

Search for lesions in mammograms: Statistical characterization of observer responses

François O. Bochud^{a)}

Institut Universitaire de Radiophysique Appliquée, Grand-Pré 1, CH-1007 Lausanne, Switzerland

Craig K. Abbey

Department of Biomedical Engineering, University of California, Davis, California 95616

Miguel P. Eckstein

Vision & Image Understanding Laboratory, Department of Psychology, University of California, Santa Barbara, California 93106

(Received 16 December 2002; revised 7 October 2003; accepted for publication 7 October 2003; published 8 December 2003)

We investigate human performance for visually detecting simulated microcalcifications and tumors embedded in x-ray mammograms as a function of signal contrast and the number of possible signal locations. Our results show that performance degradation with an increasing number of locations is well approximated by signal detection theory (SDT) with the usual Gaussian assumption. However, more stringent statistical analysis finds a departure from Gaussian assumptions for the detection of microcalcifications. We investigated whether these departures from the SDT Gaussian model could be accounted for by an increase in human internal response correlations arising from the image-pixel correlations present in $1/f$ spectrum backgrounds and/or observer internal response distributions that departed from the Gaussian assumption. Results were consistent with a departure from the Gaussian response distributions and suggested that the human observer internal responses were more compact than the Gaussian distribution. Finally, we conducted a free search experiment where the signal could appear anywhere within the image. Results show that human performance in a multiple-alternative forced-choice experiment can be used to predict performance in the clinically realistic free search experiment when the investigator takes into account the search area and the observers' inherent spatial imprecision to localize the targets. © 2004 American Association of Physicists in Medicine. [DOI: 10.1118/1.1630493]

Key words: observer performance, psychophysical analysis, anatomical backgrounds, image perception, mammography

I. INTRODUCTION

A fundamental aspect of medical images that limits the physician's ability to visually detect and classify disease is the presence of the background luminance-variations in the image. A first component of these background variations can be attributed to noise arising from the variable number of photons reaching the image-receptor (quantum noise) and being filtered by the image receptor modulation transfer function (MTF). A second component consists of the presence of normal anatomic structures in the image that may mask the occurrence of disease and that actually act as a form of noise in the detection process.

There has been a rich history of experiments studying visual detection in white Gaussian noise that is an approximation to quantum noise.¹⁻⁶ In the last 15 years, many studies have concentrated on studying this process in more complex computer generated backgrounds that attempt to mimic both the presence of quantum noise and structured backgrounds.⁷⁻¹³ Although some experiments were already realized in the early 1970s,^{14,15} it is only recently that access to digital/digitized medical images has intensified studies of visual detection in the presence of real anatomical backgrounds.¹⁶⁻²⁴ Arguably these latter studies are the most relevant to the real clinical scenario.

An important question in understanding the process by which physicians detect abnormalities in medical images is studying how detection performance varies with uncertainty about the spatial location of the signal. Many classical studies in visual signal detection have investigated the effect of uncertainty for the detection of a simple luminance increment in the absence of noise.²⁵⁻²⁷ These studies show that performance degradation with uncertainty about the spatial position of the signal is consistent with models derived in the context of signal detection theory (SDT).^{28,29} In this class of models the observer is assumed to monitor a noisy response to every possible location monitored, and then selects the location with the maximum response. The responses to each location are generally assumed to be Gaussian distributed and statistically independent of each other.^{30,31,24} Performance degrades with increasing number of possible signal locations owing to the increasing probability that at least one of the responses to the noise-only locations will exceed the response to the signal location. Such a model has been successful at predicting target detection and localization both with and without external image noise. Swensson and Judy² and Burgess and Ghandeharian⁴ investigated localization of a signal embedded in additive Gaussian white noise in one of M locations [multiple-alternative forced-choice task

(MAFC)]. They found that degradation in performance as measured by the percent of trials in which the observers correctly localize the signal was well predicted by SDT. Eckstein and Whiting¹⁹ have subsequently applied this approach to localization of simulated filling defects in arteries in x-ray coronary angiographic backgrounds finding that SDT models with the Gaussian statistical independent internal response assumption was a good approximation to the human observer.

The goal of the present study is to investigate whether the SDT models with the Gaussian-independent response assumption can be used to predict visual detection performance with a varying number of possible spatial locations for the target signal in x-ray mammographic backgrounds.

Unlike previous treatments, we investigate whether a SDT model with other non-Gaussian response distributions can better predict human-observer performance as a function of the number of possible signal locations. (Although Eckstein and Whiting¹⁹ considered the effect of rectangular and Laplace response distributions on performance as a function of number of locations, they did not evaluate how these distributions predicted the human data compared to the Gaussian internal response model.) In addition, previous studies have assumed that the unobservable internal responses of human observers are statistically independent. When independent noise samples in the different possible signal locations are used then the luminance pixel variations in one location are independent of the luminance pixel variations in the other locations and the independent internal response assumption seems reasonable. However, the possibility of response correlations seems more likely when the task involves localizing a signal within image backgrounds that have long-range spatial correlations such as power-law spectrum ($1/f^n$) images, which has been used to describe variability of natural images^{32,33} including breast tissue images by mammography.²⁴ These images contain slowly spatially varying luminance changes with significant low-pass frequency components, and therefore luminance variations in one of the possible signal locations may be correlated with those in another location within the correlation range of the tissue. In a MAFC task, a positive correlation among internal responses will decrease the probability that a response to any nonsignal location will exceed the signal location response, and thereby elevate performance over a task in which the locations are statistically independent. Under the assumption that correlations between internal responses are constant (equicorrelation), it has been shown that detection performance as a function of number of locations is indistinguishable from the case of statistical independence among responses.³⁴ (For this case human performance as a function of number of locations can be equally well fit by a model with statistically independent responses and one with a fixed correlation among response.) On the other hand, if the spatial separation among signal locations changes, then it seems unlikely that the equicorrelation model will apply. If this is the case, then the variation in performance as a function of the number of possible signal locations might depart from that predicted from the statistically independent-response model.

In this paper, we investigate the possibility that the human internal responses are correlated when possible target locations are near each other in the presence of long-range image correlations. A separate psychophysical experiment is designed to specifically test whether human observer internal responses to the possible signal locations in x-ray mammograms are correlated.

Finally, we also investigate performance in a task where the lesion might appear (with equal probability) anywhere within the image (free search). Arguably, the free search task is the most similar to the physician's task of scrutinizing areas within an image and localizing a lesion. However, this task presents the additional difficulty in that the number of locations is unknown, and so the observer responds to a continuum of possible locations. Since many of the possible locations will be very close to each other, they are most likely not statistically independent. Can performance in the free search task be predicted by human performance in the MAFC tasks? If so, then the more stylized (but simpler to analyze and model) MAFC tasks can be used to predict the more realistic location free search task. Prior to describing the experiments in detail we briefly present the theory of signal detection as applied to the MAFC tasks.

II. THEORY

A. Performance in a MAFC experiment

In a MAFC experiment the observer is presented with an image that contains the signal in one of M locations. The observer's task is to choose the location that is the most likely to contain the signal. Performance is measured with the proportion of the total trials in which the observer correctly identifies the signal location (Percent Correct, Pc). In signal detection theory, the observer is assumed to monitor a noisy internal response to each of the possible signal locations. The variability in the responses might be due to noise intrinsic to the observer or to variability in the stimuli (external noise). On each trial, the SDT model assumes that the observer chooses the location associated with the highest internal response. Performance (Pc) will degrade as a function of increasing number of possible locations owing to the increasing probability that the response to any one noise-location will exceed the response to the signal plus noise location. The probability that the observer correctly chose the location of the signal is obtained by calculating the probability that the internal response to the signal plus noise location will exceed the responses to all noise-only locations. If the observer's internal response λ to the signal plus noise location is given by the probability density function (pdf) $p(\lambda|s)$, and the internal response pdf for a noise-only location is given by $p(\lambda|n)$, then Pc can be written as

$$P_c = \int_{-\infty}^{\infty} \left(\int_{-\infty}^{\lambda'} p(\lambda|n) d\lambda \right)^{M-1} p(\lambda'|s) d\lambda', \quad (1)$$

where M is the number of possible signal locations.

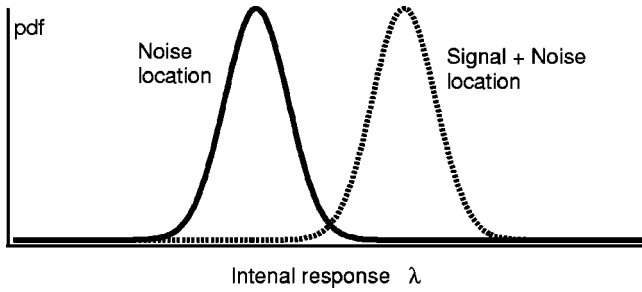


FIG. 1. Probability density function (pdf) of the internal observer response for a noise only location and for a signal+noise location.

1. The Gaussian assumption

Typically, the internal responses to the signal location and noise-only locations are assumed to be equal variance (σ_λ^2) Gaussian distributed and statistically independent.³⁵ With the Gaussian assumption (Fig. 1), the pdf for the response to the noise-only location is given by

$$p(\lambda|n) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \frac{\lambda^2}{\sigma_\lambda^2}\right), \quad (2)$$

and that for the response to the signal plus noise location is given by

$$p(\lambda|s) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} \frac{(\lambda - d')^2}{\sigma_\lambda^2}\right), \quad (3)$$

where d' is the index of detectability and is defined as the difference between the mean responses to the signal plus noise location and noise-only location divided by the square root of the average variance of the two classes:

$$d' = \frac{\langle \lambda_s \rangle - \langle \lambda_n \rangle}{\sigma_\lambda}, \quad (4)$$

where λ_s and λ_n are the responses for a signal plus noise and a noise only location, respectively, σ_λ is the standard deviation of the internal response assumed to be equal for both the signal plus noise and noise only location, and $\langle \cdots \rangle$ is the mathematical expectation operator.

For the Gaussian statistically independent pdfs Eq. (1) reduces to

$$Pc = \int_{-\infty}^{\infty} \Phi(x)^{M-1} \phi(x - d') dx, \quad (5)$$

where $\Phi(x)$ is the cumulative Gaussian function, and $\phi(x)$ is the Gaussian function.

When a series of MAFC experiments is realized with different numbers of M locations, Pc decreases as M increases even though d' is fixed. The larger the number of locations, the larger the probability that the response to any noisy location will exceed the response to the signal location, and therefore the lower Pc (see Fig. 2).

In actual human psychophysical experiments, the investigator cannot observe d' directly but rather measures Pc as a function of number of locations (M) and then infers an index of detectability (which we refer as d_{MAFC} to emphasize that it

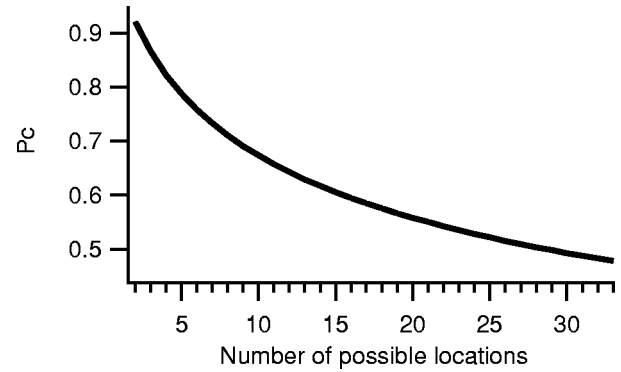


FIG. 2. Typical behavior of Pc vs the number of possible alternatives M .

is an estimate of d' obtained from a MAFC experiment) from each Pc using Eq. (5). If human performance is consistent with the SDT-Gaussian model then d_{MAFC} should be approximately constant across MAFC conditions.

2. Non-Gaussian probability density functions

The relationship between Pc and number of possible signal locations in general depends on the shape of the probability density function. Consider for instance the case of a pdf given by

$$p(\lambda) = \alpha \exp\left(-\frac{1}{2} \left(\frac{|\lambda|}{\sigma}\right)^\beta\right), \quad (6)$$

where α is a normalizing parameter and β defines the shape of the distribution. For $\beta=2$, Eq. (6) reduces to the usual Gaussian distribution. Values of β lower than 2 result in a pdf that is more spread than the Gaussian and values of β greater than 2 result in a pdf that is more compact than the Gaussian (see Fig. 3). Figure 4 shows how the degradation of Pc as a function of M varies with the different pdfs (different β values). As β increases Pc degrades less as a function of M .

On the other hand, if the human internal responses were not Gaussian distributed, and the investigator computed an index of detectability (d_{MAFC}) from the experimentally measured Pc values erroneously assuming a Gaussian distribution [through Eq. (5)], it would lead to a nonconstant value

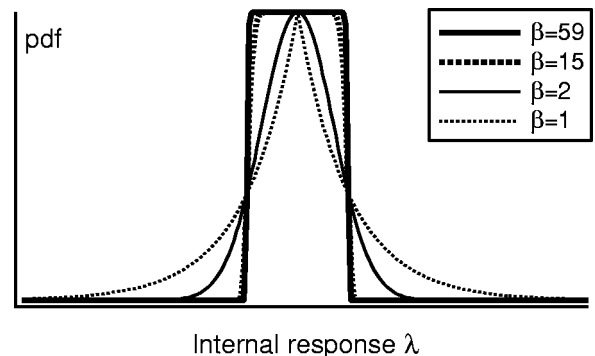


FIG. 3. Examples of different observer response pdfs defined by Eq. (6) for different values of β .

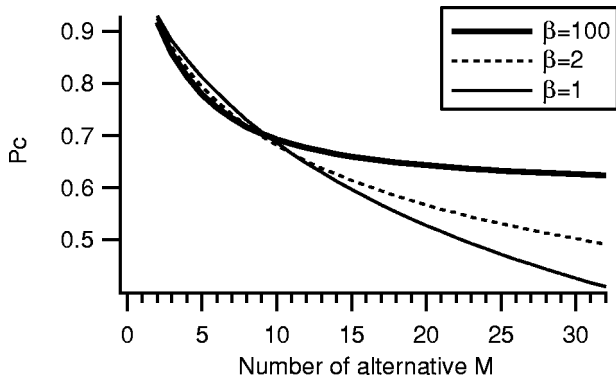


FIG. 4. Example of evolution of P_c vs the number of locations for a pdf defined according to Eq. (6) for different values of β .

of d_{MAFC} across the number of alternatives. For example, Fig. 5 shows the transformed d_{MAFC} values for P_c values in a MAFC experiment with pdfs different than Gaussian. In other words, the constancy of d_{MAFC} can be used to determine whether human performance is consistent with the Gaussian pdf assumption.

B. Internal response correlation

The standard SDT model typically assumes that the responses to the different locations are statistically independent. In the absence of external noise or when the samples of external noise at the possible signal locations are independent, this assumption is reasonable. In other instances such as the case of detecting a signal within an image containing slowly spatially varying luminance changes (low pass Gaussian or $1/f^n$ noises), the possibility of response correlations between responses to different locations arises.

The correlation between the responses to two different locations can be characterized by a correlation coefficient defined as^{34,36}

$$r_{k,m} = \frac{\langle (\lambda_{k,i} - \langle \lambda_k \rangle) (\lambda_{m,i} - \langle \lambda_m \rangle) \rangle}{\sqrt{\langle (\lambda_{k,i} - \langle \lambda_k \rangle)^2 \rangle \langle (\lambda_{m,i} - \langle \lambda_m \rangle)^2 \rangle}}, \quad (7)$$

where λ is the response, the subscripts k and m refer to two different spatial locations, $\lambda_{k,i}$ is the observer's internal re-

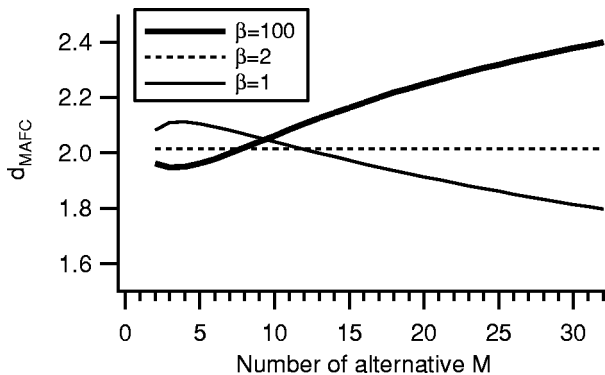


FIG. 5. Example of the variation of d_{MAFC} across the number of locations for a pdf defined according to Eq. (6) but computed from P_c as if the responses were Gaussian distributed.

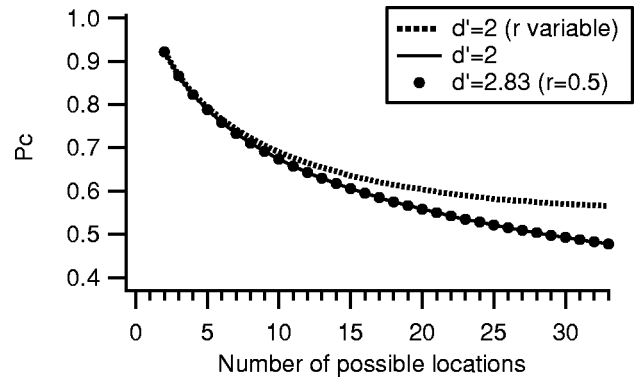


FIG. 6. Example of the variation P_c across the number of locations for a Gaussian pdf ($d' = 2$: continuous line; $d' = 2.83$ and $r = 0.5$: circles; $d' = 2$ and r varies linearly [with $r = 0$ for $M = 2$ and $r = 0.2$ for $M = 32$: dotted line]).

sponse to location k in the i th trial, $\lambda_{m,i}$ is the observer's internal response to location m in the i th trial, and the expectation $\langle \dots \rangle$ is performed among trials. When the model responses are statistically independent, then $r_{k,m} = 0$. The effect of a positive correlation between responses is to decrease the probability of any one response to the noise-only location exceeding the response to the signal plus noise location. As shown in Refs. 34 and 36, if the correlation coefficient is considered constant among the locations ($r_{k,m} = r$), then the usual relationship between P_c versus index of detectability still holds if the index of detectability, d' , is replaced by a corrected index of detectability (d'_r) defined as

$$d'_r = \frac{d'}{\sqrt{1-r}}. \quad (8)$$

As is apparent in Eq. (8), a given value of d'_r can be consistent with many combinations of d' and r . Therefore if the correlation is constant as the number of locations increases then a P_c vs M function can be identical for a variety of combinations of d' and r . Figure 6 shows how the degradation of P_c with M could be consistent on the one hand with a given d' and statistically independent responses ($r = 0$), and on the other hand with a lower d' and a positive correlation r . In other words, given a set of measured P_c as a function of M , the experimenter cannot determine whether the responses are correlated or not.

Suppose now that the correlation between internal responses changes as a function of M . For instance, suppose, as illustrated in Fig. 6, that the correlation is $r = 0$ for $M = 2$ and increases linearly to $r = 0.2$ for $M = 32$. In this case, the degradation of P_c as a function of M will depart from the statistically independent scenario. If the investigator is unaware of this increasing correlation and the measured P_c values are transformed to d_{MAFC} assuming Gaussian statistical independence then he/she would find an increasing d_{MAFC} as a function of M (Fig. 7).

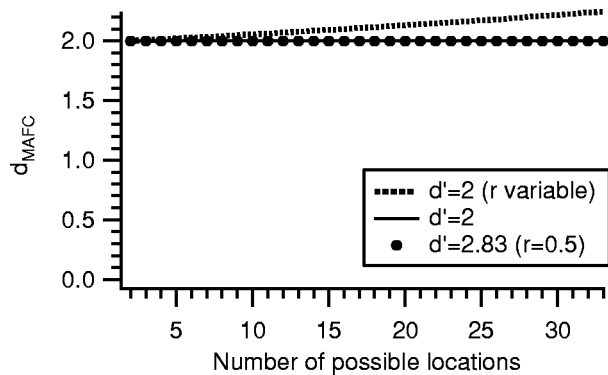


FIG. 7. Values of d_{MAFC} computed from P_c as if the pdf were Gaussian distributed with or without correlation. Symbols have the same meaning as in Fig. 6.

III. GENERAL METHODS

A. Test images

The images used in this study were breast x-ray images of patients without known pathology. They were obtained by a digital 510(K)-Bennett mammographic unit (Trex Medical Corporation, Waltham, MA) (courtesy of University of California Los Angeles Medical Center). They have a pixel size of 0.04 mm and are coded as 14 bits per pixel with a gray level proportional to the x-ray exposure.

In mammography, the radiologist is typically interested in microcalcifications, which sizes are of the order of 0.2 mm, or in tumors or masses that are about 50 times larger (1 cm). Therefore microcalcifications are generally visually searched with a magnifying glass (magnification factor of about 8). Masses are searched for by looking at the image in its original scale. In order to take this practice into account, in the present study two sets of images were created: one at the original scale for microcalcification search (0.04 mm per pixel), and the other is averaged in order to have a pixel size close to the image-display pixel size (0.3 mm).

The image size was fixed to 256×256 pixels and displayed a continuous area inside the breast (see Fig. 8). Another set of images was generated in which the possible locations could be anywhere inside the image except for a 10-pixel-wide external frame (free search condition). For the five different conditions ($M=2, 4, 9, 16$, and free search), the same set of 240 images was used. The maximum number of locations was set to 16 in order to guarantee that the observers did monitor all the possible locations. By using a higher number of alternatives in this phase of the MAFC experiments, one might have risked infringing on the underlying model of this study (that the observer chooses the maximum response out of all the possible locations) and to begin observing satisfaction of search effect.

For the tumor experiments only two and nine possible locations were considered. Each possible location was at the center of an individual 128×128 pixel image. A sample of a 9AFC image is presented in Fig. 9.

In the correlation investigation, a 2AFC experiment was set with 128×256 pixel images in which the distance be-

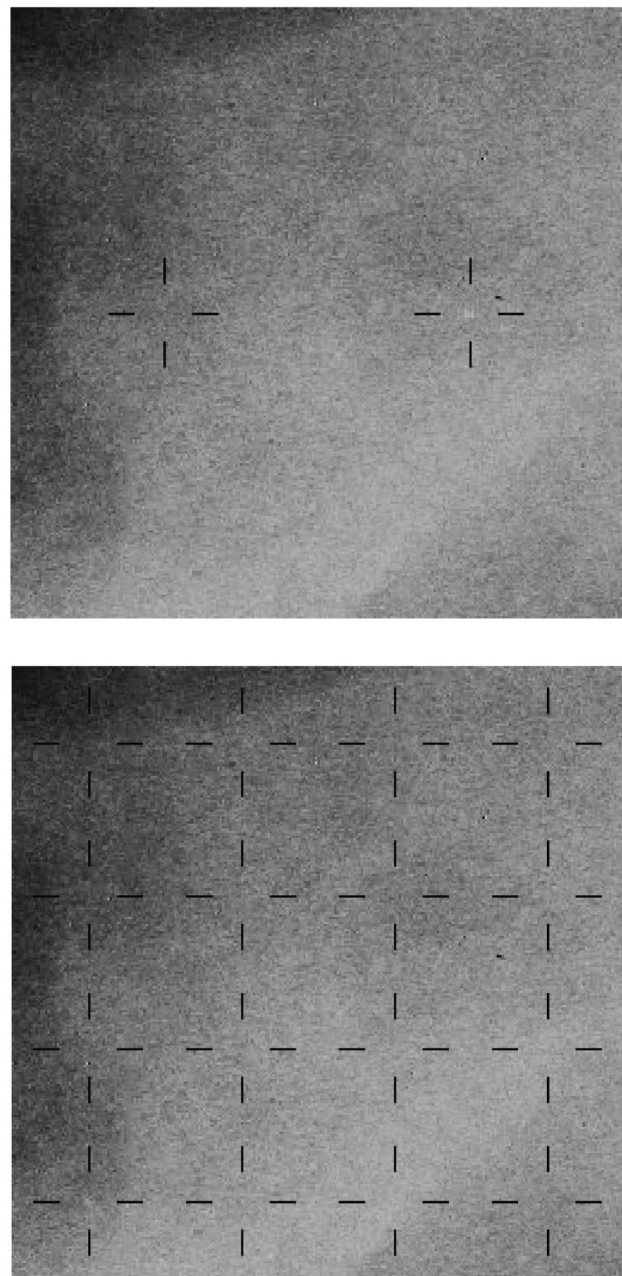


FIG. 8. Example of images used in the 2AFC and 16AFC experiments at the microcalcification scale. The groups of four black whiskers are cues added on the images in order to reduce signal location uncertainty.

tween the two possible locations could be varied (see Fig. 10).

Except for the free-search images, possible signal locations were always indicated by fixed cues. We did not provide the possibility to toggle the cues off because in these experiments eye-movement searching was not the subject of interest.

B. Signal

Because of the proportional relationship between gray level and radiation exposure of the investigated images, an

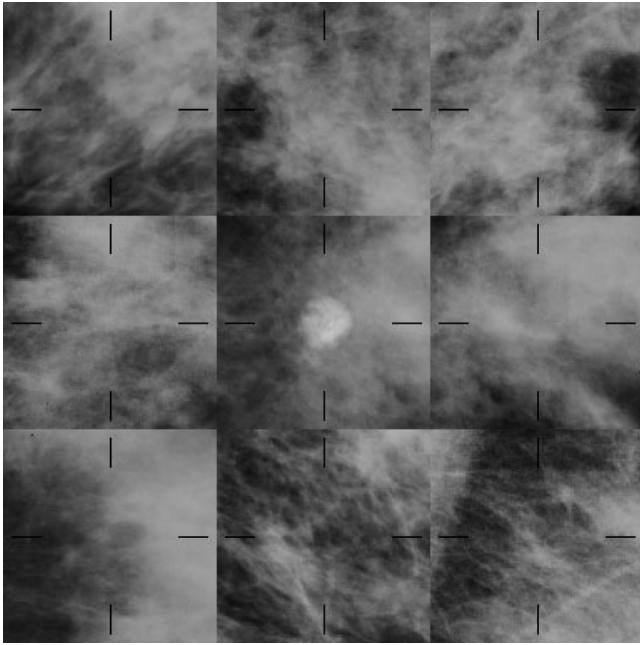


FIG. 9. Example of an image used in the 9AFC experiment at the tumor scale.

image containing the signal (\mathbf{g}_s) is simulated by multiplication of the background (\mathbf{b}) and the signal profile pixel by pixel:

$$\mathbf{g}_s = (\mathbf{S} + \mathbf{I})\mathbf{b}, \quad (9)$$

where \mathbf{I} is the identity matrix and \mathbf{S} is a diagonal matrix containing the signal profile \mathbf{s} . An element (i, j) of the matrix \mathbf{S} is expressed as $\mathbf{S}_{ij} = s_i \delta_{ij}$, where the symbol δ_{ij} is used as the Kroneker function. In this formalism, the simulated object thickness, written as the image vector \mathbf{t} , can be directly related to the signal \mathbf{s} as $s_i + 1 = \exp(-\mu t_i)$, where μ is the attenuation coefficient.

In this study, the signal is a two-dimensional-projected sphere (0.2 mm and 1.0 cm diameter for microcalcification and tumor, respectively) filtered by the modulation transfer function of the image system.

The contrast used in this work is the root mean square contrast defined as



FIG. 10. Example of an image used to investigate the presence of response correlation (interlocation distance of 208 pixels).

$$\begin{aligned} C &= \sqrt{\sum_i \frac{(g_{s_i} - b_i)^2}{b_i^2}} = \sqrt{\sum_i \frac{(s_i b_i + b_i - b_i)^2}{b_i^2}} \\ &= \sqrt{\sum_i (e^{-\mu t_i} - 1)^2}, \end{aligned} \quad (10)$$

where g_{s_i} is the pixel value of the i th image pixel of the image containing the signal, and b_i is the pixel value of the i th image pixel of the same image that only contains the background, s_i is i th component of \mathbf{s} , and t_i is i th component of \mathbf{t} . It can be seen that the images of an ensemble generated with a given value of μ all have the same contrast. For the two series of MAFC experiments (microcalcification and tumor scales), five contrast values were considered in each situation.

C. Image display

Images were presented on an Image systems M17L 0.3 mm pixel size monochrome gray-scale monitor operated by a Md2/PCI video board (Dome Imaging System, inc., Waltham, MA). This board uses a 10 bit digital-to-analog converter and allows one to adjust the relationship between digital value (e.g., gray level) and monitor output (e.g., luminance). The experiments were performed with the default-lookup table, which is very close to the perceptually linearized monitor. Images were generated as well as presented using the IDL software (Interactive Data Language, version 5.1; Research Systems, Boulder, CO).

D. Observers and procedure

Two human observers (an author and a radiologist) participated in the study. They were well trained in the experiment and both had normal/corrected vision. The mean viewing distance is estimated to be 50 cm but observers were free to adjust their distance if needed. Once trained for the detection task, the observers performed five sessions of 100 trials for each condition of interest.

For a given experiment, the contrast and the number of possible locations, or the inter-location distance were variable. Each particular condition was constant during a given session of 100 trials but the sequence of the conditions was randomly chosen. In order to minimize the effect of variable conditions from session to session, 5 “warm-up” trials were performed before each session. On each trial, observers selected with a click of the computer mouse the location they thought contained the signal. In an MAFC experiment, the location the closest to the cursor was recorded as being the correct answer. For the free search experiment, the position of the cursor was registered for each answer in image coordinates.

IV. PSYCHOPHYSICAL EXPERIMENTS

A. Experiment I: Multiple-alternative forced-choice

1. Microcalcification localization

a. Experimental design. For the microcalcifications, each MAFC condition was run for five different contrast values

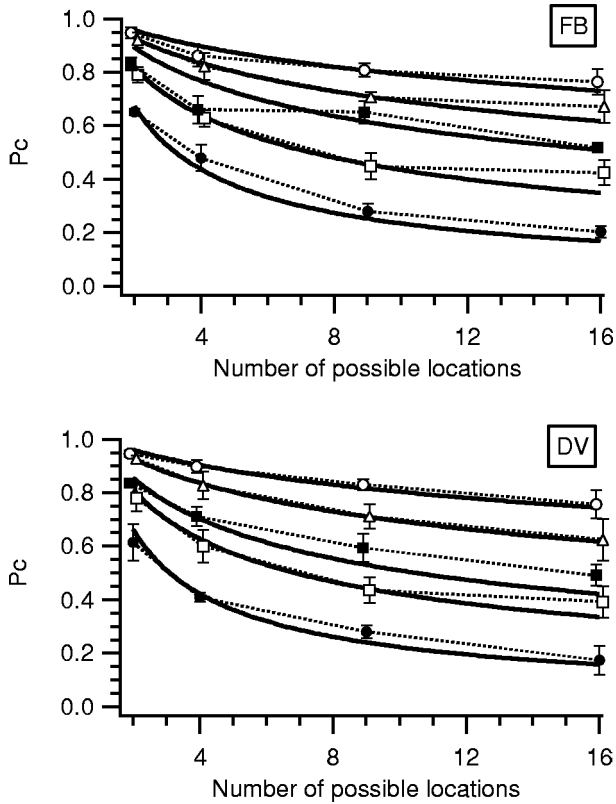


FIG. 11. Percent of correct trials (Pc) of each observer plotted against the number of possible locations at the microcalcification scale. Each dotted line corresponds to a given contrast (the lower the contrast, the lower Pc). The continuous lines correspond to the fitted performance computed with a Gaussian distributed response.

(0.2, 0.4, 0.6, 0.8, 1.0). The observer performance was first computed for each condition by calculating the percent of correct trials (Pc). The Pc value was then transformed to an index of detectability (d_{MAFC}) assuming Gaussian independent observer response pdf using Eq. (5).

b. Results. For both observers Pc is computed as a function of number of locations (M) for the five different contrasts in Fig. 11. Overall, Fig. 11 shows that performance degradation as a function of M seems to be approximated by the SDT model with the Gaussian assumption. However, if the observer response pdfs were effectively Gaussian and independent, then the values of d_{MAFC} across the number of possible locations for a given contrast should be constant. This is presented in Fig. 12 where d_{MAFC} is shown as a function of number of locations. Also plotted in Fig. 12 is the simultaneous fit of the Gaussian model to all the data for each observer. A slight tendency of increase of d_{MAFC} with the number of locations can be observed. This can be quantitatively tested through a three-way analysis of variance (ANOVA)³⁷ in which the three independent factors are the contrast, the number of possible locations, and the observers. Table I shows the result of the ANOVA with the computation of the F statistics and the critical value of F ($F_{critical}$) for $p < 0.05$. It shows that both observers are statistically equivalent and that the performance expressed in terms of d_{MAFC} is not constant across the number of possible locations. There-

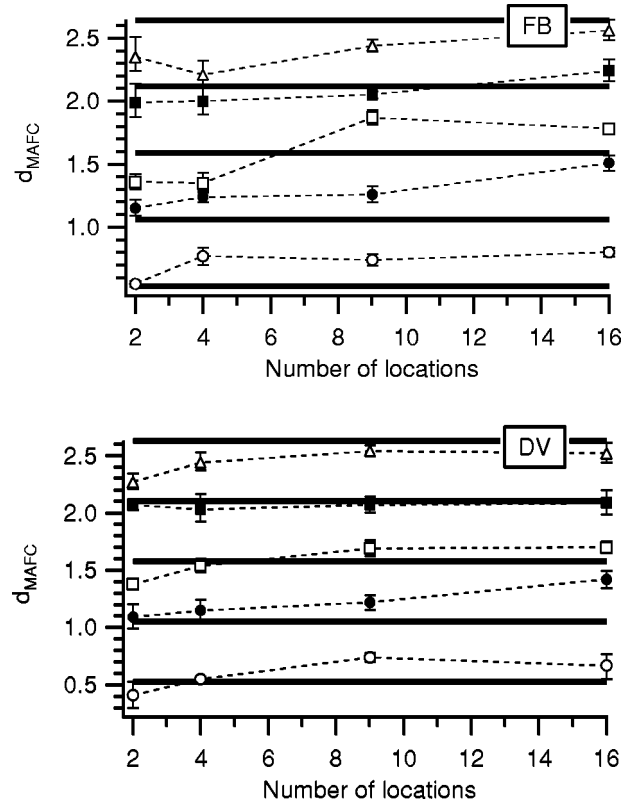


FIG. 12. Detectability of each observer plotted against the number of possible locations at the microcalcification scale. Each dotted line corresponds to a given contrast (the lower the contrast, the lower the detectability). The continuous lines correspond to the fitted performance computed with a Gaussian distributed response.

fore, the increase in performance with the number of possible locations is statistically significant.

There are two possible explanations for this finding. The first is that the response pdfs might depart from the Gaussian (Fig. 5). A second explanation might be an increase in the correlation among the human internal responses to the different locations with increasing number of locations (Fig. 7).

We first investigated whether pdfs that depart from the Gaussian might better account for human performance. The parameter β [Eq. (6)] was left to vary freely. The best fit resulted in $\beta = 59$ (observer FB) and $\beta = 15$ (observer DV). The resulting values of the best-fit d_{MAFC} are presented in Fig. 13 showing that this model captures the significant tendency of the human d_{MAFC} to increase with the number of locations. Although beta values of 15 and 59 might seem very different, this is not the case when looking at the actual shape of the pdfs. As shown in Fig. 3, both of these pdfs are very similar and much flatter than the Gaussian pdf. An estimate of their 95% confidence interval is between $\beta = 10$ and $\beta = 100$.

A second possible explanation to the increase in human observers' detectability index (d_{MAFC}) with increasing number of locations is the possibility of an increase in internal response correlations with the decreased interlocation distances associated with the higher MAFC conditions. This

TABLE I. Result of the analysis of variance (three-way ANOVA) performed on the observer results. Df is the number of degrees of freedom, $F_{critical}$ is the value of the F statistics that defines the rejected region ($p < 0.05$), and F is the computed statistics. An asterisk indicates that the constant mean hypothesis has to be rejected.

Factor	Microcalcification scale			Tumor scale		
	Df	$F_{critical}$	F	Df	$F_{critical}$	F
Contrast	4	2.37	670*	4	2.50	266*
Location	3	2.60	24.0*	1	3.97	1.86
Observer	1	3.84	1.64	1	3.97	0.12
Contrast and location	12	1.75	2.07*	4	2.50	0.49
Contrast and observer	4	2.37	1.64	4	2.50	0.38
Location and observer	3	2.60	1.10	1	3.97	0.55

possibility will be separately tested later with a psychophysical experiment.

2. Tumor localization

a. Experimental design. For the tumor localization, only two MAFC experiments were investigated: 2 and 9 possible locations for five different contrast values. Observers' percent correct performance was converted to an index of detectability using the Gaussian response pdf assumption.

b. Results. Figure 14 shows the index of detectability for each of the five contrast values for each observer separately. Analysis of variance was used to test for significant differences (Table I). The results show no statistically significant differences across the observers. In addition, and contrary to

the microcalcification scale experiment, the number of possible locations did not significantly affect the index of detectability. Therefore, the Gaussian response pdfs or the correlation free hypothesis cannot be rejected for the tumor scale detection.

As for the microcalcification experiment, a global fit of beta was also performed with the tumor experiment for each observer. This resulted in $\beta=1.1$ and $\beta=0.9$ for observers FB and DV, respectively. The uncertainties associated with values are relatively large and not straightforward to estimate. Taking into account the standard deviations of the measured P_c , possible values of β could vary between 0.5

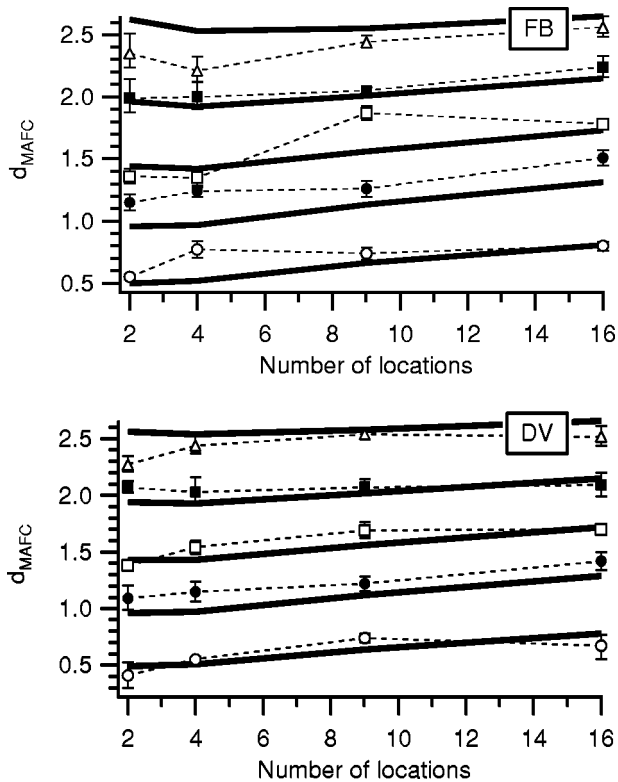


FIG. 13. Same as Fig. 12 but the continuous lines correspond to the fitted performance computed with a response function characterized by $\beta=59$ (observer FB) and $\beta=15$ (observer DV).

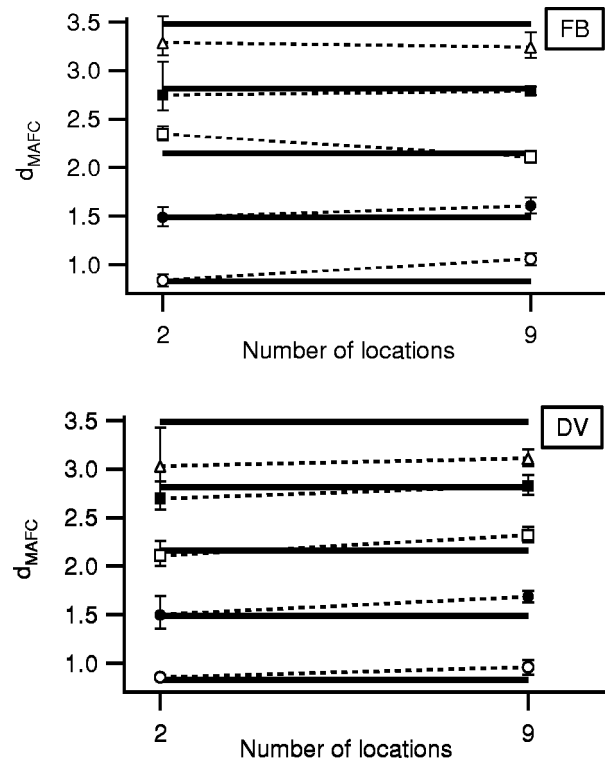


FIG. 14. Detectability of each observer plotted against the number of possible locations at the tumor scale. Each dotted line corresponds to a given contrast (the lower the contrast, the lower the detectability). The continuous lines correspond to the performance computed with a Gaussian distributed response.

and 10. Therefore, the Gaussian assumption cannot be rejected in the tumor experiment.

B. Experiment II: Internal response correlation in microcalcification detection task?

Figure 7 shows that the increase of d_{MAFC} with M could also be explained with an increase in the correlation between responses with M due to the smaller interlocation distances for high MAFC conditions than for low M conditions. The large amount of low frequencies contained in mammographic backgrounds,³⁸ might cause the correlation in the observer responses. The correlation between pixel variations increases as a function of their proximity in low-pass and power law noise.

It therefore could be possible that the increase in the human observers' d_{MAFC} is a result of the higher correlations between the human internal responses for the higher M conditions. In order to test whether internal responses are correlated, we performed a separate experiment. In this experiment we measured performance in a 2AFC microcalcification detection as a function of the distance between the two locations. The rationale is that if the correlation between the responses increased with decreasing interlocation distance, then human performance would increase. Evidence for an increase in performance in the 2AFC experiment with decreasing interlocation distance would support the idea that the increase in d_{MAFC} in the MAFC microcalcification experiment is explained with an increase in response correlation with M .

1. Experimental design

The 2AFC experiment was performed with 128×256 pixel images in which the possible target interlocation distance was variable. The possible target locations were also clearly indicated by cues (see Fig. 10) in order to minimize location intrinsic uncertainty. Four experimental conditions were investigated with interlocation distances of 12, 38, 68, and 208 pixels. A total of 251 different images were used for this experiment.

As a control experiment, the same task was first performed with the two extreme distances (12 and 208 pixels) and with uncorrelated white noise. The performance was statistically identical ($p < 0.05$) regardless of the interlocation distance. It is therefore reasonable to assume that a smaller distance does not, as such, increase the observer efficiency.

2. Results

The values of P_c obtained with different interlocation distances were transformed into $d_{2\text{AFC}}$ by assuming Gaussian pdfs. Figure 15 shows the values of $d_{2\text{AFC}}$ plotted against the interlocation distances for both observers.

The results show a small increase in performance at small distances. However, analysis of variance showed the performance does not significantly vary with interlocation distances for observer FB. Observer DV shows a statistically significant higher performance at the smallest interlocation distance.

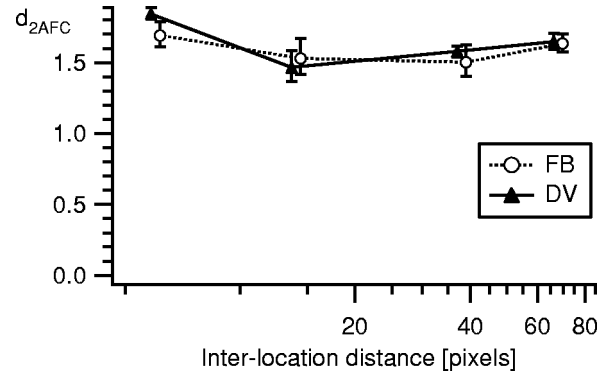


Fig. 15. Computed value of $d_{2\text{AFC}}$ vs the distance between the different possible locations.

The amount of correlation can be estimated from Eq. (8) and assuming that the responses at the largest distance are statistically independent. In this expression, r is the correlation between internal responses at the two locations, d'_r is the index of detectability in the presence of a correlation r and d' is the index of detectability when the responses are independent. The correlation r can be computed if d' and d'_r are known. If the response correlation is induced by the background low frequency content, then the value of $d_{2\text{AFC}}$ obtained at the largest distance can be considered to be independent and a good estimate of d' . All other values of $d_{2\text{AFC}}$ obtained at smaller distances are the corresponding d'_r values. Figure 16 shows that the calculated correlations for the different interlocation distances are relatively close to zero. A T-test shows that at a 95% confidence level, no points can be considered significantly different than zero. These results therefore suggest that the correlation between human internal responses is close to zero and do not increase with decreasing interlocation distance. Furthermore, the results suggest that the increase in the d_{MAFC} with increasing number of locations cannot be attributed to an increase in response correlations.

C. Experiment III: Free search

In a MAFC experiment, the number of locations as well as the exact possible location of the signal is defined and

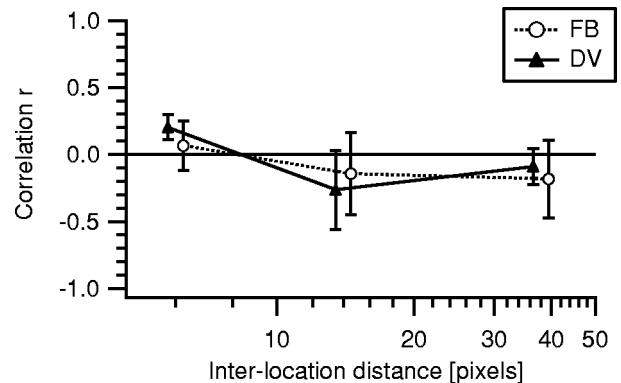


Fig. 16. Estimation of the correlation vs the interlocation distance according to Eq. (8).

known by the observer. In common practice, when the radiologist is searching the image for a pathological signal, its possible locations are generally unknown. Because the number of possible locations is not unlimited, such an experiment has to be linked to the value of d' obtained in a classical MAFC experiment.

1. Methods

In the free search experiment, the observer knows that the microcalcification is anywhere within the image except for an external border region of the image (10 pixels wide). The observers' task is to localize the microcalcification in the image by clicking with the mouse on the center of the lesion. The process of precisely spatially localizing the center of the lesion is limited by other processes distinct from visual detection such as error in perceptual judgment of the center of the lesion as well as motor error in the manual placing of the mouse. In order to estimate these latter sources of error in the observers' spatial localization of the signal (independent of sources of error due to visual detection), a training experiment was performed on a noiseless and uniform background with a signal that was always visible (equal to a high contrast microcalcification). The spatial localization error was measured as the distance (in pixels) between the signal location and the position of the cursor. We obtained a mean distance of 0.55 with a standard deviation of 0.54 and a maximum value of 2.0 pixels. Given that the signal radius is about 3 pixels, we adopted as a criterion to consider a signal localization in the free search experiment to be correct if the mouse click was located at a distance not greater than 5 pixels from the center of the lesion. Given this criterion of a 5-pixel radius circle, an image area of P pixels effectively contains approximately $M^* = P/(\pi 5^2)$ different response regions. In the present study, this ratio is equal to 559. Once the effective number of locations M^* is known, all the microcalcification data can be globally fit with beta as a free parameter.

Burgess *et al.*,²⁴ in a similar experiment, defined M^* as being more simply the ratio of the search and signal areas; but their signals were much bigger. The same definition as Burgess *et al.* in the present study would have increased M^* to 1885.

2. Results

Figure 17 shows the histograms of distances between the location clicked with the mouse and signal location for each observer and for all investigated signal contrasts on real mammograms. Superimposed on these graphs is the distribution of a hypothetical observer randomly choosing a location (with uniform distribution normalized to the number of answers above 5 pixels). For the entire experiment, the number of localization responses given by the random observer with a target to localization responses distance below 5 pixels is below 0.3%. Therefore, the probability that an observer correctly localizes the target by chance in the free search experiment is negligible. The comparison of human and random observers above the 5 pixel distance shows no significant

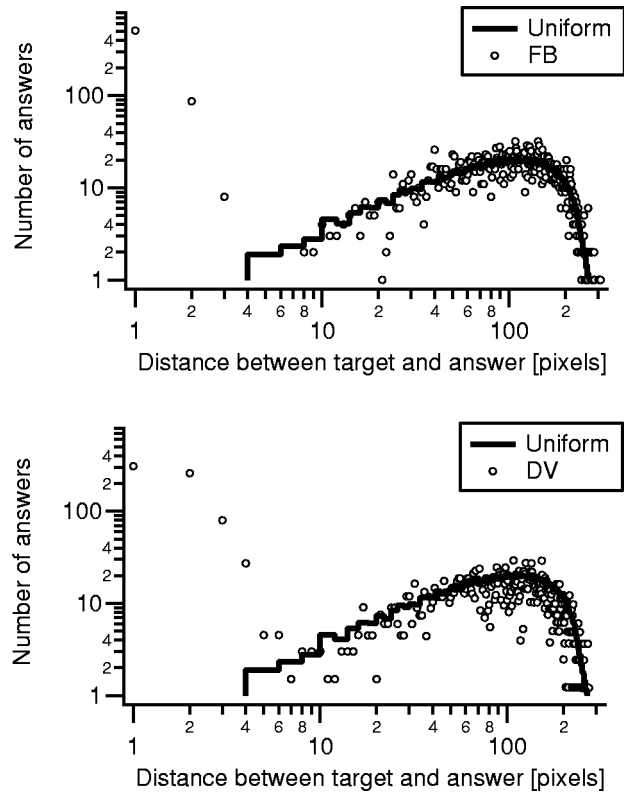


Fig. 17. Distribution of the entire target to answer distances and the uniform prediction in the free search experiment.

difference. This tends to show that when the target was missed, the selected location was not correlated with the actual signal location. Both of these findings suggest that the adopted criterion to consider localizations within a 5 pixel distance from the target center to be correct does not introduce any systematic and significant biases in our localization accuracy metric.

Figure 18 shows the experimental results of the fit of all data for each observer and each number of locations versus signal contrast. The resulting beta values are 2.8 and 2.7 for observers FB and DV, respectively.

V. DISCUSSION

A. Predicting the effect of number of locations and contrast on signal detectability

The main point of the present study is about the consistency of MAFC experiments realized with different number of locations. The usual method consists in assuming independent responses from one location to another and Gaussian pdf. Although the standard model approximates human performance, a strict analysis of the data shows that this model can be rejected for the detection of microcalcifications while it cannot be rejected for the detection of masses.

B. Correlation in the human internal responses within image MAFC tasks

The departure from the independent Gaussian pdf SDT model could have been explained by correlations in the hu-

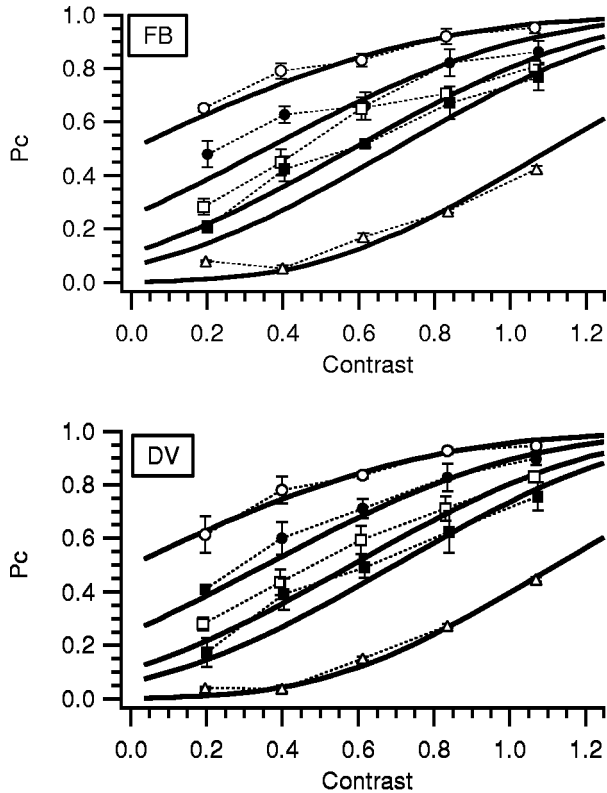


FIG. 18. Percent of correct trials vs signal contrast for each observer. The continuous lines show the global fits of all Pc with $M^*=559$ (obtained values of $\beta=2.8$ for FB; $\beta=2.7$ for DV). The dotted lines join the measured values of Pc for each number of locations in the MAFC experiments (from top to bottom: $M=2, 4, 9, 16$ and M^*).

man internal responses induced by the low frequencies in the mammographic backgrounds. Our psychophysical study specifically tested whether the correlation across internal responses increased with decreasing interlocation distance. The results suggest that the correlation does not significantly change as a function of interlocation distance. It has to be noted that this is not the case for any kind of low-pass background. Other results³⁹ showed that coronary angiographic backgrounds induce a model response correlation of about 0.4 at very small distances only. This can be explained by the fact that the amount of high frequencies is higher in mammographic than angiographic backgrounds. In conjunction with observer independent internal noise, this high frequency content can reduce any possible correlation.

C. Probability density function for observer internal responses

As no correlation could be emphasized in mammographic backgrounds, this leaves only one obvious explanation: a departure from the usual Gaussian pdf. A global fit of all the data obtained at the microcalcification scale at five signal contrasts, four numbers of locations and the free search experiment, allowed us to quantitatively estimate the pdf as being slightly more compact than the Gaussian pdf with β just below 3 [see Eq. (6)]. This is coherent with previous results obtained with x-ray coronary angiograms,¹⁹ for which

a close examination of the human performance shows a similar trend as the present results. In those experiments, d_{MAFC} computed with the independent Gaussian assumption were significantly lower for $M=2$ at the two lowest contrast values. As can be seen in Fig. 5, this trend could also be explained by a more compact observer response pdf ($\beta > 2$).

D. Free search

At first view, the free search experiment involves more complex perceptual processes than the multiple-alternative forced-choice experiments: free search is closer to the real diagnostic task, but harder to be performed and analyzed than the simple MAFC experiment. However, the results presented here show that the free search experiment can be seen as a particular MAFC experiment with an effective number of locations M^* equal to the ratio of the total number of pixels in the image divided by the signal area increased by the inherent spatial imprecision to localize the target.

A similar experiment to the one presented here was conducted by Burgess *et al.*,²⁴ in which the signal was a projected sphere (of variable size) superimposed on two types of backgrounds: real mammograms and low-pass filtered noise with the same power spectrum. The experiment was repeated twice: once in a 2AFC and once in the free search condition. In the 2AFC experiment the performance was better with real anatomical backgrounds than with filtered noise, whereas in the free search experiment, the observers performed equivalently in the real and in the filtered noise backgrounds. This characteristic is explained by the absence of location uncertainty in the 2AFC experiment: The image recognizable structures of the real mammograms seem to help the observers in the 2AFC experiment whereas in the free search experiment, recognizable structures are of no help. The important point about the results of Burgess *et al.* is their ability to predict the free search performance with the 2AFC results obtained with the filtered noise backgrounds. For that, they simply define the effective number of locations M^* as being the ratio of the search and signal areas, compute $d_{2\text{AFC}}$ with the filtered white noise backgrounds and the Gaussian assumption, and then compute Pc using Eq. (5) for M^* locations. If they had performed with the same procedure and the 2AFC results obtained with the real mammographic backgrounds (instead of the filtered white noise), they would have largely overestimated the observer response. In this case, they would have needed to increase further more M^* to fit the free search results from the 2AFC experience obtained with real mammographic backgrounds.

In the present study, the effective number of locations M^* is a bit lower than Burgess *et al.* This discrepancy may be due to the differences of the displayed backgrounds. In the present free search experiment, the detection of microcalcifications, all images are magnified by a factor 8. In the experiment of Burgess *et al.*, the study was performed at the mass or tumor scale and all images are magnified by a factor of about 2. As shown by Bochud *et al.*,²³ anatomical variations are especially perceived as a source of noise at the tumor or mass scale rather than at the microcalcification

scale. This means that the anatomical fluctuations are much more disturbing at the tumor or mass scale and that they are more prone to appear randomly as the observer is looking at the signal. Therefore, once we try to predict free search detection results from MAFC results, the scale of the background is important. At the mass or tumor scale, Burgess *et al.*²⁴ could derive their free search data with an effective number of locations M^* greater than the straightforward ratio of the search and signal areas, whereas in the present study M^* is smaller.

VI. CONCLUSIONS

The performance of an observer searching for a mass embedded in a mammographic background as a function of the number of alternatives can be described by signal detection theory with the usual Gaussian pdf. If the same experiment is conducted with a target simulating a microcalcification, we find the response pdf departs from a Gaussian to a more compact distribution. This effect is reported for only two observers, and hence requires further investigation to be extrapolated to the population of observers. However, the fact that non-Gaussians are observed suggests that the functional form of the observer response may play a role in understanding and modeling visual perception in medical imaging. Furthermore, this departure from the Gaussian distribution is consistent with a strict analysis of previous experiments utilizing angiographic x-ray images.¹⁹

We have investigated a possible alternative explanation for the differences in observed human observer performance across the number of possible signal locations. We hypothesized that the presence of internal correlations induced by variability at low spatial frequencies in the image backgrounds was responsible for the observed deviations from the predictions made under the assumption of independent Gaussian responses. However, the results presented in this study do not show any such correlations. This could be explained by the relatively large amount of high frequency noise present in the mammographic backgrounds as well as the independent observer internal noise.

Of all the tasks considered in this work, the free search experiment, in which the signal could appear anywhere in the image, is clearly the closest to clinical practice. However, our results show that the free search experiment results could be derived from more simple MAFC detection performance. In this case, the effective number of alternatives can be defined as the ratio of the total image size and the signal area increased by the inherent spatial imprecision to localize the target.

ACKNOWLEDGMENTS

Parts of these results were first presented at the Far West Image Perception Conference (28–30 May 1999, Nakoda Lodge–AB–Canada) and the Annual Meeting for Research in Visual Ophthalmology (1999, Fort Lauderdale, FL). We are very grateful to Dr. Carolyn Kimme-Smith from Univer-

sity of California Los Angeles who made the digital mammograms available for this study. We would also like to acknowledge Darko Vodopich for long sessions spent observing images. This work was funded by the Swiss National Fund for Scientific Research Fellowship to F.O.B. and National Institute of Health (NIH) RO1-HLB 53455 to M.P.E.

^{a)}Electronic mail: francois.bochud@hosvvd.ch

¹A.E. Burgess, R.F. Wagner, R.J. Jennings, and H.B. Barlow, "Efficiency of human visual signal determination," *Nature (London)* **214**, 93–94 (1981).

²P.F. Judy and R.G. Swensson, "Display thresholding of images and observer detection performance," *J. Opt. Soc. Am. A* **4**, 954–965 (1987).

³M.S. Chesters and A.H. Hay, "Quantitative relation between detectability and noise power," *Phys. Med. Biol.* **28**, 1113–1125 (1983).

⁴A.E. Burgess and H. Ghandeharian, "Visual signal detection. I. Ability to use phase information," *J. Opt. Soc. Am. A* **1**, 900–905 (1984).

⁵A.E. Burgess and H. Ghandeharian, "Visual signal detection. II. Signal-location identification," *J. Opt. Soc. Am. A* **1**, 906–910 (1984).

⁶A.E. Burgess, "Effect of quantization noise on visual signal detection in noisy images," *J. Opt. Soc. Am. A* **2**, 1424–1428 (1985).

⁷K.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).

⁸W. Gentles, T. Nguyen, W. Ho, and C. Caldwell, "Effect of spatial frequency content of the background on visual detection of a known target," *Proc. SPIE* **1652**, 341–351 (1992).

⁹J.P. Rolland, H.H. Barrett, and G.W. Seeley, "Ideal versus human observer for long-tailed point spread functions: Does deconvolution help?," *Phys. Med. Biol.* **36**, 1091–1109 (1991).

¹⁰J.P. Rolland and H.H. Barrett, "Effect of random background inhomogeneity on observer detection performance," *J. Opt. Soc. Am. A* **9**, 649–658 (1992).

¹¹H.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).

¹²A.E. Burgess, "Visual signal detection with two-component noise: Low-pass spectrum effect," *J. Opt. Soc. Am. A* **16**, 694–704 (1999).

¹³A.E. Burgess, "Comparison of non-prewhitening and Hotelling observer models," *Proc. SPIE* **2436**, 2–9 (1995).

¹⁴G. Revesz, H.L. Kundel, and M.A. Graber, "The influence of structured noise on the detection of radiologic abnormalities," *Invest. Radiol.* **9**, 479–486 (1974).

¹⁵H.L. Kundel and G. Revesz, "Lesion conspicuity, structured noise, and film reader error," *AJR, Am. J. Roentgenol.* **126**, 1233–1238 (1976).

¹⁶P.F. Judy, R.G. Swensson, R.D. Nawfel, and K.H. Chan, "Contrast detail curves for liver CT," *Med. Phys.* **19**, 1167–1174 (1992).

¹⁷S.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).

¹⁸F.O. Bochud, F.R. Verdun, C. Hessler, and J.F. Valley, "Detectability of radiological images: The influence of anatomical noise," *Proc. SPIE* **2436**, 156–164 (1995).

¹⁹M.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).

²⁰E. Buhr and D. Hoeschen, "Bildrauschen und Diagnose von Rundherden in Thoraxaufnahme," *Z. Med. Phys.* **6**, 80–86 (1996).

²¹E. Samei, M.J. Flynn, and W.R. Eyler, "Simulation of subtle lung nodules in projection chest radiography," *Radiology* **202**, 117–124 (1997).

²²M.P. Eckstein and J.S. Whiting, "Why do anatomical backgrounds reduce detectability?," *Invest. Radiol.* **33**, 203–208 (1998).

²³F.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).

²⁴A.E. Burgess, F. Jacobson, and P.F. Judy, "Human observer detection experiments with mammograms and power-law noise," *Med. Phys.* **28**, 419–437 (2001).

²⁵W.P. Tanner, "Physiological implications of psychophysical data," *Ann. N.Y. Acad. Sci.* **89**, 752–761 (1961).

- ²⁶D.J. Lasley and T.E. Cohn, "Detection of a luminance increment: Effect of temporal uncertainty," *J. Opt. Soc. Am. A* **71**, 845–850 (1981).
- ²⁷T.E. Cohn and D.J. Lasley, "Detectability of a luminance increment: Effect of spatial uncertainty on foveal luminance increment detectability," *J. Opt. Soc. Am. A* **2**, 820–825 (1985).
- ²⁸D.M. Green and J.A. Swets, *Signal Detection and Psychophysics* (Krieger, New York, 1966).
- ²⁹K. Fukunaga, *Introduction to Statistical Pattern Recognition* (Academic, Boston, 1990).
- ³⁰P. Xue and D.L. Wilson, "Effects of motion blurring in x-ray fluoroscopy," *Med. Phys.* **25**, 587–599 (1998).
- ³¹R. Aufrichtig, "Comparison of low contrast detectability between a digital amorphous silicon and a screen-film based imaging system for thoracic radiography," *Med. Phys.* **26**, 1349–1358 (1999).
- ³²A. Van der Schaaf and J.H. Van Hateren, "Modelling the power spectra of natural images: statistics and information," *Vision Res.* **36**, 2759–2770 (1996).
- ³³N. Brady, "Spatial scale interactions and image statistics," *Perception* **26**, 1089–1100 (1997).
- ³⁴M.P. Eckstein, C.K. Abbey, and F.O. Bochud, "Visual signal detection in structured backgrounds. IV. Figure of merit for model performance in multiple-alternative forced-choice detection task with correlation response," *J. Opt. Soc. Am. A* **17**, 206–217 (2000).
- ³⁵P.B. Elliott, "Appendix I: Tables of d' ," in *Signal Detection and Recognition by Human Observers: Contemporary Readings*, edited by J.A. Swets (Wiley, New York), pp. 651–685.
- ³⁶C.K. Abbey and M.P. Eckstein, "Derivation of a detectability index for correlated responses in multiple-alternative forced-choice experiments," *J. Opt. Soc. Am. A* **17**, 2101–2104 (2000).
- ³⁷S. Weinberg and K. Goldberg, *Statistics for the Behavioral Sciences* (Cambridge University Press, Cambridge, 1990).
- ³⁸F.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).
- ³⁹F.O. Bochud, C.K. Abbey, and M.P. Eckstein, "Correlated human responses for visual detection in natural images," in *Proceedings of the Annual Meeting for Research in Visual Ophthalmology*, Fort Lauderdale, FL, USA, Proc. IOVS **40**, 350 (1999).