Mass detection on mammograms: signal variations and performance changes for human and model observers

C. Castella[a,b], K. Kinkel[c], M. P. Eckstein[d], C. K. Abbey[d], F. R. Verdun[a], R. S. Saunders[e], E. Samei[e], and F. O. Bochud[*a]

[a] University Institute for Radiation Physics, DUMSC, CHUV and University of Lausanne, CH-1007 Lausanne, Switzerland;

[b] LPHE, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland;

[c] Clinique des Grangettes, CH-1224 Chêne-Bougeries, Switzerland;

[d] Dept. of Psychology, University of California, Santa Barbara, CA, USA 93106-9660;

[e] Duke Advanced Imaging Laboratories, Duke University, Durham, NC, USA 27705

ABSTRACT

We studied the influence of signal variability on human and model observer performances for a detection task with mammographic backgrounds and computer generated clustered lumpy backgrounds (CLB). We used synthetic yet realistic masses and backgrounds that have been validated by radiologists during previous studies, ensuring conditions close to the clinical situation. Four trained non-physician observers participated in two-alternative forced-choice (2-AFC) experiments. They were asked to detect synthetic masses superimposed on real mammographic backgrounds or CLB. Separate experiments were conducted with sets of benign and malignant masses. Results under the signal-known-exactly (SKE) paradigm were compared with signal-known-statistically (SKS) experiments. In the latter case, the signal was chosen randomly for each of the 1,400 2-AFC trials (image pairs) among a set of 50 masses with similar dimensions, and the observers did not know which signal was present. Human observers' results were then compared with model observers (channelized Hotelling with Difference-of-Gaussian and Gabor channels) in the same experimental conditions. Results show that the performance of the human observers does not differ significantly when benign masses are superimposed on real images or on CLB with locally matched gray level mean and standard deviation. For both benign and malignant masses, the performance does not differ significantly between SKE and SKS experiments, when the signals' dimensions do not vary throughout the experiment. However, there is a performance drop when the SKS signals' dimensions vary from 5.5 to 9.5 mm in the same experiment. Noise level in the model observers can be adjusted to reproduce human observers' proportion of correct answers in the 2-AFC task within 5% accuracy for most conditions.

Keywords: Image perception, model observers, observer performance evaluation

1. INTRODUCTION

Detection tasks are foundational to many medical imaging applications. In radiology, these tasks involve first determining whether a signal is present in the image, then determining its specifications like exact location, size, shape, and malignancy. While modeling the full detection process is still out of the scope of current psychophysical studies, numerous authors have reported results and models in order to better understand the processes behind various detection tasks. Most of the time, these experiments only approximate the clinical reality, using statistically or exactly known backgrounds, signals, or signal localizations, in order to facilitate the data collection and the robustness of the analysis.

When studying mass detection in a typical radiological task, use of real backgrounds and masses is an option for getting closer to reality, but the collection of hundreds or thousands of similar images can be difficult and time-consuming. Therefore, mammographic non-stationary backgrounds are most of the time replaced by 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], phantoms elements [9], Gaussian or Gabor functions [11-13]. Signal location uncertainty in the clinical task can be

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

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 [16] recently showed that trained naïve (non-physician) observers' performance was very close to that of radiologists.

In order to get closer to 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] in order to synthesize second-generation CLB reproducing visual and statistical properties of mammograms. On the other hand, Saunders *et al.* [17,18] recently developed an algorithm capable of generating benign or malignant breast masses signals, based on the analysis of real masses characteristics. In this study, we merged the two approaches to generate fully synthetic images containing breast masses.

Another aspect of clinical relevance that was introduced in our study is the signal uncertainty. While most studies concentrated on Signal-Known-Exactly (SKE) tasks, where the signal presented to the observers is known and do not vary throughout the entire experiment, little is known about more realistic conditions. In the latter case, the signal should be different for each realization, or at least should not be known exactly by the observer. In order to model experiments involving various signals, Signal-Known-Exactly but Varying (SKEV) and Signal-Known-Statistically (SKS) paradigms have been introduced [19-22] in M-AFC tasks. In the first case, a pool of signals is used throughout the detection experiment. The observer knows exactly which signal is present at one of the M possible locations, but the signal itself is different from one trial (presentation of M images) to the next. This prevents the observer to adapt his strategy specifically to a given signal, while keeping further analysis and comparison with model observers simple. In SKS tasks, the signal is also chosen for each trial out of a pool of different candidates, but the observer does not get any information about which one is used. This situation is the closest to a real radiological task, but it complicates the results analysis, compared to SKE or SKEV.

In order to reproduce or predict human performance in detection tasks, model observers have been developed and successfully adapted to SKE experiments [14,16,23] Later, models have been adapted to SKS tasks [19] and Eckstein [20,21] and Zhang [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, much closer to clinical tasks than "simple" SKE tasks.

The purpose of the present study was to use real and realistic [8] mammographic backgrounds, and benign and malignant mass signals [18] in 2-AFC experiments, for comparing SKE and SKS performance of human observers. We wanted to test the hypothesis that the detection performance is the same with CLB and real images when local statistics are matched, and the effects of signal size and shape uncertainty on the observers. Additionally, we wanted to compare the performances of linear model observers to humans', and determine if they could predict human results in the same experimental conditions.

2. MATERIAL AND METHODS

2.1 2-AFC setup

Four non-physician 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 of each trial. Signal-present images were constructed with a distinct set of backgrounds than signal-absent images, in order to avoid potential bias (same image presented with and without signal in a given trial, or in consecutive trials, for example).

The observers had to determine which image was the most likely to contain the signal. Fiduciary cues were provided in order 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 after every 25 trials.

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 [24]. Pixel size of the display screen was 0.25 mm. Observers' performance was given by the proportion of correct answers P_c, computed from the N_T=1,400 trial outcomes o_i as:

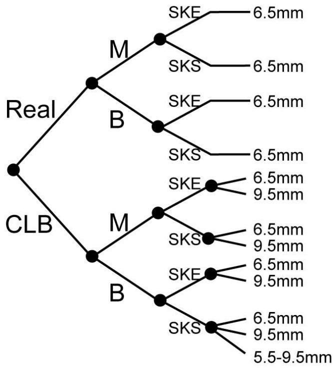

Fig. 1. Backgrounds, mass type, and signal conditions for the 13 psychophysical experiments. B= benign masses, M = malignant masses. Here, SKE stands for an experiment with a single signal of given shape and size, and SKS an experiment with a signal with variable shape but a given size, except for the last experiment where both shape and size were variable.

$$P_c = \frac{1}{N_T} \sum\nolimits_{i=1}^{N_T} o_i \, , \tag{1}$$

where $o_i = 1$ if the answer to trial i was correct, and 0 if it was incorrect. Standard error for P_c was derived by computing P_c for 28 mutually disjoints subsets of 50 consecutive trials.

2.2 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]. 2,800 256 by 256 pixels square ROIs were manually selected, and resampled to 154 by 154 pixels, in order to emulate a magnification factor of 1.5 on the display screen, reproducing typical clinical settings [9]. The CLB had been designed in order to mimic digital mammograms ROIs, and their statistical and visual properties assessed by radiologists in a previous study [8].

In order to obtain as comparable conditions as possible between these two image types, 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 by 40 pixels central area of the displayed images. This corresponds to the area covered by the fiduciary cues, over which the human observers focused. Over this area, the mean gray level was set to 128, and the standard deviation to 20, ensuring that the rescaled images lied in the middle of the display screen dynamic range.

2.3 Signals

The signals were synthetic masses developed by Saunders *et al.* [18]. Based on the analysis of real breast masses properties, they provided excellent signals for conducting experiments with highly realistic masses. We chose to use two kind of simulated breast masses: benign, oval circumscribed masses, and malignant masses with an irregular shape and ill-defined borders [25]. 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.

The masses were added linearly to the backgrounds. The amplitude of the signal, defined as the maximum intensity of the mass, was set after preliminary experiments by one of the authors, in order to obtain a P_c between .70 and .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 *Pc* of at least .70 for the actual experimental contrast conditions. Depending on the observers and series conditions, this training phase lasted from 100 to about 500 trials.

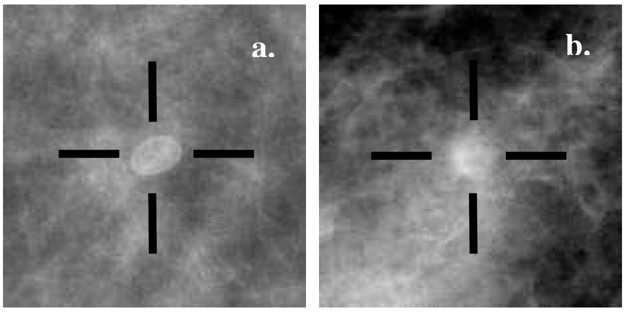

Fig. 2 (a) Example of CLB with a digitally embedded benign mass. (b) Real mammogram ROI with a digitally embedded malignant mass. Signal contrast has been highly increased for printing purpose.

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 trained with the same signal that would be presented during the actual experiment. For SKS, the signal was chosen randomly for each trial among a pool of 50 similar candidates from 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. As for SKE experiments, the 50 similar signals used for the SKS conditions were the same 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: 10 masses per size constituted the pool of SKS signals. The aim was to compare SKS results when the mass size was kept constant and its shape and orientation only 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.

2.4 Model observers

Known SKE model observers were implemented and tested against human results: Channelized Hotelling (CH) observer [14,15,28] with dense Difference-of-Gaussian (DDOG) channels and Gabor functions channels [9,29]. These models have been already extensively described in the literature, and will only be briefly reviewed here.

The Hotelling observer is a linear observer whose response λ to an image **g** is given by :

$$\lambda = \mathbf{w}_{\mathbf{Hot}}^{\mathbf{T}}\,\mathbf{g}, \tag{2}$$

where both $\mathbf{w}_{\mathbf{Hot}}$, the model template, and the image **g** are seen as 1-D vectors. The template itself is derived from the covariance matrix of the backgrounds $\mathbf{K_b}$ as:

$$\mathbf{w}_{\mathbf{Hot}} = \mathbf{K_b}^{-1}[<\mathbf{g_s}>-<\mathbf{g_n}>], \tag{3}$$

where $<\mathbf{g_s}>$ and $<\mathbf{g_n}>$ are respectively the means of the images containing the signal, and the background-only images. The covariance matrix inversion in Eq. (3) is often impractical to implement, since for N by N pixels images, the size of this matrix is N^2 by N^2. Moreover, the large number of independent images needed for getting a non-singular estimate of the covariance matrix is rarely reached in a typical experimental study. In order to overcome these computational issues, the Hotelling observer may be approximated by reducing the images to a small set of variable responses channels [5,14,15,28,29]. The CH observer template is then given by:

$$\mathbf{w}_{\mathbf{CH}} = (\mathbf{K}_{\mathbf{b,c}} + \mathbf{K}_{\boldsymbol{\varepsilon}})^{-1}\,\mathbf{s} \tag{4}$$

In Eq. 4, $\mathbf{K}_{\mathbf{b,c}}$ is the channelized covariance matrix, which is computed from the background images as $\mathbf{K}_{\mathbf{b,c}}= <(\mathbf{T}^{\mathbf{t}}\mathbf{g_n}-<\mathbf{T}^{\mathbf{t}}\mathbf{g_n}>)(\mathbf{T}^{\mathbf{t}}\mathbf{g_n}-<\mathbf{T}^{\mathbf{t}}\mathbf{g_n}>)^{\mathbf{t}}>$, where the column vectors of the matrix **T** each represent the spatial profile of a channel. The noiseless covariance matrices $\mathbf{K}_{\mathbf{b,c}}$ were estimated by sampling, using the 1,400 signal-absent images of the 2-AFC experiments. **s** is the expectation of the signal seen through the channels: $\mathbf{s}=\mathbf{T}^{\mathbf{t}}[<\mathbf{g_s}>-<\mathbf{g_n}>]$. $\mathbf{K}_{\boldsymbol{\varepsilon}}$ is the covariance matrix of the internal noise. In this work, internal noise was assumed to be zero mean, independent in each channel, with variance proportional to the variance of the background noise in each channel, with a proportionality factor β. The noisy channelized background covariance matrix $\mathbf{K}_{\mathbf{b,c,n}}$ could thus be computed as:

$$\begin{cases} (\mathbf{K}_{\mathbf{b,c,n}})_{i,j} = (\mathbf{K}_{\mathbf{b,c}})_{i,j} \text{ if } i \neq j \\ (\mathbf{K}_{\mathbf{b,c,n}})_{i,j} = \beta(\mathbf{K}_{\mathbf{b,c}})_{i,j} \text{ if } i=j \end{cases} \tag{5}$$

The two kinds of channels used for this study are the DDOG channels as described by Abbey and Barrett [29], and channels using Gabor functions. The Gabor channels were constructed as:

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

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 five orientations, 8 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 to $\Lambda_{\max}$=192 pixels in discrete steps spaced by a multiplicative factor of 1.4.

These models were also adapted to the SKS task using the sum of likelihood rule [30]. This involved computing one template per possible signal, and the distributions of the response λ for each template and with the 1,400 signal-present and 1,400 signal-absent images used in the experiments. Details of the whole procedure can be found in the paper by Zhang *et al.* [30].

The performance of these models were estimated by the direct use of the corresponding templates in 2-AFC experiments, using the same backgrounds and signals as the human observers. The obtained P_c*'s* were then compared to human observers performance.

3. RESULTS

3.1 Robustness of the results

Potential sources of bias in the human observers 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 effect was also tested, by comparing the proportion of correct answers of the first and last 200 trials to P_c for the whole 1,400 trial for each experiment and observer. Again, no significant deviation from random differences between the beginning, the end, and the experiment as a whole was shown for any observer, suggesting that they 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 mutually disjoint subsets of 50 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 [31]. The correlation coefficient was not significantly different from 0 for any observer. This means 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.

3.2 Human observers results

3.2.1 Benign masses

P_c values averaged over the four human observers for each of the 13 experimental conditions described in Fig. 1 are given in Table 1 in section 3.3. As an example of typical individual experimental results, human observers' P_c for the 6.5 mm benign masses are given in Fig. 3(a) for SKE tasks, and (b) for SKS. For SKE experiments, P_c averaged over the four observers was 77.7±0.5% for the real backgrounds, and 78.3±0.6% for the CLB (p=.43). For the SKS tasks, the difference between real images (P_c=79.0±0.5%) and CLB (77.4±0.7%) was statistically significant (p=.04), but this effect is only due to Obs. #3 results, while the others performed similarly with the two background types.

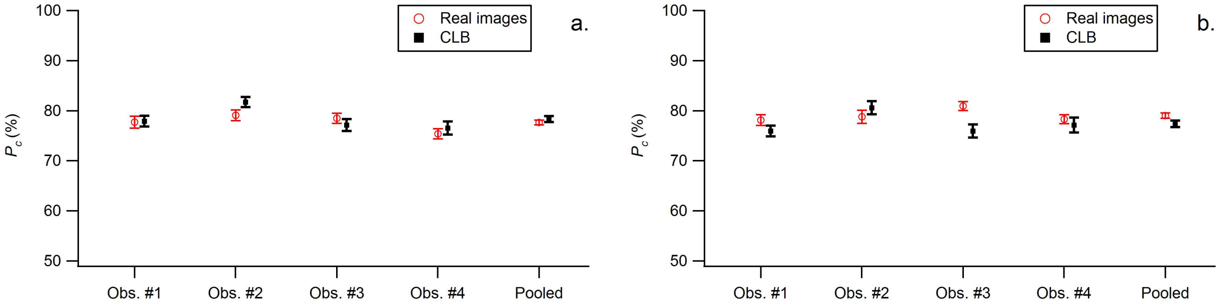

Fig. 3. (a) Human observers' proportions of correct answers (P_c) for the SKE tasks with the 6.5 mm benign masses. The rightmost values for each figure were obtained by pooling all observers data. The error bars represent the standard error (b) Same for the SKS tasks.

When comparing SKE and SKS tasks for the 6.5 mm masses, there is no significant difference neither for real backgrounds (p=.09) nor CLB (p=.24).

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 4.7% in P_c's, p<10^{-4}), but smaller in the SKS task (1.8%, p=.026). This trend is also visible in the size uncertainty experiment with signal size ranging from 5.5 to 9.5 mm (Fig. 4). In this case again, ANOVA analysis, performed with the pooled human observer data, indicated that the mass size dependence was significant (F=2.46, df=4, p=.04).

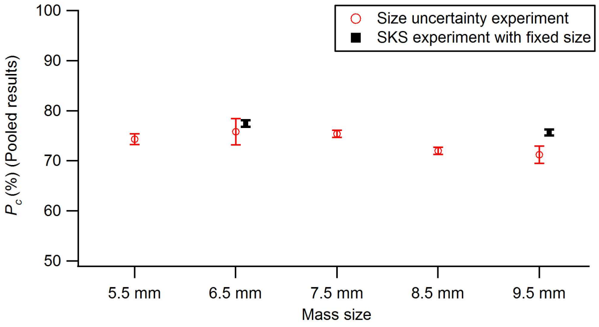

Fig. 4. Pooled observers results in the size uncertainty experiment (open circles). For comparison, the performance in the SKS experiments with fixed size signals are shown (black squares).

3.2.2 Malignant masses

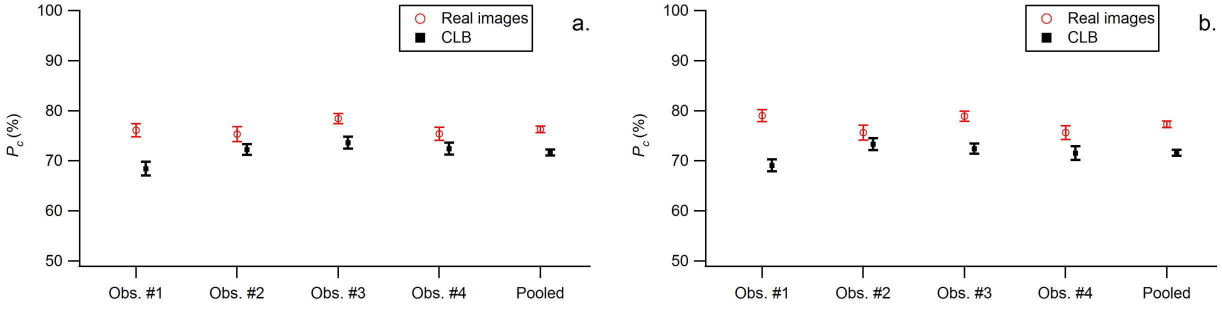

Fig. 5. (a) Human observers results for the SKE tasks with the 6.5 mm malignant masses. The rightmost values for each figure were obtained by pooling all observers data. (b) Same for the SKS tasks.

Human observers' P_c's for the 6.5 mm malignant masses are shown in Fig. 5 (a) (SKE) and (b) (SKS). For these masses, pooled P_c is significantly higher with the real images than with the CLB for both SKE (4.6%, $p<10^{-4}$), and SKS tasks (6.7%, $p<10^{-4}$).

As for benign masses, the performance is very close between SKE and SKS tasks for both real backgrounds (1%, p=.21) and CLB (0.9%, p=.92).

The mass size effect is different than for benign masses. 9.5 mm malignant masses are significantly better detected than 6.5 mm masses: the difference is clear in both SKE (3.5%, p=.003) and SKS (4.8%, $p<10^{-4}$) tasks.

3.3 Model observers

Table 1 presents P_c values obtained with the model observers in the same 2-AFC conditions as human observers. Bold values indicate deviations from human observer results smaller than 5% in units of P_c.

Table 1. Comparison of human and model observers performances for the different experimental conditions. Bold values indicate absolute differences between human and model results smaller than 5% in units of P_c.

Benign	Conditions	Human	DDOG	GABOR
masses	Real,SKE,6.5	77.7%	**81.1%**	**79.1%**
	CLB,SKE,6.5	78.3%	**81.4%**	**78.3%**
	Real,SKS,6.5	79.0%	**81.1%**	**77.5%**
	CLB,SKS,6.5	77.4%	**80.8%**	**76.3%**
	CLB,SKE,9.5	73.6%	**75.9%**	88.4%
	CLB,SKS,9.5	75.6%	**75.9%**	89.3%
	CLB,SKS,5.5-9.5	73.8%	**71.6%**	**78.1%**

Malignant	Conditions	Human	DDOG	GABOR
masses	Real,SKE,6.5	76.3%	**79.7%**	**78.0%**
	CLB,SKE,6.5	71.7%	78.0%	**75.2%**
	Real,SKS,6.5	77.3%	**79.6%**	**77.0%**
	CLB,SKS,6.5	71.6%	78.6%	**74.0%**
	CLB,SKE,9.5	74.2%	**72.9%**	**77.6%**
	CLB,SKS,9.5	76.4%	**73.5%**	**78.0%**

CH observer with DDOG channels was implemented without adding internal noise, whereas β factor in Eq. 5 was set empirically to 2 (benign masses) and 20 (malignant masses) for the CH observer with Gabor channels. This way, these two models fit human results within 5% accuracy in most experimental conditions.

Both models are good predictors for human response to most experimental conditions. For benign masses, the models P_c's do not differ by more than 1% when the only change in conditions is real backgrounds versus CLB, and by 1.5% between SKE and SKS tasks. The decrease in performance with larger masses is reproduced by the CH observer with DDOG channels, whereas the performance of CH observer with Gabor channels increases with the size of the signal. For this reason, results for the experiment with benign masses ranging from 5.5 to 9.5 are less precisely reproduced by the latter. For malignant masses, the performance drop between real and CLB images is predicted by both models, although less strongly than for human observers. Again differences between SKE and SKS tasks are less than 1.5% in units of P_c.

4. DISCUSSION

4.1 Influence of the background and the local statistics

In a previous study [9], we had shown 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 a dissociation 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 for ensuring comparable conditions. For this reason, we tried to follow a more local approach in the current study, and matched the statistics specifically on the central part of the images, where the observers focused.

For benign masses, human observers achieved very close performance for both backgrounds. Individual observers' differences in P_c did not exceed ±2.6%, except for one observer in the SKS task (Obs. #3, 5%). For these signals with sharp edges, comparable to the one 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: The systematically lower performance with CLB suggests that their random, stationary nature is more likely to hide signals with smooth edges than nonstationary images, in a way similar to Zhang *et al.* [32] findings with highly nonstationary backgrounds. Since the malignant masses look like blobs, they might have been more difficult to detect on the CLB, which contain clusters of blobs by construction.

Finally, as in the previous study, the best performance with CLB images was generally obtained by Obs. #2, one of the co-authors (CC). However, this observer was regularly outperformed by one or several others in the experiments with real images. This distinctive knowledge of CLB could have been an unconscious help for this observer when performing the experiments.

4.2 Influence of the signal shape uncertainty

Quite surprisingly, human observers performed as well for the SKS tasks as for the corresponding SKE tasks, when the signal size was constant over the experiment and the only uncertainty was its shape. This shows that, although after being trained with high contrast versions of the same signals as in the actual experiments, human observers were not able to develop and use 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 our study, SKS experiments were performed with a constant signal size, whereas Zhang *et al.* used projected ellipsoid signals ranging from 1x1 to 7.5x3 mm, introducing important size uncertainty. Furthermore, our last experiment with signal size ranging from 5.5 to 9.5 mm confirmed that introducing size uncertainty lowered the detection performance of the observers, compared to the SKS experiments with fixed size (comparison points in Fig. 4, p=.25 for 6.5 mm masses, p=.02 for 9.5mm). 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 disks signals on correlated noise, or to more general findings with contrast-detail experiments with power-law noise approximating mammograms [4,5]. Judy [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, in order to maintain a constant non-prewhitening matched filter observer performance, whereas we used a fixed signal contrast for all mass sizes.

In our study, the effect of signal shape uncertainty has been investigated with CLB and real images, while the size uncertainty experiment was only performed with the CLB. However, results from the previous section with benign masses 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 between both background types.

Our findings suggest that human observers are much more sensitive to signal size uncertainty than signal shape uncertainty. Following this idea, assessing human observer detection performance for such non-trivial 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.

4.3 Model observers versus human observers

Excellent agreement between model and human observers results for nearly all conditions (Table 1) indicate that CH observer with 10 DDOG channels, or more sophisticated CH observer with 80 Gabor channels are good predictors of the effect of size and shape uncertainty over human detection performance. It is remarkable to note that circularly symmetric DDOG channels lead to such good results, even with signals that are as highly non-symmetric in the spatial domain as malignant masses. The single effect that is only partly predicted by this model observer is the human performance difference between real backgrounds and CLB with malignant masses. Otherwise, all other human results fit within 3.5% accuracy in units of P_c: SKE versus SKS performance, 6.5 versus 9.5 mm signals, experiment with signal size uncertainty.

The noiseless CH observer with 80 Gabor channels was close to a perfect observer, since its P_c ranged from 95 to 100% for all different conditions. However, as our aim was to reproduce human observer results, it was necessary to degrade its performance by adding internal noise. A good match with human results was obtained with channel internal noise [29], which is consistent with a recent study that compared channel internal noise to variable decision noise for this model observer [33]. The values for the noise proportionality constant β (see Eq. 5) are empirical, and were chosen in order to fit most human results for benign and malignant masses. Since human observers' precise strategies for detecting these two different kinds of signals are still unclear, it might be questionable whether introducing two different noise levels (β=2 for benign masses, β=20 for malignant masses) is appropriate, but it was the only way to fit human results for both signal types. This way, and as for CH observer with DDOG channels, human observers results and trends are remarkably well fitted by the CH observer with Gabor channels. The only exception is for the large benign masses, where this model outperforms human observers. This is the source of the difference for the size uncertainty experiment.

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 did not differ significantly between SKE and SKS experiments, when the signal size was kept constant. However, human observers were sensitive to signal size uncertainty, and failed maintaining their performance between fixed-size and size-varying experiments.

Human observers performed similarly with real backgrounds and CLB for detecting benign masses. However, malignant masses were more difficult to detect on CLB than on real images. This difference, and the strategies used for detecting the two mass types on the two different backgrounds will be further investigated, using Human Linear Template [9,34-37]. Among other goals, we hope to be able to determine if the strategies used for benign and malignant masses are based on the same approach, or if human observers change their strategies between sharp-edged benign masses and malignant masses with smoother edges.

Excellent agreement with human observers was obtained for CH observers using circularly symmetric DDOG channels, or Gabor channels with an appropriate internal noise level. They provide a very good way to reproduce or predict human observers results in the same experimental conditions, and objectively optimize image quality.

Finally, one has to keep in mind that the SKS approach (or SKEV, since both have been proven to lead to correlated results [19,21,22]) is still far from the actual clinical world. Many other factors influence the radiologists' ability to correctly detect masses on mammograms: much wider search space, signal location uncertainty, and extremely low prevalence on the order of 7 per 1000 cases [38,39], for example. Moreover, real clinical strategies also include comparison with the other breast and global breast architecture analysis. However, introducing all these elements in psychophysical experiments would lead to overly hard to interpret results, correlated in many ways. 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 observers detection strategy and performance.

ACKNOWLEDGMENTS

The authors would like to thank Cécile Althaus, Gaëlle Bocksberger, and Valentine Clémence for their participation to the detection experiments. This work was supported by Swiss National Science Foundation grant 320000-113863/1

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," Optics Express **4**, 33-43 (1999).

[8] C. Castella, K. Kinkel, F. Descombes, M. P. Eckstein, P.-E. Sottas, F. R. Verdun, F.O. Bochud, "Mammographic texture synthesis using genetic programming and clustered lumpy background," Proc. SPIE **6146**, 238-249 (2006).

[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, 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 Series in Pure and Applied Optics, Hoboken, 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", *Handbook of medical imaging. Volume 1. Physics and psychophysics*, SPIE press, Bellingham, 2000

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

[24] NEMA, *Digital Imaging and Communications in Medicine (DICOM) Part 14: Grayscale Display Standard function*, Rosslyn, 2000

[25] C. J. D'Orsi, *Illustrated Breast Imaging Reporting and DATA system (BIRADS)*, American College of Radiology, Reston, 1998

[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 *et al.*, "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] 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).

[29] 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).

[30] 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).

[31] R. Lowry, "VassarStats: Web Site for Statistical Computation," http://faculty.vassar.edu/lowry/VassarStats.html

[32] 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).

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

[34] A. J. Ahumada, Jr, "Classification image weights and internal noise level estimation," Journal of Vision **2**, 121-131 (2002).

[35] J. A. Solomon, "Noise reveals visual mechanisms of detection and discrimination," Journal of Vision **2**, 105-120 (2002)

[36] R. F. Murray, P. J. Bennett, and A. B. Sekuler, "Optimal methods for calculating classification images: Weighted sums," Journal of Vision **2**, 79-104 (2002).

[37] 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).

[38] Perry N, Broeders M, de Wolf C, Törnberg S, Holland R, von Karsa L (eds). *European Guidelines for quality assurance in breast cancer screening and diagnosis, 4th ed.*, Office for Official Publications of the European Communities, Luxembourg, 2006

[39] 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).