Design of a Model Observer to Evaluate Calcification Detectability in Breast Tomosynthesis and Application to Smoothing Prior Optimization

Koen Michielsen and Johan Nuyts*
Department of Imaging and Pathology, division of Nuclear Medicine
& Molecular Imaging, KU Leuven, Leuven 3000, Belgium and
Medical Imaging Research Center, KU Leuven, Leuven 3000, Belgium

Lesley Cockmartin, Nicholas Marshall, and Hilde Bosmans
Department of Radiology, University Hospitals Leuven, Leuven 3000, Belgium and
Medical Imaging Research Center, KU Leuven, Leuven 3000, Belgium

(Dated: November 18, 2016)

Purpose: In this work we design and validate a model observer that can detect groups of microcalcifications in a four alternative forced choice (4-AFC) experiment and use it to optimize a smoothing prior for detectability of microcalcifications.

Methods: A channelized Hotelling observer (CHO) with eight Laguerre-Gauss channels was designed to detect groups of five microcalcifications in a background of acrylic spheres by adding the CHO log-likelihood ratios calculated at the expected locations of the five calcifications.

This model observer is then applied to optimize the detectability of the microcalcifications as a function of the smoothing prior. We examine the quadratic and total variation (TV) priors, and a combination of both. A selection of these reconstructions was then evaluated by human observers to validate the correct working of the model observer.

Results: We found a clear maximum for the detectability of microcalcification when using the total variation prior with weight $\beta_{TV} = 35$. Detectability only varied over a small range for the quadratic and combined quadratic-TV priors when weight β_Q of the quadratic prior was changed by two orders of magnitude.

Spearman correlation with human observers was good except for the highest value of β for the quadratic and TV priors. Excluding those, we found $\rho = 0.93$ when comparing detection fractions, and $\rho = 0.86$ for the fitted detection threshold diameter.

Conclusions: We successfully designed a model observer that was able to predict human performance over a large range of settings of the smoothing prior, except for the highest values of β which were outside the useful range for good image quality.

Since detectability only depends weakly on the strength of the combined prior, it is not possible to pick an optimal smoothness based only on this criterion. On the other hand, such choice can now be made based on other criteria without worrying about calcification detectability.

Medical Physics 2016.

I. INTRODUCTION

Digital breast tomosynthesis (DBT) is a recent three dimensional (3D) breast imaging technique with increasing clinical use. Compared to 2D mammography, reconstruction of the breast's anatomy in thick slices with high in-plane resolution allows better visualization of low contrast lesions due to the removal of interference from overlapping tissues. According to the meta-analysis by Lei et al.[1] single view DBT results in better sensitivity and specificity than two view digital mammography for diagnosing benign and malignant breast lesions, even though there remains a risk of underclassifying malignant lesions that present as microcalcifications[2].

Despite being commercially available for a few years already, current breast tomosynthesis systems[3] show a large variety in reconstruction methods and acquisition parameters such as angular range and number of projections, indicating that there is no obvious optimal practical implementation of a breast tomosynthesis system. Further optimization within the constraints of existing hardware could therefore potentially increase the clinical performance of these devices, but this process requires relevant quality metrics and efficient methods to evaluate them. A good candidate is the detection performance of simple geometric shapes in a structured background, because for this task human performance can be estimated by model observers[4–8].

When model observers are applied in breast tomosynthesis, they are typically used to evaluate either the projection geometry or the 3D reconstruction technique. In many instances a channelized Hotelling observer (CHO) is used to evaluate detectability of low contrast mass-type lesions. Chawla et al.[9] used this method to examine the effect of the number of projection images and the total angular range of those projections on detectability in both the projection and the reconstructed image domains. Possible sources of discrepancies between different model observer implementations were examined by Young et al.[10] who showed the need to take inter-projection correlations into account when applying the CHO in the projection domain and by Park et al.[11] who demonstrated that the choice of different 2D and 3D observer channels resulted in different preferences in system geometry. Focusing on the evaluation of 3D reconstruction techniques, Van de Sompel et al.[12] used the CHO to compare variants of filtered backprojection (FBP), the simultaneous algebraic reconstruction technique (SART), and maximum likelihood (ML) reconstruction, while Zeng et al.[13] found that the choice of reconstruction method did not influence the ranking of different acquisition geometries.

Some authors used alternative model observers to the CHO, such as the channelized non-prewithening (CNPW) observer from the works of Gifford et al.[14] and Lau et al.[15] which they used to examine the effect of the number of projection views on detectability. This observer was shown to predict human observer performance when combined with a separate holistic search step[15].

In cases where the background structure and noise are stationary, it is possible to use frequency domain observers instead of the CHO and CNPW which are applied in the image domain. Reiser and Nishikawa[16] presented a prewithening observer which was used to evaluate detectability of low contrast masses as a function of the number of projections, scan angle, and quantum noise. Wang et al.[17] used the same observer type to optimize a slice thickness filter in FBP and Gang et al.[18] evaluated performance of five frequency domain model observers for a wide range of scan angles and found reasonable correspondence with human observers.

While most authors focus on the detection of mass-like lesions, a few also concentrate on the detection of microcalcifications, an area where current tomosynthesis implementations do not have an advantage over digital mammography[19]. Das et al. used the same visual search CNPW as Lau et al.[15] and applied it to compare the detectability of microcalcifications in FBP and penalized ML reconstructions[20] and to evaluate the effect of the cutoff frequency of a Butterworth filter in FBP reconstruction[21]. Hu and Zhao[22] applied a frequency domain prewithening observer to demonstrate the effects of angular dose distribution on the detection of microcalcifications, and Sidky et al.[23] used the same type of model observer and task to optimize a total variation smoothing prior.

Here we focus on this last task by designing a channelized Hotelling observer that can predict human observer performance in a microcalcification detection task, and then apply this model observer to optimize the smoothing prior in the model based ML iterative reconstruction we presented previously[24].

II. MATERIALS AND METHODS

A. Phantom & Reconstruction

We used a prototype phantom designed to compare the performance of 2D full-field digital mammography and digital breast tomosynthesis systems[25]. It contains 3D printed masses and microcalcification particles ($CaCO_3$, 2.7 g/cm^3, produced by Leeds Test Objects Ltd, Boroughbridge, United Kingdom) placed within a structured background consisting of acrylic (PMMA) spheres with diameters of 15.9 mm, 12.7 mm, 9.5 mm, 6.4 mm, 3.2 mm, and 1.6 mm, placed in water, which together show statistical properties close to these of patient images[35]. A photograph and a reconstructed slice of the phantom are shown in figure 1. In this work we only consider the microcalcification targets.

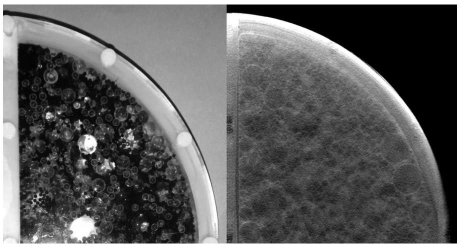

FIG. 1. A photograph (left) and reconstructed slice (right) of the spheres-phantom.

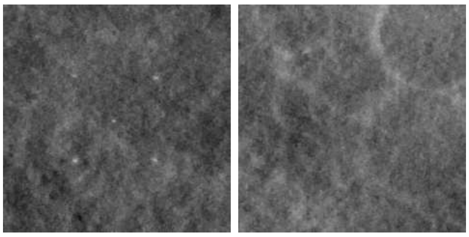

FIG. 2. Two 20×20 mm^2 regions from the phantom, one with the microcalcification target, and one without.

The phantom includes five microcalcification groups fixed on a 2 mm thick PMMA plate, placed at a height of 20 mm in the phantom, and 50 mm away from the chest wall side. Each group consists of five calcifications arranged such that four lie in the corners of a square (side 7.1 mm), with the fifth at the center, as shown in figure 2. The different groups contain calcifications with diameters in the following ranges: 90–100 μm, 112–125 μm, 140–160 μm, 180–200 μm, and 224–250 μm. The thin plate itself is not visible in the reconstructions and the limited angle acquisition causes the structured background to be visible inside the volume of the thin plate in the form of out-of-plane artifacts.

Fifteen sets of projection data were acquired on the Siemens Mammomat Inspiration tomosynthesis system for each of three exposure settings: the one determined by the automatic exposure control (AEC), and at half and double the AEC dose level. The detector has a pixel spacing of 85 μm and each acquisitions consists of 25 projections spread over 50°. Between every set of three acquisitions at the different dose levels, the phantom was shaken and placed back on the detector in order to generate a different background structure with the same statistical properties by displacing the spheres in the phantom.

Phantom images for all three dose levels were reconstructed using the Siemens system filtered backprojection (FBP) which includes 2×2 pixel binning of the projection data, based on the work of Mertelmeier et al.[26] and Orman et al.[27]. Images for the AEC dose level were also reconstructed with the MLTR$_{\mathrm{pr}}$ method[24,28] starting from full resolution projection data.

MLTR$_{\mathrm{pr}}$ is an iterative reconstruction algorithm that maximizes the posterior likelihood L in equation (1). It depends on the measured data y_i, the forward model $\hat{y}_i$ which includes an acquisition dependent resolution model to compensate for blur introduced by the motion of the x-ray tube during image acquisition, and includes additional constraints in the form of a quadratic[29] smoothing prior with weight β_Q and a total variation TV_{l_1}[30,31] smoothing prior with weight β_{TV}. The two smoothing priors further depend on the reconstruction volume $\vec{\mu}$ (indexed by j and k) and neighbor weights w_{jk}.

$$L = \sum_i y_i \ln \hat{y}_i - \hat{y}_i - \frac{\beta_Q}{4} \sum_{j,k} w_{jk} (\mu_j - \mu_k)^2 - 4\beta_{TV} \sum_{j,k} w_{jk} \left|\mu_j - \mu_k\right| \tag{1}$$

All reconstructions were performed with voxel sizes of 85×85×1000 μm^3 and from these reconstructions regions of 236×236×30 voxels, roughly corresponding to 20×20×30 mm^3 , were extracted for use in the observer studies described below.

B. Human Observer Study

Images acquired from the phantom were used to evaluate the detectability threshold of the included groups of calcifications for a specific combination of system settings by means of a human observer study. The study took the form of a four-alternative forced choice (4-AFC) where all four choices were shown to the observer at the same time, since this has been shown to provide the best results when working with inexperienced observers such as students or interns[32]. In such a study each observer is shown four 3D regions of interest (ROI) as a stack of sequential reconstructed slices, one of which contains a group of microcalcifications, and then selects the ROI thought to contain the lesion. In practical terms, the 4-AFC cases were presented in groups of 15, with matching target size and reconstruction type within each group. The order in which these groups were evaluated was randomized independently for each observer. By evaluating this experiment for different calcification diameters and observers for each reconstruction, we can determine the smallest diameter for which at least 62.5% of the microcalcification groups remain visible by a two parameter psychometric curve fit[33] to the correctly detected fractions $d(\phi)$ for each target diameter ϕ:

$$d(\phi) = 0.25 + \frac{0.75}{1 + \left(\frac{\phi}{\phi_{tr}}\right)^{-f}}. \tag{2}$$

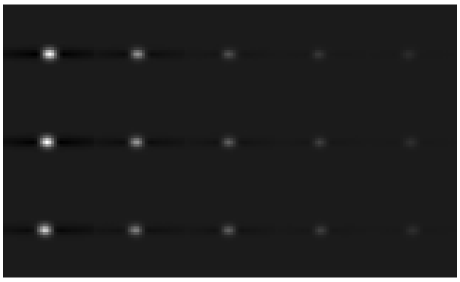

FIG. 3. Templates of the five target diameters for an FBP reconstruction, showing the focus plane in the middle together with the planes above (top) and below (bottom).

The free parameters are the 62.5% detection threshold diameter ϕ_{tr}, and the slope of the curve f.

C. Model Observer Design

In the design and tuning of the model observer, we relied on independent datasets as much as possible, such that it was not tuned and evaluated on the same or similar data[34]. The common aspect that was shared for all data was the system geometry of the Siemens Mammomat Inspiration which means additional validation will be needed before applying the MO to other system geometries.

1. *Templates and Channels*

The first task in setting up the model observer is generating a set of signal templates for the reconstructed microcalcification targets in the phantom. Because the target sizes were chosen in order to create a wide range of detection levels (from non-visible to subtle to obvious), it was not feasible to get good signal templates from the measured data. Therefore simulated projection data were used to create the signal templates.

Using the system geometry of the Siemens Mammomat Inspiration tomosynthesis system, we simulated noise- and scatter-free projection data of the microcalcification targets in a homogeneous background, and of the same homogeneous background without the targets. The targets were simulated at an isotropic voxel size of 5 μm, and the background at 85 μm. The detector pixels were supersampled by a factor of 5 to model partial volume effects, and eight source positions were simulated to model the tube motion during the 120 ms exposure time. The target templates were then obtained by subtracting the reconstruction of the background from the reconstruction with the target included. Examples are shown in figures 3 and 4 for FBP and MLTR$_{\mathrm{pr}}$ with $\beta_Q = 2 \cdot 10^4$ and $\beta_{TV} = 2$ respectively.

The channels for the CHO were selected by evaluating the performance of the first 2 to 16 Laguerre-Gauss (LG) channels[36,37], with width σ of the Gaussian part set to between 1 and 5 pixels of 85 μm in steps of 0.2 pixels, on

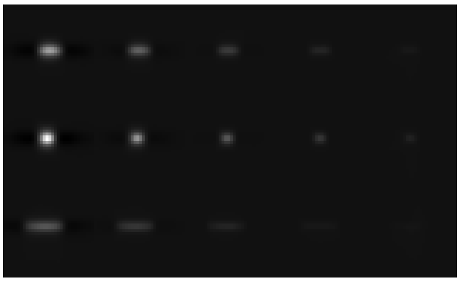

FIG. 4. Templates of the five target diameters for an MLTR$_{\mathrm{pr}}$ reconstruction, showing the focus plane in the middle together with the planes above (top) and below (bottom).

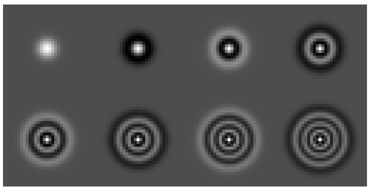

FIG. 5. Laguerre-Gauss channels with $\sigma = 187\ \mu$m, at the same scale as figures 3 and 4.

a fully simulated dataset from a previous study[38]. This dataset consisted of clusters of spherical microcalcifications between 100 μm and 200 μm set in random locations in a background of filtered white noise. Projections for these data were simulated as described above for the templates, after which Poisson noise was added and the projections were reconstructed using precursors of the FBP and MLTR$_{\mathrm{pr}}$ methods[28,39] in this paper. From this evaluation, the first eight LG channels with width σ set to 187 μm (2.2 pixels of 85 μm were selected because that setting resulted in the highest free-search ROC (FROC) area under the curve (AUC) determined by the weighted JAFROC method[40].

The efficient LG channels were selected over the anthropomorphic difference-of-Gaussian or Gabor channels[6,8,41], because the search step included in the model observer (see section II C 3) to account for the uncertain target locations in the phantom, will decrease the initial performance. Therefore it seemed more prudent to start from the higher scoring LG channels, where this drop in performance could be tuned to match human performance, rather than start from the anthropomorphic channels, where this further drop in performance could end up lower than the actual human performance.

The selected settings create the channels that are shown in figure 5, which is at the same scale as figures 3 and 4 so that the relative sizes of the channels and targets can be compared.

Since the targets all have diameters of 250 μm or less, and reconstructions have a typical plane separation of 1 mm we chose to use a single-plane observer on the in-focus reconstructed plane rather than a multi-plane observer. With this implementation we also avoid problems that would be caused by the different appearance of the out-of-plane artifacts of the calcifications in the different reconstruction types. This can be seen in the top and bottom rows of figure 3 and 4: the center of the out-of-plane artifact is shifted slightly compared to the central plane target with the direction of the shift depending on the location of the target in the phantom. Additionally the out-of-plane artifacts look very different in both reconstruction techniques. These two observations are the reason why the same channels, centered at the same location in the planes above and below the focus plane, cannot be used.

2. The 4-AFC Experiment

With these signal templates and channels, we can apply a channelized Hotelling observer (CHO) to the individual calcifications in each group. This is however not the approach adopted by the human observers, who examined the image for the entire group of five calcifications, rather than each calcification individually. In order to perform the same 4-AFC experiment as the human observers, the model observer has to calculate a single likelihood ratio for the presence of five calcifications c_i at locations ℓ_i with $i \in \{1, 2, 3, 4, 5\}$ for each presented ROI. This means calculating $p(L_1, L_2, L_3, L_4, L_5|c_i = 0)$ and $p(L_1, L_2, L_3, L_4, L_5|c_i=1)$, with c_i=0 shorthand for c_1=c_2=c_3=c_4=c_5=0 and L_i the 8-element channel output of image location ℓ_i, calculated as follows:

$$L_{ik} = \sum_j C_{jk}\ell_{ij} \tag{3}$$

with k indexing the eight selected channels C, and j indexing the 32×32 voxels in each channel and in the region around location ℓ_i to which they are applied.

For simplicity, we assume these five locations are independent. This is sufficiently accurate for the data noise because after reconstruction, these correlations are mostly oriented perpendicular to the planes, and the noise correlations within the plane are very small and of very short range, because these locations do not share any projection lines and are essentially reconstructed from different detector pixels. This approximation is less accurate for the anatomical noise which is created the PMMA spheres in the phantom which have diameters between 1.6mm and 16mm, and thus will create anatomical noise correlations between locations in that range. Because there are either five calcifications present or none, we can say:

$$\begin{aligned} & p(L_1, L_2, L_3, L_4, L_5|c_i = 1) && (4)\\ & \quad = p(L_1|c_i = 1) \cdot p(L_2|c_i = 1) \ldots \cdot p(L_5|c_i = 1) && (5)\\ & \quad = p(L_1|c_1 = 1) \cdot p(L_2|c_2 = 1) \ldots \cdot p(L_5|c_5 = 1), && (6) \end{aligned}$$

which means we can add the log-likelihood ratios of the five locations to obtain the log-likelihood ratio $q(A)$ for region of interest A.

$$\begin{aligned} q(A) &= \ln \frac{p(A|\bar{A}_1)}{p(A|\bar{A}_0)} && (7)\\ &= \ln \frac{p(L_1, L_2, L_3, L_4, L_5|c_i = 1)}{p(L_1, L_2, L_3, L_4, L_5|c_i = 0)} && (8)\\ &= \ln \frac{p(L_1|c_1 = 1)}{p(L_1|c_1 = 0)} + \ldots + \ln \frac{p(L_5|c_5 = 1)}{p(L_5|c_5 = 0)}. && (9) \end{aligned}$$

Each individual log-likelihood ratio q was then calculated as follows[42]:

$$q(L) = \left(\bar{L}_1 - \bar{L}_0\right)' C_A^{-1} L - \frac{1}{2}\left(\bar{L}_1' C_A^{-1} \bar{L}_1 - \bar{L}_0' C_A^{-1} \bar{L}_0\right), \quad (10)$$

with L the channel output of location ℓ in ROI A for which the score is being calculated, $\bar{L}_1$ the template of signal present, $\bar{L}_0$ the template for signal absent, and C_A^{-1} the 8×8 inverse covariance of the channel output calculated specifically in the evaluated ROI A. This covariance matrix was itself calculated by applying the selected channels to approximately 10 000 partially overlapping regions of 32×32 voxels, each one shifted by 8 pixels in-plane from the previous region and with regions extracted from every other plane in the selected ROI (with a size of 236×236×30 voxels).

3. *Calcification Group Geometry*

In ideal circumstances, when the exact location of each calcification is known, scores for those five locations could just be added to get the score of each ROI. However, in reality the exact location of each calcification in the phantom was not known because the targets in this prototype phantom were positioned by hand. Therefore the target regions for the 4-AFC study were extracted by using their relative position to the location of the largest calcification group, which was clearly visible in all images.

To account for inaccuracies due to deviations from the expected target geometry described in section II A, the model observer performs a search through five planes centered around the expected position of the calcifications in each ROI. The central target is searched for in a disk with a diameter of 32 pixels (2.72 mm). The search regions for the four calcifications on the corner positions were then determined relative to this variable position of the central calcification. The peripheral microcalcifications are allowed to be within a disc with diameter of 24 pixels (2.04 mm) centered at their expected locations 43 pixels (3.66 mm) to the left or right and 43 pixels to top or bottom from the central calcification, as shown in figure 6. All five calcifications also need to be contained within two adjacent planes at most. The final score for each image stack is now the maximum score that falls within these geometric constraints.

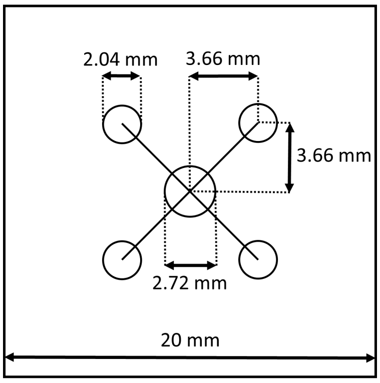

FIG. 6. Calcification group geometry.

4. *Reference Data*

Before applying the model observer to its intended task of optimizing a smoothing prior as described in the next section, we compared it to a small set of reference data to make sure the model observer was working as intended. This dataset consisted of the three FBP reconstructions of the low, AEC, and high dose level phantom acquisitions, which had been evaluated by five human observers[25], and one additional MLTR$_{\text{pr}}$ reconstruction with $\beta_Q = 2 \cdot 10^4$ and $\beta_{TV} = 2$ of the AEC dose level acquisition which was evaluated by a different group of five observers.

An initial comparison between the results from the 4-AFC evaluations performed by the model observer and those of the human observers found that using a single, global covariance matrix to describe the correlations for all ROI reconstructed with the same method resulted in a large mismatch. This problem was solved by determining an individual, local covariance matrix for each evaluated ROI, as in equation (10). The reference data were then used to set the geometric constraints described in section II C 3. Even though no internal noise was included in the model observer, these constraints can actually be seen as performing the same role as internal noise, since both can be tuned such that the model observer matches human observer results. In this instance, allowing more flexibility in the constraints, i.e. allowing larger search areas for the model observer, as in figure 6, will lower the performance of the model observer, because it increases the chance of finding high scoring noise structures in the background images.

Figure 7 shows the 4-AFC scores and psychometric curves for the human observer and the final implementation of the model observer. Because it is relatively hard to compare the results from both observers, a direct comparison of the 20 detected fractions (5 diameters for 4 reconstructions) from the 4-AFC evaluations is shown in figure 8, and the comparison of the fitted detection threshold diameters ϕ_{tr} is shown in figure 9.

Figure 8 shows that the detected fraction of the the smallest microcalcification group (90–100 μm) in the

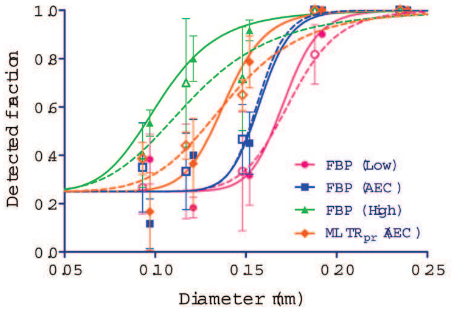

FIG. 7. 4-AFC results and fitted psychometric curves for human (hollow symbols & dashed lines) and model (full symbols & lines) observers. The symbols are slightly shifted for better visibility.

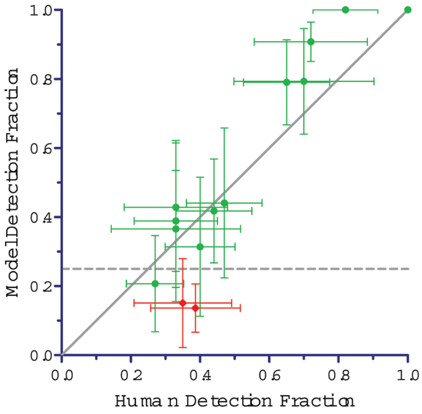

FIG. 8. Detection fractions from 20 4-AFC experiments evaluated by human and model observers, with 7 overlapping points at (1,1). The scores of the 90–100 μm targets for AEC dose level FBP and MLTR$_{\text{pr}}$ are shown in red. The error bars represent 95% confidence intervals.

FBP and MLTR$_{\text{pr}}$ reconstructions of the AEC dose acquisition are underestimated by the model observer, with one of the points even scoring significantly below the guessing level of 25%. Because the reference dataset was too small to examine this behavior further, the performance of the smallest target size (90–100 μm) was examined in more detail in the results of the application of the model observer on the smoothing prior optimization task.

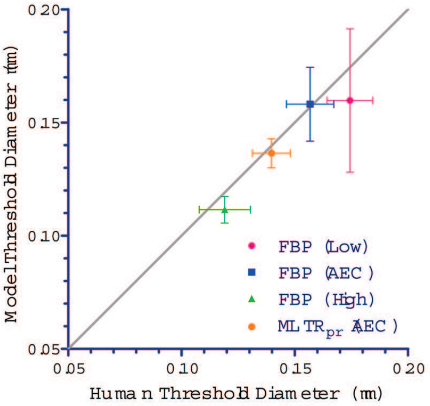

FIG. 9. The fitted detection threshold diameter ϕ_{tr} for human and model observers. The error bars represent 95% confidence intervals.

D. Smoothing Prior Optimization & Model Observer Validation

We applied the model observer presented in section II C to the optimization of the smoothing prior weights β_Q and β_{TV} in the MLTR$_{\text{pr}}$ reconstruction of the AEC dose level phantom measurements. For this task, we examine the quadratic smoothing prior for strength β_Q between $2 \cdot 10^3$ and $5 \cdot 10^5$, the total variation with ℓ_1 norm for strength β_{TV} between 1 and 50, and the combined prior with β_Q between $2{\cdot}10^3$ and $2{\cdot}10^5$ for $\beta_{TV} = 2$, and β_Q between $2 \cdot 10^4$ and $2 \cdot 10^5$ for $\beta_{TV} \in \{4, 6, 8\}$. With these ranges of the prior weight, the reconstructed images vary from noisier to smoother than the images used in clinical practice.

After finishing the optimization study, nine prior settings were selected for evaluation by human observers in order to validate the results obtained by the model observer. The 4-AFC experiments for the selected reconstructions were performed by five human observers. Spearman's ρ, which does not make assumptions on the variable distributions, is used to calculate the correlations between human and model observer detection fractions and fitted threshold diameters.

III. RESULTS

A. Smoothing Prior Optimization

Figure 10 shows the fitted threshold diameter to the 4-AFC experiment results for the quadratic, total variation, and combined priors. The smallest threshold diameter of 120 μm was reached for the total variation prior with $\beta_{TV} = 35$. Threshold diameters only decrease slowly with increasing prior strength for the quadratic

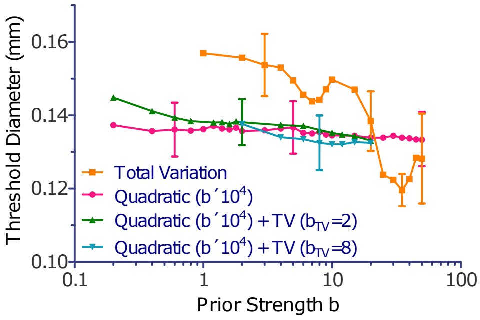

FIG. 10. Fitted threshold diameter from model observer results for the quadratic, total variation, and combined priors, with 95% confidence intervals for the cases selected for human reading.

Prior Type	Prior Strength (β_Q, β_{TV})
Quadratic	$(6 \cdot 10^3, 0)$; $(5 \cdot 10^4, 0)$; $(5 \cdot 10^5, 0)$
Combined	$(2 \cdot 10^4, 2)$; $(8 \cdot 10^4, 8)$
Total Variation	$(0, 3)$; $(0, 20)$; $(0, 35)$; $(0, 50)$

TABLE I. Prior settings selected for verification by human readers.

and combined smoothing priors, with threshold diameters between 133 μm and 138 μm for $\beta_Q > 10^4$. Results of the combined prior with $\beta_{TV} = 4$ and $\beta_{TV} = 6$ are not shown because they overlap with the results of the quadratic and combined priors.

B. Model Observer Validation

The nine prior settings that were selected for validation by human observers are listed in table I. Figure 11 shows a calcification group with 180–200 μm targets for each of these priors to illustrate the wide range of settings that was chosen.

Human and model observer detection scores for these reconstruction settings are shown in figure 12, with the two outliers shown in red: the scores of the 112–125 μm target for $\beta_Q = 5 \cdot 10^5$ and the 140–160 μm target for $\beta_{TV} = 50$, i.e. the highest values for β for the quadratic and total variation priors respectively. The scores of the 90–100 μm targets are not included, and examined separately in section III C. Correlation coefficients (Spearman ρ) and corresponding 95% confidence intervals between human and model observers for these scores were 0.917 [0.844–0.956] and 0.928 [0.862–0.963] with and without the outliers respectively and these are significant in both instances (p<0.001).

Figures 13 and 14 show the fitted detection threshold diameter ϕ_{tr} for human and model observers, with the same two outliers ($\beta_Q = 5 \cdot 10^5$ and $\beta_{TV} = 50$) as in figure 12. Correlation coefficients (Spearman ρ) between human and model observers for these thresholds were 0.857 (p=0.024) without the outliers, and 0.550 (p=0.133) with outliers included. There were too few points to calculate a reliable confidence interval.

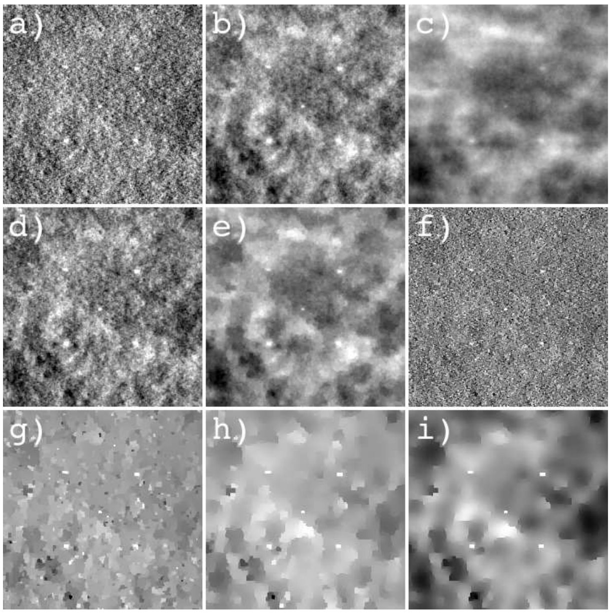

FIG. 11. Reconstruction of the 180–200 μm targets for: a) $\beta_Q = 6 \cdot 10^3$, b) $\beta_Q = 5 \cdot 10^4$, c) $\beta_Q = 5 \cdot 10^5$, d) $\beta_Q = 2 \cdot 10^4$ and $\beta_{TV} = 2$, e) $\beta_Q = 8 \cdot 10^4$ and $\beta_{TV} = 8$, f) $\beta_{TV} = 3$, g) $\beta_{TV} = 20$, h) $\beta_{TV} = 35$, and i) $\beta_{TV} = 50$.

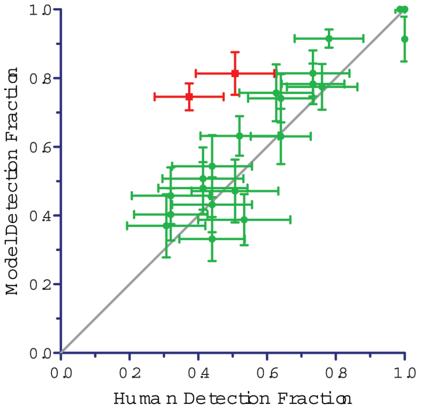

FIG. 12. Model observer detection fraction as a function of the human detection fraction for all target diameters except 90–100 μm. Two outliers (112–125 μm for $\beta_Q = 5 \cdot 10^5$ and 140–160 μm for $\beta_{TV} = 50$) are shown in red. The error bars represent 95% confidence intervals.

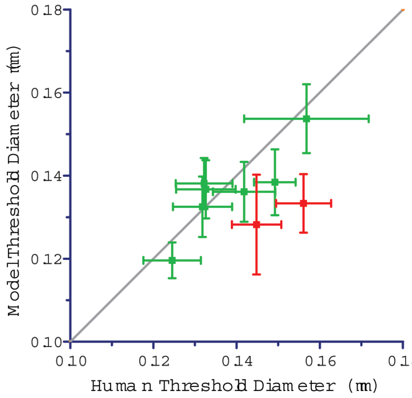

FIG. 13. The fitted detection threshold diameter ϕ_{tr} for human and model observers. Two outliers ($\beta_Q = 5 \cdot 10^5$ and $\beta_{TV} = 50$) are shown in red. The error bars represent 95% confidence intervals.

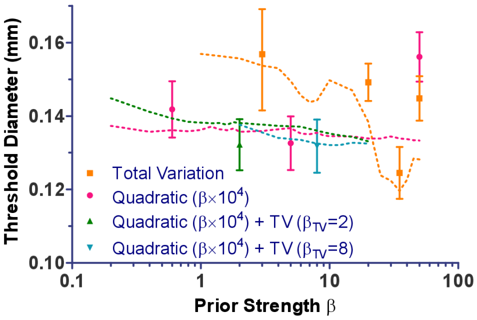

FIG. 14. Fitted threshold diameter from human observer results for the quadratic, total variation, and combined priors, with 95% confidence intervals. The dashed lines are the corresponding model observer results from figure 10.

C. Evaluation of the 90–100 μm Target

Scatter plots of the detection fraction of the nine prior settings listed in table I are shown in figure 15 for human and model observers and for the smallest target size. Our assumption for this target is that it is too small to be seen in DBT reconstructions, which was verified through visual inspection with graphical aids pointing to the correct locations. Therefore both human and model observers are expected to result in a detection fraction consistent with the theoretical guess rate of 0.25 in a 4-AFC experiment for this target size. However, we find that the average detection fraction of the smallest target is 0.36 for human readers, which is significantly ($p<0.001$) above the guess level. The model observer on the other hand scores 18% correct, significantly below the guess level ($p=0.004$).

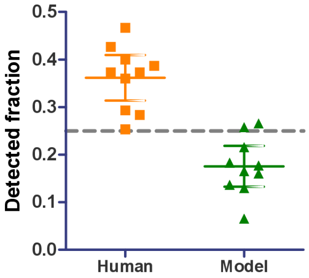

FIG. 15. Scatter plot with mean and 95% confidence interval of the human and model observer scores of the 90–100 μm targets for the reconstructions listed in table I.

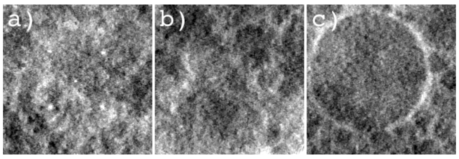

FIG. 16. Different phantom backgrounds: a) with target present, b) without target, but resembling target background, c) without target, not resembling target background.

Feedback from the observers indicated that the guess rate might have increased because some background types were correctly assumed to never have a target present. This can be seen in figure 16: 16a shows the signal in its typical background, and 16b and 16c show two types of normal backgrounds. Because the microcalcification targets in the phantom are mounted on a thin PMMA plate, the presence of the sphere in 16c means that there cannot be a microcalcification group at the same locations. Thus if the background in 16c appeared in the 4-AFC experiment, it would be rightly discarded as a candidate, and the observer would then choose between the three remaining images, effectively reducing the 4-AFC to 3-AFC and thus increasing the guessrate accordingly.

With this information we performed an additional experiment to check if the background types that were quickly discarded by the human observers might have a reversed effect on the model observer, resulting in the lower than expected performance. For this we selected one of the nine reconstructions ($\beta_Q = 2 \cdot 10^4$, $\beta_{TV} = 2$) and visually inspected all background images to see if there was a single large sphere visible in the central plane. Images where no sphere was clearly in focus (as in figure 16b) were included in group A, images where such a sphere was clearly visible (as in figure 16c) were included in group B.

The model observer log-likelihood ratios q for these groups are plotted in figure 17. The mean scores are 11.49 for the abnormal cases, 11.52 for background cases in group A, and 11.92 for background cases in group B. Restricting the normal cases in the 4-AFC experiment to cases from group A results in a detection fraction of 0.22 (95% CI:[0.11–0.33]), up from 0.14 (95% CI:[0.06–0.21]).

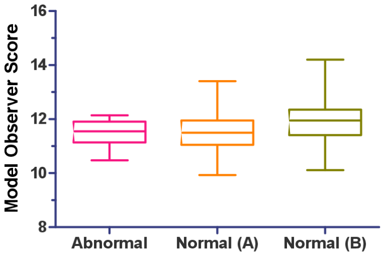

FIG. 17. Boxplot with whiskers from minimum to maximum of the model observer log-likelihood ratio of the smallest diameter targets, normal cases with background similar to target backgrounds, and normal cases with background not similar to target backgrounds.

The new score of 0.22 is not significantly different from 0.25 (p=0.550), while the old score of 0.14 was significantly lower (p=0.006).

IV. DISCUSSION

In this work we set out to accomplish two main goals: design a model observer that can predict human observer performance in a calcification detection task, and apply this model observer to maximize calcification detectability by optimizing the weights of the smoothing prior. This optimization was succesfull and maximal detectability was found for the total variation prior with sharp peaks in both detection threshold diameter and detected fraction at $\beta_{TV} = 35$. Figure 14 shows that this optimum was found for both human and model observers, and that the ranking of the different settings was the same for both observer types. Threshold diameters for the quadratic and combined quadratic and total variation prior only changed by a small amount over the examined range, with the threshold improving slowly with β_Q increasing two orders of magnitude, and images ranging from too noisy to too smooth (as in figures 11a and 11c).

The sharp minimum at $\beta_{TV} = 35$, rather than a more gradual change is probably a consequence of the nonlinear behavior of the total variation prior. Since the threshold diameter is most strongly influenced by the detectability of the 140–160 μm target, this value of beta results in the largest possible noise suppression that does not suppress the targets at this crucial diameter. Below this strength, irregular speckle noise (as seen in figure 11g) still confounds detectability, while at this specific prior strength, the background is reduced to a featureless piecewise-constant area, with few noise specks that could be confused with the calcification targets, while the targets themselves are not yet suppressed by the prior. Above this strength both noise and targets are suppressed, and thus the threshold diameter increases again.

Unfortunately the prior with the highest detection rate produces reconstructions that are not clinically useful since most small scale and low contrast information has been removed from the image, and thus detection and characterization of mass lesions in clinical images would be practically impossible. This means that the presented optimization process would not have been successful in choosing an appropriate setting for our smoothing prior if the goal had been to select a clinically acceptable reconstruction. Although the detection of small calcifications is a necessary condition for choosing a good smoothing prior, that addresses a weaker point of current tomosynthesis systems, it seems it is not a sufficient condition to guarantee overall optimal performance.

This means we would need either a replacement task or one or more additional tasks in the optimization process that would be more sensitive to oversmoothing than the current detection task. Possible options could be distinguishing between different orientations of a capital letter 'E' or 'C', i.e. the tumbling E or Landolt C used to measure visual acuity, or more applied to mammography, distinguishing between smooth and irregular microcalcifications. With these alternative or additional tasks, it seems more likely that the quadratic or mixed smoothing prior will provide the best compromise, since the detection performance remains stable over a large range of β, unlike the total variation prior, where a relatively small change in β would result in performance worse than the quadratic prior.

Even though the chosen detection task was not sufficient for the clinical optimization of a smoothing prior, the model observer itself managed to predict human observer results accurately over a large range of prior settings, except for the highest β_Q and β_{TV} values. Considering that these highest prior strengths are clearly not clinically relevant, we find that the model observer is a useful tool in the evaluation of the microcalcification detection task specified in this phantom, and can in fact be used instead of human observers.

Though care was taken to make sure the data used to design and train the model observer was independent from the data used in the evaluation of the smoothing prior, some aspects were shared between design and application. First, all data, both simulated and measured, were acquired using the geometry of the Siemens Mammomat Inspiration system, and as such additional validation will be needed when expanding the model observer to systems of different vendors. Second, there was a limited overlap in the datasets used in the design and application of the MO. In specific, the MLTR_{pr} reconstruction used in the design was one of the reconstructions considered in the task of optimizing the smoothing prior. However, due to the presence of three additional FBP reconstructions in the design stage, we think there was not much chance of overfitting the observer to the data for the application. And last, the MO was trained and evaluated specifically on a prototype version of the phantom described in section II A, and thus additional validation will be needed when switching to an updated design to alleviate the problems described in section III C, or application to a different phantom altogether. Any data obtained in these additional validations will also be useful to increase the statistical strength of the calculated Spearman correlation between human and model observers, which is currently limited by the low number of cases, especially for the correlation of the threshold diameters.

When evaluating the detectability for the smallest target diameter of 90–100μm we found that the conflicting performance between human and model observers was caused by scoring differences for two types of backgrounds. Human observers correctly considered the backgrounds that contained a large in focus sphere (as in figure 16c) unlikely to contain the target, while the model observer scored these as more likely to contain the target. This behavior results in a performance increase for the human observers, and a decrease for the model observer, both of which are probably present for all target diameters, but most clearly seen for the smallest target. The working hypothesis for the behavior of the model observer is that the large low structure-noise area in background regions with a central big PMMA sphere results in a covariance matrix with low values, which means that any false signal found there is multiplied by the inverse of this matrix, giving higher likelihood ratios, while in truth, the presence of this solid PMMA sphere excludes the possibility of a signal being present. This hypothesis is consistent with the results of the additional experiment, where removing the background type that caused this problem was excluded, resulted in the model observer performing as expected. Although this problem was only examined for the smallest target size, it is likely to have affected the performance at other diameters too, although to a lesser extent since once the targets become visible, less guessing is required in the 4-AFC experiment.

The cause of finding these two different background types can be seen in the design of the phantom, described in section II A. It is the thin PMMA plate to which the calcification targets are attached that prevents the larger PMMA spheres being present at the same location, and thus any instance where such sphere is present must mean no targets are present. In order to make sure that all backgrounds presented during the 4-AFC study are of the same type, it would be possible to separate them by visual inspection, but a better solution would be to adapt the phantom in such a way that all backgrounds resemble the ones in figures 16a and 16b. Since the calcifications need a supporting structure, the most straightforward solution is to extend the thin PMMA plate through the entire phantom, and thus avoiding any background regions in which large PMMA spheres are present at the same height as the calcification targets.

V. CONCLUSION

We successfully designed a model observer that was able to predict human performance over a large range of settings of the smoothing prior, except for the highest values of β which were outside the useful range for good image quality.

Based on the model observer results, we were able to choose a smoothing prior that optimizes the detection of microcalcifications in our iterative reconstruction and a human observer study confirmed that this prior yielded good results for the calcification detection task. Unfortunately, this 'optimal' prior applies strong smoothing, and tends to erase small scale and low contrast information in the reconstructed images, which makes it unsuitable in a clinical setting. Therefore we must also conclude that this detection task lacks the complexity or subtlety required to optimize for the task of reading clinical images.

When focusing on the combined quadratic and total variation prior, we find detectability only changes slightly for different amounts of smoothing, and thus an optimal strength for this prior can be selected based on other criteria without worrying about calcification detectability.

ACKNOWLEDGMENTS

The authors would like to thank Esmaeel Rangraz, Ahmadreza Rezaei, Georg Schramm, Tao Sun, and Anna Turco for participating in the 4-AFC experiment, and Siemens Healthcare for their financial support.

[*] johan.nuyts@uzleuven.be

[1] J. Lei, P. Yang, L. Zhang, Y. Wang, and K. Yang, "Diagnostic accuracy of digital breast tomosynthesis versus digital mammography for benign and malignant lesions in breasts: a meta-analysis," Eur. Radiol. **24**, 595–602 (2014).

[2] A. Tagliafico, G. Mariscotti, M. Durando, C. Stevanin, G. Tagliafico, L. Martino, B. Bignotti, M. Calabrese, and N. Houssami, "Characterisation of microcalcification clusters on 2D digital mammography (FFDM) and digital breast tomosynthesis (DBT): does DBT underestimate microcalcification clusters? results of a multicentre study," Eur. Radiol. **25**, 9–14 (2014).

[3] I. Sechopoulos, "A review of breast tomosynthesis. part I. the image acquisition process," Med. Phys. **40**, 014301 (2013).

[4] H.H. Barrett, J. Yao, J.P. Rolland, and K.J. Myers, "Model observers for assessment of image quality," Proceedings of the National Academy of Sciences of the United States of America **90**, 9758–9765 (1993).

[5] J. Beutel, H.L. Kundel, and R.L. Van Metter, *Handbook of Medical Imaging, Volume 1. Physics and Psychophysics* (SPIE Press, 2000).

[6] C.K. Abbey and H.H. Barrett, "Human- and model-observer performance in ramp-spectrum noise: effects of regularization and object variability," J. Opt. Soc. Am. A

18, 473–488 (2001).

[7] A.E. Burgess, "Visual perception studies and observer models in medical imaging," Semin. Nucl. Med. **41**, 419–436 (2011).

[8] X. He and S. Park, "Model observers in medical imaging research," Theranostics **3**, 774–786 (2013).

[9] A.S. Chawla, J.Y. Lo, J.A. Baker, and E. Samei, "Optimized image acquisition for breast tomosynthesis in projection and reconstruction space," Med. Phys. **36**, 4859–4869 (2009).

[10] S. Young, S. Park, S.K. Anderson, A. Badano, K.J. Myers, and P. Bakic, "Estimating breast tomosynthesis performance in detection tasks with variable-background phantoms," Proc. SPIE **7258**, 72580O (2009).

[11] S. Park, G.Z. Zhang, R. Zeng, and K.J. Myers, "Comparing observer models and feature selection methods for a task-based statistical assessment of digital breast tomsynthesis in reconstruction space," Proc. SPIE **9037**, 90370M (2014).

[12] D. Van de Sompel, M. Brady, and J. Boone, "Task-based performance analysis of FBP, SART and ML for digital breast tomosynthesis using signal CNR and channelised hotelling observers," Med. Image Anal. **15**, 53–70 (2010).

[13] R. Zeng, S. Park, P. Bakic, and K.J. Myers, "Evaluating the sensitivity of the optimization of acquisition geometry to the choice of reconstruction algorithm in digital breast tomosynthesis through a simulation study," Phys. Med. Biol. **60**, 1259–1288 (2015).

[14] H.C. Gifford, C.S. Didier, M. Das, and S.J. Glick, "Optimizing breast-tomosynthesis acquisition parameters with scanning model observers," Proc. SPIE **6917**, 69170S (2008).

[15] B.A. Lau, M. Das, and H.C. Gifford, "Towards visual-search model observers for mass detection in breast tomosynthesis," Proc. SPIE **8668**, 86680X (2013).

[16] I.S. Reiser and R.M. Nishikawa, "Task-based assessment of breast tomosynthesis: Effect of acquisition parameters and quantum noise," Med. Phys. **37**, 1591–1600 (2010).

[17] X. Wang, J.G. Mainprize, G. Wu, and M.J. Yaffe, "Task-Based evaluation of image quality of filtered back projection for breast tomosynthesis," in *Digital Mammography*, edited by J. Marti, A. Oliver, and R. Marti (Springer, Berlin, 2010), pp. 106–113.

[18] G.J. Gang, J., J.W. Stayman, D.J. Tward, W. Zbijewski,J.L. Prince, and J.H. Siewerdsen, "Analysis of fourier-domain task-based detectability index in tomosynthesis and cone-beam CT in relation to human observer performance," Med. Phys. **38**, 1754–1768 (2011).

[19] J.A. Baker and J.Y. Lo, "Breast tomosynthesis: state-of-the-art and review of the literature," Acad. Radiol. **18**, 1298–1310 (2011).

[20] M. Das and H.C. Gifford, "Comparison of model-observer and human-observer performance for breast tomosynthesis: effect of reconstruction and acquisition parameters," Proc. SPIE **7961**, 796118 (2011).

[21] M. Das, C. Connolly, S.J. Glick, and H.C. Gifford, "Effect of postreconstruction filter strength on microcalcification detection at different imaging doses in digital breast tomosynthesis: human and model observer studies," Proc. SPIE **8313**, 831321 (2012).

[22] Y.-H. Hu and W. Zhao, "The effect of angular dose distribution on the detection of microcalcifications in digital breast tomosynthesis," Med. Phys. **38**, 2455–2466, 2011.

[23] E.Y. Sidky, Y. Duchin, I. Reiser, C. Ullberg, and X. Pan, "Optimizing algorithm parameters based on a model observer detection task for image reconstruction in digital breast tomosynthesis," in *IEEE Nuclear Science Symposium Conference Record*, edited by D. Townsend (IEEE, Piscataway NJ, 2011), pp. 4230–4232.

[24] K. Michielsen and J. Nuyts, "Multigrid reconstruction with block-iterative updates for breast tomosynthesis," Med. Phys. **42**, 6537–6548 (2015).

[25] L. Cockmartin, N.W. Marshall, G. Zhang, K. Lemmens, E. Shaheen, and H. Bosmans, "Comparison of detection performance between 2D digital mammography and breast tomosynthesis using a structured physical phantom," presented at the Annual Meeting of the Radiological Society of North America (2015).

[26] T. Mertelmeier, J. Orman, W. Haerer, and M.K. Dudam, "Optimizing filtered backprojection reconstruction for a breast tomosynthesis prototype device," Proc. SPIE **6142**, 61420F (2006).

[27] J. Orman, T. Mertelmeier, and W. Haerer, "Adaptation of image quality using various filter setups in the filtered backprojection approach for digital breast tomosynthesis," in *Digital Mammography*, edited by S.M. Astley, M. Brady, C. Rose, and R. Zwiggelaar (Springer, Berlin, 2006), pp. 175–182.

[28] K. Michielsen, K. Van Slambrouck, A. Jerebko, and J. Nuyts, "Patchwork reconstruction with resolution modeling for digital breast tomosynthesis," Med. Phys. **40**, 031105 (2013).

[29] E. Mumcuoğlu, R.M. Leahy, and S.R. Cherry, "Bayesian reconstruction of PET images: methodology and performance analysis," Phys. Med. Biol. **41**, 1777–1807 (1996).

[30] A. Sawatzky, C. Brune, F. Wubbeling, T. Kosters, K. Schafers, and M. Burger, "Accurate EM-TV algorithm in PET with low SNR," in *IEEE Nuclear Science Symposium Conference Record*, edited by P. Selin (IEEE, Piscataway NJ, 2008), pp. 5133–5137.

[31] A. Beck and M. Teboulle, "Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems," IEEE Trans. Image Process. **18**, 2419–2434 (2009).

[32] F. Jäkel and F.A. Wichmann, "Spatial four-alternative forced-choice method is the preferred psychophysical method for naïve observers," J. Vis. **6**, 1307–1322 (2006).

[33] N. Karssemeijer and M. Thijssen, "Determination of contrast-detail curves of mammography systems by automated image analysis," in *Digital Mammography*, edited by K. Doi, M.L. Giger, R.M. Nishikawa, R.A. Schmidt (Elsevier, Amsterdam, 1996), pp. 155–160.

[34] J.G. Brankov, "Evaluation of the channelized Hotelling observer with an internal-noise model in a train-test paradigm for cardiac SPECT defect detection," Phys. Med. Biol. **58**, 7159–7182 (2013).

[35] L. Cockmartin, H.Bosmans, and N.W. Marshall, "Comparative power law analysis of structured breast phantom and patient images in digital mammography and breast tomosynthesis," Med. Phys. **40**, 081920 (2013).

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

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

[38] K. Michielsen, F. Zanca, N. Marshall, H. Bosmans, and J. Nuyts, "Two complementary model observers to evaluate reconstructions of simulated micro-calcifications in digital breast tomosynthesis," Proc. SPIE **8673**, 86730G (2013).

[39] J. Ludwig, T. Mertelmeier, H. Kunze, and Wo. Härer, "A novel approach for filtered backprojection in tomosynthesis based on filter kernels determined by iterative reconstruction techniques," in *Digital Mammography*, edited by E.A. Krupinski (Springer, Berlin, 2008), pp. 612–620.

[40] D. Chakraborty, "Analysis of Location Specific Observer Performance Data: Validated Extensions of the Jackknife Free-Response (JAFROC) Method," Acad. Radiol. **13**, 1187–1193 (2006).

[41] Y. Zhang, B.T. Pham, and M.P. Eckstein, "The effect of nonlinear human visual system components on performance of a channelized Hotelling observer in structured backgrounds," IEEE Trans. Med. Imaging **25**, 1348–1362 (2006).

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