, since the number of different combinations of free parameters is considerably large. An alternative to these time-consuming

---

* francois.bochud@chuv.ch; phone +41 21 623 34 34; www.chuv.ch/ira

sessions of repetitive psychophysical experiments is to use objective model observers that mimic human decisions and that can be run on computers with large sets of data.

Model observers, which have been developed and tested with success in various applications for mammography, projection radiography, or computed tomography [8,9], are still in the early stage of development in tomosynthesis. A recent study by Gifford *et al.* investigated a scanning noiseless channelized Hotelling Observer and compared different number of projections and angular span combinations [10]. Reiser *et al.* compared filtered-backprojection and iterative maximum-likelihood expectation maximization reconstruction methods with a prewhitening observer in a simplified detection task with a spherical signal in a homogeneous phantom [11]. In another study, Pineda *et al.* used a channelized Hotelling and non-prewhitening model observers with and without eye filters for optimizing a tomosynthesis system for the detection of lung nodules [12].

The purpose of this study is to adapt and validate model observers in a realistic BT framework, and compare the relative performance of the models with digital mammography (DM) and BT. For this, we use images consisting of realistic masses digitally embedded on real patient anatomical backgrounds, for which human observers performance has already been characterized [4], and adapt specific model observers that have been used for modeling DM tasks in previous studies [13,14].

In the original study with these hybrid images, Ruschin *et al.* [4] conducted 4-alternative forced-choice (4-AFC) detection experiments, and determined which signal contrast level led to the same performance between DM and BT. The authors concluded that significantly less (about one fourth) signal contrast was needed for BT, suggesting a good potential for dose reduction in BT, compared to DM. The present study aims at verifying whether models can predict human detection performance in the same conditions.

2. MATERIAL AND METHODS

2.1 Hybrid images: patients backgrounds and realistic signals

The digital mammography (DM) and breast tomosynthesis (BT) hybrid image set was the same as in the original study with human observers [4]. They were constructed by digitally embedding realistic breast masses [15,16] into signal-absent background images.

Following approval by the local radiation protection committee and informed consent by the patients involved in this study, thirty patients underwent breast examinations with both DM and BT. This way, the case database was made up of identical patients for both modalities.

The BT images were acquired with a prototype unit adapted from the DM Mammomat Novation (Siemens, Erlangen, Germany) [17]. The beam quality in tomosynthesis mode was the same as the one determined by the automatic exposure control device in DM mode. For tomosynthesis acquisition, 25 projection images per examination were taken over an angular span of about 50 degrees, resulting in a total tube load twice as high as for a single-view DM acquisition. Out of the 25 projections, 13 only were used for reconstructing the breast volume, in order to keep the total dose to the breast contributing to the final images approximately the same between BT and DM. The detector pixel size was 70 μm for DM, and 85 μm for BT.

The simulated tumors were adapted from 2D lesions used in previous studies [15,16]. Twenty different lesions with mean x and y dimensions of 8.4 mm (range: 7.4-8.8 mm) and 6.6 mm (5.3-7.8 mm) were generated for the study. Each 2D lesion was mapped to an ellipsoid with a length in z dimension of 5 mm to generate the 3D tumor. An example of simulated lesion is given in Fig. 1. The radial spatial and frequency profiles have been expressed in mm and mm^{-1}, respectively, in order to allow for comparing the profiles despite the different pixel sizes.

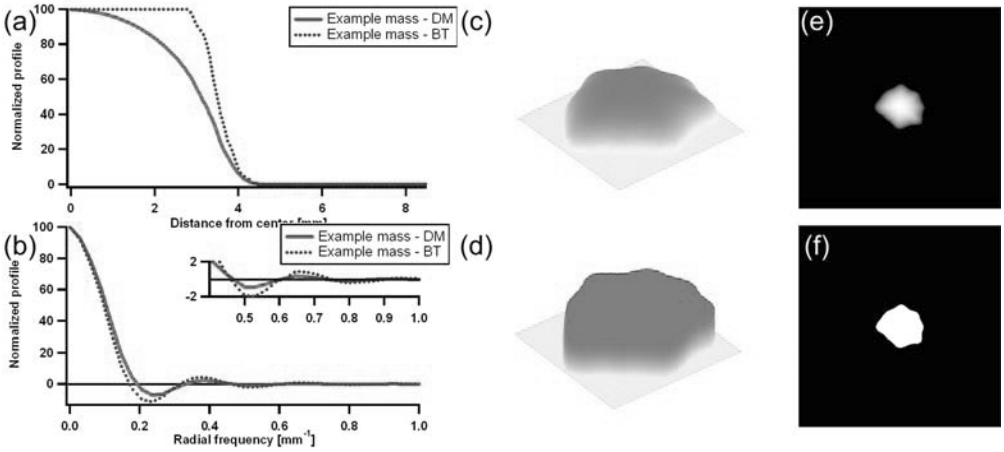


Fig. 1. Left. Example of signal used in the psychophysical tasks. (a) Radial profile in the spatial domain (b) Radial profile in the frequency domain. Center. 3D views of the pixel value intensity of the same lesion in digital mammography (DM) (c) and breast tomosynthesis (BT) (d) images. Right. 2D view of the corresponding reference signals in the DM (e), and BT (f) psychophysical detection tasks.

The complete description of the embedding method for DM and BT can be found in the original study with human observers [4]. Assuming a constant difference in attenuation coefficient $\Delta\mu$ between the lesion and the surrounding breast tissue, the individual BT projections and the DM projection on the detector plane were altered by modeling the attenuation of the primary x-ray beam using geometrical descriptions of the imaging units. Each pixel (i, j) value of the *k-th* altered projection was thus given by:

$$\mathrm{Im}_{a,k}(i,j) = \mathrm{Im}_{o,k}(i,j)\exp\left(-t_k(i,j)\Delta\mu\right), \tag{1}$$

where $\mathrm{Im}_{o,k}$ and $\mathrm{Im}_{a,k}$ are respectively the original and altered *k-th* projections, and t_k is the integrated tumor thickness along the x-ray trajectories for the (i, j) pixel of the *k-th* projection. The BT altered projections were then used for reconstructing the breast volume, while the final signal-present images for DM were the images obtained by Eq. (1).

The signal intensity of a simulated lesion, S, was defined for both modalities as the relative increase in total attenuation resulting from the addition of the lesion to the central (k=0) projection. It was computed as the root mean square of the projected signal:

$$S = \sqrt{\frac{1}{NM}\sum_{i=1}^{M}\sum_{j=1}^{N}\left(t_{k=0}(i,j)\Delta\mu\right)^2} = \Delta\mu\sqrt{\frac{1}{NM}\sum_{i=1}^{M}\sum_{j=1}^{N}\left(t_{k=0}(i,j)\right)^2} \tag{2}$$

In Eq.(2), M and N are the numbers of elements in the x and y directions, respectively. $\Delta\mu$ values were computed from Eq. (2) in order to yield a desired value of S, and then used in Eq. (1) to alter the detector pixel values.

On the original patient DM and BT data (without tumor addition), 20 non-overlapping signal-absent regions of interest (ROI) per patient were manually selected within the breast volume and the contra-lateral breast, avoiding any suspicious location within the breast. The ROI were square 34x34 mm areas, which translates to 400x400 pixels for BT, and 486x486 pixels for DM, due to the difference in detector pixel size. These images served as signal-absent ROI in the psychophysical studies. For signal-present images and for computational time reasons, a set of 60 embedded signals (30 patients x 2 masses) was created for the BT study, while the embedding of the synthetic masses could be performed in real-time during the psychophysical study for DM. The signal-present ROI were centered on the simulated lesions in *x*, *y*, and *z* directions in the reconstructed breast volumes (BT), or on the *x-y* lesion position (DM).

Before being displayed to the observers, we applied a logarithmic function to the DM ROI, and the look-up table was inverted, such that denser regions appeared brighter [18]. Finally, BT and DM ROI were windowed so that the mean value of the images corresponded to the middle of the dynamic range of the 8-bit scale of the display screen.

2.2 4-Alternative Forced-Choice tasks with human and model observers

Using the signal-present and signal-absent images described in section 2.1, 4-AFC psychophysical experiments were designed in order to compare BT and DM modalities. The tasks consisted in 60 trials per experimental condition. On each trial, four ROI that had been randomly chosen from four different patients (in order to minimize possible correlations across locations) were presented to the human and used for the model observers. Out of these four ROI, three were signal-absent and one was one of the 60 signal-present cases. The observers were asked to indicate the image that they estimated as the most likely to contain the signal. In addition to the four ROI, the observers were given at each trial a high-contrast reference image (see Fig. 1-e and –f) of the actual signal. This approach is known as Signal-Known-Exactly but variable (SKEV) task. It allows for testing the observers' responses to a variety of signals, while keeping the analysis as straightforward as a Signal-Known-Exactly (SKE) task.

Nine observers participated to the human observer study: four radiologists and five medical physicists [4]. Due to the time needed to generate and reconstruct the tomosynthesis images, one contrast was studied (S=0.010), while four conditions (S=0.036, 0.042, 0.048, 0.054) were tested for DM.

We obtained model observer performance from sample driven Monte-Carlo simulations of 4-AFC, using the exact same images, signal contrasts, and ROI selection procedure as for human observers. For the BT task, the 60-trial experiment at S=0.010 was repeated 100 times with the signal-present cases being randomly associated with different signal-absent ROI. For the DM task, the real-time addition of the lesions to the backgrounds at the desired intensity allowed to repeat 100 realizations of the 60-trial task for values of S equal to 0.010, 0.025, 0.036, 0.042, 0.048, and 0.054.

Figure of merit for both human and model observers was the detectability, *d'*. First, the percentage of correct answers P_c for a given observer and contrast was calculated from the corresponding 60-trial experiment outcomes o_i as:

$$P_c = \frac{1}{60}\sum_{i=1}^{60} o_i \, , \tag{3}$$

where $o_i = 1$ if the signal-present image had been chosen by the observer at the *i*-th trial, and 0 otherwise. P_c was then converted to an empirically obtained index of detectability *d'*, by generating a look-up table for P_c versus *d'* from the usual cumulative Gaussian relationship, under the assumption that the variances of the responses to the signal-present and signal absent locations are identical [9,19]:

$$P_c(d', M) = \int_{-\infty}^{\infty} \phi(x - d')[\Phi(x)]^{M-1} dx \tag{4}$$

In Eq.(4), M=4 is the number of alternatives in the AFC task, $\phi(x) = \frac{1}{\sqrt{2\pi}}\exp(-x^2/2)$, and $\Phi(x) = \int_{-\infty}^{x} \phi(y)dy$ is the cumulative Gaussian distribution function. The transformation from P_c to *d'* under the assumption of statistically independent responses can be justified by the fact that the four backgrounds presented at each trial had been extracted from four different patients.

2.3 Model observers

The decision variable of a general linear observer to an image $\mathbf{g}_i$ is given by the product between a template $\mathbf{w}$ and the image, both being expressed as 1-D vectors, with an optional internal noise term ε:

$$\lambda_i = \mathbf{w}^T \mathbf{g}_i + \varepsilon \tag{5}$$

For this study, we implemented various linear observers that have been previously used for modeling human observers' decisions for detection tasks in DM: non-prewhitening matched-filter (NPW) [8], NPW with an eye filter (NPWE) [20], channelized Hotelling (CHO) observer [9,21] with dense difference-of-Gaussians (DDOG) [14, 22] and Gabor functions channels [13,14,23]. All these models have been extensively described in the literature, and below we briefly summarize their analytical expressions.

For a given signal $\mathbf{s}$, the NPW and NPWE templates are defined respectively by:

$$\mathbf{w}_{\mathbf{NPW}} = \mathbf{s} \tag{6}$$

and

$$\mathbf{w}_{\mathbf{NPWE}} = \mathbf{E}^T \mathbf{E} \mathbf{s} \tag{7}$$

In Eq. (7), $\mathbf{E}(\rho) = \rho^n \exp(-c\rho^2)$ is an eye filter that accounts for human eye different sensitivity to radial frequency ρ. The parameters used for the eye filter (n=1.3, c=0.0041) are from Burgess [20].

The CHO templates are derived from the covariance matrix of the backgrounds seen through the channels $\mathbf{K}_{\mathrm{b,c}}$ as:

$$\mathbf{w}_{\mathbf{Hot}} = \mathbf{K}_{\mathrm{b,c}}^{-1} \mathbf{s} \tag{8}$$

The covariance matrices for BT and DM were estimated by sampling from the 600 signal-absent ROI $\mathbf{g_n}$ for each modality. They were computed as $\mathbf{K}_{\mathrm{b,c}} = \left\langle (\mathbf{T}^t \mathbf{g}_\mathbf{n} - <\mathbf{T}^t \mathbf{g}_\mathbf{n}>)(\mathbf{T}^t \mathbf{g}_\mathbf{n} - <\mathbf{T}^t \mathbf{g}_\mathbf{n}>)^t \right\rangle$, where the column vectors of the matrix T each represent the spatial profile of a channel. The DDOG channels were circularly symmetric functions defined in the Fourier domain as the difference of two Gaussians (12 channels in total), while Gabor channels were oriented Gabor functions in the spatial domain (5 orientations, 35 channels for DM, 40 for BT). The number of channels was chosen in order to maximize the models' performance, based on the same performance estimation procedure as described in section 2.2.

3. RESULTS

3.1 Model observer templates

Typical examples of 2D model observer templates are given in Fig. 2 for the BT task (upper row), and DM task (lower row). The templates are those derived for the signal given in Fig. 1. They are 2-D representations of **w** in Eq. (5).

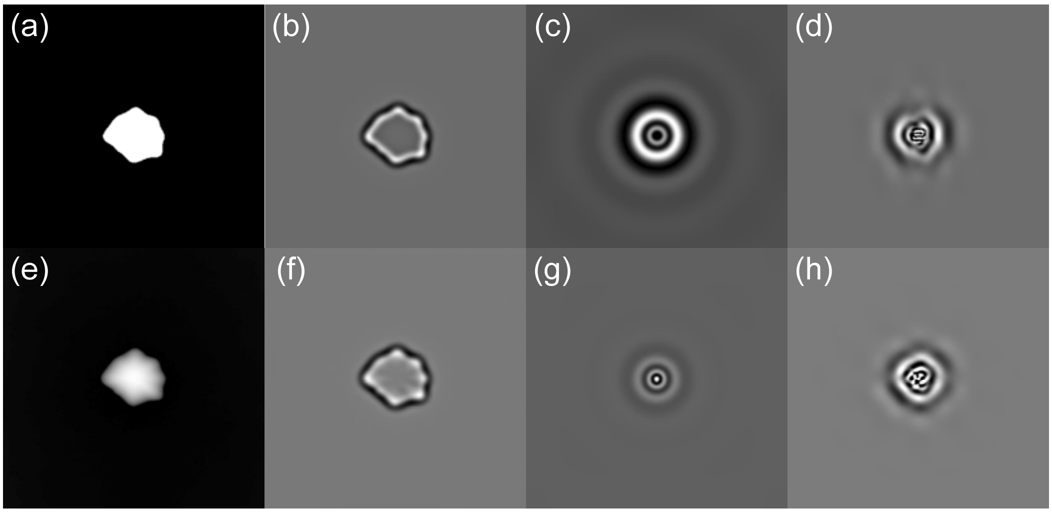


Fig. 2. Examples of model observers templates in the spatial domain, computed for the BT task (upper row) and DM task (lower row). (a, e): NPW, (b, f): NPWE, (c, g): CHO with DDOG channels, (d, h): CHO with Gabor channels

The radially averaged frequency profiles of the model observer templates shown in Fig. 2 are given in Fig. 3 for DM (left) and BT (right).

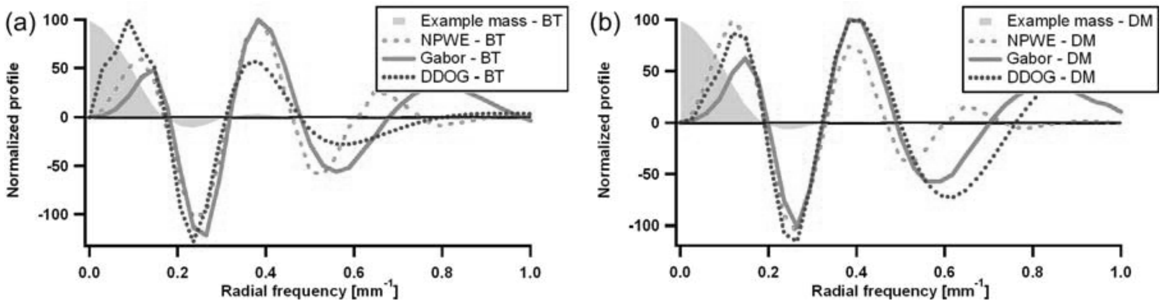
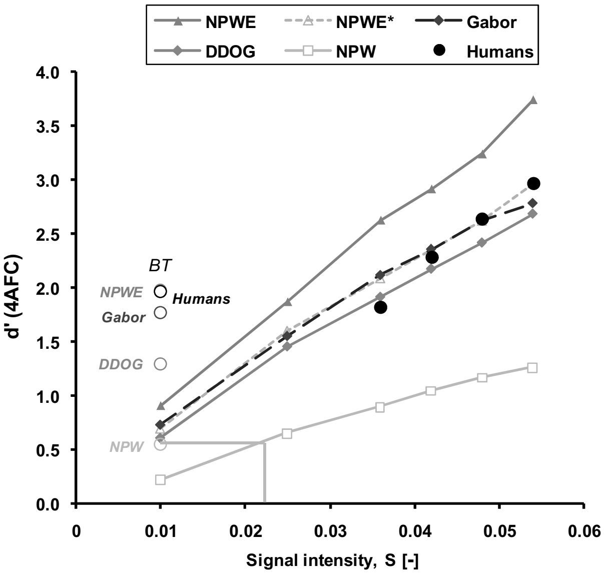


Fig. 3. Examples of radial frequency profiles of model observer templates for the BT (a) and DM (right) tasks. The signal profiles are drawn (filled blue areas) for comparison. The example mass is the same as in Fig. 1.

3.2 Models performance

Fig. 4 compares the performance of the model observers for the BT and DM tasks to that of human observers. Since most models appeared to have a performance level that was not better than that of the human observers for the studied detection tasks, the optional noise term ε in Eq.(5) was set to zero for nearly all conditions. There was one exception, however, for the NPWE model in DM. As its performance was constantly higher than humans, internal noise was added in this model response as a zero-mean Gaussian distributed random variable, which variance was proportional to the variance of the template response to the background-only images [8,9]. The variance was computed by sampling the response over the 600 DM backgrounds. The proportionality constant factor, p_n, was chosen in order to minimize the root mean square error (RMSE) over the four DM tasks between the model with internal noise and the human observers [14]. The NPWE model with such internal noise is indicated NPWE*, with a corresponding p_n of 0 for BT, and 0.6 for DM.

From the linear fits of the performance data presented in Fig. 4, the signal intensity ratio needed for obtaining the same performance in DM as in the BT task with a signal intensity of S=0.010 was computed.



Observer	R^2 correlation coefficient	Signal intensity ratio
NPWE	0.997	2.7
NPWE*	0.996	2.9
Gabor	0.991	3.0
DDOG	0.995	2.3
NPW	0.993	2.4
Humans	0.994	3.8

Fig. 4. 4-AFC digital mammography (DM) task performance of the model observers as a function of signal intensity S. For comparison, human results are indicated with filled circles, and performance in breast tomosynthesis (BT) task with open circles. For each model, the correlation coefficient for linear fits is indicated, as well as the signal intensity ratio needed in DM for matching the BT performance with a signal intensity of S=0.010 (see example given for NPW). NPWE* values correspond to the NPWE model with noise intensity that minimizes the error over the DM tasks performed by the human observers. Standard error bars are about the size of the symbols and have not been represented in the figures for clarity reasons.

For BT, the statistical significance of the differences between the models and the human observers for the S=0.010 detection task was assessed by a two-sided t-test (unequal sample size, unequal variances) on the average performance difference:

$$t = \frac{\left| < d'_h > - < d'_m > \right|}{\sqrt{\frac{\sigma_h^2}{\mathrm{n}_h} + \frac{\sigma_m^2}{\mathrm{n}_m}}}, \tag{9}$$

Where n_h=9 is the number of observers who participated to the psychophysical experiments, and n_m=100 is the number of 60-trial 4AFC repetitions performed for testing a given model. The number of degrees of freedom *d.f.* was given by:

$$d.f. = \frac{(s_h^2 / n_h + s_m^2 / n_m)^2}{(s_h^2 / n_h)^2 / (n_h - 1) + (s_m^2 / n_m)^2 / (n_m - 1)} \tag{10}$$

Note that using Eq. (9) and (10) assumes that all the variability is statistical, and that there is no bias due to the limited number of cases. Also, the inter-observer variance σ_h^2 was found to be about twice as large as σ_m^2 : with n_m about ten times larger than n_h, these two equations are thus dominated by the human variance terms.

For DM, the performance was compared using a two-way analysis of variance (ANOVA) performed for each model against human observers and using the four studied contrast conditions (S=0.036, 0.042, 0.048, and 0.054).

Additionally, the RMSE for the different models for BT (one task), DM (four tasks), and both modalities pooled together (five tasks), are given in Fig. 5. The RMSE provides an overall parameter for comparing human and models over several tasks.

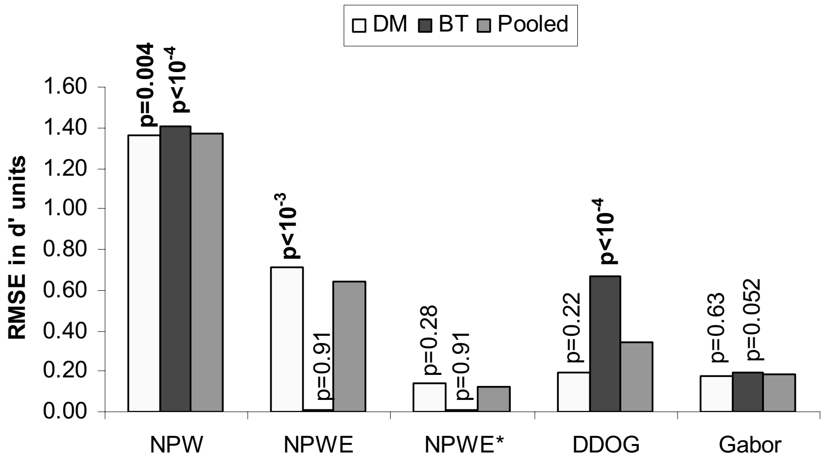


Fig. 5. Root mean square error (RMSE) between models and human observers for the DM and BT tasks. The last column for each model represents the pooled error for the BT task and the four DM tasks with different contrasts. NPWE* values correspond to the NPWE model with the noise intensity chosen to best match human performance.

4. DISCUSSION

4.1 Model observer templates

The model observers frequency profiles shown in Fig. 3 illustrate the fact that the models use very similar information in the Fourier domain for BT and DM. NPW template (which is by definition the signal to be detected in the psychophysical task) information is mainly concentrated at low frequencies, where anatomical noise is the highest [2], and human efficiency is limited [20]. The other models follow the first signal oscillations in the Fourier domain, and adapt their template to the different signals. Before discussing the features of each model observer template, it should be noted that the analysis of the circularly averaged radial profiles is by essence a simplification for such non-symmetrical signals.

The NPWE template acts as an edge-enhancement filter, as it is particularly visible in Fig. 2. This explains why BT and DM templates are so similar in the spatial domain, since their edge information (the projected shape of the area covered by the signal in the central projection) is the same. In the Fourier domain, slight differences are visible between the two modalities, reflecting the differences in the signal oscillations. Compared to NPW, the low frequencies are greatly filtered by the eye filter.

The channelized Hotelling model with DDOG channels has a clearly limited adaptability to the different signals, due to the use of circularly symmetric basis functions. While templates for DM and BT both exhibit stimulation/inhibition transitions around signal edges, the transition is much more marked for BT, since the reference signal is here a binary image with infinitely sharp edges. The stimulation area of the templates for DM is more concentrated at the very center of the signal, thus more prone to decision errors resulting from local high pixel values in the central part of the images. In the Fourier domain, the oscillations of the radial frequency profile of the templates follow those of the NPWE up to about 0.5 cyc/pixel. At higher frequencies, the limited number of basis functions introduces artifacts that are non-related to the signal.

The use of asymmetric Gabor channels in the channelized Hotelling model allows for a better but still limited ability to adapt to specific signals, as can be seen in Fig. 2. In the Fourier domain, the profile is again similar to that of the NPWE model for frequencies up to 0.5 cyc/mm, with less emphasis on the low frequencies due to the choice of the Gabor channels frequencies.

4.2 Models performance and comparison with human observers

The poor performance of the NPW model was expected and is particularly visible on Fig. 4 and Fig. 5. This basic model, although optimal in white noise, fails to correctly model human decision processes for detection tasks involving more clinically realistic backgrounds [9, 14,22].

With the addition of the eye filter and appropriate internal noise, the NPWE matches human performance level very well for BT (no significant difference, p=0.91) and DM (p=0.28) tasks. The RMSE computed over the five psychophysical conditions is the lowest of all models (Fig. 5), and comparable to the values obtained in a previous study by Castella *et al.* with benign and malignant masses embedded into digital mammograms [14]. Without internal noise, the NPWE significantly outperforms the human observers in the DM task ($p<10^{-3}$). This was not a surprise, since previous studies with synthetic signals embedded on filtered white noise [22], angiograms [24], or mammograms [13] had shown the same trend. The reason why NPWE model does not have a better performance than humans for the BT task may be due to the fact that the reference signal given to the models for the BT task was a binary image, physically equivalent to an infinitesimally thin slice through the signal, instead of the actual thickness of the reconstructed tomosynthesis image, 1mm [17]. This lack of information concerning the actual signal edges might have been misleading for the models, and have degraded their performance for the BT task. This should be further investigated in future work.

The two CHO models have different abilities for reproducing human observer results. With DDOG channels, the systematic underperformance can probably be attributed to the strong restriction caused by the choice of circularly symmetric channels for such complex signals. The difference is especially large for the BT task (0.67 in *d'* units, $p<10^{-4}$), while systematic but not significant for DM (p=0.22). With Gabor channels, the model leads to a better match for DM (p=0.63) and an excellent overall RMSE, although performing nearly significantly below human observers in BT (p=0.052). This underperformance in BT may be due, like for the NPWE model, to the imprecision of the signal edges used to the model, or to a suboptimal choice of channels. All Gabor functions having rather smooth variations, the model is not as efficient as the NPWE for optimally matching the stimulation/inhibition regions around the signal edges in BT.

4.3 Relative performance comparison between BT and DM

One key point of the present study was to assess the ability to correctly reproduce the BT potential that had been observed in the human observer studies. For all these models, the signal intensity needed in DM for reaching the performance of models with the BT task at S=0.010 was 2.3 to 3 times higher. For the best two models, in particular (NPWE with internal noise, and CHO with Gabor channels), the ratio is 2.9 and 3.0, respectively. This result is slightly lower than what had been observed with the human observers (the signal intensity ratio was found to be about 3.8 [4]). However, given the uncertainties due to the fact that one condition only had been used for BT in both studies, this suggests that the potential of future developments in the BT field like reconstruction techniques, effect of angle span, or number of projections, may be assessed by model observer studies.

4.4 Limitations of the study

As mentioned in the human observer study by Ruschin *et al.* [4], the encouraging results obtained by this comparison of DM and BT for a database of matched clinical patient background images and synthetic 3D breast masses are still subject to further work in order to enhance clinical realism of the hybrid images, and to test more conditions.

First, signals in our approach have by definition a constant $\Delta\mu$ in Eq. (1). This facilitates the inclusion of synthetic signals on the DM or BT projections but could lead to unrealistic looking lesions, especially in regions containing heterogeneous areas consisting of fatty and glandular tissues. A more elaborate and clinically realistic approach would be to consider signal with constant μ instead, by adapting Eq. (1). This could be done using a model of the compressed breast, for example.

Second, this study is based on a restricted set of tomosynthesis reconstructed images only. As signals cannot be embedded "on-the-fly" for BT, the number of studied conditions is limited. Since the observers are presented the same signal-present images, this may lead to unwanted correlations, particularly when estimating the variance of the performance in the detection task. For future studies, it would be useful to generate other signal-present images, and/or test other signal sizes and contrasts.

Finally, one has to remember that we used ad hoc DM ROI-based processing before displaying the images to the observers. This approach is similar to that of Burgess [18], but may alter their performance. One alternative could be to add the signals to the pixel values of the mammograms, then process the mammograms as a whole with the manufacturer algorithm, before selecting the ROI.

5. CONCLUSIONS

The present study compared the relative performance of model observers in detection tasks in DM and BT with a set of realistic hybrid images created with the same set of patients and signals. Our results showed that the equivalent detection performance at reduced signal intensity (or, equivalently, reduced contrast) observed in a previous human observers study [4] could be reproduced these model observers, with an especially good match with human observer data for the NPWE model with internal noise, and CHO model with Gabor function channels.

These results confirm the potential advantage of BT compared to DM for improving the detection of subtle lesions at equivalent breast dose. Given that BT is still a technique under development, our results also show that model observers could be valuable for testing the effects of the technical parameters involved in the image acquisition on the detection performance.

ACKLOWLEDGMENTS

This work was supported by Swiss National Science Foundation grant 320000-113863/1.

REFERENCES

[1] F. O. Bochud, J. F. Valley, F. R. Verdun, C. Hessler, and P. Schnyder, "Estimation of the noisy component of anatomical backgrounds," Med. Phys. **26**, 1365-1370 (1999).

[2] A. E. Burgess, F. L. Jacobson, and P. F. Judy, "Human observer experiments with mammograms and power-law noise," Med. Phys. **28**, 419-437 (2001).

[3] H.-P. Chan, J. Wei, B. Sahiner, E. A. Rafferty, T. Wuo, M. A. Roubidoux, R. H. Moore, D. B. Kopans, L. M. Hadjiski, and M. A. Helvie, "Computer-aided Detection System for Breast Masses on Digital Tomosynthesis Mammograms: Preliminary Experience", Radiology **237**, 1075-1080 (2005).

[4] M. Ruschin, P. Timberg, T. Svahn, I. Andersson, B. Hemdal, S. Mattsson, M. Bath, and A. Tingberg, "Improved in-plane visibility of tumors using breast tomosynthesis," Proc. SPIE **6510**, 65101J (2007).

[5] S. Suryanarayanan, A. Karellas, S. Vedantham, S. J. Glick, C. J. d'Orsi, S. P. Baker, and R. L. Webber, "Comparison of tomosynthesis methods used with digital mammography", Acad. Radiol **7**, 1085-1097 (2000).

[6] X. Gong, S. J. Glick, B. Liu, A. A. Vedula, and S. Thacker, "A computer simulation study comparing lesion detection accuracy with digital mammography, breast tomosynthesis, and cone-beam CT breast imaging," Med. Phys. **33**, 1041-1052 (2006).

[7] J. T. Dobbins III and D. J. Godfrey, "Digital x-ray tomosynthesis: current state of the art and clinical potential," Phys. Med. Biol. **48**, R65-R106 (2003).

[8] H. H. Barrett and K. J. Myers. *Foundations of Image Science*, Wiley Series in Pure and Applied Optics, Hoboken, 2004

[9] M.P. Eckstein, C. K. Abbey and F. O. Bochud, "A practical guide to model observers for visual detection in synthetic and natural noisy images", *Handbook of medical imaging. Volume 1. Physics and psychophysics*, SPIE press, Bellingham, 2000

[10] H. C. Gifford, C. S. Didier, M. Das, and S. J. Glick, "Optimizing Breast-Tomosynthesis Acquisition Parameters with Scanning Model Observers," Proc. SPIE **6917**, 69170S (2008).

[11] I. Reiser, B. A. Lau, and R. M. Nishikawa, "Effect of Scan Angle and Reconstruction Algorithm on Model Observer Performance in Tomosynthesis," Proc. IWDM 2008, LNCS **5116**, 601-611 (2008).

[12] A. R. Pineda, S. Yoon, D. S. Paik, and R. Fahrig, "Optimization of a tomosynthesis system for the detection of lung nodules," Med. Phys. **33**, 1372-1379 (2006).

[13] C. Castella, C. K. Abbey, M. P. Eckstein, F. R. Verdun, K. Kinkel, and F. O. Bochud, "Human linear template with mammographic backgrounds estimated with a genetic algorithm," J. Opt. Soc. Am. A **24**, B1-B12 (2007).

[14] C. Castella, M. P. Eckstein, K. Kinkel, C. K. Abbey, F. R. Verdun, R. S. Saunders, E. Samei, and F. O. Bochud, "Mass detection on mammograms: influence of signal shape uncertainty on human and model observers," *to appear in J. Opt. Soc. Am. A*

[15] M. Ruschin, A. Tingberg, M. Bath, M. Hakansson, and I. Andersson, "Using simple mathematical functions to simulate pathological structures—input for digital mammography clinical trial," Radiat. Prot. Dosim. **114**, 424-431 (2005).

[16] R. S. Saunders, E. Samei, J. Baker, and D. Delong, "Simulation of mammographic lesions," Acad Radiol **13**, 860-870 (2006).

[17] M. Bissonnette *et al.*, "Digital breast tomosynthesis using an amorphous selenium flat panel detector," Proc. SPIE **5745**, 529-540 (2005).

[18] A. E. Burgess, "On the noise variance of a digital mammography system," Med. Phys. **31**, 1987-1995 (2004).

[19] F. O. Bochud, C. K. Abbey, M. P. Eckstein, "Visual signal detection in structured backgrounds. III. Calculation of figures of merit for model observers in statistically nonstationary backgrounds," J. Opt. Soc. Am. A **17**, 193-205 (2000).

[20] A. E. Burgess, "Statistically defined backgrounds: Performance of a modified non-pre whitening observer," J. Opt. Soc. Am. A 11, 1237-1242 (1994).

[21] K. J. Myers and H. H. Barrett, "Addition of a channel mechanism to the ideal-observer model," J. Opt. Soc. Am. A 4, 2447-2457 (1987).

[22] C. K. Abbey and H. H. Barrett, "Human- and model-observer performance in ramp-spectrum noise: effects of regularization and object variability," J. Opt. Soc. Am. A **18**, 473-487 (2001).

[23] Y. Zhang, B. P. Pham, M. P. Eckstein, "Task-based model/human observer evaluation of SPIHT wavelet compression with human visual system-based quantization," Acad. Radiol. **12**, 324-336 (2005)

[24] Y. Zhang, B. T. Pham, and M. P. Eckstein, "Automated optimization of JPEG 2000 encoder options based on model observer performance for detecting variable signals in X-ray coronary angiograms," IEEE transactions on Medical Imaging **23**, 459-474 (2004).

4. Conclusions and perspectives

By providing a systematic and objective approach for breast cancer lesions detection, this work extended the traditional psychophysical studies toolbox, in order to better understand the processes that govern such tasks. Various mathematical, signal processing, or vision modeling methods such as texture analysis, Bayesian classifiers, genetic algorithms, or human HLT estimation, were adapted to clinically realistic detection tasks in x-ray breast imaging.

Among the original contributions of this work, an objective and reproducible way for assessing breast density was developed, mammographic backgrounds with highly realistic visual and statistical properties (second-generation CLB) were synthesized and validated, clinically realistic breast masses were used in various psychophysical studies, the HLT was estimated in several realistic conditions, and a task-based relative comparison of mammography and tomosynthesis with a matched cases database was carried out. Additionally, a systematic comparison between human observers and adapted model observers was conducted for each studied detection task, in order to assess the potential of these models to predict human observers' detection performance in the same conditions.

The breast BI-RADS density classifiers, combined or not with an algorithm for automatically selecting the regions of interest, could readily be incorporated into mammographic units in order to reduce the variability inherent to the definition of the four BI-RADS density classes, or as an educational or training tool to achieve the same goal. The classifiers could also help optimizing the image display conditions, for example by performing different image processing according to the density.

The second-generation CLB significantly improved the visual and statistical quality of existing 2D breast texture synthesis methods. They were used in the subsequent psychophysical studies of this work. The simplicity of the method and the virtually unlimited number of images that can be produced, make them perfect candidates for psychophysical studies focused on mammography. Adaptation of the optimization method to other kinds of medical or nonmedical images that could be assimilated to superposition of blobs is straightforward, as long as a sufficiently large database of reference images exists.

The psychophysical studies with a simplified spherical signal and simulated breast lesions provided valuable information about human observer's perception of this kind of clinically realistic tasks. Results showed that the observers processed the real and synthetic images the same way and confirmed that local anatomical noise (in the immediate area around the signal) could have a major effect on the performance. They also showed that the observer's performance did not differ in SKE or SKS tasks as long as the signal size was kept constant, but that the observers were sensitive to signal size uncertainty. All these findings should be taken into account for optimally designing future psychophysical experiments in mammography. In particular, they should guide backgrounds preprocessing and signals choices.

Excellent results with both mammography and tomosynthesis tasks were obtained with three model observers: the NPWE model, the CHO with Gabor channels, and the HLT. The potential of the first two to correctly predict human observers' performance had already been proven in various psychophysical conditions with geometrical signals in mammography or with correlated backgrounds. Our series of studies with clinically realistic conditions completed the previous ones, with a particular emphasis on optimal ways for adding internal noise in order to best match human results. The results with the HLT allowed to not only study the performance of the human observers, but also their detection strategy. They suggest that the observers focused essentially on the signal edges detection, rather than looking for bright areas, similarly to the way the NPWE observer processes the signals.

The good match between human's and model observer's results obtained for the various clinically realistic conditions under study confirms that these models provide a reliable tool for objectively assessing image quality. In both mammography and tomosynthesis, costly and time-consuming studies involving radiologists for testing new imaging technology developments could thus be avoided and replaced by objective model observers studies.

There are still several open questions and potential applications of the tools developed in this work. It would be particularly appealing to perform similar psychophysical studies with separate background databases corresponding to the different breast density classes. According to Burgess and Judy's results obtained with correlated noise [Burgess, 2007], there should be dissociations in performance among the different density classes. It would thus be interesting to study the ability of the model observers to predict these dissociations, and to estimate the HLT for each density class in order to find out whether human observers adapt their detection strategy in such conditions.

Additionally, discrimination studies could be conducted, with the same signals and backgrounds as the ones used in this work. In such tasks, instead of being requested to detect a low-intensity signal, the observers would have to discriminate benign from malignant signals. This is a typical aspect of clinical tasks that has not been investigated here, and it would be of interest to test the models in such conditions.

Concerning the breast tomosynthesis study, more signal conditions could be generated (e.g. various sizes, contrasts, or lesion types), in order to improve the significance of our results.

As a final word, it should be reminded that the AFC studies conducted in this work are simplifications of the actual detection processes performed by the radiologists. The search for signal candidates in the whole breast, for example, has not been investigated, and many studies with eye trackers have shown that it is a complex and multistep procedure. Further work could focus on trying to reproduce it with non-Location Known Exactly psychophysical studies.

Acknowledgments

While reaching the end of my PhD journey, I realize that this work would never have been that accomplished without the presence of many colleagues, friends, and family members around me. I definitely believe more in cooperation than in competition, and I think that it paid off during the last four years…

First, I would like to thank my co-supervisor François Bochud for all the energy and patience he dedicated to this project and to my questions. During my first day at IRA, I remember François giving me H.H. Barrett's huge book about Image Science, and telling me that "It would be a good thing to start with chapter 11, 12, and 13. And maybe the first ten chapters as well". That was about 400 pages in total, and that day was one of the longest of my life. But after that somewhat surprising beginning, during the years I spent at IRA, François was really the ideal supervisor: always there when I was seeking for advice, amazingly efficient to read and understand everything I was writing, giving me freedom when I wanted to try some alternatives that had not been planned, and helping me getting back to the project schedule when I needed the deadlines were approaching. François gave me time for writing the papers presented in this thesis and always helped me during their redaction, and I owe him a lot for having given this opportunity to me.

Thank you also to my other thesis supervisor Aurelio Bay. Although not being from the field of medical imaging, Aurelio is the kind of person who could simply listen, understand, and come up with new ideas every time we discussed the project together. Working under his supervision was a very motivating opportunity.

Then, I would like to thank all co-authors of the papers and the conference proceedings. Everyone put a lot of effort into it, and I was very lucky to be able to work with all these specialists. I think in particular to Miguel Eckstein, who never counted the hours when we emailed him drafts, conference proceeding manuscripts, papers, or anything to read that was remotely related to the project. François, who had made a postdoc with Miguel in Los Angeles a few years ago, had been enthusiastic about him and his ability to make projects progress, and today I understand why. I would like also to thank Craig Abbey, in particular for the time he spent in San Diego last February with me, the day before my oral presentation at SPIE, patiently explaining that there was something slightly wrong with the way I was presenting some of the results. The night was short, but thanks to Craig, everything went fine the day after… I also thank Francis R. Verdun for his constructive comments about the physics and the clinical aspects of medical imaging, and for his great help when we were looking for tomosynthesis images.

During the course of this project, I have been lucky enough to meet specialists and have valuable exchanges with them every time I was facing a new milestone. I am thinking in particular to Robert S. Saunders, and to Mark Ruschin, two medical physicists from Duke and Toronto. They answered all my numerous questions about synthetic masses and created for us the lesion signals we needed for our experiments. Thanks to Rob and Mark, I could gain a lot of time in my PhD and get much closer to clinically realistic tasks than I initially thought when I started the project. Personal contacts with Art Burgess have also contributed to solidify many concepts in my mind.

Valuable input has also been brought by Karen Kinkel, M.D., from the Clinique des Grangettes. It is tricky, when one is only working with physicists, to stay close to clinical practice, and Dr. Kinkel was always there to share with us her medical expertise about the images we were planning to work with. Her input was particularly valuable during the validation of our synthetic backgrounds.

At IRA, I have spent four years with great people. I would like to thank all of them for the moments we shared, the answers they gave to my questions, the passionate unending debates during lunch that made me laugh most of the time, and the good moments during or after work. I am thinking in particular to my colleagues from the medical imaging group, with whom I learned every day, and to my friends from the “bureau des assistants”.

Very special thanks also to Cécile Althaus, Gaëlle Bocksberger, Valentine Clémence, and Roman Naegeli, the observers who spent hours in the dark looking for my nearly invisible signals. Sometimes I felt bad about how long they had to destroy their eyes, but they never got discouraged, and many of our results are due to their patience. Thanks also to Samuel De Laere for compiling the mammogram database we used throughout this work.

But work is not everything! Over the years, countless friends have been there for me, and have contributed to make me live some of the best years in my life. In order to be sure not to forget any name, I will simply thank here my friends from the music bands I have been playing in, those from EPFL, the fans of our traditional gourmet excursions to the Café des Bouchers, the friends from my beloved Jura and my future homeland Valais, my climbing pals, and all the incredible people I have met during my around-the-world trip. Being surrounded by such great friends is a gift, and I have no doubt every one of them contributed in some way to keep me motivated and positive during my PhD.

Finally, I would like to thank my family with all my heart, for supporting me during these years. Every day, I could count on them to listen to me, to make me smile, or simply to be there for me. I also think to my “new” family, they will recognize themselves…

This list would of course not be complete without the one who changed my life, who became my friend, my girlfriend, and is now my fiancée. The one who has been sharing all my moments of happiness and doubts for the last three years, who stood by my side every day, and who is the reason why I can often be caught, lost in my thoughts, smiling. Lysianne, there are no words to tell you how much I care for you. Or maybe there are, but I will keep them for April 4th!

This work has been funded by grants from the Swiss National Science Foundation #3252BO-104273, and 320000-113863/1.

References

[Abbey, 2001a] C. K. Abbey and M. P. Eckstein, "Maximum-Likelihood and Maximum-A-Posteriori estimates of human-observer templates," Proc. SPIE Medical Imaging **4324**, 114-122 (2001).

[Abbey, 2001b] C. K. Abbey and H. H. Barrett, "Human- and model-observer performance in ramp-spectrum noise: effects of regularization and object variability," J. Opt. Soc. Am. A **18**, 473-488 (2001).

[Abbey, 2002] C. K. Abbey, M. P. Eckstein, S. S. Shimozaki, A. H. Baydush, D. M. Catarious, and C. E. Floyd, "Human-observer templates for detection of a simulated lesion in mammographic images," Proc. SPIE Medical Imaging **4686**, 25-36 (2002).

[Banks, 2004] E. Banks, G. Reeves, V. Beral, D. Bull, B. Crossley, M. Simmonds, E. Hilton, S. Bailey, N. Barrett, P. Briers, R. English, A. Jackson, E. Kutt, J. Lavelle, L. Rockall, M. G. Wallis, M. Wilson, J. Patnick, "Influence of personal characteristics of individual women on sensitivity and specificity in mammography in the Million Women Study: cohort study," BMJ **329**, 477-482 (2004).

[Barrett, 1998] H. H. Barrett, C. K. Abbey, B. Gallas, and M. P. Eckstein, "Stabilized estimates of Hotelling-observer detection performance in patient-structured noise," Proc. SPIE Medical Imaging **3340**, 27-43 (1998).

[Barrett, 1998] H. H. Barrett, C. K. Abbey, and E. Clarkson, "Objective assessment of image quality. III. ROC metrics, ideal observers, and likelihood generating functions," J. Opt. Soc. Am. A **15**, 1520–1535 (1998).

[Barrett, 2004] H. H. Barrett and K. J. Myers. *Foundations of Image Science* (Wiley Series in Pure and Applied Optics, Hoboken, 2004).

[Baydush, 2001] A. H. Baydush, D. M. Catarious Jr, J. Y. Lo, C. K. Abbey, and C. E. Floyd Jr, "Computerized classification of suspicious regions in chest radiographs using subregions Hotelling observers," Med. Phys. **28**, 2403-2409 (2001).

[Bochud, 1999a] F. O. Bochud, J. F. Valley, F. R. Verdun, C. Hessler, and P. Schnyder, "Estimation of the noisy component of anatomical backgrounds," Med. Phys. **26**, 1365-1370 (1999).

[Bochud, 1999b] F. O. Bochud, C. K. Abbey, and M. P. Eckstein, "Statistical texture synthesis of mammographic images with clustered lumpy backgrounds," Opt. Express **4**, 33-43 (1999).

[Bochud, 2000] F. O. Bochud, C. K. Abbey, and M. P. Eckstein, "Visual signal detection in structured backgrounds. III. Calculation of figures of merit for model observers in statistically nonstationary backgrounds," J. Opt. Soc. Am. A **17**, 193-205 (2000).

[Boyd, 1995] N. F. Boyd, J. W. Byng, R. A. Long, E. K. Fishell, L. E. Little, A. B. Miller, G. A. Lockwood, G. L. Tritchler, and M. J. Yaffe, "Quantitative classification of mammographic densities and breast cancer risk: Results from the Canadian National Breast Screening study," J. Natl. Cancer Inst. **87**, 670–675 (1995).

[Boyle, 2003] P. Boyle, "Mammographic breast cancer screening: after the dust has settled," The Breast **12**, 351-356 (2003).

[Brettle, 2007] D. S. Brettle, E. Berry, and M. A. Smith, "The effect of experience on detectability in local anatomical noise," Br. J. Radiol. **80**, 186-193 (2007).

[Bruce, 2003] V. Bruce, P. R. Green, M. A. Georgeson. Visual perception, 4th edition, (Psychology Press, Hove & London, 2003).

[Burgess, 1981] A. E. Burgess, R. F. Wagner, R. J. Jennings, and H. B. Barlow, "Efficiency of human visual signal discrimination," Science **214**, 93-84 (1981).

[Burgess, 1984] A. E. Burgess and H. Ghandeharian, "Visual signal detection. I. Ability to use phase information," J. Opt. Soc. Am. A **1**, 900-905 (1984).

[Burgess, 1985] A. E. Burgess, "Visual signal detection. III. On Bayesian use of prior knowledge and cross correlation," J. Opt. Soc. Am. A **2**, 1498-1507 (1985).

[Burgess, 1994] A. E. Burgess, "Statistically defined backgrounds: performance of a modified nonprewhitening observer," J. Opt. Soc. Am. A **11**, 1237–1242 (1994).

[Burgess, 1995] A. E. Burgess, "Comparison of receiver operating characteristic and forced choice observer performance measurement methods," Med. Phys. **22**, 643-655 (1995).

[Burgess, 1997] A. E. Burgess, X. Li, and C. K. Abbey, "Visual signal detectability with two noise components: anomalous masking effects," J. Opt. Soc. Am. A **14**, 2420-2442 (1997).

[Burgess, 2001] A. E. Burgess, F. L. Jacobson, and P. F. Judy, "Human observer detection experiments with mammograms and power-law noise," Med. Phys. **28**, 419-437 (2001).

[Burgess, 2007] A. E. Burgess and P. F. Judy, "Signal detection in power-law noise: effect of spectrum exponents," J. Opt. Soc. Am. A **24**, B52-B60 (2007).

[Carton, 2006] A. G. Carton, J. Li, M. Albert, S. Chen, and A. D. Maidment, "Quantification for contrast-enhanced digital breast tomosynthesis," Proc. SPIE Medical Imaging **6142**, 111-121 (2006).

[Castella, 2007a] C. Castella, K. Kinkel, M. P. Eckstein, P.-E. Sottas, F. R. Verdun, F. O. Bochud, "Semiautomatic Mammographic Parenchymal Patterns Classification Using Multiple Statistical Features," Academic Radiology **14**, 1486-1499 (2007).

[Castella, 2007b] C. Castella, C. K. Abbey, M. P. Eckstein, F. R. Verdun, K. Kinkel, and F. O. Bochud, "Human linear template with mammographic backgrounds estimated with a genetic algorithm," J. Opt. Soc. Am. A **24**, B1-B12 (2007).

[Castella, 2008] C. Castella, K. Kinkel, F. Descombes, M. P. Eckstein, P. Sottas, F. R. Verdun, and F. O. Bochud, "Mammographic texture synthesis: second-generation clustered lumpy backgrounds using a genetic algorithm," Opt. Express **16**, 7595-7607 (2008).

[Castella, 2009a] C. Castella, M. P. Eckstein, C. K. Abbey, K. Kinkel, F. R. Verdun, R. S. Saunders, E. Samei, and F. O. Bochud, "Mass detection on mammograms: influence of signal shape uncertainty on human and model observers," J. Opt. Soc. Am. A **26**, 425-436 (2009).

[Castella, 2009b] C. Castella, M. Ruschin, M. P. Eckstein, C. K. Abbey, K. Kinkel, F. R. Verdun, A. Tingberg, and F. O. Bochud, "Masses detection in breast tomosynthesis and digital mammography: a model observer study," *to appear in Proc. SPIE Medical Imaging (2009).*

[Chan, 2005] H.-P. Chan, J. Wei, B. Sahiner, E. A. Rafferty, T. Wuo, M. A. Roubidoux, R. H. Moore, D. B. Kopans, L. M. Hadjiski, and M. A. Helvie, “Computer-aided Detection System for Breast Masses on Digital Tomosynthesis Mammograms: Preliminary Experience,” Radiology **237**, 1075-1080 (2005).

[Chen, 2004] Y. Chen, J. Y. Lo, and J. T. Dobbins III, “Gaussian frequency blending algorithm with matrix inversion tomosynthesis (MITS) and filtered backprojection (FBP) for better digital breast tomosynthesis reconstruction,” Proc. SPIE Medical Imaging **6142**, 122-130 (2006).

[Chawla, 2007] A. S. Chawla, E. Samei, R. S. Saunders, C. K. Abbey, and D. Delong, “Effect of dose reduction on the detection of mammographic lesions: A mathematical observer model analysis,” Med. Phys. **34**, 3385-3398 (2007).

[Colditz, 1995] G. A. Colditz, S. E. Hankison, D. J. Hunter, W. C. Willett, J. E. Manson, M. J. Stampfer, C. Hennekens, B. Rosner, and F. E. Speizer, “The use of estrogens and progestins and the risk of breast cancer in postmenopausal women,” N. Engl. J. Med. **332**, 1589–1593 (1995).

[Curado, 2007] M. P. Curado, B. Edwards, H. R. Shin, H. Storm, J. Ferlay, and M. Heanue, *Cancer Incidence in Five Continents, Vol. IX, IARC Scientific Publications No. 160* (IARCPress, Lyon, 2007).

[Deck, 2006] W. Deck and R. Kakuma, “Screening mammography: a reassessment,” Agence d'Evaluation des Technologies et des Modes d'Intervention en Sante (AETMIS) **97** (2005).

[de Koning, 2003] H. J. de Koning, “Mammographic screening: evidence from randomized controlled trials,” Ann. Oncol. **14**, 1185-1189 (2003).

[Dobbins, 2003] J. T. Dobbins and D. J. Godfrey, “Digital x-ray tomosynthesis: current state of the art and clinical potential,” Phys. Med. Biol. **48**, R65-R106 (2003).

[Eberhard, 2006] J. W. Eberhard, P. Staudinger, J. Smolenski, J. Ding, A. Schmitz, J. McCoy, A. Al-Khalidy, M. A. Rumsey, W. Ross, C. E. Landberg, P. L. Carson, M. M. Goodsitt, H. P. Chan, M. Roubidoux, J. Thomas, J. Osland, “High-speed large-angle mammography tomosynthesis system,” Proc. SPIE Medical Imaging **6142**, 100-110 (2006).

[Eckstein, 1998] M. P. Eckstein, C. K. Abbey, and J. S. Whitning, “Human vs. model observers in anatomic backgrounds,” Proc. SPIE Medical Imaging **3340**, 16-26 (1998).

[Eckstein, 2000] M. P. Eckstein, C. K. Abbey, and F. O. Bochud, “A practical guide to model observers for visual detection in synthetic and natural noisy images,” in *Handbook of medical imaging. Volume 1. Physics and psychophysics* (SPIE press, Bellingham, 2000).

[Ferlay, 2001] J. Ferlay, F. Bray, D. M. Parkin, and P. Pisani, *Globocan 2000: Cancer Incidence and Mortality Worldwide, IARC Cancer Bases No. 5* (IARCPress, Lyon, 2001).

[Ferlay, 2007] J. Ferlay, P. Autier, M. Boniol, M. Heanue, M. Colombet, and P. Boyle, “Estimates of the cancer incidence and mortality in Europe in 2006,” Ann. Oncol. **18**, 581-592 (2007).

[Fitzgibbons, 2000] P. L. Fitzgibbons, D. L. Page, D. Weaver, A. D Thor, D. C. Allred, G. M. Clark, S. G. Ruby , F. O'Malley, J. F. Simpson, J. L. Connolly, D. F. Hayes, S. B. Edge, A. Lichter, S. J. Schnitt, “Prognostic factors in breast cancer. College of American Pathologists Consensus Statement 1999,” Arch. Pathol. Lab. Med. **124**, 966–978 (2000).

[Fletcher, 2003] S. W. Fletcher, J. G. Elmore, “Clinical practice: mammographic screening for breast cancer,” N. Eng. J. Med. **348**, 1672-1680 (2003).

[Gallas, 2003] B. D. Gallas and H. H. Barrett, "Validating the use of channels to estimate the ideal linear observer," J. Opt. Soc. Am. A **20**, 1725-1738 (2003).

[Gallas, 2007] B. D. Gallas, G. A. Pennello, and K. J. Myers, "Multireader multicase variance analysis for binary data," J. Opt. Soc. Am. A **24**, B70-B80 (2007).

[Gifford, 2008] H. C. Gifford, C. S. Didier, M. Das, and S. J. Glick, "Optimizing Breast-Tomosynthesis Acquisition Parameters with Scanning Model Observers," Proc. SPIE Medical Imaging **6917**, 69170S (2008).

[Gong, 2006] X. Gong, S. J Glick, A. A. Vedula, and S. Thacker, "A computer simulation study comparing lesion detection accuracy with digital mammography, breast tomosynthesis, and cone-beam CT breast imaging," Med. Phys. **33**, 1041-1052 (2006).

[Gotzsche, 2000] P. C. Gotzsche and O. Olsen, "Is screening for breast cancer with mammography justifiable?," Lancet **355**, 129-134 (2000).

[Green, 2003] B. B. Green, S. H. Taplin, "Breast Cancer Screening Controversies," J. Am. Board Fam. Pract. **16**, 233-241 (2003).

[Heine, 2001] J. J. Heine and P. Malhotra, "Mammographic tissue, breast cancer risk, serial image analysis, and digital mammography: Part 1. Tissue and related risk factors," Acad. Radiol. **9**, 298–316 (2001).

[Huda, 2006] W. Huda, K. M. Odgen, E. M. Scalzetti, D. R. Dance, and E. A. Bertrand, "How Do Lesion Size and Random Noise Affect Detection Performance in Digital Mammography?," Acad. Radiol. **13**, 1355-1366 (2006).

[Humphrey, 2002] L. L. Humphrey, M. Helfand, B. K. S. Chan, and S. H. Woolf, "Breast cancer screening: a summary of the evidence for the U.S. preventive services task force," Ann. Intern. Med. **137**, 347-360 (2002).

[IARC, 2002] *Breast Cancer Screening. International Agency for Research on Cancer (IARC) Handbooks of Cancer Prevention. Vol. 7* (IARC Press, Lyon, 2002)

[ICRU, 2008] ICRU report 79, "Receiver Operating Characteistic Analysis in Medical Imaging," Journal of the ICRU **8** (Oxford University Press, 2008).

[Judy, 1997] P. F. Judy, M. F. Foley Kijewski, and R. G. Swensson, "Observer detection performance loss: target size uncertainty," Proc. SPIE Medical Imaging **3036**, 39-47 (1997).

[Karssemeijer, 1998] N. Karssemeijer, "Automated classification of parenchymal patterns in mammograms," Phys Med Biol **43**, 365–378 (1998).

[Kelsey, 1993] J. L. Kelsey, M. D. Gammon, and E. M. John, "Reproductive factors and breast cancer," Epidemiol. Rev. **15**, 36–47 (1993).

[Lai, 2005] C.-J. Lai, C. C. Shaw, G. J. Whitman, D. A. Johnston, W. T. Yang, V. Selinko, E. Arribas, B. Dogan, and S. Cheenu Kappadath, "Visibility of simulated microcalcifications-A hardcopy-based comparison of three mammographic systems," Med. Phys. **32**, 182-194 (2005).

[Maidment, 2006] A. D. A. Maidment, C. K. Ullberg, K. Lindman, L. Adelöw, J. Egerström, M. Ecklund, T. Franke, U. Jordung, T. Kristoffersson, L. Lindqvist, D. Marchal, H. Olla, E. Penton, J. Rantanen, S.

Solokov, N. Weber, and H. Wersterberg, “Evaluation of a photon-counting breast tomosynthesis imaging system,” Proc. SPIE Medical Imaging **6142**, 89-99 (2006).

[Marcelja, 1980] S. Marcelja, “Mathematical Description of Simple Cortical Cells,” J. Opt. Soc. Am. A **4**, 1297-1300 (1980).

[Mertelmeier, 2006] T. Mertelmeier, W. Haerer, J Orman, and M. K. Dudam, “Optimizing filtered backprojection for a breast tomosynthesis prototype device,” Proc. SPIE Medical Imaging **6142**, 131-142 (2006).

[Movshon, 1978] J. A. Movshon, I. D. Thomson, D. J. Tolhurst, “Spatial summation in the Receptive Fields of Simple Cells in the Cat’s Striate Cortex,” J. Physiol. London **283**, 53-77 (1978).

[Myers, 1985] K. J. Myers, H. H. Barrett, M. C. Borgstrom, D. D. Patton, and G. W. Seeley, “Effect of noise correlation on detectability of disk signals in medical imaging,” J. Opt. Soc. Am. A **2**, 1752-1759 (1985).

[Myers, 1987] K. J. Myers and H. H. Barrett, “The addition of a channel mechanism to the ideal-observer model,” J. Opt. Soc. Am. A **4**, 2447–2457 (1987).

[Obenauer, 2005] S. Obenauer, K. P. Hermann, E. Grabbe, “Applications and literature review of the BI-RADS classification,” Eur. Radiol. **15**, 1027-1036 (2005)

[Olsen, 2001] O. Olsen and P. C. Gotzsche, “Cochrane review on screening for breast cancer with mammography,” Lancet **358**, 1340-1342 (2001).

[Park, 2007] J. M. Park, E. A. Franken, M. Garg, L. L. Fajardo, and L. T. Niklason, “Breast Tomosynthesis: Present Considerations and Future Applications,” Radiographics **27**, S231-S240 (2007).

[Pineda, 2006] A. R. Pineda, S. Yoon, D. S. Paik, and R. Fahrig, “Optimization of a tomosynthesis system for the detection of lung nodules,” Med. Phys. **33**, 1372-1379 (2006).

[Pisano, 2005] E. D. Pisano, C. Gatsonis, E. Hendrick, M. Yaffe, J. K. Baum, S. Acharyya, E. F. Conant, L. L. Fajardo, L. Bassett, C. D'Orsi, R. Jong, M. Rebner, “Diagnostic performance of digital versus film mammography for breast-cancer screening,” N. Eng. J. Med. **353**, 1773-1783 (2005).

[Reiser, 2008] I. Reiser, B. A. Lau, and R. M. Nishikawa, “Effect of Scan Angle and Reconstruction Algorithm on Model Observer Performance in Tomosynthesis,” Proc. IWDM 2008, LNCS **5116**, 601-611 (2008).

[Rolland, 1992] J. P. Rolland and H. H. Barrett, “Effect of random background inhomogeneity on observer detection performance,” J. Opt. Soc. Am. A **9**, 649-658 (1992).

[Ruschin, 2005] M. Ruschin, A. Tingberg, M. Bath, A. Grahn, M. Hakansson, B. Hemdal, and I. Andersson, “Using simple mathematical functions to simulate pathological structures-input for digital mammography clinical trial,” Rad. Prot. Dosim. **114**, 424-431 (2005).

[Ruschin, 2007a] M. Ruschin, P. Timberg, T. Svahn, I. Andersson, B. Hemdal, S. Mattsson, M. Bath, and A. Tingberg, “Improved in-plane visibility of tumors using breast tomosynthesis,” Proc. SPIE Medical Imaging **6510**, 65101J (2007).

[Ruschin, 2007b] M. Ruschin, P. Timberg, M. Bath, B. Hemdal, T. Svahn, R. S. Saunders, E. Samei, I. Andersson, S. Mattsson, D. P. Chakraborty, and A. Tingberg, “Dose dependence of mass and microcalcification detection in digital mammography: Free response human observer study,” Med. Phys. **34**, 400-407 (2007).

[Samei, 1998] E. Samei, M. J. Flynn, W. R. Eyler, and E. Peterson, “The Effect of Local Background Anatomical Patterns on the Detection of Subtle Lung Nodules in Chest Radiographs,” Proc. SPIE Medical Imaging **3340**, 44-54 (1998).

[Samei, 2007] E. Samei, R. S. Saunders, J. A. Baker, and D. M. Delong, “Digital mammography: Effects of Reduced Radiation Dose on Diagnostic Performance,” Radiology **243**, 396-404 (2007).

[Saunders, 2006] R. S. Saunders, E. Samei, J. A. Baker, D. M. Delong, "Simulation of Mammographic Lesions," Acad. Radiol. **13**, 860-870 (2006).

[Skaane, 2005] P. Skaane, C. Balleyguier, F. Diekmann, S. Diekmann, J.-C. Piguet, K. Young, and L. T. Niklason, “Breast Lesion Detection and Classification: Comparison of Screen-Film Mammography and Full-Filed Digital Mammography with Soft-copy Reading-Observer Performance Study,” Radiology **237**, 37-44 (2005).

[Smith, 2005] A. Smith, “Full Field Breast Tomosynthesis,” Radiol. Manage. **27**, 25-31 (2005).

[Suryanarayanan, 2000] S. Suryanarayanan, A. Karellas, S. Vedantham, S. J. Glick, C. J. D’Orsi, S. P. Baker, and R. L. Webber, “Comparison of tomosynthesis methods used with digital mammography,” Acad. Radiol. **7**, 1085-1097 (2000).

[Suryanarayanan, 2001] S. Suryanarayanan, A. Karellas, S. Vedantham, S. J. Glick, C. J. D’Orsi, S. P. Baker, and R. L. Webber, “Evaluation of linear and nonlinear tomosynthetic reconstruction methods in digital mammography,” Acad. Radiol. **8**, 219-224 (2001).

[Suryanarayanan, 2005] S. Suryanarayanan, A. Karellas, S. Vedantham, S. M. Waldrop, and C. J. D’Orsi, “Detection of Simulated Lesions on Data-compressed Digital Mammograms,” Radiology **236**, 31-36 (2005).

[Tanner, 1958] W. P. Tanner Jr and T. G. Birdsall, “Definitions of d' and μ as Psychophysical Measures,” J. Acoust. Soc. of Am. **30**, 922-928 (1958).

[Van Gils, 1999] C. H. van Gils, J. H. Hendriks, R. Holland, N. Karssemeijer, J. D. Otten, H. Straatman, and A. L. Verbeek, “Changes in mammographic breast density and concomitant changes in breast cancer risk,” Eur. J. Cancer Prev. **8**, 509–515 (1999).

[Wu, 2004] T. Wu, R. H. Moore, E. A. Rafferty, and D. B. Koppans, “A comparison of reconstruction algorithms for breast tomosynthesis,” Medical Physics **31**, 2636-2647 (2004).

[Zhang, 2004a] Y. Zhang, B. Pham, and M. P. Eckstein, “Automated Optimization of JPEG 2000 Encoder Options Based on Model Observer Performance for Detecting Variable Signals in X-Ray Coronary Angiograms,” IEEE Transactions on Medical Imaging **23**, 459- 474 (2004).

[Zhang, 2004b] Y. Zhang, B. Pham, and M. P. Eckstein, “Evaluation of JPEG 2000 Encoder Options: Human and Model Observer Detection of Variable Signals in X-Ray Coronary Angiograms,” IEEE Transactions on Medical Imaging **23**, 613-632 (2004).

[Zhang, 2005] Y. Zhang, B. T. Pham, and M. P. Eckstein, "Task-Based Model/Human Observer Evaluation of SPHIT Wavelet Compression With Human Visual System-based Quantization," Acad. Radiol. **12**, 324-336 (2005).

[Zhang, 2006] Y. Zhang, C. K. Abbey, and M. P. Eckstein, "Adaptative detection mechanisms in globally statistically nonstationary –oriented noise," J. Opt. Soc. Am. A **23**, 1549-1558 (2006).

[Zhang, 2007] Y. Zhang, B. T. Pham, and M. P. Eckstein, "Evaluation of internal noise methods for Hotelling observers," Med. Phys. **34**, 3312-3322 (2007).

[Ziv, 2003] E. Ziv, R. Smith-Bindman, and K. Kerlikowske, "Mammographic breast density and family history of breast cancer," J. Natl. Cancer Inst. **95**, 556–558 (2003).

APPENDIX: Screening mammography from a practical point of view

In Switzerland, systematic screening mammography programs are organized in all French-speaking cantons. In the other cantons, women may be referred by their general practitioner or gynaecologist for a mammography examination. In the near future, these inequalities between cantons may change, with systematic screening programs introduced in all remaining cantons as well.

The systematic screening programs (Switzerland or Worldwide) usually include the following steps. First, letters of invitations are sent from a screening center to the eligible women, with informations about the clinical procedures, risks and benefits, and the local radiographic institutes that are accredited for screening mammography.

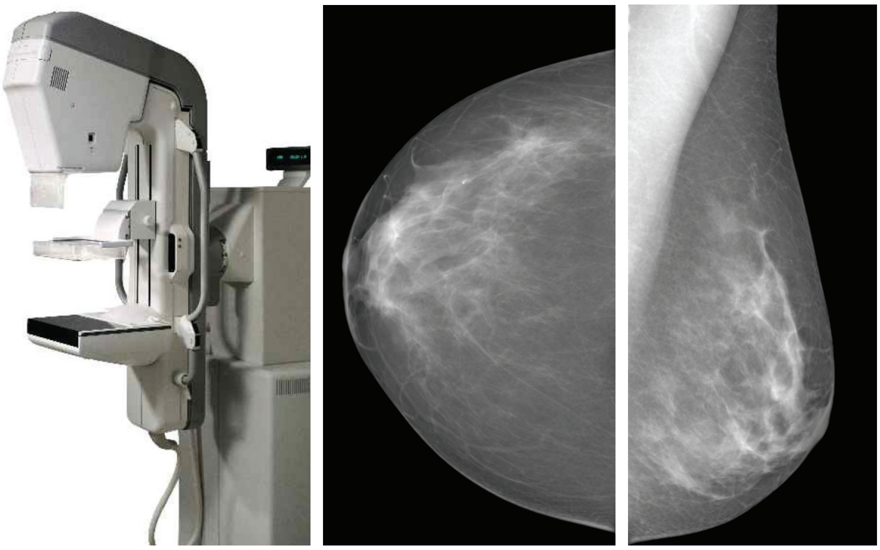


Fig. 8. (Left) Mammographic unit used in this study GE Senograph 2000D (GE Healthcare, Waukesha, USA) and examples of usual views: Cranio-Caudal (Center) and Mediolateral Oblique(Right)

During the examination, two mammograms per breast are acquired, corresponding to different projections (see Fig. 8). This allows to lower the issues caused by tissue superposition. The first reading of the mammograms is done at the institute.

In order to improve the screening performance, a second reading is done at the screening centre. If the two readers disagree, a third reading is performed at the screening center. The results are then communicated to the patient.

If the results are positive, follow-up examinations are organized under the responsibility of the patient’s physician. These additional examinations can include a diagnostic mammography, magnetic resonance imaging, an ultrasound, and/or a biopsy with histopathological analysis. If the results of the follow-up examinations or the initial screening mammography readings are negative, the patient will be invited again two years later.

Curriculum Vitae

Cyril Castella

Avenue d’Echallens 36

1004 Lausanne

Switzerland

+41 (0)76 324 32 82

cyril.castella@gmail.com

Key features

MSc in Physics, Swiss Federal Institute of Technology (EPFL)

PhD in medical physics

Languages: English, French, German

Education

March 2009	PhD degree, EPFL
2004 - 2009	PhD student, EPFL and University Hospital, Lausanne
2004 -	Swiss Society of Radiobiology and Medical Physics training, exam planned in 2009
2004	Radiation protection specialist certification, IRA, Lausanne
1999 - 2004	MSc degree in Physics, EPFL, Lausanne
1999	High school degree, Lycée Cantonal, Porrentruy

Areas of Expertise

Medical Imaging

Radiography, fluoroscopy, mammography, tomosynthesis, CT scan. Associated quality assurance. Specialist in image processing and detection model observers.

Radiation protection

Radiation protection instruction: radiographer students, physicians performing high-dose examinations, screening mammography lectures.

Computer science

Office tools, LateX, HTML. Programming languages : IDL, Fortran, C, Matlab, Igor

Projects

MSc project

The goal of this project was to assess the potential use of semiconductor radiation detectors for microbeam radiation therapy. The measurements were conducted at the European Synchrotron Radiation Facility in Grenoble (ESRF, France), and then validated using Monte Carlo simulations with the stochastic code MCNP. Thanks to an optimized use of the beam time at ESRF, the results allowed the project to be continued after my graduation.

PhD project

From currently used digital mammography to emerging breast tomosynthesis, this research project was centered on objectively characterizing the quality of the images used by radiologists for breast cancer detection. The performance of trained observers was tested using psychophysical experiments, and detection model observers were developed and optimized using a genetic algorithm. Thanks to several dynamic international collaborations that further complemented our local competences, the clinical realism of the detection tasks considered in this project was significantly improved compared to previous studies in the field. Additionally, several scientific papers were published during the PhD.

Languages

French	Native language
English	Very good spoken and written skills
German	Good spoken and written skills

List of publications, talks, and posters

Publications in peer-reviewed journals

[Castella, 2009a] C. Castella, M. P. Eckstein, C. K. Abbey, K. Kinkel, F. R. Verdun, R. S. Saunders, E. Samei, and F. O. Bochud, “Mass detection on mammograms: influence of signal shape uncertainty on human and model observers,” J. Opt. Soc. Am. A **26**, 425-436 (2009).

C. Castella, K. Kinkel, F. Descombes, M. P. Eckstein, P.-E. Sottas, F. R. Verdun, and F. O. Bochud, “Mammographic texture synthesis: second-generation clustered lumpy backgrounds using a genetic algorithm,” Opt. Express **16**, 7595-7607 (2008).

C. Castella, K. Kinkel, M. P. Eckstein, P.-E. Sottas, F. R. Verdun, and F. O. Bochud, “Semiautomatic Mammographic Parenchymal Patterns Classification Using Multiple Statistical Features,” Academic Radiology **14**, 1486-1499 (2007).

C. Castella, C. K. Abbey, M. P. Eckstein, F. R. Verdun, K. Kinkel, and F. O. Bochud, "Human linear template with mammographic backgrounds estimated with a genetic algorithm," J. Opt. Soc. Am. A **24**, B1-B12 (2007).

Publications in conference proceedings

C. Castella, M. Ruschin, M. P. Eckstein, C. K. Abbey, K. Kinkel, F. R. Verdun, A. Tingberg, and F. O. Bochud, “Masses detection in breast tomosynthesis and digital mammography: a model observer study,” *to appear in Proc. SPIE Medical Imaging 2009*

C. Castella, K. Kinkel, M. P. Eckstein, C. K. Abbey, F. R. Verdun, R. S. Saunders, E. Samei, and F. O. Bochud, “Mass detection on mammograms: signal variations and performance changes for human and model observers,” Proc. SPIE Medical Imaging **6917**, 69170K (2008).

C. Castella, K. Kinkel, F. R. Verdun, M. P. Eckstein, C. K. Abbey, and F. O. Bochud, “Mass detection on real and synthetic mammograms: human observer templates and local statistics,” Proc. SPIE Medical Imaging **6515**, 65150U (2007).

C. Castella, K. Kinkel, F. Descombes, M.P. Eckstein, P.-E. Sottas, F. R. Verdun, and F. O. Bochud, "Mammographic texture synthesis using genetic programming and clustered lumpy background", Proc. SPIE Medical Imaging **6146**, 61460U (2006).

Talks and posters

FBM Research Day 2009, poster

C. Castella, M. Ruschin, M. P. Eckstein, C. K. Abbey, K. Kinkel, F. R. Verdun, A. Tingberg, and F. O. Bochud, “Breast tomosynthesis and digital mammography: a comparative study for breast masses detection,” Faculty of Biology and Medicine Research Day, Lausanne (2009).

SSRMP 2008, talk

C. Castella, K. Kinkel, M. P. Eckstein, C. K. Abbey, F. R. Verdun, R. S. Saunders, E. Samei, and F. O. Bochud, “Mass detection on mammograms: how do humans and models deal with signal uncertainty?,” Swiss Society of Radiobiology and Medical Physics meeting, Chur (2008).

SPS 2008, talk

C. Castella, “Mass detection on mammograms: variable signals and related performance changes for human and model observers,” Swiss Physical Society meeting, Geneva (2008).

EPFL Research Day 2008, poster

C. Castella, C. K. Abbey, M. P. Eckstein, F. R. Verdun, K. Kinkel, and F. O. Bochud, “Human observer template in mammography,” EPFL Research Day, Lausanne (2008).

FBM Research Day 2008, poster

C. Castella, K. Kinkel, F. Descombes, M. P. Eckstein, P.-E Sottas, F. R. Verdun, and F. Bochud, “Human observer template in mammography," Faculty of Biology and Medicine Research Day, Lausanne (2008).

SSRMP 2007, poster

C. Castella, C. K. Abbey, M. P. Eckstein, F. R. Verdun, K. Kinkel, and F. O. Bochud, “Human observer template for the detection of a synthetic mass on mammographic background,” Swiss Society of Radiobiology and Medical Physics meeting, Bern (2007).

SSR 2006, invited talk

C. Castella, "La Tomosynthèse", Swiss Radiology Society meeting, Lausanne (2006).

HECV Santé 2006, invited talk

C. Castella, F. Descombes, "Création d'images mammo synthétiques : un exemple de collaboration IRA – HECV Santé Filière TRM", HECV Santé Research Day (2006).

SSRMP 2005, talk

C. Castella, K. Kinkel, M. P. Eckstein, P.-E. Sottas, F. R. Verdun, F.O. Bochud, “Automatic mammographic parenchymal patterns classification using multiple statistical features”, Swiss Society of Radiobiology and Medical Physics meeting, Lausanne (2005).

Extra-professional activities

Harmonie Shostakovich windband, Switzerland– president (2005-2009) and musician-www.shosta.tk

Rock climbing, mountaineering-www.around-annapurna.com

Personal details

28, Swiss nationality, single