Title

Classification images for localization performance in ramp-spectrum noise

Journal

Medical Physics, 45(5)

ISSN

0094-2405

Authors

Abbey, Craig K
Samuelson, Frank W
Zeng, Rongping
et al.

Publication Date

2018-05-01

DOI

10.1002/mp.12857

Peer reviewed

Classification images for localization performance in ramp-spectrum noise

Craig K. Abbey[a)]
Department of Psychological and Brain Sciences, University of California Santa Barbara, Santa Barbara, CA 93106, USA

Frank W. Samuelson and Rongping Zeng
Division of Imaging Diagnostics and Software Reliability, United States Food and Drug Administration, White Oaks, MD 20993, USA

John M. Boone
Departments of Radiology and Biomedical Engineering, University of California Davis, Sacramento, CA 95817, USA

Miguel P. Eckstein
Department of Psychological and Brain Sciences, University of California Santa Barbara, Santa Barbara, CA 93106, USA

Kyle Myers
Division of Imaging Diagnostics and Software Reliability, United States Food and Drug Administration, White Oaks, MD 20993, USA

(Received 5 January 2018; revised 22 February 2018; accepted for publication 25 February 2018; published 11 April 2018)

Purpose: This study investigates forced localization of targets in simulated images with statistical properties similar to trans-axial sections of x-ray computed tomography (CT) volumes. A total of 24 imaging conditions are considered, comprising two target sizes, three levels of background variability, and four levels of frequency apodization. The goal of the study is to better understand how human observers perform forced-localization tasks in images with CT-like statistical properties.
Methods: The transfer properties of CT systems are modeled by a shift-invariant transfer function in addition to apodization filters that modulate high spatial frequencies. The images contain noise that is the combination of a ramp-spectrum component, simulating the effect of acquisition noise in CT, and a power-law component, simulating the effect of normal anatomy in the background, which are modulated by the apodization filter as well. Observer performance is characterized using two psychophysical techniques: efficiency analysis and classification image analysis. Observer efficiency quantifies how much diagnostic information is being used by observers to perform a task, and classification images show how that information is being accessed in the form of a perceptual filter.
Results: Psychophysical studies from five subjects form the basis of the results. Observer efficiency ranges from 29% to 77% across the different conditions. The lowest efficiency is observed in conditions with uniform backgrounds, where significant effects of apodization are found. The classification images, estimated using smoothing windows, suggest that human observers use center-surround filters to perform the task, and these are subjected to a number of subsequent analyses. When implemented as a scanning linear filter, the classification images appear to capture most of the observer variability in efficiency ($r^2 = 0.86$). The frequency spectra of the classification images show that frequency weights generally appear bandpass in nature, with peak frequency and bandwidth that vary with statistical properties of the images.
Conclusions: In these experiments, the classification images appear to capture important features of human-observer performance. Frequency apodization only appears to have a significant effect on performance in the absence of anatomical variability, where the observers appear to underweight low spatial frequencies that have relatively little noise. Frequency weights derived from the classification images generally have a bandpass structure, with adaptation to different conditions seen in the peak frequency and bandwidth. The classification image spectra show relatively modest changes in response to different levels of apodization, with some evidence that observers are attempting to rebalance the apodized spectrum presented to them. [https://doi.org/10.1002/mp.12857]

Key words: classification images, noise, noise power spectrum, observer performance

1. INTRODUCTION

X-ray computed tomography (CT) has become a mainstay of diagnostic imaging in many areas of medicine because of its ability to render internal structures of the body with high accuracy. A cross-sectional image of a tomographic reconstruction synthesizes information across multiple angular views to resolve the superposition of structures in "raw" projection data. Another result of the reconstruction process is that acquisition noise — quantum noise in x-ray production

and electronic detector noise — adopts a characteristic texture in which noise power increases approximately linearly with frequency, if left unchecked.[1–3] The result is so-called ramp-spectrum noise, and it presents a challenge to the presentation of high-frequency information in CT images. Traditionally, in standard filtered-backprojection reconstructions, noise is controlled by the use of apodization filters that modulate high spatial frequencies.[2,4] These filters suppress high-frequency noise, but at the cost of limiting the resolution of the resulting images to some extent as well.

The presumption of apodization filters is that they help the end-user — typically a radiologist — most efficiently access the diagnostic information in the image. However, relatively little is known about how image readers actually extract this information to perform even simple tasks in ramp-spectrum noise, with or without apodization. This is the motivation for this work, which applies psychophysical methods derived from vision science to directly assess mechanisms of task performance. We see this goal as related to, but somewhat distinct from, a substantial effort in the field to develop model observers,[5–10] which have also been applied to tomographic images with ramp-spectrum noise.[11–16] At a fundamental level, model observers are predictive in nature rather than explanatory, with a goal of comparing and optimizing imaging systems without time-consuming and costly observer performance studies. While we hope that our results may be used to inform the development of anthropomorphic model observers, the focus of this study is to investigate human observers.

Our studies analyze observer performance in terms of classification images,[17–20] a technique that utilizes the noise fields of the image stimuli along with decision outcomes to directly estimate perceptual filters underlying task performance. We also use the concept of task efficiency with respect to the ideal observer[20–27] as a way to quantify how well observers are able to utilize the relevant information in the images to perform the task. This combination allows us to assess how much information is lost due to errors in the tuning of perceptual filters, and how much is lost from other effects like internal noise or inefficient search. The task we investigate is forced localization, which requires the observer to localize (by a pointer-click) the most likely position of the target within an image. Thus, the task forces the subject to search the image and discriminate the target from a variable background that includes effects of ramp-spectrum noise and background variability with a power-law spectrum that simulates the effect of normal anatomy in the image.[28,29] A total of 24 different imaging conditions are considered that comprise two target sizes, three levels of background variability, and four levels of apodization.

The classification image technique and the ideal observer analysis used here have recently been developed for forced-localization tasks.[20] In order to meet the Gaussian assumptions needed for both of these, we use simulation images drawn from stationary Gaussian random processes with statistical properties that simulate ramp-spectrum of CT acquisition noise, the power-law spectrum of background variability in CT images, and effects of different apodization filters. From the forced-localization experiments for these images, we show how the various imaging conditions affect task efficiency and the perceptual tuning of observer localization filters.

2. IMAGE STIMULI

The images used in these forced-localization tasks consist of a target profile embedded in a stationary Gaussian random field as a background. Both the target and the background have properties that are qualitatively similar to components of axial CT images. Target profiles include effects of blurring due apodization and an intrinsic system transfer function, while the backgrounds include ramp-spectrum noise and power-law-spectrum background variability — so-called anatomical noise — that are both impacted by apodization filters.

We simulate images as 256 × 256 pixel arrays that are 12.8 cm on a side, with a 128 × 128 region in the center of the image that serves as potential locations for the target. The putative pixel size, $\Delta = 0.5$ mm, is roughly consistent with clinical CT scanners that are often used to discretize a 24-cm square axial field of view into a 512 × 512 array. The use of a Gaussian process to generate the image backgrounds allows us compute the ideal observer for forced-localization tasks and, therefore, to derive efficiency of the human observer. The Gaussian process is also an assumption of prior work using the classification image technique in forced-localization tasks.[20]

2.A. Target profiles

The target profiles are derived from one of two disks of higher intensity, meant to simulate a "lesion" of increased x-ray attenuation in CT. We investigate a smaller disk with a diameter of 1 mm, and a larger disk with a diameter of 4 mm. The 1 mm disk is relatively small for lesions detected by CT, but this size was chosen to put more emphasis on higher spatial frequencies in the task. The disk object is created at 4 × 4 oversampling, and then down-sampled to the image size by averaging. This allows for partial volume effects and reduces aliasing.

The disk may be thought of as the underlying object being imaged. In the process of becoming a target profile in the image domain, the disk is smoothed by filters representing the intrinsic modulation-transfer function of the imaging system and by the apodization kernel implemented in the CT reconstruction. We model the system MTF by a cosine roll-off function that goes from 1 at DC to 0 at the Nyquist frequency ($f_{\text{Nyq}} = 1/2\Delta = 1.0$ cyc/mm),

$$\text{MTF}(f) = \begin{cases} \frac{1}{2}\left(1 + \cos\left(\pi \frac{f}{f_{\text{Nyq}}}\right)\right) & f \le f_{\text{Nyq}} \\ 0 & f_{\text{Nyq}} < f \end{cases}, \tag{1}$$

where f represents the 2D radial frequency, $f = \sqrt{f_x^2 + f_y^2}$, for f_x and f_y representing spatial frequencies in the horizontal and vertical imaging directions, respectively. In a typical axial display, the horizontal direction is lateral in the body (right to left), and the vertical direction is anterior–posterior.

Apodization is implemented using a so-called Shepp–Logan filter[4] with different frequency cutoffs. For a given frequency cutoff, f_C, the functional form of the filter is

$$A(f) = \begin{cases} \mathrm{Sinc}\left(\pi \frac{f}{f_C}\right) & f \le f_C \\ 0 & f_{Nyq} < f \end{cases}. \quad (2)$$

The four apodization levels tested consist of no apodization (Apodization Level 1), $f_C = 2f_{Nyq}$ (Apodization Level 2), $f_C = f_{Nyq}$ (Apodization Level 3), $f_C = 0.75f_{Nyq}$ (Apodization Level 4). These apodization parameters range from nothing (Level 1) to a fairly aggressive filtering (Level 4) that cuts off the spectrum well below Nyquist. Figure 1(a) plots the apodization various apodization functions used (no apodization is plotted as a constant). The total filtering of the disk object is the product of the system MTF and the apodization function. Plots of total signal filters are shown in Fig. 1(b). Figure 2 illustrates a panel of target profiles, showing both target sizes and the four levels of apodization.

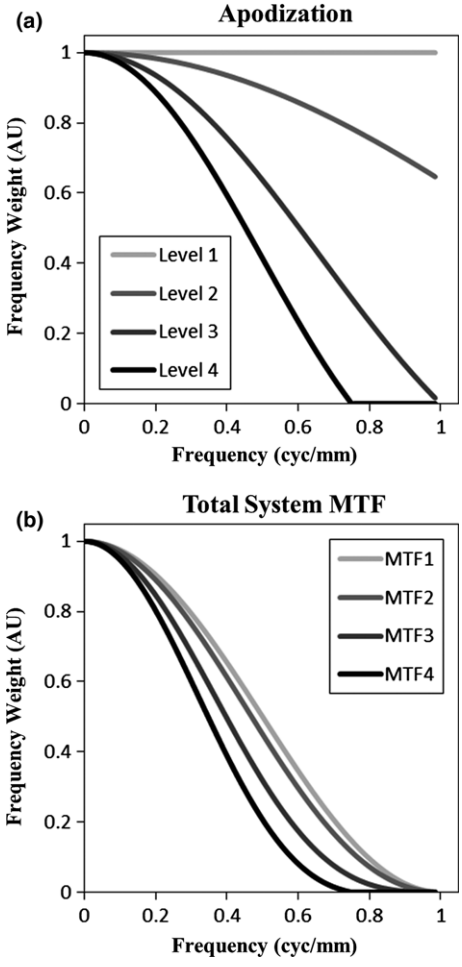

FIG. 1. Signal transfer components. The four levels of frequency apodization are shown (a) and the total system MTF (b), composed of the product of the apodization function and the intrinsic system MTF. Note that the plot of MTF1 represents the intrinsic system MTF since apodization is constant at this level.

2.B. Noise power spectra

All noise fields used in this work are samples from stationary Gaussian random fields that are generated by filtering white noise. The statistical properties of the noise fields are, therefore, characterized by their noise power spectra (NPS). Each NPS, in turn, is determined by three components: An acquisition noise component, a background variability component, and the apodization level applied to the images.

Acquisition noise in tomographic reconstructions arising from Poisson nature of x-ray production and attenuation as well as electronic noise in the x-ray detector[30,31] are well known to approximate a "ramp" spectrum in which noise power is directly proportional to radial frequency.[1–3] We use a ramp function over the frequency range out to the Nyquist frequency, except at the lowest frequencies where we prevent noise power from going to zero using a quadratic tail. This tail is parameterized by the low-frequency transition point, $f_0 = 0.05f_{Nyq}$. The acquisition noise power spectra is given by

$$S_{Aq}(f) = \begin{cases} C_{Aq}(f_0/2 + f^2/2f_0) & 0 \le f \le f_0 \\ C_{Aq} f & f_0 \le f \le f_{Nyq} \\ 0 & f_{Nyq} < f \end{cases} \quad (3)$$

The acquisition noise normalization constant, C_{Aq}, is set so that unapodized acquisition noise will lead to a pixel

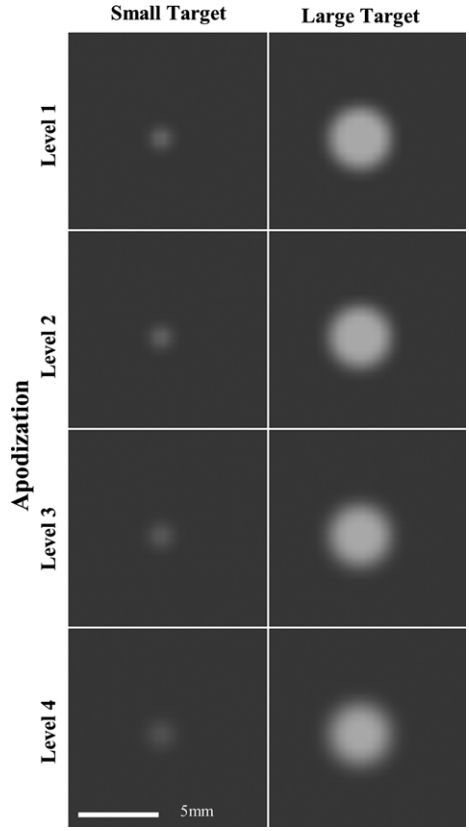

FIG. 2. Target panel. The panel shows the two target sizes (1 mm and 4 mm diameter) at the four levels of apodization used in the experiments at 100% object contrast. Note that for the smaller target, partial volume effects reduce the target contrast, particularly at higher levels of apodization.

standard deviation (SD) of 20 Hounsfield Units (HU), which is in the range of clinical scanning.

Background variability, describing variability due to the arrangement of normal anatomy in the field of view, is typically characterized by a power-law spectrum.[28,32–34] We use a power-law with an exponent of −2, which is consistent with some of the limited data available for CT imaging.[29,35] A small constant, $\varepsilon = 1/N\Delta$, is added to the radial frequency to prevent instability at low spatial frequencies. The resulting functional form of the spectrum of background variability is given by

$$S_{\mathrm{Bck}}(f) = \begin{cases} C_{\mathrm{Bck}}(f+\varepsilon)^{-2} & 0 \le f \le f_{\mathrm{Nyq}} \\ 0 & f_{\mathrm{Nyq}} < f \end{cases}. \quad (4)$$

The normalization constant, C_{Bck}, is set to one of three possible values, which determines the level of background variability. For uniform backgrounds (i.e., no background variability), $C_{\mathrm{Bck}} = 0$. We define a low level of background variability to occur when the image pixel SD due to background variability is 50% of the SD due to (unapodized) acquisition noise. We define a high level of background variability to occur when SD due to the background variability is 100% of the SD due to acquisition noise.

The total power spectrum of a set of images includes the effects of apodization as well, resulting in a functional form of

$$S_{\mathrm{Tot}}(f) = |A(f)|^2 \left(S_{\mathrm{Aq}}(f) + S_{\mathrm{Bck}}(f)\right). \quad (5)$$

Figure 3(a) shows plots of noise power spectra for conditions with a uniform background (i.e., $S_{\mathrm{Bck}}(f) = 0$). These plots show the influence of apodization on the acquisition noise power spectrum. Figure 3(b) shows power spectra from unapodized noise fields at the three levels of background variability. Note the log-scale y-axis of this plot, which is needed to effectively convey the large range of image power across these conditions. Figure 4 shows a panel of the 12 noise textures used in the experiments here.

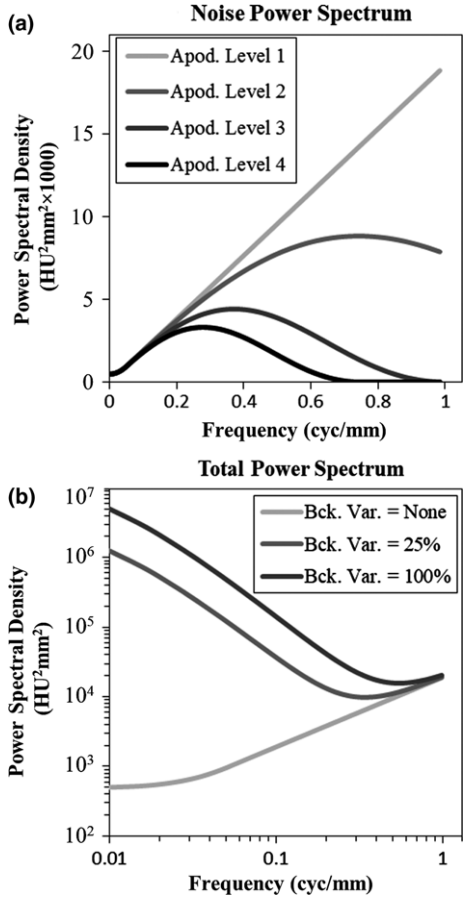

FIG. 3. Noise components. The effect of apodization on the Ramp-noise power spectrum is shown (a), along with unapodized (level 1) power spectra showing the effect of background variability in addition to noise (b).

2.C. Forced-localization stimuli

The targets and noise fields described above are used to generate the forced-localization stimuli. Each combination consists of a target size (2 possible), apodization level (4 possible), and level of background variability (3 possible), for a total of 24 experimental conditions. A 256 × 256 pixel stimulus is generated by adding a randomly shifted target to a sample noise field. A target is generated from a sampled disk object, as described above, that has been scaled to have a given amplitude in HU. The disk is then filtered by the system MTF described in Eq. (1), and the apodization function described in Eq. (2).

Let the array, $T[n, m]$ with $n = 0, \cdots, 255$ and $m = 0, \cdots, 255$, represent the resulting target for a given experimental condition. And let $N[n, m]$ represent the noise field. A stimulus in trial t is given by

$$G_t[n,m] = T\left[n - n_t^{\mathrm{True}}, m - m_t^{\mathrm{True}}\right] + N_t[n,m] \quad (6)$$

where n_t^{True} and m_t^{True} are the horizontal and vertical positions of the target in the trial. When $n - n_t^{\mathrm{True}}$ or $m - m_t^{\mathrm{True}}$ are negative in Eq. (6), we add 256 to this difference thereby wrapping around to the other side of the image. The true locations are sampled independently from a uniform distribution over the central 128 pixels of the horizontal and vertical dimensions (i.e., n_t^{True} and $m_t^{\mathrm{True}} \sim U\{64 : 191\}$).

The first step in the generation of a noise field is making the noise filter in the frequency domain. Let k and l represent the discrete frequency indices corresponding to n and m, and let $v_{x,k}$ and $v_{y,l}$ be the corresponding spatial-frequency variables defined by

$$v_{x,k} = \begin{cases} \frac{k}{\Delta 256} & 0 \le k \le 128 \\ \frac{k-256}{\Delta 256} & 129 < k \le 255 \end{cases}, \quad (7)$$

where Δ is the pixel size. There is a corresponding definition for $v_{y,l}$, with the resulting radial frequency for the indices k and l being $\rho_{k,l} = \sqrt{v_{x,k}^2 + v_{y,l}^2}$. The discrete filter, $F_{\mathrm{Stim}}[k, l]$, used to generate the image stimuli for an experimental condition is given by

$$F_{\mathrm{Stim}}[k,l] = \sqrt{S_{\mathrm{Tot}}(\rho_{k,l})}. \quad (8)$$

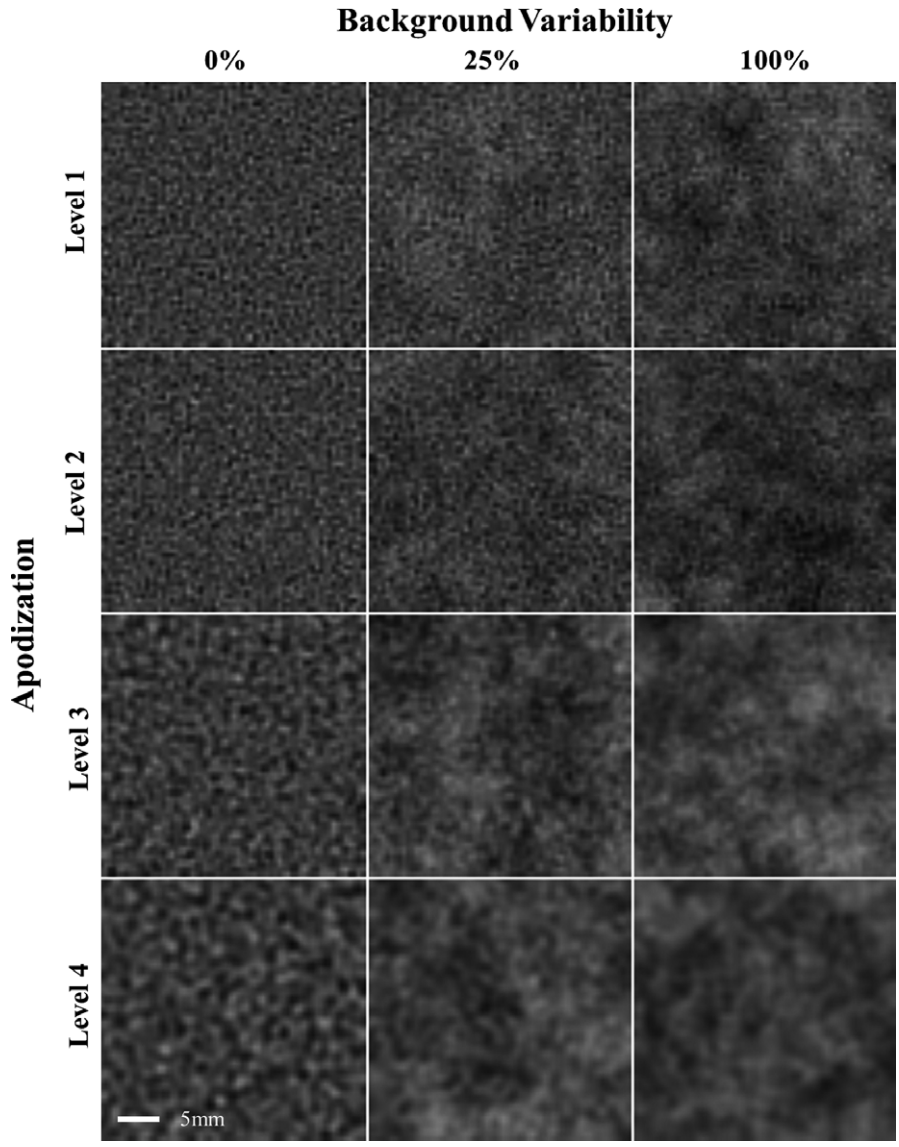

FIG. 4. Image background panel. The panel shows the various background image textures used in the experiments, which are affected by the amount of background variability as well as the level of apodization.

The noise field is generated by sampling an array of zero-mean unit-variance white noise. This initial sample is input to a 2D FFT, algorithm, multiplied by $F_{\text{Stim}}[k, l]$ from Eq. (8), and then inverse transformed back to the spatial domain to generate $N_t[n, m]$ in Eq. (6). Example noise textures from the procedure may be seen in Fig. 4.

3. PSYCHOPHYSICAL METHODS

All psychophysical studies were carried out under an IRB approved human subject protocol at UC Santa Barbara.

3.A. Display procedure

Noisy stimuli, generated according to Eq. (6), are used in a display procedure to acquire the localization response data. One of the important roles of the display procedure is to window and level the stimuli so that they are displayed within the dynamic range of the monitor used. The window and level setting were based in the noise field and set so that the mean background level of the stimuli is 100 gray levels (GL) on the display, and the pixel standard deviation is 20 GL. In order to make the localization response less susceptible to localization errors, the image is magnified by a factor of 2 on the monitor, giving it an effective isotropic display pixel size of 0.66 mm on the monitor used for the experiments.

The display program renders the stimuli in the center of a full-screen window, with the mean background set to 100 GL. For reference, an image of the noiseless target is also displayed above the stimuli. Finally, hash marks indicating the region of possible locations are added to the image to focus the reader on the appropriate part of the image. Figure 5 shows a part of the display window, cropped around the stimulus, with a high-amplitude large target (right side of image at 4 o'clock) for illustration.

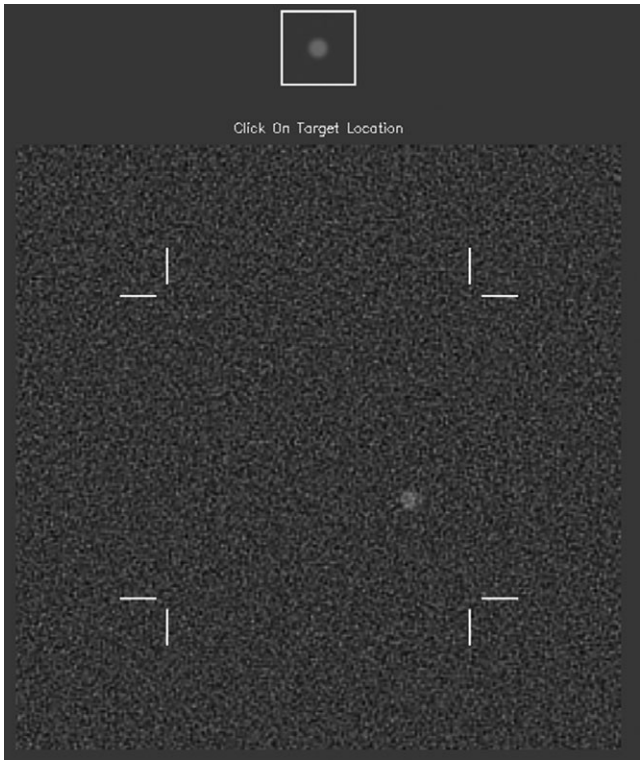

FIG. 5. Stimulus display. This image shows the central portion of the image display window that subjects use in the psychophysical experiments. The reference image of the noiseless target can be seen at the top of the figure, as well as hash marks indicating the search region. The gray border extends out further in the actual display.

To respond, a subject makes a pointer-click on the location believed to be the center of the target. The localization response is correct if it falls within a tolerance region of five pixels or less from the target center. This somewhat generous tolerance region (larger than the targets) was chosen based on prior publications showing little effect of larger regions,[36,37] and our intention that incorrect localizations used to generate the classification images not be impacted by the target profile. The display briefly gives feedback indicating whether the localization was correct or not.

The display procedure is used in two ways for each experimental condition. It is first used to estimate the threshold target amplitude needed to get a proportion of correct responses (P_C) that is approximately 80%, similar to previous work using this methodology.[20,38] The target amplitude is then fixed at this amplitude for the remainder of the study, which forms the basis for estimating the classification image. A threshold of P_C near 80% ensures that the task is hard enough that subjects will have some incorrect responses, which are essential for the classification image procedure we use, without being so difficult that subjects become unstable in their decision making. The staircase procedure we use also gives subjects some experience and training with the localization task in each condition.

Threshold amplitudes are estimated using a 3-down-1-up adaptive staircase procedure. In this procedure, the target amplitude starts at a high contrast, similar to Fig. 5. After three correct responses, the target amplitude is reduced by 15%. When a single incorrect decision is made, target amplitude is increased by 15%. After an initial "burn-in" period, this process will cause the chain of target amplitudes to fluctuate around the 80% correct level.[39] The staircase terminates after 12 reversals (i.e., when the chain goes from decreasing to increasing or vice versa). The 80% correct threshold is estimated from the geometric mean of the staircase amplitudes from the 4th reversal (to account for burn-in) to the termination point. The staircase procedure is repeated six times, with the first run considered a practice run, and the threshold estimates from the remaining five runs averaged as the final estimate of threshold amplitude.

The adaptive staircase runs are followed by a total of 40 runs of 50 trials each (2,000 trials total) at the threshold amplitude. We refer to these as the template runs, with each subject completing a template run in each of the 24 imaging conditions. Each subject contributed a total of 48,000 trials across all conditions, with 240,000 trials across the five subjects. These data are the basis for the estimation of efficiency and the classification images we report below. In principle, P_C should be 80% as a result of the threshold estimation procedure. We shall see below that observed P_C values are usually close to 80%, but there are some residual differences.

3.B. Ideal observer and efficiency

Task efficiency is computed for each subject in each experimental condition using the subject's threshold amplitude and the observed P_C for the template runs. Let A_{Sub} represent the threshold amplitude for a given subject. Let A_{IO} be the target amplitude that causes the ideal observer to perform the forced-localization task with the same P_C as the subject. We briefly describe how this threshold is obtained below. Task efficiency for the subject is estimated by the ratio[22,40]

$$\eta_{\text{Sub}} = \left(\frac{A_{\text{IO}}}{A_{\text{Sub}}}\right)^2. \tag{9}$$

This is an estimate since P_C is an estimate of proportion correct. Note that a 95% confidence interval on efficiency is obtained by transforming a 95% confidence integral on P_C for the subject, which is determined from the standard error over 40 runs of 50 trials each.

The ideal observer P_C is determined from Monte-Carlo studies as described previously,[20] which we will summarize briefly here. For a sample image, the ideal observer computes the posterior probability of the target being located at each point in the search region. This is accomplished by convolving the image with a prewhitened match filter and exponentiating the result. The posterior probability is then convolved with a disk profile that is five pixels in radius representing the criterion for a correct response. The location of the maximum point from this convolution is the ideal observer's localization response for the image.

Ideal observer P_C is compiled into a look-up table at several target amplitudes for the same 2,000 stimuli used in the psychophysical study. For an input P_C value, the threshold

amplitude is found using linear interpolation between points on the table.

3.C. Classification image procedure

The classification image procedure used here is similar to a previous study.[20] For a given subject in a given experimental condition, noise fields from stimuli in the template runs that are incorrectly localized (i.e., localization response is more than five pixels from the target location) are used to generate the classification image. We expect approximately 400 such trials for an 80% correct amplitude threshold. Let $N_i^{\mathrm{Inc}}[n,m]$ represent the noise field of the ith incorrect trial in an experiment ($i = 1, \ldots, I$). The first step in generating a classification image is filtering these noise fields with the inverse of the noise power spectrum

$$q_i[n,m] = \mathrm{FFT}^{-1}\left(\frac{\hat{N}_i^{\mathrm{Inc}}[k,l]}{S_{\mathrm{Tot}}[k,l]+\varepsilon}\right), \tag{10}$$

where $\hat{N}_i^{\mathrm{Inc}}[k,l]$ is the FFT of $N_i^{\mathrm{Inc}}[n,m]$, and ε is a small constant (less than 0.131% of max S_{Tot}) that prevents instability errors when $S_{\mathrm{Tot}}[k,l]$ is near zero. The estimated classification image, $w[n, m]$, is obtained by aligning the q_i to the localization response $\left[n_i^{Loc}, m_i^{Loc}\right]$ for the subject and averaging,

$$w[n,m] = \frac{1}{I}\sum_{i=1}^{I} q_i\left[n-n_i^{Loc}, m-m_i^{Loc}\right], \tag{11}$$

where it is understood that negative values of $n - n_i^{Loc}$ or $m - m_i^{Loc}$ "wrap around" to the other side of the array.

We are also interested in the frequency weighting of observer localization templates. Since target and noise spectra are radially symmetric, we represent these by a radial average. Let $\hat{Q}_i[k,l]$ be the FFT of $q_i[n-n_i^{Loc}, m-m_i^{Loc}]$, and let $\rho_{k,l}$ represent the radial distance from the origin defined in Section 2.C. We define the radial average, $\hat{R}_i[v]$ as the average of $\hat{Q}_i[k,l]$, for all points in which $|\rho_{k,l} - v| < \varepsilon$ with ε being half the frequency sampling of the stimuli. Note that because of conjugate symmetry of the FFT, this average will always be real valued. We estimate the average frequency weighting across incorrect images as

$$R_w[v] = \frac{1}{I}\sum_{i=1}^{I} R_i[v]. \tag{12}$$

Uncertainty in the average is estimated by the standard error of this average.

3.D. Localization error correction

A potential difficulty with fine-grain localization studies of the sort we use here is the potential for distortions from subjects' motor errors in response. We would like for the localization response to indicate the perceived location of the target center, and the subjects are instructed to exercise care selecting this point. However, it is known that localization responses are subject to positioning errors.[41,42] Piloting studies (data not shown) suggest that these errors are relatively small, typically with a root-mean-squared error less than two pixels. However, given that the alignment step in Eq. (11) is an important part of the classification image technique, we use a two-part scheme to adjust for motor error in the localization responses.

We begin by generating a classification image using uncorrected responses in Eq. (11). This initial classification image has a spatial window applied to it that extends out to just over twice the radius of the target (HWHM 1.05 mm for the small target and 4.2 mm for the large target with cosine roll-off). After spatial windowing, the classification image is smoothed by applying a window in the frequency domain (HWHM of 0.4 cyc/mm for the small target and 0.17 cyc/mm for the large target with cosine roll-off). In addition, the imaginary component of the FFT is set to zero, which symmetrizes the classification image about the origin.

This initial, highly filtered classification image is then applied to each location within two pixels of the subject's response. The adjusted localization response is taken to be the location with the maximum filter value. This process can be iterated, using the updated localization responses to generate the initial classification image. As we shall see below, analysis of correct localizations show that the adjusted localizations are generally closer to the actual target centers, but there is little difference after the first adjustment. Hence, a single iteration is used for classification image results.

4. RESULTS

Subjects performed the 24 experimental conditions in order of signal size and background variability and randomized over the level of apodization. For each condition, subjects ran the adaptive staircase procedure six times, with the average of the final five runs used to obtain the fixed amplitude threshold for the classification image runs, which were started immediately afterward. The classification image was broken into 40 runs of 50 trials each for a total of 2,000 trials total in each condition.

As we shall see below, the threshold estimation procedure generally produced results close to 80% correct. However, in one case, a subject made several early mistakes in the adaptive threshold runs for one condition, leading to a substantial overestimate of the threshold amplitude. This in turn led to the subject getting 99.5% correct in the template runs. In this one case, the subject re-ran the condition (at a later time) achieving a more reasonable threshold and subsequent localization performance. The second run for that subject in that condition is reported here.

4.A. Performance results

Performance results for the forced-localization experiments are shown in Fig. 6 as a function of the experimental conditions (signal size, level of background variability, and level of apodization) averaged across the five subjects (and

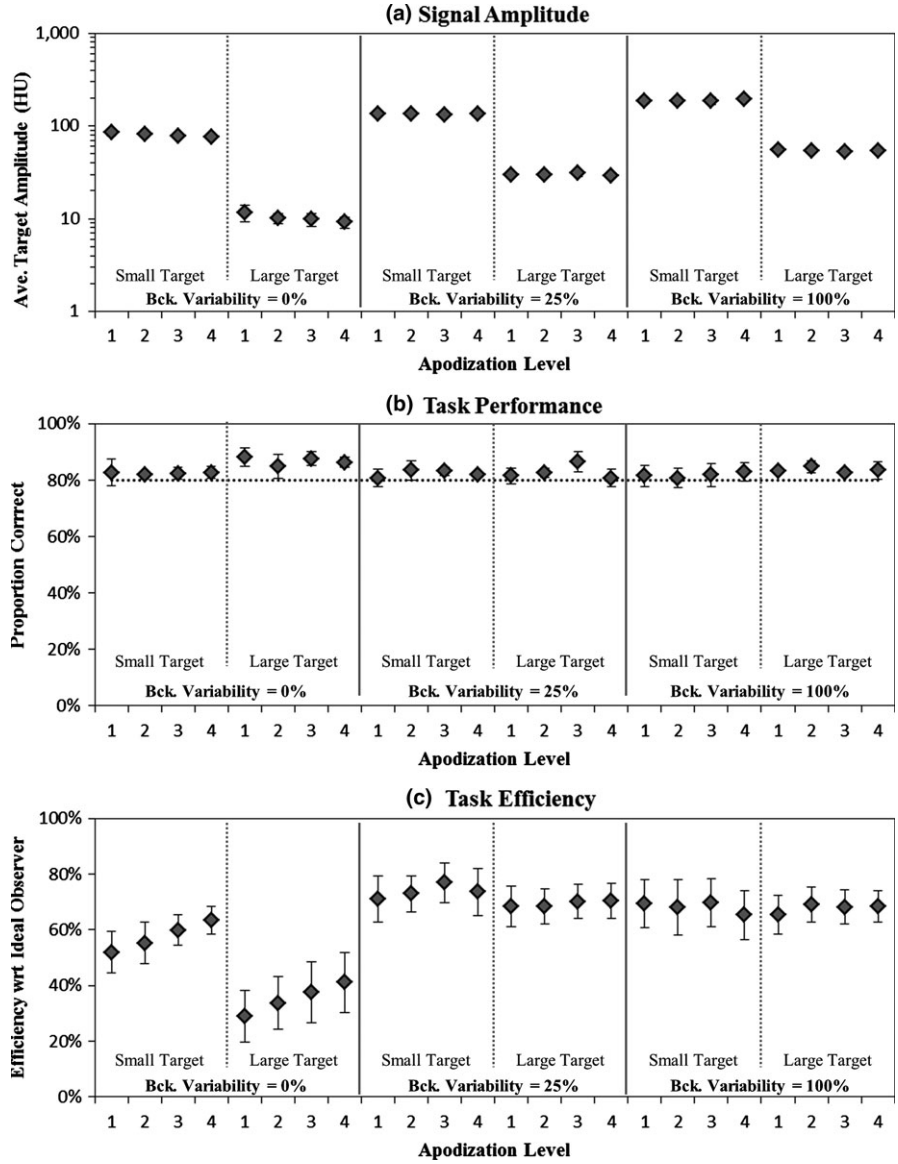

FIG. 6. Task performance data. Performance data from the 24 experimental conditions are shown with 95% confidence intervals across subjects, as a function of target size, amount of background variability, and level of apodization. The target amplitude data (a) are the result of the threshold estimation procedure. The task performance data (b) show that the observed proportion of correct localizations was reasonably close to the 80% threshold (dotted line) used for threshold estimation. Task efficiency with respect to the ideal observer (c) shows a considerable dependence across conditions with notably large confidence intervals due to inter-subject variability.

95% confidence intervals derived from the standard error across subjects). The amplitude thresholds [Fig. 6(a)] determined from the adaptive staircase procedure vary considerably across signal size and background variability. For a given signal size and level of background variability, less variability is seen across apodization levels. But it should be noted that a logarithmic scaling is used in this plot which may compresses smaller differences.

Figure 6(b) shows the proportion of correct localizations in the subsequent classification image studies using the fixed target amplitudes from the staircase runs. This serves as a check that the adaptive staircase is producing a target amplitude threshold that is approximately 80% correct. The figure suggests that most scores are close to this targeted level, although there appears to be some tendency for higher performance. The average PC across subjects and conditions is 83.2% and ranges from 74% to 94% for individuals. We expect observer task efficiency and performance mechanisms to be relatively stable to small changes in target amplitude. As a result, we do not expect the relatively small differences between observed PC and the targeted 80% correct threshold seen in Fig. 6(b) to have a significant effect on our results. For example, if we observe a P_C of 87% from a given subject for a given condition, we expect that the task efficiency of the subject and the classification image will be similar to what we would observe if we had used a slightly lower target amplitude that resulted in 80% correct.

Task efficiency, our primary endpoint of task performance, is plotted in Fig. 6(c). Here, we find considerable variation across signal size and background variability, particularly between uniform-background conditions and the conditions with some background variability. We also see some variability across apodization levels. Of interest, here is that variability due to different levels of apodization appears to be most pronounced in the uniform-background conditions.

This observation is confirmed by ANOVA modeling of efficiency for each of the five subjects at each of the six combinations of signal size and background variability, as shown in Table I. Each cell of the table is the *P*-value from a one-way ANOVA with four apodization levels as the main effects and the five blocks of 400 trials as replications. A total of nine *P*-values remain significant after a Bonferroni correction, and these are all found in cells corresponding to a uniform background.

4.B. Localization accuracy

Since the localization response is fundamental to the classification image procedure in these tasks, we begin by examining the accuracy of localization responses for correct trials (i.e., within the 5-pixel tolerance range). This allows us to see how the refinement procedure changes the response relative to the actual target center. We can evaluate bias in the localization response in x and y from the average positional error between the response and the true signal location in each dimension, μ_x and μ_y. We take the magnitude of both terms as the bias magnitude, $M_{\text{Bias}} = \sqrt{\mu_x^2 + \mu_y^2}$. We evaluate variance of the localization responses, σ_x^2 and σ_y^2, and use these to define the magnitude of localization variability, $M_{\text{Var}} = \sqrt{\sigma_x^2 + \sigma_y^2}$. The total error in the localization responses is characterized by the root-mean-squared error (RMSE), defined as

$$\text{RMSE} = \sqrt{M_{\text{Bias}}^2 + M_{\text{Var}}^2}.$$

Plots of average RMSE are shown in Fig. 7. We expect that it may be harder to localize a larger target, so we have separated the experimental condition into 12 small-target conditions and 12 large-target conditions. For each subject, the RMSE values from the 12 experimental conditions in each group are averaged. Subject averages are plotted as a function of the number of passes in the localization correction algorithm (0 passes is the original unmodified data). Consistent with our expectations, the larger target conditions have a larger RMSE in the unmodified localization data. In both cases, the RMSE drops with the first pass of the correction procedure. Additional passes seem to have little effect.

The plotted RMSE values are mostly due to variability rather than bias. The magnitude of variability explains 83% of the mean-squared error (RMSE^2) for the small target, and 97% of mean-squared error for the large target. Our results suggest that the correction procedure is reducing the effects of pointing errors by the subjects, and so, we use one pass of the procedure for our classification image results.

4.C. Classification images

The experiments generated a total of 120 classification images (five subjects, 24 conditions) in this work. In the interest of brevity, we will summarize the classification image results by presenting averages across subjects here. This is consistent with our interest in average subject performance at this stage, rather than individual differences between subjects. Individual differences in previous works involving classification images have been found to be fairly substantial,[43–45] and may explain some of the individual differences in performance, but we consider this topic to be the subject of future work.

Average classification image results are shown in Fig. 8. The classification images themselves are estimated for each subject from incorrect localization noise fields using Eq. (11), with one pass of localization correction. To reduce the effects of estimation error, the classification image estimates include windowing and filtering steps. The small-target classification images are shown after applying a spatial window that is constant out to four pixels (2 mm), with a cosine roll-off out to eight pixels. The result is then smoothed using a filter that is constant out to 0.5 cyc/mm with a cosine roll-off out to 1.0 cyc/mm (Nyquist). The resulting classification images (averaged across subjects) are shown in Fig. 8(a). The large-target classification images are shown after applying a

TABLE I. Significance of apodization on efficiency.

	Sm. Sig. BV = 0%	Lg. Sig. BV = 0%	Sm. Sig. BV = 25%	Lg. Sig. BV = 25%	Sm. Sig. BV = 100%	Lg. Sig. BV = 100%
Subject 1	<0.0001*	0.0286	0.02320	0.05341	0.44533	0.00482
Subject 2	<0.0001*	<0.0001*	0.09308	0.42095	0.00415	0.08555
Subject 3	<0.0001*	<0.0001*	0.04256	0.17757	0.01812	0.00671
Subject 4	<0.0001*	<0.0001*	0.06389	0.09195	0.02334	0.01535
Subject 5	0.0014*	<0.0001*	0.22903	0.47279	0.09309	0.40386

P-values for the significance of an apodization effect from a one-way ANOVA for each signal size and level of background variability and for each subject. The asterisk (*) indicates results that are significant after Bonferroni correction for a familywise error rate of 5%.

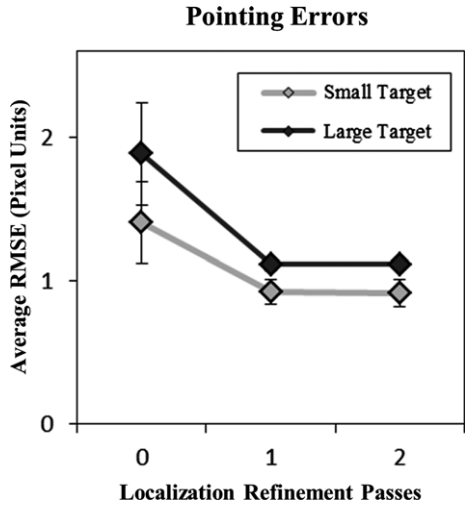

FIG. 7. Pointing errors. The average RMSE across subjects is plotted as a function of the number of localization refinement passes (0 passes represent the original unmodified localization responses). The experimental conditions are separated into two groups based on the size of the target. Confidence intervals (95%) are derived from the standard error across subjects.

spatial window that is constant out to 10 pixels (5 mm), with a cosine roll-off out to 20 pixels. The result is then smoothed using a frequency filter that is constant out to 0.2 cyc/mm with a cosine roll-off out to 0.4 cyc/mm. The resulting classification images (averaged across subjects) are shown in Fig. 8(b).

The classification images in the panel generally show a region of central positive weights with varying degrees of negative weights in the nearby surrounding areas. The patterns clearly scale with target size, and the magnitude of the negative surround appears to be getting stronger as the background variability increases (across columns). It is less clear what the differences are across the different levels of apodization, although careful inspection seems to indicate that there is some estimation error in the higher apodization levels. This is consistent with the fact that apodization reduces high spatial frequencies, making the classification image estimates noisier at these frequencies. Further assessments of the classification images are described in the next section.

5. DISCUSSION

The efficiency and classification image results shown in the previous section give some indications of how subjects perform the localization task. Here, we explore some of the implications of these results. Specifically, we are interested in better understanding how much of the variance in task efficiency can be explained by the observed classification images, how to characterize spatial weighting in the localization tasks, and to what extent any changes in the template with different apodization levels represents undoing the smoothing applied by the apodization filter. Each of these questions is treated below, with an additional sub-section reviewing the limitations of this study.

5.A. Explaining variance in task efficiency

Figure 6(c) shows that there is considerable variability in efficiency across tasks in this study, with average efficiencies ranging from 29% to 77% across the different conditions. It is of interest to know how much of this efficiency can be explained by the estimated classification images. To address this question, we have implemented each of the classification images in Fig. 8 as the kernel of a scanning linear model. The model chooses a location based on the maximum response of the kernel. The location is considered correct if it is within the same five-pixel distance of the target center that was used for the human-observer experiments. Since we seek to isolate the role of the classification images, we have not added any source of internal noise or any limitation of the search process, which are almost surely present in the human-observer data to some extent. So we might expect the efficiency of the scanning models to be somewhat higher than for humans. We have evaluated scanning template models on all 24 experimental conditions using a set of 2,000 images in each condition that are independent from those used in the human-observer experiment (and subsequently used in the construction of the classification images). Target amplitude has been adjusted to match the average proportion correct of the subjects plotted in Fig. 6(b). Efficiency is computed in the same way as human-observer efficiency.

A scatterplot of average human-observer efficiency and classification image efficiency is shown in Fig. 9. There is a clear association between the average classification image efficiency and human-observer efficiency. A linear regression of classification image efficiency as the independent variable and human-observer efficiency as the dependent variable has a highly significant slope ($P < 10^{-6}$) and a significant intercept ($P = 0.01$) as well. The estimated slope is 1.09 with standard error of 0.09, so the slope is not significantly different from 1. Thus, these data are well described by a simple offset model in which efficiency of the human observers is modeled as the efficiency of the classification image minus 12.8%. The R^2 value of the offset model is 0.86 showing that it explains most of the variance in the average human-observer efficiency data.

This strong association leads us to believe that the classification images are capturing most of the observer performance effects of the different statistical properties across these tasks. Other effects, such as internal noise or visual search patterns, have a relatively small consistent influence on efficiency.

5.B. Characterizing subject classification images

Models of observer performance are often based on features derived from the spatial-frequency domain. This motivates us to consider the spectral components of the classification images as a way to characterize them. In Fig. 10, we plot radial averages of the Discrete Fourier Transform of the classification images, implemented via the Fast Fourier Transform (FFT), normalized so that the peak value

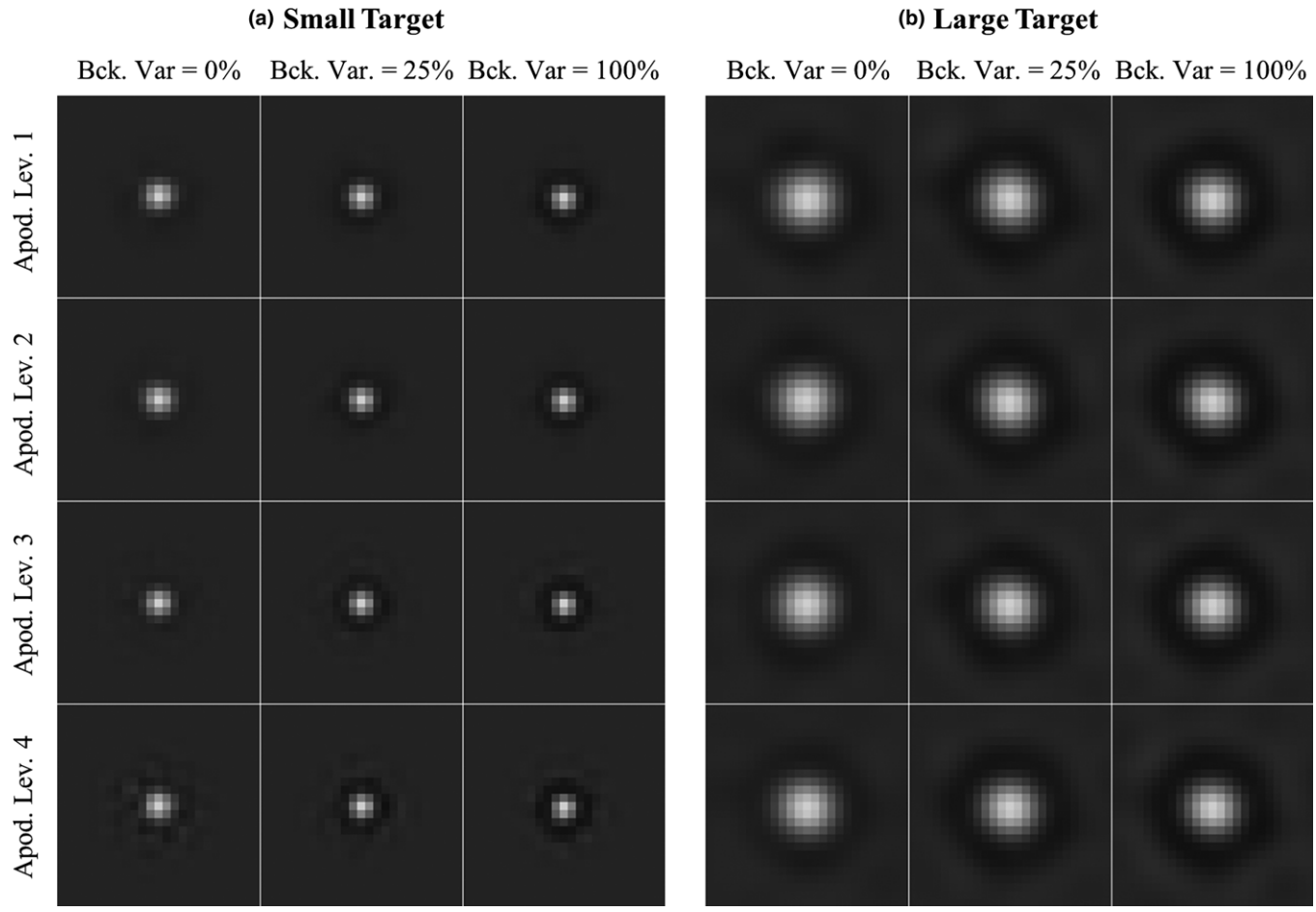

FIG. 8. Classification images averaged across subjects. Subject averaged classification images for large (a) and small (b) targets, estimated according to the procedure in Eq. (11), are shown for each of the 24 tasks.

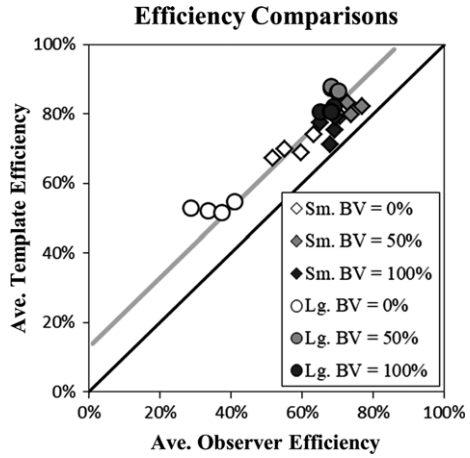

FIG. 9. Scatterplot of observer efficiency and template efficiency. The average efficiency of human observers in each task is plotted against efficiency of the average template in each condition. Note the legend indicates the signal size (Sm. or Lg.) and the relative magnitude of background variability (BV = 0%, 50% or 100%). The data are reasonably well fit ($r^2 = 0.86$) by an offset of 12.4% in efficiency (gray line).

is 1. We can think of these plots as representing the spatial-frequency weights used by subjects (on average) in the small-target (Fig. 10a) and large-target (Fig. 10b) tasks. Even though FFTs are generally a complex quantity, these plots are real since the conjugate property of the FFT will cancel any imaginary component in the radial average. The data used for these plots have been subjected to spatial windowing as described in Section 4.C (which smooths the spectrum), but they have not had the frequency window applied. Each plot shows the spectra for all four apodization levels at a given level of background variability and target size.

It is clear from these plots that the frequency weights generally form a bandpass structure. In the conditions involving a uniform object, the band falls off only modestly on the low-frequency side of the peak. As the level of background variability increases, there is more low-frequency suppression. This low-frequency suppression is related to the negative peripheral weights in Fig. 8. For the small-target conditions [Fig. 10(a)], there is also some evidence of a shift to higher spatial frequencies with increased apodization. This shift is most apparent in the two conditions with background variability, where the band moves to higher frequency at the higher apodization levels.

As a way to quantify these observations, we plot the fractional bandwidth of the classification image frequency weights as a function of the peak frequency in Fig. 11. The peak frequency is found from the argmax of a parabola fit by

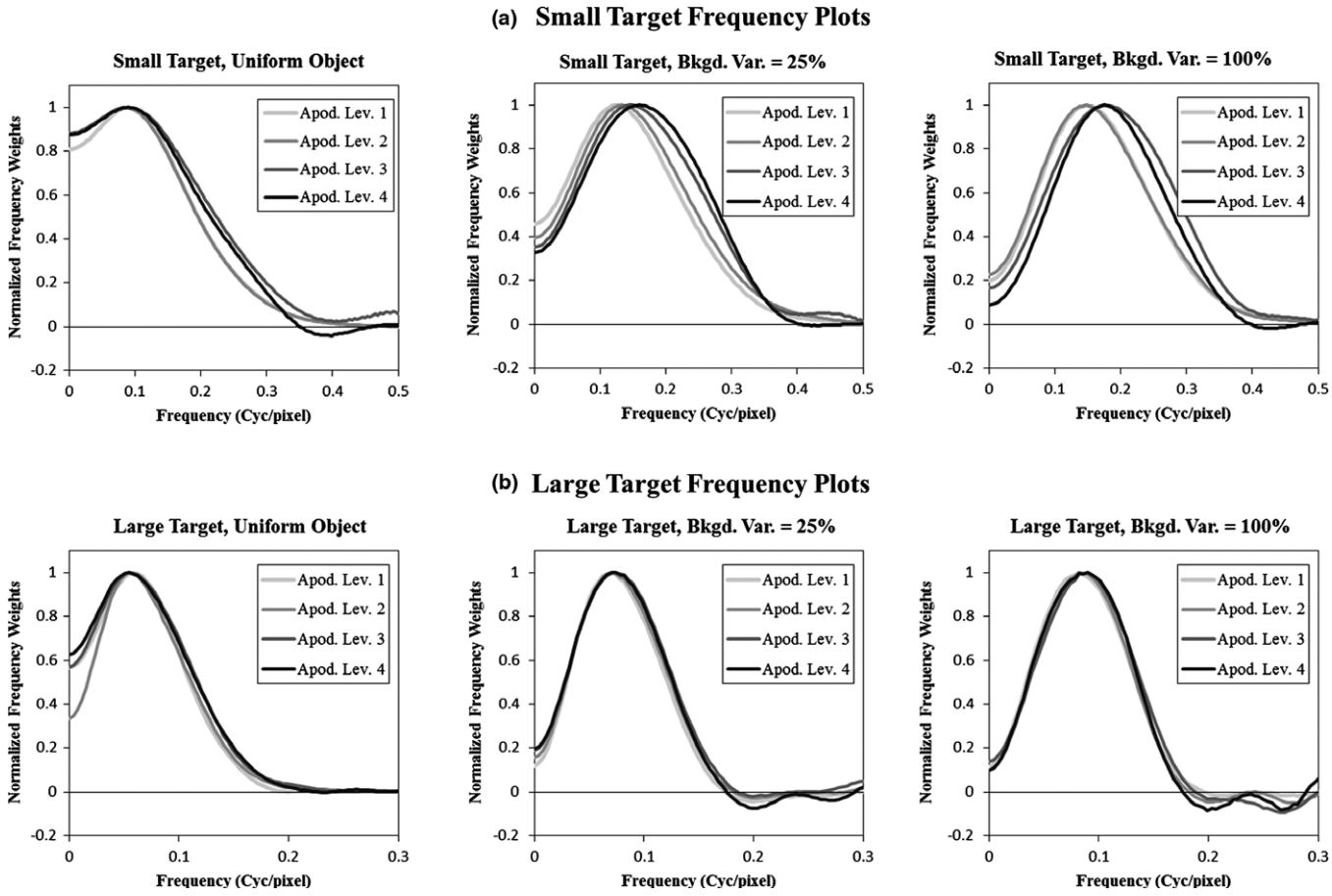

FIG. 10. Classification image spectra. The plots show radial averages of the classification images for large (a) and small (b) targets and for each level of background variability. These use the same data as was used to generate the spatial profiles shown in Figure 8, except that no frequency window is applied.

least-squares to the points greater than 0.95 on each plot. The bandwidth is the frequency range for values of 0.5 or more on each plot (i.e., full-width at half-max). Each plotted line represents a given target size and apodization level over the 3 levels of background variability.

For all plots, increasing background variability also increases the peak frequency of the classification images. For the large targets, peak frequencies increase by an average of 51% over the range of background variability. For the small targets, peak frequencies increase by 85% with the most substantial increases in the two highest apodization levels. Bandwidths remain relatively constant, changing by 20% or less as background variability increases. For the large targets, the different levels of apodization seem to have little effect, except for uniform backgrounds where there appears to be some variation in fractional bandwidth. Both Figs. 10 and 11 would suggest that apodization does not have much effect on spatial weighting for the large target. This is perhaps not surprising since the signal energy of the large target is not impacted much by the various levels of apodization. For small targets, which have more energy in the spatial frequencies affected by apodization, increasing apodization leads to higher peak frequencies with relatively little effect on bandwidth.

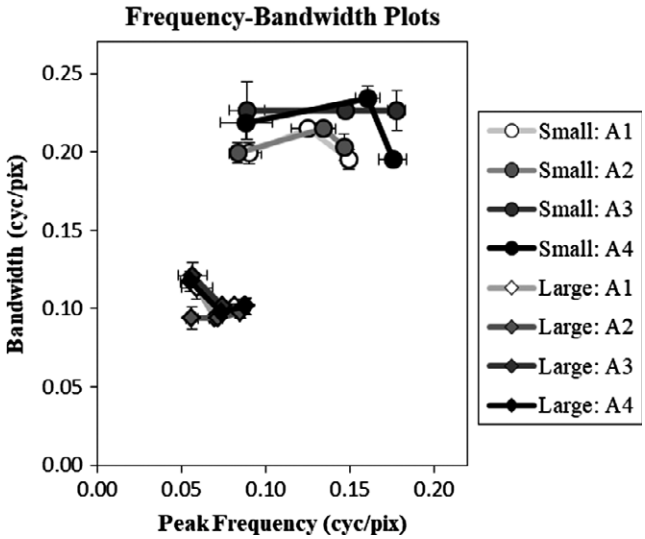

FIG. 11. Peak frequency vs bandwidth plots. For each combination of target size (Small, Large) and apodization level (A1–A4), a plot shows template bandwidth as a function peak spatial frequency, plotted across the three levels of background variability. Increasing background variability generally leads to higher peak frequency and lower bandwidth. Error bars represent ±1SE derived from bootstrapping across subjects (200 resamples).

5.C. Effect of apodization

Figures 10 and 11 suggest that as high spatial frequencies of the target are suppressed through apodization, human observers show signs of increasing their weights for these frequencies, potentially undoing the apodization to some degree. To focus on this possibility, more generally, we investigate a different approach to the classification images.

We evaluate classification images using subject localization responses to the apodized images in each condition, but when creating the classification images, we use unapodized images in Eq. (10) (along with the power spectrum of the unapodized images as S_{Tot}). This makes any apodization effect akin to a perceptual effect, since the observer responses have effects of apodization in them, but the images do not. An observer that is invariant to apodization (i.e., is able to "undo" the effects of apodization) should give the same response (up to internal noise and pointing errors), and therefore generate approximately the same classification image for any level of apodization. Alternatively, if the observer cannot undo the effects of apodization, we would expect these classification images to differ.

In contrast, the standard classification image technique (using apodized noise fields) would predict the opposite. If the observer is able to undo the effects of apodization (by increasing weights of high spatial frequencies as the apodization level increases), then the standard classification image technique should be able to capture this, since it gives weights for the apodized stimuli. For the standard classification image technique, an observer that does not adapt to apodization should produce the same weights across apodization levels.

These two descriptions represent the extremes of perfect undoing of apodization compared with no undoing of apodization. The reality will likely be somewhere in the middle. One way to get at partial effects is to evaluate the differences between spatial templates across different levels of apodization. We use a measure of difference between spatial weights that is based on the difference between normalized classification images. Let $w_{A_1}[n,m]$ and $w_{A_2}[n,m]$ be two classification images from different levels of apodization. Our measure of difference between w_{A_1} and w_{A_2} is given by

$$D(w_{A_1}, w_{A_2}) = \sqrt{\sum_{n,m}\left(\frac{w_{A_1}[n,m]}{\|w_{A_1}\|} - \frac{w_{A_2}[n,m]}{\|w_{A_2}\|}\right)^2}, \qquad (13)$$

where $\|\mathrm{L}\|$ represents the standard Euclidean norm.

We have evaluated Eq. (13) for each subject, using the standard classification images estimated using apodized noise fields (Fig. 8) and with unapodized noise fields (data not shown). Spatial windowing and frequency filtering were applied as described in Section 4.C. Apodization has such a small effect on the large-target conditions that the templates are virtually identical regardless of whether apodized or unapodized noise fields are used. This is consistent with Figs. 10 and 11, where very little effect of apodization is seen for the large targets. As a result, we will focus on results for the small target, as shown in Fig. 12.

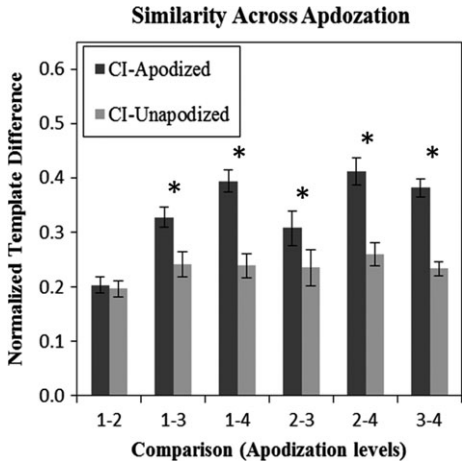

FIG. 12. Template Similarity. Normalized template differences are plotted for six possible comparisons of the four apodization levels. Results are averaged across background variability in the small-target conditions and across subjects. Error bars are the standard error across subjects. Significant differences (paired *t*-test) after correction for multiple comparisons are indicated with an asterisk (*).

These plots show the template difference averaged across levels of background variability and across subjects for each of the six possible comparisons of the four apodization levels. Normalized differences are generally smaller between the classification images generated from unapodized noise fields. Paired comparisons across subjects find significant differences in five of the six comparisons, which persist with a Bonferroni correction for multiple (i.e., a total of six) comparisons. This result suggests that in the small-target task, part of what the subjects are doing is adapting to the apodization applied to the images.

5.D. Limitations of the study

We believe that the results we have presented are useful for understanding how human observers perform tasks that are limited by noisy stimuli with statistical properties similar to CT images. However, we have made a number of assumptions and simplifications that may limit the generality of our results for practical purposes. Our task, which involves forced localization of a fixed target profile embedded in stationary Gaussian noise, is considerably simplified relative to practice in radiology or other specializations. The basic assumption here is that these simplifications, which allow us to compute task efficiency and subject classification images, reveal basic observer processes that are at work in more complex applications as well. We also have used non-clinical subjects as our readers, under the assumption that this sort of simple task is not heavily dependent on clinical training.

6. CONCLUSIONS

The classification image approach used in these experiments allows a closer look at how our human subjects perform forced-localization tasks in images with statistical

properties that resemble axial CT images. We find that the average statistical efficiency of subject performance in our experiments varies substantially (29% to 77%) depending on components of the images (target size, background variability, and apodization). The lowest efficiencies were observed in the uniform-background conditions. These were the only conditions in which apodization appeared to have an effect on performance. When the tasks included background variability, efficiency increased with a loss of any significant dependency on apodization.

We find that classification images consistently adopt "center-surround" profiles where positive central weights are bordered by more peripheral negative weights. This results in a bandpass structure to the classification images in the spatial-frequency domain. The classification images explain most of the variance in efficiency (86%) using a simple constant-offset model, which suggests a relatively modest and predictable role for other components of task performance such as internal noise and visual search. Within the confines of bandpass center-surround profiles, the classification images do appear to be changing. These changes appear, to some extent, to undo the effects of apodization.

At a broader level, our results are consistent with the idea that human observers are able to partially adapt to the statistical structure of images in performing the forced-localization task. They also suggest that bandpass filters may be a productive direction for modeling spatial weighting by human observers, where the center frequency and bandwidth of the filter are determined by statistical properties of the images.

ACKNOWLEDGMENTS

The authors acknowledge support from the US Food and Drug Administration (IPA 1457) and the National Institutes of Health (R01-EB018958 and R01-CA181081).

CONFLICT OF INTEREST

None declared.

[a]Author to whom correspondence should be addressed. Electronic mail: ckabbey@ucsb.edu.

REFERENCES

1. Barrett HH, Gordon SK, Hershel RS. Statistical limitations in transaxial tomography. *Comput Biol Med*. 1976;6:307–323.
2. Riederer SJ, Pelc NJ, Chesler DA. The noise power spectrum in computed x-ray tomography. *Phys Med Biol*. 1978;23:446.
3. Kijewski MF, Judy PF. The noise power spectrum of CT images. *Phys Med Biol*. 1987;32:565.
4. Shepp LA, Logan BF. The Fourier reconstruction of a head section. *IEEE Trans Nucl Sci*. 1974;21:21–43.
5. Myers KJ, Barrett HH, Borgstrom MC, Patton DD, Seeley GW. Effect of noise correlation on detectability of disk signals in medical imaging. *J Opt Soc Am A*. 1985;2:1752–1759.
6. Myers KJ, Barrett HH. Addition of a channel mechanism to the ideal-observer model. *J Opt Soc Am A*. 1987;4:2447–2457.
7. Rolland JP, Barrett HH. Effect of random background inhomogeneity on observer detection performance. *J Opt Soc Am A*. 1992;9:649–658.
8. Barrett HH, Yao J, Rolland JP, Myers KJ. Model observers for assessment of image quality. *Proc Natl Acad Sci USA*. 1993;90:9758–9765.
9. Burgess AE. Statistically defined backgrounds: performance of a modified nonprewhitening observer model. *J Opt Soc Am A Opt Image Sci Vis*. 1994;11:1237–1242.
10. Burgess AE, Li X, Abbey CK. Visual signal detectability with two noise components: anomalous masking effects. *J Opt Soc Am A Opt Image Sci Vis*. Sep 1997;14:2420–2442.
11. Abbey CK, Barrett HH. Human- and model-observer performance in ramp-spectrum noise: effects of regularization and object variability. *J Opt Soc Am A Opt Image Sci Vis*. 2001;18:473–488.
12. Gifford HC, King MA, Pretorius PH, Wells RG. A comparison of human and model observers in multislice LROC studies. *IEEE Trans Med Imaging*. 2005;24:160–169.
13. Platisa L, Goossens B, Vansteenkiste E, et al. Channelized Hotelling observers for the assessment of volumetric imaging data sets. *J Opt Soc Am A Opt Image Sci Vis*. 2011;28:1145–1163.
14. Packard NJ, Abbey CK, Yang K, Boone JM. Effect of slice thickness on detectability in breast CT using a prewhitened matched filter and simulated mass lesions. *Med Phys*. 2012;39:1818–1830.
15. Fan J, Tseng H-W, Kupinski M, Cao G, Sainath P, Hsieh J. Study of the radiation dose reduction capability of a CT reconstruction algorithm: LCD performance assessment using mathematical model observers. In: *SPIE Medical Imaging*; 2013, pp. 86731Q–86731Q-8.
16. Leng S, Yu L, Zhang Y, Carter R, Toledano AY, McCollough CH. Correlation between model observer and human observer performance in CT imaging when lesion location is uncertain. *Med Phys*. 2013; 40:081908.
17. Eckstein MP, Ahumada AJ Jr. Classification images: a tool to analyze visual strategies. *J Vis*. 2002;2:1x.
18. Abbey CK, Eckstein MP. Classification image analysis: estimation and statistical inference for two-alternative forced-choice experiments. *J Vis*. 2002;2:66–78.
19. Murray RF. Classification images: a review. *J Vis*. 2011;11:1–25.
20. Abbey CK, Eckstein MP. Observer efficiency in free-localization tasks with correlated noise. *Front Psychol*. 2014;5:1–13.
21. Burgess AE, Wagner RF, Jennings RJ, Barlow HB. Efficiency of human visual signal discrimination. *Science*. 1981;214:93–94.
22. Kersten D. Spatial summation in visual noise. *Vis Res*. 1984;24:1977–1990.
23. Tjan BS, Braje WL, Legge GE, Kersten D. Human efficiency for recognizing 3-D objects in luminance noise. *Vis Res*. 1995;35:3053–3069.
24. Park S, Clarkson E, Kupinski MA, Barrett HH. Efficiency of the human observer detecting random signals in random backgrounds. *J Opt Soc Am A Opt Image Sci Vis*. 2005;22:3–16.
25. Abbey CK, Zemp RJ, Liu J, Lindfors KK, Insana MF. Observer efficiency in discrimination tasks simulating malignant and benign breast lesions imaged with ultrasound. *IEEE Trans Med Imaging*. 2006;25:198–209.
26. Park S, Gallas BD, Badano A, Petrick NA, Myers KJ. Efficiency of the human observer for detecting a Gaussian signal at a known location in non-Gaussian distributed lumpy backgrounds. *J Opt Soc Am A Opt Image Sci Vis*. 2007;24:911–921.
27. Liu B, Zhou L, Kulkarni S, Gindi G. The efficiency of the human observer for lesion detection and localization in emission tomography. *Phys Med Biol*. 2009;54:2651–2666.
28. Burgess AE, Jacobson FL, Judy PF. Human observer detection experiments with mammograms and power-law noise. *Med Phys*. 2001;28:419–437.
29. Metheany KG, Abbey CK, Packard N, Boone JM. Characterizing anatomical variability in breast CT images. *Med Phys*. 2008;35:4685–4694.
30. Hsieh J. Computed tomography: principles, design, artifacts, and recent advances; 2009.
31. Bushberg JT, Boone JM. *The Essential Physics of Medical Imaging*. Philadelphia, PA: Lippincott Williams & Wilkins; 2011.
32. Bochud FO, Valley JF, Verdun FR, Hessler C, Schnyder P. Estimation of the noisy component of anatomical backgrounds. *Med Phys*. 1999;26: 1365–1370.

33. Heine JJ, Deans SR, Cullers DK, Stauduhar R, Clarke LP. Multiresolution statistical analysis of high-resolution digital mammograms. *IEEE Trans Med Imaging*. 1997;16:503–515.
34. Heine JJ, Deans SR, Velthuizen RP, Clarke LP. On the statistical nature of mammograms. *Med Phys*. 1999;26:2254–2265.
35. Chen L, Abbey CK, Boone JM. Association between power law coefficients of the anatomical noise power spectrum and lesion detectability in breast imaging modalities. *Phys Med Biol*. 2013;58: 1663–1681.
36. Haygood T, Ryan J, Brennan P, et al. On the choice of acceptance radius in free-response observer performance studies. *Br J Radiol*. 2013;86:42313554–42313554.
37. Gifford H, Kinahan P, Lartizien C, King M. Evaluation of multiclass model observers in PET LROC studies. *IEEE Trans Nucl Sci*. 2007;54:116–123.
38. Abbey CK, Eckstein MP. High human-observer efficiency for forced-localization tasks in correlated noise. In: *SPIE Medical Imaging*; 2010, pp. 76270R–76270R-8.
39. Garcıa-Pérez MA. Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties. *Vis Res*. 1998;38:1861–1881.
40. Kersten D. Statistical efficiency for the detection of visual noise. *Vis Res*. 1987;27:1029–1040.
41. Trommershäuser J, Maloney LT, Landy MS. Statistical decision theory and trade-offs in the control of motor response. *Spat Vis*. 2003;16:255–275.
42. Trommershäuser J, Gepshtein S, Maloney LT, Landy MS, Banks MS. Optimal compensation for changes in task-relevant movement variability. *J Neurosci*. 2005;25:7169–7178.
43. Gold JM, Murray RF, Bennett PJ, Sekuler AB. Deriving behavioural receptive fields for visually completed contours. *Curr Biol*. 2000;10:663–666.
44. Abbey CK, Eckstein MP. Classification images for detection, contrast discrimination, and identification tasks with a common ideal observer. *J Vis*. 2006;6:335–355.
45. Abbey CK, Eckstein MP. Classification images for simple detection and discrimination tasks in correlated noise. *J Opt Soc Am A Opt Image Sci Vis*. 2007;24:B110–B124.