Model and human observer reproducibility for detection of microcalcification clusters in digital breast tomosynthesis images of three- dimensionally structured test object

Dimitar Petrov
Nicholas Marshall
Kenneth Young
Guozhi Zhang
Hilde Bosmans

Model and human observer reproducibility for detection of microcalcification clusters in digital breast tomosynthesis images of three-dimensionally structured test object

Dimitar Petrov,^{a,*} Nicholas Marshall,^{a,b} Kenneth Young,^c Guozhi Zhang,^b and Hilde Bosmans^{a,b}

^aKU Leuven, Department of Medical Physics and Quality Assessment, Leuven, Belgium

^bUZ Leuven, Department of Radiology, Leuven, Belgium

^cRoyal Surrey County Hospital, NCCPM, Guildford, United Kingdom

Abstract. We compare the reproducibility of the human observers and a channelized Hotelling observer (CHO), when reading digital breast tomosynthesis (DBT) images of a physical phantom containing a breast simulating structured background and calcification clusters at three dose levels. The phantom is scanned 217 times on a Siemens Inspiration DBT system. Volumes of interest, with and without the calcification targets, are extracted and the human observers' percentage of correct (PC) scores is evaluated using a four-alternative forced choice method. A two-layer CHO is developed using the human observer results. The first layer consists of a localizing CHO that identifies the most conspicuous calcifications using two Laguerre–Gauss channels. Then a CHO with eight Gabor channels estimates the PC score for the calcification cluster. Observer reproducibility is estimated by bootstrapping, and the standard deviation (SD) is used as a figure of merit. The CHO closely approximated the human observer results for all the three dose levels with a correlation of >0.97 . For the larger calcification cluster sizes, both observers have similar reproducibility, whereas the CHO is more reproducible for the smaller calcifications, with a maximum of 5.5 SD against 13.1 SD for the human observers. The developed CHO is a good candidate for automated reading of the calcification clusters of the structured phantom, with better reproducibility than the human readers for small calcifications. © 2019 Society of Photo-Optical Instrumentation Engineers (SPIE) [DOI: [10.1117/1.JMI.6.1.015503](https://doi.org/10.1117/1.JMI.6.1.015503)]

Keywords: digital breast tomosynthesis; channelized Hotelling observer; reproducibility; calcifications; human; model; observer.

Paper 18169RR received Aug. 6, 2018; accepted for publication Mar. 5, 2019; published online Mar. 23, 2019.

1 Introduction and Purpose

Digital breast tomosynthesis (DBT) is an imaging technique that provides some three-dimensional (3-D) information on the breast from a series of projection images. This modality promises improved low-contrast lesion detectability compared to conventional two-dimensional (2-D) mammography in the presence of breast anatomical structure,^{1,2} due to the reduction in overlapping fibroglandular tissues. The ability of DBT systems to detect microcalcifications (MC), however, remains a topic of interest, as some studies show conflicting results.^{3–6} The gold standard for such image quality estimation is a clinical trial, where the human observer's efficiency to detect calcification clusters in anatomical images is evaluated. It is very difficult to find an alternative to the large-scale clinical studies, where many real-world aspects of clinical studies are evaluated in the measured performance (e.g., sensitivity and specificity) of a given modality for some imaged population. There is rapid progress being made in the field of virtual studies,^{7,8} and these promising methods are likely to augment or help to orient practical clinical studies at some point in the future. If a practical estimate of imaging performance for a modality in the field is required at a fixed time point (e.g., for QC purposes), then we are left with test-object-based evaluations. A work by Warren et al.⁹ has shown that well-defined stimuli

within a test object can correlate with the human observer's performance in reading clinical images, suggesting that there is value in carefully controlled evaluations using test objects.

Quality control guidelines in mammography^{10–12} often specify the minimum detectability level that must be reached for some specified targets if the system in question is to be suitable for clinical use in a breast-screening context. In practice, these stimuli are often generated by test objects containing circular or speck-like objects positioned within a homogeneous background, such as the CDMAM¹³ or the ACR¹¹ phantoms. Regarding the CDMAM test object, human readers initially performed the reading to assess the object detectability contrast threshold or threshold gold thickness, for the specific exposure condition used clinically, on a given imaging system. Over time, human reading has been replaced by computerized or automated reading.^{14,15}

It is important that the evaluation process is reproducible, i.e., producing a stable detection threshold, such that the system performance can be tracked reliably for QC purposes over time. The variability of threshold contrast estimated using the CDMAM phantom has been compared for human and automatic readouts,¹⁶ showing that some of the automated readout/processing methods gave reliable threshold contrasts over certain target diameters. However, for DBT there is still no consensus on how to estimate the minimally required detectability level, as its benefits regarding 3-D information of the breast

*Address all correspondence to Dimitar Petrov, E-mail: dimitar.petrov@uzleuven.be

and the reduced overlapping glandular tissue cannot be expressed in the same terms as in 2-D mammography (e.g., CDMAM). Testing the detectability of specific targets within test objects against specified (regulatory) detectability levels also has implications regarding the reproducibility of the measurements made, either by human readers or by automated read-out software. It is important that the uncertainty on the measured thresholds is known so that meaningful decisions regarding system performance can be made.

The recent work by Cockmartin et al.¹⁷ introduced a 3-D structured test phantom suitable for evaluating DBT imaging, with scoring again performed by human readers. This phantom is the starting point for this study, with the aim of investigating human reader reproducibility and testing computerized scoring alternatives. Recent studies show that model observers (MOs) can be used for such scoring tasks in DBT, showing good correlation with human results.^{18–20} The calcification clusters within the phantom are of special interest, given that the detection of small objects such as MC is likely to be limited by system quantum noise²¹ even in breast simulating structured backgrounds. This study first describes the scoring of the calcification clusters by human observers to derive detectability score and the associated reproducibility. A MO is then developed using the human reader output and the reproducibility of both observers are compared. The MO presented here is a development of an earlier channelized Hotelling observer (CHO)²² that modeled human observer performance for the task of detecting calcifications using internal noise. This approach of adjusting CHO performance to match that of human readers; however, is prone to overfitting and will not likely generalize well to new data.²³ In the previous work,^{24,25} we had found that the internal noise approach could be replaced by careful channel tuning. Here, the CHO channel parameters are selected to include sufficient information from the images such that human observer performance is matched. This method is shown to approximate human performance for six different models of DBT system, covering a wide range of scanning parameters, without the need for further tuning. Nevertheless, using this approach to detect calcification clusters compromised the localization of the separate calcification particles. To overcome these problems, this study describes a two-stage CHO that utilizes separate localization and classification stages.²⁶ The concept of a two-stage CHO was introduced as visual search observers (VSOs) in prior work by the group of Gifford et al.,^{27–30} with an initial localization stage in which a MO found candidates potentially belonging to a given calcification cluster. This was followed by a candidate deletion step and then the second stage applied another MO at the derived locations to estimate the cluster detectability. The work of Das and Gifford²⁷ using a VSO did not require that all MCs in the cluster be detected and this requirement will also be employed in the study described here. There are some differences from the visual search-based methods,^{27–30} notably that our implementation focuses on channel tuning at the classification stage. Furthermore, the initial localization stage implements a CHO that utilizes prior information about the target size, shape, and location to detect the most probable locations of the calcification particles forming the cluster and in the classification stage—a tuned CHO is applied at the derived locations to estimate detectability. This way, the sensitive area of the algorithm is always aligned to the same locations within the images, regardless of the classification scores and tuning settings, thus ensuring that the detectability

changes reflect only the influence of the tuning parameters used and not misalignments. Finally, in this study we extend the reproducibility evaluation over three dose levels, as the noise level may have some influence on both human and CHO reproducibility.

2 Methods

2.1 Phantom Properties and Image Acquisition

The images used in this study were generated using a 3-D structured phantom,¹⁷ made of an acrylic semicircular container filled with equal volumes of acrylic spheres of six different diameters (15.9, 12.7, 9.52, 6.35, 3.18, and 1.58 mm) with the remaining volume filled with water (Fig. 1). Five calcification clusters were inserted in the phantom, with calcification particles of a given group lying, respectively, within diameter ranges of 90 to 100 μm (mean 95 μm), 112 to 125 μm (mean 119 μm), 140 to 160 μm (mean 150 μm), 180 to 200 μm (mean 190 μm), and 224 to 250 μm (mean 237 μm) (Fig. 1). Each of the five groups consisted of five individual calcifications placed in an X shape within the phantom, with one calcification located at the center and four at the extreme positions.

The phantom was scanned a total of 217 times on a Siemens Inspiration tomosynthesis system (Siemens AG Healthcare, Erlangen, Germany) in order to generate a sufficiently large image database to assess reproducibility. The first acquisition was acquired under automatic exposure control (AEC) to establish typical tube voltage, anode/filter (A/F), and tube current-time product [milliamperes-second (mAs)] for the phantom on this system. The following 216 DBT scans were made at three dose levels: 72 acquisitions with fixed parameters set as close as possible to the AEC values (30 kV, W/Rh A/F, and 200 mAs), 72 acquisitions at a low-dose setting (30 kV, W/Rh A/F, and 112 mAs), and 72 acquisitions at a high-dose setting (30 kV, W/Rh A/F, and 400 mAs), all reconstructed with EMPIRE reconstruction algorithm.³¹ The phantom was shaken after each acquisition to produce different realizations of the background. The images of each dose group were split into

Fig. 1 (a) A photograph of the phantom. (b) On the left, a DBT slice of the phantom and, on the right, the positions where the regions of interest (ROIs) are cropped from. Green squares depict the positions of the signal-absent ROIs and red depicts the signal-present ROIs. (c) Example images of the cropped ROIs with five calcifications forming a cluster (four in the corners of a square and one in the center).

six sets, each containing 12 scans, of which the first 5 (60 DBT scans) were used for human and MO reading, whereas the sixth set was used for training the MO template. Volumes of interest (VOIs) with size of $20 \times 20 \times 30 \text{ mm}^3$ ($236 \times 236 \times 30$ voxels) were cropped for the observer studies [Fig. 1(b)]. From each scan, 15 partially overlapping signal-absent VOIs were extracted from the signal-free volume within the phantom, along with five unique signal-present VOIs centered at each of the five calcification clusters, corresponding to the five cluster sizes. Thus, the 12 scans generated image reading sets that contained 180 signal-absent VOIs and 60 signal-present VOIs.

2.2 Human Observer Study

Six medical physicists participated in a four-alternative forced choice (4AFC) image reading study. Three signal-absent VOIs along with one signal-present VOI were simultaneously visualized and the observer task was to indicate the signal-present image. If the score was correct, then this was defined as a “hit.” After each reading session, the percentage of correctly detected signals (PC) was calculated by the equation

$$PC = \frac{\text{Number of hits}}{\text{Number of trials in a session}} \cdot 100\%. \quad (1)$$

The 4AFC study was conducted using an in-house software tool³² and the images were read on a diagnostic monitor (5MP Barco MDNG-6121) at an ambient light level of 3 lux. The observers were instructed to score the sessions at a consistent distance from the screen equal to around 50 cm. No magnification and window-level adjustments were allowed; only scrolling through the VOI was permitted (in fact, this was necessary). No time constraints were imposed for the reading sessions and the whole study was performed within a time period of 1 week. It is common for the observers to read a training dataset before participating in such reading studies; however, at the time of our study, our observers were participating in other reading studies with the same phantom, thus a training set was not required. To study the reproducibility, standard deviation (SD) was estimated for each of the six observers by bootstrapping 100 times: 12 random 4AFC scores were drawn without replacement from the 60 scores that were available by combining the five datasets at a given dose level and calcification size.

For each condition, the smallest SD among the human observers was used to quantify the human reproducibility, with the assumption that the smallest SD in the responses represented the most precise human reader in our group with the least external distraction. This way the greater part of the variability would be caused by the reading task, and the effect of observer strain or distraction will be diminished. The mean PCs from the five image datasets with error bars equal to the bootstrapped SD were plotted along with the observer scores for each session and dose level.

2.3 Model Observer

Signal detection theory³³ utilizes the concept of a decision variable, which is an internal response generated by the observer that is related to the presence of a stimulus (e.g., a calcification here) and this forms the basis of the decision made by the observer for each AFC trial. In this study a CHO was implemented, with a decision variable calculated using the following equation:³⁴

$$\lambda(v) = w_{\text{CHO}}^t v, \quad v = U^t g; \quad w_{\text{CHO}} = s^t K_v^{-1}. \quad (2)$$

Here, v is the channel output vector, which is a product of the transposed channel matrix U and the image vector g . In the same expression, w_{CHO} is the observer template, formed from the expected signal output vector s and the inverse of the covariance matrix K_v^{-1} . The covariance matrix is estimated from signal-absent images and the expected signal is approximated using a Gaussian function with full width at half maximum equal to the average diameter of individual calcifications within a calcification group. Thus, only signal-absent training images are needed for our application. To minimize the CHO bias, the training of the observer template is performed on a separate image set of 12 DBT acquisitions that are not used for reading. From the training image dataset, 180 signal-absent VOIs are extracted, where for each VOI the five central slices are taken for the covariance matrix formation, making 900 signal-absent training images.

The Gaussian approximation of the expected signal was needed, due to the insufficient signal-present images for training. This way the sensitive region of the CHO included only a single Gaussian function, corresponding to a single calcification particle. As a consequence, the detection of the calcification cluster within a VOI requires the exact location of the calcification particles within the cluster of 5. Owing to the physical manufacturing tolerances, the exact location of the targets within the physical phantom is not known and thus the CHO algorithm was divided into two stages—localization followed by classification.²⁶

2.3.1 Localization channelized Hotelling observer

The localization stage was inspired from the work of Michielsen et al.¹⁸ in which the calcification clusters were detected via scanning five volumes within the VOI and the localization and the classification were done in a single stage. In the current study, the central calcification was detected by scanning an area in the center of the ROI with the size of $1.36 \times 1.36 \text{ mm}^2$ (Fig. 2), i.e., 16×16 pixels for the Siemens system. This produced a decision variable map, where the position of the highest observer response was then taken as the most probable calcification position. From this location, the expected positions of the four extreme calcifications were located 5.2 mm from the center calcification. Each of the four corner calcifications was detected by the same scanning area that was used to find the

Fig. 2 Visualization of the searching mechanism. On the left: A decision variable map showing the observer performance within the central 16×16 pixels of a signal-present image. The position of the maximum observer performance is taken as the real location of the center calcification. On the right, from this position the extreme calcifications are detected within areas with the same size.

center calcification ($1.36 \times 1.36 \text{ mm}^2$). This was repeated in the four adjacent planes (two above and two below the central plane) and, for each calcification, the most probable position within the VOI was stored for the classification stage. This process was performed for both the signal-present and the signal-absent images.

The localization CHO was implemented using two Laguerre–Gauss (LG)³⁵ channels:

$$C_j(r) = \frac{\sqrt{2}}{a_u} \exp\left(\frac{-\pi r^2}{a_u^2}\right) L_j\left(\frac{2\pi r^2}{a_u^2}\right);$$

$$L_j(x) = \sum_{k=0}^j (-1)^k \binom{j}{k} \frac{x^k}{k!}, \quad (3)$$

where $a_u = \sqrt{2\pi}\sigma_u$ is the SD of the Gaussian and σ_u is used as a tuning factor. The rank j of the Laguerre function was varied from 0 to 1 forming the two channels.

The number of channels and the tuning factor were varied in order to pick the channels that maximize the localization abilities of the algorithm. This was performed by visually inspecting the decision variable map (Fig. 2), with the final channel set selected as the one that generated a narrow peak (ideally at only one pixel within the scanned area).

2.3.2 Classification channelized Hotelling observer

At the classification stage, another CHO was used to compute five observer responses for each signal present and signal absent image, later on used to estimate the final observer performance. Given that the CHO performance was to be compared with human reader results, a CHO with anthropomorphic channel properties was used.³⁶ Eight Gabor channels³⁷ were selected:

$$C_{i,j,k}(x,y) = e^{-\frac{4 \ln(2)}{w_i^2} \frac{x^2+y^2}{2}} \cos[2\pi f(x \cos \theta_j + y \sin \theta_j) + \phi_k], \quad (4)$$

where $w_i = \frac{t_w}{(e^i+2)^{p_{xs}}}$, $f = \frac{t_f}{w_i}$, $\theta_j = \frac{\pi j}{t_t}$, and $\phi_k = 45 \text{ k}$.

The parameters w_i , f , θ_j , and ϕ_k controlled the width, (spatial) frequency, orientation, and phase of the channels, respectively. In this Gabor channel implementation, the frequency was set as a function of channel width, resulting in a channel set with just one sinusoidal maximum at the point selected by x and y spatial locations in the equation. In generating the eight channels here, one phase ($i = 0$), two orientations ($j = [0,1]$), and four frequencies ($k = [0,3]$) were selected. Our implementation, therefore, had three tuning factors, t_w , t_f , and t_t , influencing channel width, frequency, and orientation, respectively. In this study, before the comparison of the human and MOs, the Gabor parameters were tuned to give the closest approximation of human observer performance. The tuning was performed by evaluating the CHO performance at the AEC dose level with different combinations of the three tuning parameters, where the $t_w \in [0.9, 20.0]$, $t_f \in [3, 160]$, $t_t \in [5 \text{ deg}, 175 \text{ deg}]$. Candidates for channel parameter sets were selected using three criteria between CHO and human observer scores: the Pearson's correlation (r), linear regression slope (α), and the mean error (ME), with ideal scores of $r = 1$, $\alpha = 1$ and $\text{ME} = 0$. Initial candidate parameter values had to fulfill the following criteria: correlation $r > 0.98$; the linear slope $\alpha \in [0.98, 1.02]$; and the absolute $\text{ME} < 2 \text{ PC}$, where these

criteria were combined with equal weight. These sets of candidate parameters were then tested on the low- and high-dose levels and the best performing parameter set for all dose levels was chosen, i.e., the set with the highest correlation, slope closest to 1.0, and smallest ME combined with equal weight for all dose levels. To avoid overfitting, the tuning was performed on a separate image dataset, different from the images used for reading in the reproducibility study.

To calculate the CHO response for the signal-present and signal-absent VOIs, the Gabor channels were generated for each VOI at the most probable calcification locations estimated from the localization stage. This gave five observer responses corresponding to the five calcification particles forming the cluster. As the human observers internally produced a single observer response for the VOI and scored the calcification cluster as a single target, the CHO should mimic this operation.¹⁸ If we assume the five CHO responses to be independent, the overall image observer response can be calculated as a summation of the observer responses from the individual calcifications:

$$\lambda(v|H_i)_k = \sum_{j=1}^3 \lambda_{jk}, \quad (5)$$

where $(v|H_i)$, $i \in \{0,1\}$ refers to the image with signal-present or signal-absent ground truth, k refers to the five calcification groups sizes, and j refers to the five calcification positions within the cluster. In the equation $j \in \{1,2,3\}$, as the human observers reported, the detection of the three calcifications in the expected positions is enough to score the image positively. In other words, the MO decision variable for the calcification cluster is a sum of the three highest decision variables calculated for the individual calcifications.

To facilitate comparison with the human results, the CHO scores were expressed as the PC in a 4AFC test. This was done by comparing a decision variable estimated from a signal-present image with three decision variables from three signal-absent images. The outcome was scored a “hit” when the decision variable for the signal-present image was the greatest among the four. The PC was calculated using the same formula used for the human observer scores. Finally, CHO reproducibility was assessed by calculating the SD from bootstrapping the results 100 times, where each sample consisted of 12 4AFC trials. The mean and SD estimates were plotted along with the CHO results for each reading session.

3 Results

3.1 Channelized Hotelling Observer Tuning

The tuning parameter for the localization CHO (σ_u) was set at $\sigma_u = 2$; this way, the first stage CHO produces a decision variable map consisting of a single peak (Fig. 2). Figure 3 shows the tuning ability of the Gabor channels used in the classification stage with visualization of the channels. Each graph represents the effect in the CHO scores when tuning only a single parameter, while keeping the others constant. It can be seen that the channel tuning parameters clearly influence CHO performance, with a different effect for each calcification diameter. Table 1 contains the tuning parameters chosen for this study, with the criteria scores shown in Table 2. Figure 4 shows the model against human observer results, where the CHO results are produced using the aforementioned tuning parameters.

Fig. 3 (a) CHO performance expressed as PC plotted against Gabor channel tuning parameter values. The title of each graph indicates the parameter varied, while the other tuning parameters were kept constant at the values indicated by the two vertical dashed lines in the other graphs. The tuning parameters specified by the dashed lines describe the channels used for the reproducibility study. (b) The three images below each graph visualize the exact channels produced by setting the parameter at the lowest, the median, and the highest values within the specific studied range. The visualizations are formed from eight rectangles (60 × 236 px) cropped from the eight used channels.

Table 1 CHO channels tuning parameters.

First layer CHO (localization)	Second layer CHO (classification)		
	θ_t (deg)	t_f	t_w
σ_u			
2	50	57	4.1

Table 2 MO against human observer results criteria.

Dose level	Pearson's correlation (r)	Linear regression slope (a)	Mean error (ME) (PC)
Low	0.994	1.13	4.44
AEC	0.991	1.05	0.69
High	0.977	1.13	5.89

3.2 Observer Reproducibility

Scatter plots of the model and human observer reproducibility results are presented in Fig. 5. For each mean calcification diameter, the graphs show PC values for each of the five image datasets (12 scans in each dataset) in gray and the mean PC with

the SD estimates are indicated with black lines/bars. Table 3 gives SD values for each dose level, observer, and mean calcification diameter.

The reproducibility results showed that the CHO was approximately a factor of 2 times more reproducible than human reading for the smaller calcification sizes. For the human observer, the lowest reproducibility was found for the 119 μm calcification cluster in the low-dose level dataset with $\text{SD}_{\text{Human}, 119 \mu\text{m}} = 13.1$ PC and the most reproducible were the two largest calcification clusters, which were 100% correctly read by the humans with no variation, and thus SD equal to 0.0. For the MO, the AEC dose level 95- μm calcification cluster had the lowest reproducibility, with a $\text{SD}_{\text{Model}, 95 \mu\text{m}} = 5.5$ PC and the most reproducible were the two largest clusters for low- and high-dose levels, which had SD equal to 0.0 PC. In the case of the AEC dose level, the CHO in fact exhibited some variability in the PC for the two largest calcification clusters, with SD equal to 2.2 and 0.9, compared to the value of 0.0 seen for the human observers.

4 Discussion

The aim of this study was to assess the human and MO reproducibility when reading calcification clusters in DBT images containing a structured background, generated using a physical test object. Three dose levels were studied to check how observer reproducibility varies as the system (x-ray) noise

Fig. 4 MO PC results plotted against the human observer PC results. (a) Low-dose level, (b) AEC dose level, and (c) high-dose level.

Fig. 5 Observer results for (a) low-dose level, (b) AEC dose level, and (c) high-dose level. For each dose level, the top graph shows the human observer reproducibility results and the bottom graph shows the MO reproducibility results. On the graphs, the observer scores for each of the five sessions; the mean and the SD estimates are depicted by black and by gray, respectively.

Table 3 SD of the observer results measured in PC.

Dose level	Observer	95 μm	119 μm	150 μm	190 μm	237 μm
Low	SD _{CHO}	5.5	4.8	5.2	1.1	0.0
	SD _{Human}	12.9	13.1	9.6	0.0	0.0
AEC	SD _{CHO}	4.7	5.4	3.2	2.2	0.9
	SD _{Human}	11.2	12.4	7.0	0.0	0.0
High	SD _{CHO}	4.2	5.4	1.2	0.0	0.0
	SD _{Human}	11.9	3.6	3.7	0.0	0.0

level is varied between images. The initial work on this topic²² did not have the localization stage described in this study, but instead used an internal noise algorithm to adjust the CHO performance so that the human observer PC results were matched. Using the internal noise step, we found algorithm evaluation results of $r = 0.997$; $a = 0.92$, and $ME = -0.57$ PC at the AEC dose level only.²² Using internal noise, however, compromised the localization abilities of that algorithm, which led to a poor performance when applied to the two additional dose levels studied in the present study. Furthermore, the guessing rate of the previous CHO implementation exceeded 40%. These arguments were the trigger to further improve the algorithm. The CHO detection was split into two stages: an initial localization step followed by a classification step. The localization stage used two LG channels and did not include any observer performance like area under the curve (AUC), detectability index (d'), or PC, as such an estimate could compromise the localization performance of the algorithm. The most probable calcification locations forming the calcification cluster were estimated only by the first stage MO responses calculated from five volumes

within the VOIs. To avoid bias, the localization stage was performed on both signal-present and signal-absent images. The second stage, that of classification, then generated eight (tuned) Gabor channels at the most probable calcification positions for the signal-present and the signal-absent images. The single decision variable produced for each image was then used in the 4AFC evaluation. Using this technique, the guessing rate in terms of the PC of the modified CHO method was 23 ± 4 PC, which is close to the expected value of 25 for a 4AFC method.

Both the VSOs introduced by Gifford et al.^{27–30} and our method use two consecutive MOs. However, the impetus behind our implementation is solely to allow channel tuning. In general, the VSO algorithms find as many candidates as possible within the image in the first step, followed by a candidate deletion, where a candidate is dismissed, if its performance is under a certain threshold or their location is considered mistaken. In contrast, the two-stage CHO presented here is spatially restricted in its first stage to five areas known *a-priori*, where in each of them a single calcification particle is expected. In fact, once the localization channel parameters are set, we expect that the algorithm will always pick the same one position for each of the five particles. This way, i.e., without candidate deletion, the classification stage can utilize a tuned channel set to approximate human performance. In the case of the VSO method, it seems that the opposite is observed in that the performance matching mostly occurs in the localization and candidate deletion stage.

The results in Fig. 3 clearly show that the tuning of the CHO can be done by selecting the channel parameters instead of the addition of internal noise, which is prone to overfitting.²³ The graph on the left side (Fig. 3), showing the observer performance against the orientation of the Gabor filters, notes an improvement in performance at around 90-deg spatial orientation for the calcification particles with a size of 150 μm. This suggests that the imaged calcifications might not be ideally

circular and a more realistic expected signal estimate could be used in the future instead of a Gaussian blob. The graph in the middle (Fig. 3), showing the observer performance against the frequency tuning parameter, shows a drastic drop in performance for $t_f > 99$, caused due to a too high frequency to include the full calcification size, making the signal-present and signal-absent decision variables in a closer value range. The same effect is reached, when the channel width is too small, depicted on the right graph (Fig. 3) for $t_w < 2.5$. On the same graph, by making the channels much wider than the object size (i.e., increasing the t_w value), the performance is gradually lowered, as an effect of the inclusion of too much background and decreasing the strength of the signal in the decision variable.

The results in Fig. 4 show that the parameters used for the second (classification) stage can be tuned so that CHO performance predicts the human reader results and the single parameter set applies at all three dose (noise) levels for this system. The two-stage CHO gives a good approximation to human reader results with $r > 0.98$, $a < 1.1$, and $ME < 5.9$ PC, without the use of internal noise. An increase in microcalcification detection rate with increasing dose is clearly visible in the data. After tuning and training, the MO achieves the same scores as human readers and therefore the impact of dose on the scores is similar for both MO and human reading. In an earlier study using human scoring of the same phantom,¹⁷ the phantom has already been used to evaluate a series of DBT systems operated in 2-D and DBT mode, using human reading. A combined analysis using 2-D and DBT data is required before any initial estimates of “optimal” dose levels can be given for patient studies, for the different of DBT systems. Detailed dose optimization studies are outside the scope of the present study. It must be stressed that the human reader data are used to train the CHO and therefore it gives similar PC results as the human readers by design. Nevertheless, the reading results shown in Fig. 4 represent the CHO performance on separate image set and differ within error bars from the reading results from the tuning process.

The reproducibility data show that the MO is more reproducible than human observers for the smaller calcification sizes. This is an encouraging result for the application of the test object for quality assurance purposes using MOs. For example, the increased reproducibility of the CHO would lead to a statistically significant difference between AEC and low-dose levels for the 119 and 150 μm calcification clusters with, respectively, $p < 0.001$ and $p < 0.001$ for the CHO, as opposed to $p = 0.068$ and $p = 0.088$ for the human observers. For the larger calcification particle sizes (190 and 237 μm) at AEC dose level, despite the good agreement between the CHO and the human observer scores, the CHO reproducibility is notably poorer, with values of 2.2 and 0.9 PC for the CHO compared to 0.0 and 0.0 PC found for human readers. This could indicate that some further tuning might be needed or that a larger image dataset is required.

Young et al.¹⁶ compared the human reader reproducibility with that of four different methods of processing contrast-detail curves for the CDMAM mammography test object, which also presents a 4AFC detection tasks to the observers. The four processing methods were applied to the output of the CDCOM module automatic detection algorithm developed by Karssemeijer and Thijssen.¹⁴ The study found lower contrast thresholds for the CDCOM-based methods, yet higher reproducibility for the human observers than for the automatic scoring

methods. In contrast, the data in our study show better reproducibility for the CHO by a factor of ~ 2 . Higher reproducibility for human reading of CDMAM over the CDCOM method could result from the task being disk detection in a homogeneous background containing only x-ray quantum mottle. If a test object with targets set in a homogeneous background (e.g., the ACR digital mammography phantom) is used for constancy testing of DBT systems, then the in-focus plane can be established for a given system and this plane, which is used for constancy testing, is always scored by the same reader. This scenario is likely to give low uncertainty on the threshold diameters obtained; however, it will not assess the detection performance of the system for targets in structured background (which is likely to be related to aspects such as the angular range and reconstruction algorithm used).³⁸ One could argue that a simpler phantom could be used for constancy testing, where the aim is to find changes in system performance.

This CHO implementation has certain limitations. Our reliance on tuning to match the detection performance of a certain group of human observers could be considered a limitation in this study. For anthropomorphic MOs, one could argue that the channel functions should be rather closely linked to the task and thus tuning away from the expected (task) function to alter MO performance is somewhat arbitrary and a move away from a pure task-based observer. Future work could explore the implementation of a full VSO; however, we will be constrained by the test object. In fact, our CHO is specifically developed for the evaluation of the 3-D structured test object, in the place of human observers in the Medical Physics department. We are tuning to a specific group of observers (obviously we could tune to a different set of observers); however, we want a tool that produces similar output as the observers at our center. The localization stage requires a prior knowledge of the approximate position of the calcification particles forming the cluster. This is needed in order to avoid scanning large area within the phantom as this would require significant amount of computational time. In our case, the exact locations of the targets rarely exceed 5 to 6 pixels from the expected locations; thus, 8 pixels from the center is a preventative measure. Although the CHO algorithm can be used on images derived with different acquisition parameters (kV, image reconstruction, etc.) or on different DBT systems, the channel parameters for both stages will need to be validated to achieve a good agreement with human observers. Nevertheless, our previous study²⁴ on detectability of the nonspiculated simulating masses in the same phantom background showed that channels generated in the real space perform well in different conditions like different DBT systems, dose levels, and reconstruction algorithms. The channel parameters described in this study might be used as a starting point for a finer tuning to specific applications.

5 Conclusions

This study has demonstrated that the CHO is approximately a factor of 2 times more reproducible for smaller calcification diameters than the human observers in this study. This makes the combination of structured phantom and CHO reading a promising approach for quality assurance and related activities.

Disclosures

No conflict of interest is declared by the authors.

Acknowledgments

This work was a part of Optimam II project of Cancer Research UK (Grant No. C30682/A17321). We wish to thank our readers for their time and help and Frederik De Keyzer for his assistance with the statistical approach.

References

1. T. M. Svahn et al., "Breast tomosynthesis and digital mammography: a comparison of diagnostic accuracy," *Br. J. Radiol.* **85**(1019), e1074–e1082 (2012).
2. P. Skaane et al., "Prospective trial comparing full-field digital mammography (FFDM) versus combined FFDM and tomosynthesis in a population-based screening programme using independent double reading with arbitration," *Eur. Radiol.* **23**(8), 2061–2071 (2013).
3. Z. Li et al., "Comparison of microcalcification detectability in FFDM and DBT using a virtual clinical trial," *Proc. SPIE* **10577**, 105770D (2018).
4. M. L. Spangler et al., "Detection and classification of calcifications on digital breast tomosynthesis and 2D digital mammography: a comparison," *Am. J. Roentgenol.* **196**(2), 320–324 (2011).
5. A. Hadjipanteli et al., "Detection of microcalcification clusters in 2D mammography and digital breast tomosynthesis and the relation to the standard method of measuring image quality," in *XIV Med. Conf. on Med. and Biol. Eng. and Comp.*, E. Kyriacou, S. Christofides, and C. Pattichis, Eds., Springer, Cham, pp. 217–221 (2016).
6. D. Kopans et al., "Calcifications in the breast and digital breast tomosynthesis," *Breast J.* **17**(6), 638–644 (2011).
7. P. R. Bakic et al., "Realistic simulation of breast tissue microstructure in software anthropomorphic phantoms," *Lect. Notes Comput. Sci.* **8539**, 348–355 (2014).
8. U.S. Food and Drug Administration, "VICTRE: virtual imaging clinical trials for regulatory evaluation," 2018, <https://www.fda.gov/MedicalDevices/ScienceandResearch/ResearchPrograms/ucm477418.htm>.
9. L. M. Warren et al., "Effect of image quality on calcification detection in digital mammography," *Med Phys.* **39**(6), 3202–3213 (2012).
10. R. E. van Engen et al., "A European protocol for technical quality control of breast tomosynthesis systems," *Lect. Notes in Comp. Sci.* **8539**, 452–459 (2014).
11. American College of Radiology, *Digital Mammography Quality Control Manual*, 2016, https://www.kansasradiationphysics.com/uploads/2/2/4/6/22462248/2016_digital_mammo_qc_manual__8.1.16.pdf.
12. U.S. Food and Drug Administration, *Mammography Quality Standards Act Regulations*, 6 February 2002, <https://www.fda.gov/Radiation-EmittingProducts/MammographyQualityStandardsActandProgram/Regulations/ucm110906.htm> (accessed 6 July 2018).
13. J. A. Thomas et al., "Contrast-detail phantom scoring methodology," *Med. Phys.* **32**(3), 807–814 (2005).
14. N. Karssemeijer and M. A. O. Thijssen, "Determination of contrast-detail curves of mammography systems by automated image analysis," in *Digital Mammography*, K. Doi et al., Eds., pp. 115–160, Elsevier, Amsterdam (1996). <https://pdfs.semanticscholar.org/4ca7/9902a77995baa090ff5c6c9a56673dccc48e5.pdf> (accessed 5 April 2018).
15. K. C. Young et al., "Evaluation of software for reading images of the CDMAM test object to assess digital mammography systems," *Proc. SPIE* **6913**, 69131C (2008).
16. K. C. Young et al., "Comparison of software and human observers in reading images of the CDMAM test object to assess digital mammography systems," *Proc. SPIE* **6142**, 614206 (2006).
17. L. Cockmartin et al., "Design and application of a structured phantom for detection performance comparison between breast tomosynthesis and digital mammography," *Phys. Med. Biol.* **62**(3), 758–780 (2017).
18. K. Michielsen et al., "Design of a model observer to evaluate calcification detectability in breast tomosynthesis and application to smoothing prior optimization," *Med. Phys.* **43**(12), 6577–6587 (2016).
19. L. Cockmartin, N. W. Marshall, and H. Bosmans, "Comparison of SNDR, NPWE model and human observer results for spherical densities and microcalcifications in real patient backgrounds for 2D digital mammography and breast tomosynthesis," *Lect. Notes in Comp. Sci.* **8539**, 134–141 (2014).
20. Y.-H. Hu and W. Zhao, "The effect of angular dose distribution on the detection of microcalcifications in digital breast tomosynthesis," *Med. Phys.* **38**(5), 2455–2466 (2011).
21. F. O. Bochud et al., "Estimation of the noisy component of anatomical backgrounds," *Med. Phys.* **26**(7), 1365–1370 (1999).
22. D. Petrov et al., "Model and human observer reproducibility for detecting microcalcifications in digital breast tomosynthesis images," *Proc. SPIE* **10577**, 105770B (2018).
23. 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).
24. D. Petrov et al., "Real space channelization for generic DBT system image quality evaluation with channelized Hotelling observer," *Proc. SPIE* **10136**, 101360N (2017).
25. D. Petrov et al., "Systematic approach to a channelized Hotelling model observer implementation for a physical phantom containing mass-like lesions: application to digital breast tomosynthesis," *Phys. Med.* **58**, 8–20 (2019).
26. H. L. Kundel et al., "Holistic component of image perception in mammogram interpretation: gaze-tracking study," *Radiology* **242**(2), 396–402 (2007).
27. 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).
28. H. C. Gifford, Z. Liang, and M. Das, "Visual-search observers for assessing tomographic x-ray image quality," *Med. Phys.* **43**(3), 1563–1575 (2016).
29. H. C. Gifford et al., "Visual-search models for location-known detection tasks," *Proc. SPIE* **10136**, 1013612 (2017).
30. Z. Karbaschi and H. C. Gifford, "Assessing CT acquisition parameters with visual-search model observers," *J. Med. Imaging* **5**(02), 0225501 (2018).
31. S. Abdurahman et al., "Optimizing high resolution reconstruction in digital breast tomosynthesis using filtered back projection," *Lect. Notes Comput. Sci.* **8539**, 520–527 (2014).
32. G. Zhang, L. Cockmartin, and H. Bosmans, "A four-alternative forced choice (4AFC) software for observer performance evaluation in radiology," *Proc. SPIE* **9787**, 97871E (2016).
33. D. M. Green and J. A. Swets, *Signal Detection Theory and Psychophysics*, Wiley & Sons, Inc., New York (1966).
34. H. H. Barrett et al., "Model observers for assessment of image quality," *Proc. Natl. Acad. Sci. U. S. A.* **90**(21), 9758–9765 (1993).
35. B. D. Gallas and H. H. Barrett, "Validating the use of channels to estimate the ideal linear observer," *J. Opt. Soc. Am. A* **20**(9), 1725–1738 (2003).
36. H. H. Barrett et al., "Task-based measures of image quality and their relation to radiation dose and patient risk," *Phys. Med. Biol.* **60**(2), R1–R75 (2015).
37. A. B. Watson, "Detection and recognition of simple spatial forms," NASA Technical Memorandum 84353, pp. 100–114 (1983). <http://hdl.handle.net/2060/19830015903>.
38. I. Sechopoulos, "A review of breast tomosynthesis. Part II. Image reconstruction, processing and analysis, and advanced applications," *Med. Phys.* **40**(1), 014302 (2013).

Biographies of the authors are not available.