Anthropomorphic model observer performance in three-dimensional detection task for low-contrast computed tomography

Alexandre Ba
Miguel P. Eckstein
Damien Racine
Julien G. Ott
Francis Verdun
Sabine Kobbe-Schmidt
François O. Bochud

Anthropomorphic model observer performance in three-dimensional detection task for low-contrast computed tomography

Alexandre Ba,^a Miguel P. Eckstein,^b Damien Racine,^a Julien G. Ott,^a Francis Verdun,^a Sabine Kobbe-Schmidt,^c and François O. Bochud^{a,*}

^aLausanne University Hospital, Institute of Radiation Physics, Lausanne, Switzerland

^bUniversity of California Santa Barbara, Department of Psychological and Brain Sciences, Santa Barbara, California 93106, United States

^cLausanne University Hospital, Department of Radiology, Lausanne, Switzerland

Abstract. X-ray medical imaging is increasingly becoming three-dimensional (3-D). The dose to the population and its management are of special concern in computed tomography (CT). Task-based methods with model observers to assess the dose-image quality trade-off are promising tools, but they still need to be validated for real volumetric images. The purpose of the present work is to evaluate anthropomorphic model observers in 3-D detection tasks for low-contrast CT images. We scanned a low-contrast phantom containing four types of signals at three dose levels and used two reconstruction algorithms. We implemented a multislice model observer based on the channelized Hotelling observer (msCHO) with anthropomorphic channels and investigated different internal noise methods. We found a good correlation for all tested model observers. These results suggest that the msCHO can be used as a relevant task-based method to evaluate low-contrast detection for CT and optimize scan protocols to lower dose in an efficient way. © 2015 Society of Photo-Optical Instrumentation Engineers (SPIE) [DOI: 10.1117/1.JMI.3.1.011009]

Keywords: three-dimensional analysis technique; computed tomography; image quality; model observer; observer performance evaluation; volumetric imaging.

Paper 15163SSPR received Aug. 17, 2015; accepted for publication Nov. 23, 2015; published online Dec. 29, 2015.

1 Introduction

For several decades, x-ray medical imaging has been evolving into 3-D, in particular with the advent of computed tomography (CT). The expansion to volumetric body data has resulted in more accurate diagnosis^{1,2} but inevitably induced a higher radiation dose to the patient compared to equivalent two-dimensional (2-D) techniques. Furthermore, the number of CT devices and examinations performed on patients is increasing, raising concerns about the mean population dose and necessitating increased management of the radiological process.^{3,4} However, the radiological process can be optimized only with a simultaneous consideration of image quality. To lower the radiation dose while maintaining useful image quality in CT, manufacturers have proposed new techniques such as iterative reconstruction (IR).^{5–7} It is a promising tool, but its contributions to image quality need to be objectively evaluated.

Image quality should be directly related to diagnostic function.⁸ In this context, image quality assessment should include a diagnostic task such as detection, localization, or classification of a pathology, but also an observer, a representative set of images, and a relevant metric.⁹ When defining image quality at the task level, the observer can either be a human, such as a radiologist, or a mathematical model that attempts to predict human observer performance. Unlike psychophysical evaluations with humans, model observers require less time for image quality evaluations with respect to many system parameters.

Model observers have been widely used with 2-D medical imaging modalities and validated for different backgrounds and signals.^{10–15} One successful class of model is the channelized Hotelling observer (CHO),^{16–18} which starts by processing the image through channels and then applies the best linear strategy (under Gaussian assumptions) on the channel outputs in order to perform the task. Depending on the characteristics of the chosen channels, the performance of the CHO can be made to approach the ideal observer or to mimic the human observer.^{11,14,19,20}

The CHO has been used to evaluate CT with 2-D detection or location tasks^{21–24} of circular objects mimicking lesions in homogeneous backgrounds. However, CT data are 3-D and radiologists use the whole volume available to make their diagnosis. 2-D detection tasks do not take into account the effect of additional information provided by adjacent image slices. Thus, a more thorough evaluation of task-based image quality assessment for CT should include the volumetric aspect of the data and 3-D model observers. Furthermore, the dimensionality (2-D versus 3-D) of the task can lead to different correlations between model and human observers,^{25,26} emphasizing the importance of volumetric data when using a numerical observer for optimizing protocols.

Frameworks have already been proposed to evaluate 3-D imaging modalities with 3-D detection tasks and model observers, particularly in nuclear medicine.^{26–29} More recently, Platiša et al.²⁵ described a general design for a model observer adapted to 3-D data sets based on the CHO. This research investigated three general mathematical expressions for model observers

*Address all correspondence to: François O. Bochud, E-mail: Francois.Bochud@chuv.ch

dedicated to the detection of a signal. The single-slice channelized Hotelling observer (ssCHO) considers a 2-D representative image slice extracted from the 3-D stack. This model is equivalent to the conventional use of the CHO for a 2-D detection task. The multislice CHO (msCHO) computes 2-D CHO responses in several 2-D slices and combines them in order to quantify an overall response. The volumetric CHO (vCHO) directly computes a CHO response on the 3-D image as if it were a common 2-D image and takes into account correlations across slices, unlike the ssCHO. The authors conclude that the vCHO with channels attempting to mimic the ideal observer (e.g., Laguerre–Gauss channels) could be useful for hardware optimization. Alternatively, they proposed the use of the msCHO with appropriate anthropomorphic channels, which are close to the response properties of the human visual system,³⁰ as a surrogate for human observers in image quality assessment.

Previous studies have used simulated images and 3-D signals to evaluate image quality,^{25,31} but, to our knowledge, the msCHO has not yet been tested on real CT images. The purpose of the present work was to investigate 3-D detection tasks with msCHO in low-contrast real CT images and to evaluate their correlation with human observers. We implemented two strategies of the anthropomorphic msCHO.²⁵ They first used slice-specific templates (msCHOa). The model used as many templates as slices present in the 3-D image and applied them to each corresponding slice. The second strategy (msCHOb) used a single template calculated from the central slice of the 3-D image and applied it to every slice in the data set.

Human observer perceptual decisions are also limited by inherent variability in the sensory processing of information, which leads to variability in their decisions. Many of the anthropomorphic model observers take this source of internal variability into account^{17,32,33} by incorporating internal noise into the models. Two methods are commonly used. The first method modifies the scalar decision variable coming from the

integration of channel responses for each considered location and trial by adding an independent random variable. The second method is applied to each individual channel's response. Because the channel internal noise is statistically independent, it modifies the covariance of the channel responses and influences the effective linear template (i.e., the linear combination of channels) of the msCHO model. Both internal noise methods for CHO have been studied for a single-slice 2-D model observer,³⁴ showing that the channel-based internal noise method better predicts human performance when the channels are anthropomorphic. To our knowledge, there has not been an evaluation of the internal noise method for multislice model observers.

In this study, we evaluated four configurations of the msCHO: (1) msCHOa and (2) msCHOb with Gabor channels and (3) msCHOa and (4) msCHOb with dense difference of Gaussian (DDOG) channels. These models were implemented to predict human observer performance in 3-D medical CT images reconstructed using one of the two different algorithms: IR and filtered backprojection (FBP). We considered the detection of low-contrast spherical signal known exactly (SKE) at three dose levels.

2 Materials and Methods

2.1 Image Data Sets

A chest phantom (QRM, Mochrendorf, Germany, 30 cm×20 cm cross-section) was used to simulate the attenuation of an average-sized adult [Fig. 1(a)]. The phantom contained a core module with four spherical signals of 6- and 8-mm diameters to mimic lesions [Fig. 1(b)]. An additional module with no signal was also imaged in order to obtain signal-free image samples. The signals had low-contrast CT numbers of 10 and 20 HU relative to the background at 120 kVp. Image samples were obtained by scanning the phantom 50 different times on a

Fig. 1 (a) Chest phantom with modulus containing low-contrast spherical signals of 10 and 20 HU and 6 and 8 mm. An additional modulus without signals was scanned in order to produce signal-absent ROI (not represented). (b) Scan of the phantom at a high dose level (600 mAs, 120 kVp) and a contrast of 20 HU in order to make the signals visible.

64-slice CT scanner (HD750, GE Healthcare) in helical mode with constant x-ray tube current of 40, 90, and 150 mAs. Corresponding radiation dose levels were 2.5, 5.7, and 9.5 mGy in terms of volume computed tomography dose index (CTDI_{vol}), respectively. The scanning parameters, 1-s rotation time, 0.98 pitch in helical mode, and 64×0.625 mm collimation, were the same as those used in the routine protocol for clinical practice at the Lausanne University Hospital. Acquisitions were reconstructed with the traditional FBP and also with a commercial implementation of a model-based IR algorithm⁵ available on the scanner (VEO, GE Healthcare). Reconstructed slice thickness was 0.625 mm with a 0.625 mm interval. In total, we investigated 24 experimental conditions (2 signal sizes $\times$ 2 signal contrasts $\times$ 3 dose levels $\times$ 2 reconstruction algorithms). Note that the combination of 20 HU contrast and 8 mm signal sizes is not reported in Sec. 3 because these were used for training.

For each condition, we cropped 32×32 -pixel regions of interest (ROIs) around the signal locations. We did the same with the images obtained with the signal-absent modules. To generate multislice ROIs, this operation was reproduced with 17 consecutive slices centered on the signal central slice. We extracted 200 ROIs with signal-present and 200 ROIs with signal-absent.

Due to the positioning of the signals in the phantom (Fig. 1), ROIs were limited to 32×32 pixels (1 pixel = 0.58 mm); therefore, the original images were magnified from 32×32 pixels to 100×100 pixels (1 pixel = 0.58 mm) using a bilinear interpolation method.

2.2 Human Psychophysical Study and Performance Estimation

The image data sets were used to conduct four-alternative forced-choice experiments (4AFC) where the signal position is exactly known with three observers familiar with low-contrast detection. The participants were all male, aged 25, 26, and 30 at the time of the study, and all had experience with 4AFC experiments because they had participated in previous detection

studies with the same phantom. The observer study consisted of four sessions corresponding to the four signal contrast and size combinations (8 mm/20 HU, 6 mm/20 HU, 8 mm/10 HU, 6 mm/10 HU). The first session was for training only, with 8 mm/20 HU signals at each dose level. This training session consisted of 150 trials for each reconstruction algorithm and was performed to familiarize the observer with the signal and background characteristics specific to FBP and IR reconstructed images. Feedback was given after each trial, allowing the observer to learn from their strategies. The following three sessions were for testing, with 6 mm/20 HU, 8 mm/10 HU, and 6 mm/10 HU signals, which consisted of 200 trials per experimental condition corresponding to the number of signal-present samples. For a given 4AFC trial, three signal-absent images without replacement were chosen but were replaced after the trial. As we produced 200 signal-absent images for each dose and reconstruction algorithm, each 4AFC trial can be considered a unique set, and correlation on observer performance should be minimal.

As in clinical practice, the observer could freely scroll throughout the stack of image slices. To reduce in-plane spatial uncertainties,^{35,36} horizontal cues were displayed on each slice and two additional cues were displayed vertically in the central slice (Fig. 2). All images were scaled to 8 bit with the same fixed window level and width chosen from the mean maximum and mean minimum intensities over all images for each reconstruction algorithm. Observers could not zoom or pan the images, and all sessions were performed in a dark room with low ambient light (<10 lux). The images were displayed on a grayscale monitor (RadiForce MX210, EIZO) calibrated to the DICOM Grayscale Display Function and TG18 standards.³⁷

To evaluate human observer performance, the percentage of correct decision (P_c) was computed from all trials for each observer. As an estimation of the average observer performance, we used multireader multicase variance analysis method for binary data.³⁸ For each image category, each observer evaluated the same number of cases. For each trial, signal-present and signal-absent cases were chosen randomly. Therefore, from

Fig. 2 Screenshot of the experimental interface used in the human observer study (a) when an off-center slice is displayed and (b) when a central slice is displayed.

a statistical point of view, each case was considered independently of the others.

2.3 Model Observers

2.3.1 Multislice channelized Hotelling model observers

Most model observers in medical imaging are linear and can be reduced to the dot product of a linear template and the image data. The CHO is defined by its template, which is derived from the inverse of the covariance matrix of the image's background and the ensemble means under the signal-present and -absent classes. The CHO includes the preprocessing of the image by a set of P channels. Each channel applied to the image $\mathbf{g}$ produces one scalar output. These P channel outputs are then combined linearly in order to maximize the task performance. In practice, the linear process consists of prewhitening the channel outputs and applying a filter that matches the signal of interest as seen through the channels. Detailed explanations of the mathematical foundations of the CHO can be found in Refs. 9, 16, 18, and 39.

In order to take into account the multiple slices contained in the 3-D CT images, we followed the publication of Platiša et al.²⁵ The model observer decision variable was obtained in two steps. First, a conventional 2-D CHO was applied to each of the N slices composing the volumetric image. This provides N decision variables associated with each slice. Second, these decision variables are processed by an HO in order to compute a single decision variable. Formally, this means that each slice $\mathbf{g}_n$ ($n = 1, \dots, N$) of the volumetric image $\mathbf{g}$ is first processed by a set of 2-D channels $\mathbf{u}_p$ ($p = 1, \dots, P$), resulting in a vector of channelized slices $\mathbf{v} = [\mathbf{v}_1, \dots, \mathbf{v}_n]$, with $\mathbf{v}_n = \mathbf{U}^T \mathbf{g}_n$ and where $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_n]$.

We considered two implementations of the msCHO.²⁵ The first one is the msCHOa, which takes into account the second-order statistics of each slice individually. The second implementation is the msCHOb, which defines its strategy according to the slice of the image where the signal is centered and applies it indiscriminately to each individual slice.

Thus, the msCHOa computes independent templates $\mathbf{w}_n = [\mathbf{w}_1, \dots, \mathbf{w}_n]$, where $\mathbf{w}_n = \mathbf{K}_{v_n}^{-1}[\langle \mathbf{v}_{n,2} \rangle - \langle \mathbf{v}_{n,1} \rangle]$ and $\mathbf{K}_{v_n}$ is the covariance matrix describing variance and covariance

of channel responses for each slice present in the multislice data set. Subscripts 1 and 2 refer to the signal-present and signal-absent hypotheses, respectively. In a similar manner, the msCHOb computes a single template $\mathbf{w}$ from the central slice of the multislice data set. Then $\mathbf{v}$ is used to compute a vector of decision variables $\lambda_{\text{stack}} = [\lambda_{\text{slice},1}, \dots, \lambda_{\text{slice},n}]$, where $\lambda_{\text{slice},n} = \mathbf{w}_n^T \mathbf{v}_n$ for msCHOa or $\lambda_{\text{slice},n} = \mathbf{w}^T \mathbf{v}_n$ for msCHOb.

The second step integrates the resulting vector of decision variables λ_{stack} with the HO to compute a single scalar decision variable $\lambda_{\text{final}} = \mathbf{w}_{\text{HO}}^T \lambda_{\text{stack}}$, where $\mathbf{w}_{\text{HO}} = \mathbf{K}_{\text{slice}}^{-1}[\langle \lambda_{\text{slice},2} \rangle - \langle \lambda_{\text{slice},1} \rangle]$. The covariance matrix $\mathbf{K}_{\text{slice}}$ describes the variance and covariance of the training vector decision variables. To estimate the HO template with respect to the model, the template is applied on images from the training set, resulting in training vectors of decision variables, which are used to estimate $\mathbf{K}_{\text{slice}}$.

The msCHO framework for both msCHOa and msCHOb is described in Fig. 3.

2.3.2 Channels

Gabor⁴⁰ and DDOG¹⁷ channels were chosen because of their ability to mimic the visual human process and successfully predict human observer performance.

The Gabor channels function is defined by

$$V(x,y) = \exp \left[-4 \ln(2) \frac{(x^2 + y^2)}{w_s^2} \right] \cos[2\pi f(x \cos \theta + y \sin \theta) + \beta], \quad (1)$$

where f is the spatial frequency, θ is the orientation, w_s is the width, and β is the phase.

The Gabor channels parameters proposed in this study used five orientations, seven frequencies, and one phase, resulting in 35 channels. The channel parameters used are $\beta = 0$ and $w_s = 0.56/f$ for a bandwidth of one octave. Orientations were chosen with values ranging from 18 deg to 305 deg in steps spaced by 71.6 deg. Spatial frequencies were chosen with values ranging from 0.5 to 5 cycles/deg in steps spaced by a multiplicative factor of 1.4.

The DDOG channels are those proposed by Abbey and Barrett¹⁷ with radial frequency profile functions defined by

Fig. 3 The msCHO framework illustrating the derivation of the single decision variable from multislice stack.

$$C_j(\rho) = \exp\left[-\frac{1}{2}\left(\frac{\rho}{Q\sigma_j}\right)^2\right] - \exp\left[-\frac{1}{2}\left(\frac{\rho}{\sigma_j}\right)^2\right], \quad (2)$$

where $\sigma_j = \sigma_0\alpha^j$ is the channel standard deviation of the j 'th channel and σ_0 is the initial standard deviation. The number of channels used is 10, with the following parameters: $\sigma_0 = 0.005$, $\alpha = 1.4$, and $Q = 1.66$.

2.3.3 Training and testing sets for model observers

The image data set was divided into two independent sets of images: we used 1/3 of the available images for training and 2/3 for testing. The training set was used to estimate the model's template components⁴¹ (covariance matrix K_p , mean channelized signal $\langle v_n \rangle$, and mean training vectors $\langle \lambda_{\text{slice}} \rangle$). Specifically, we used the "hold-out" method,⁹ which requires independent sets of images for training and testing. This method involves finding a trade-off between the distribution of the number of images in the training and testing sets. To assess the models with the same number of images as the human observers for testing, we decided the minimum number of images for training would be 1/3 of the available images. The testing data set was used to estimate decision variables λ_{final} .

2.3.4 Model observer performance calculation

Model and human observer performances were directly compared through the measure P_c computed from a 4AFC experiment. The advantage of P_c over the more commonly used d' is that we do not need to hypothesize regarding the distribution of the decision variables. We therefore conducted a 4AFC experiment with the model observers and calculated P_c as the figure of merit. From the testing data set, we randomly selected one signal-present ROI and three signal-absent ROIs. The model observers computed a decision variable for each of the four images: if a signal-present image corresponded to the highest decision variable, the trial was considered a success; if not, the trial was considered a fail. P_c was estimated after all signal-present ROI had been presented.

The variability in the testing set sample was estimated with the bootstrap resampling technique.⁴² Each model observer performed 100 4AFC experiments for each condition. For each 4AFC experiment, the number of trials was fixed at 140, which is equal to the number of signal-present samples ($2 \times 1/3$ of the testing set), and for each trial, three signal-absent images were randomly selected without replacement, but were replaced after the trial. This allowed us to estimate the standard error from the standard deviation of the resample P_c over the 100 repetitions.

2.3.5 Internal noise

Internal noise methods for the CHO were previously investigated with a simulated signal embedded in real x-ray coronary angiogram backgrounds.³⁴ Two methods were proposed: (1) the decision variable internal noise method modifies the trial's decision variable by adding a random variable

$$\lambda_n = \lambda_e + \varepsilon, \quad (3)$$

where ε is sampled from a normal distribution with zero mean and a standard deviation proportional to the signal-absent decision variable's standard deviation $N(0; c\sigma_\lambda)$, where c is

a proportional constant integer used for calibrating internal noise and σ_λ is the signal-absent decision variable's standard deviation; (2) the channel internal noise method modifies the model template by adding an independent random variable at each channel response. The trial decision variable is a composition of the weighted sum of the channel responses and a scalar random variable sample ε from $N(0; \sqrt{\mathbf{K}_{\text{int}}})$, where $\mathbf{K}_{\text{int}} = c\text{diag}(\mathbf{K}_e)$ is the internal noise covariance matrix and $\mathbf{K}_e$ is the covariance matrix describing variance and covariance of channel output due to image noise

$$\lambda_n = \sum_{p=1}^P w_p (\mathbf{v}_{n,p} + \varepsilon_p), \quad (4)$$

where $\mathbf{v}_{n,p}$ is the slice n seen through channel p . To fit with human observers, model observer performance with the testing data set is degraded by adjusting the level of internal noise, which minimizes the root mean square error (RMSE) between human and model observer P_c

$$\text{RMSE} = \sqrt{\frac{1}{12} \sum_{\text{cond}=1}^{12} (P_{c(\text{cond}; \text{human})} - P_{c(\text{cond}; \text{msCHO with } \varepsilon)})^2}, \quad (5)$$

where cond stands for the experimental condition investigated.

3 Results

Figures 4 and 5 show model and human observer performances for 6 mm/10 HU, 6 mm/20 HU, and 8 mm/10 HU signals for FBP and IR reconstructed images, respectively, with both the decision variable and channel internal noise methods. Model observers were fitted to human observers, finding the internal noise value that minimized the RMSE. The internal noise value was the same for all signals and doses but with a different value of internal noise according to the reconstruction method. To estimate variability in the testing set, model observer performances were obtained by averaging the P_c through 100 repetitions of the 4AFC experiment with 140 trials each corresponding to the 140 signal-present images in the testing set ($2 \times 1/3$ of the testing set). Error bars correspond to P_c standard error. For FBP reconstructed images (Fig. 4), all model observer performances seem to be less accurate at predicting human observer performances for any signal or internal noise insertion method compared to IR reconstructed images (Fig. 5; see also Table 1). There is no single type of model and internal noise insertion method combination that invariably best predicts human observer performance across all investigated conditions.

Table 1 summarizes the RMSE of the model compared to human observers for msCHOa and msCHO with Gabor and DDOG channels and with both decision variable and channel internal noise methods. For comparison, RMSE was also calculated for overall conditions with the same value of internal noise for all signals and reconstruction algorithms and without internal noise insertion. As expected, inserting internal noise into the model observers makes them closer to the human observers, and there is an overall good agreement between human and model observer performances when internal noise is added. The statistical significance in the difference between human and model observer performances for each method is shown in Table 2. All F -values did not reach significance when internal noise was

Fig. 4 Comparison between human and model observer performances for FBP reconstructed images: (a), (c), (e) model observer performance with decision variable internal noise and (b), (d), (f) model observer performance with channel internal noise.

Fig. 5 Comparison between human and model observer performances for IR reconstructed images: (a), (c), (e) model observer performance with decision variable internal noise and (b), (d), (f) model observer performance with channel internal noise.

Table 1 RMSE in P_c between model observer and average human observer performances for both internal noise insertion methods. For each method, RMSE was calculated for all signals and doses investigated, including reconstruction algorithm conditions in two ways: (1) internal noise was separately adjusted for IR and FBP images and (2) internal noise was adjusted for overall reconstruction algorithms using a single internal noise value. The “Without internal noise” column refers to the case where no internal noise was added to model observers for overall signals, dose, and reconstruction algorithms investigated. Average human observer performance values are presented in Figs. 4 and 5.

		Decision variable internal noise			Channel internal noise			Without internal noise
		IR	FBP	Overall	IR	FBP	Overall	
msCHOa	DDOG	3.68	6.49	10.42	6.12	11.09	11.99	22.23
	Gabor	2.84	5.86	9.90	3.24	5.94	9.79	12.23
msCHOb	DDOG	3.59	5.33	10.69	1.44	6.22	9.83	22.11
	Gabor	3.27	5.69	9.80	2.83	5.43	11.24	13.52

Table 2 F -value to evaluate the difference between model observer and average human observer performances. The F -test was estimated comparing the mean square error between models and human observers and the variance across human observers. Significant differences in performance levels ($p < 0.05$) are indicated in bold.

		Decision variable internal noise			Channel internal noise			Without internal noise
		IR	FBP	Overall	IR	FBP	Overall	
msCHOa	DDOG	0.21	0.64	1.65	0.57	1.86	2.18	7.49
	Gabor	0.12	0.52	1.49	0.16	0.53	1.45	2.27
msCHOb	DDOG	0.20	0.43	1.73	0.03	0.59	1.46	7.41
	Gabor	0.16	0.49	1.46	0.12	0.45	1.91	2.77

added [$p > 0.05$; $F(11, 2) = 3.98$]. When no internal noise was added, the differences between human and model performance reached significance for msCHOa and msCHOb with DDOG channels.

Figure 6 shows FBP versus IR reconstructed images P_c for human and model observer with msCHOb and DDOG channels. The performance of the model observer was fit to the human performance using different internal noise levels for each of the reconstruction algorithms. For 6 mm/10 HU and 8 mm/10 HU signals, IR reconstructed images show higher P_c values than FBP reconstructed images at any tube current level. All differences were significant ($p > 0.05$ with a double sided t -test). All other models implemented led to similar results, but only one model was presented for brevity. Table 3 shows the c values used to calibrate internal noise for msCHOa and msCHOb with DDOG and Gabor channels. The c parameters reported correspond to values leading to the lowest RMSE, which best predict human observer performance.

The decision variable and channel internal noise methods provide similar c values for most types of channel and model combinations except for msCHOb with DDOG channels, which require a higher amount of internal noise to fit the human observers. There are differences with respect to the type of channels—Gabor channels systematically required a lower internal noise level than DDOG channels.

4 Discussion

The results of our study suggest that the msCHO can predict human observer performances for real CT images obtained

with a dedicated phantom. When msCHO performance is adjusted with respect to the reconstruction algorithm, RMSEs are lower than or equal to 6% in terms of P_c (Table 1) and down to 1% for the best case, showing a good agreement between model and human observers.

In the present work, we evaluated the correlation between msCHO and human observers for different model components: template derivation, type of channels, and internal noise insertion methods. Our study suggests that using a slice-specific templates method (msCHOa) or a template from the central slice applied to other slices (msCHOb) provides similar results. We wanted to evaluate both methods in order to have some insight about the human template when detecting a signal spread over multiple slices. The msCHOb approach could be justified if the signal characteristics are not changing significantly from one image slice to another and the slice thickness is small. In this case, it seems reasonable to use a single template independently of the slice position. The msCHOa strategy computes individual templates according to the position of the slice and seems, at first, to be the most relevant at mimicking the human template; we expected that slice-specific templates (msCHOa) would best capture the human observer strategy for multislice images. However, our results show that a single template for all slices has the same efficiency. This is at least the case in this particular situation of SKE on a uniform background. We were cautious about choosing diagnostic-representative parameters, but other inter- and intraslice thicknesses and signal sizes may lead to different performances of the msCHOa and msCHOb.

Fig. 6 Comparison between IR and FBP reconstructed images PC for human and model observers. The model observer used for illustration is the msCHO_b and DDOG channels.

Both types of tested channels (Gabor and DDOG) appear to be as efficient at mimicking human observers when internal noise is added. A similar comparison has been performed by Tseng et al.,²³ but with ssCHO instead of msCHO and with both Gabor and DDOG channels to assess low-contrast detection in CT. Their results are similar to ours. Nevertheless, our results show that without internal noise, DDOG channels display a systematically higher RMSE than Gabor channels. This demonstrates that DDOG channels are more efficient than Gabor channels in terms of signal-to-noise ratio in the

context of circularly symmetric signals. Indeed, the level of internal noise needed to match the human performance was systematically much lower for Gabor than for DDOG channels and comparable to the values in previous studies.^{32,43} This may be a hint that Gabor channels are closer than DDOG channels to capture the human visual filtering process in this kind of 3-D detection task.

Both internal noise methods showed good and similar results at predicting human observer performances. Adding internal noise directly to the channel output is equivalent to modifying

Table 3 Calibration values for msCHO_a and msCHO_b with DDOG and Gabor channels.

		Decision variable internal noise			Channel internal noise		
		IR	FBP	Overall	IR	FBP	Overall
msCHO _a	DDOG	11	15	14	5	17	14
	Gabor	1	5	4	1	3	3
msCHO _b	DDOG	11	15	15	52	100	103
	Gabor	1	5	5	1	30	15

the template, and this may be more representative of the human strategy than adding internal noise at the end of the process on the decision variable. Indeed, for ssCHO, Zhang et al.³⁴ concluded that channel internal noise provides a better correlation with human observers because of the human-like channels feature of the ssCHO. However, in the present study with msCHO, both internal noise methods performed the same.

In our study, we observed no difference between a model applying a single template to each image slice (msCHOa) or templates specific to each slice (msCHOb), as well as no difference in whether the internal noise is added on the template or on the channels. On the other hand, the type of reconstruction algorithm (FBP or IR) seems to have an influence. Better correlations between human and model observers were found for IR reconstructed images (lower RMSE, lower variability in terms of P_c , lower F -values) than for FBP images. As shown in Table 3, the c value used to adjust internal noise varies significantly between IR and FBP images, meaning different correlations between model and human observers according to the reconstruction algorithm. Furthermore, adjusting the internal noise simultaneously for IR and FBP images led to less efficient results, and any model accurately predicted the advantage of IR over FBP (Fig. 6) with the same internal noise value for both algorithms. This suggests that human observers are applying different templates depending on the texture of the image.^{44,45} Indeed, one distinctive aspect of IR reconstructed images is that they are preprocessed in order to reduce their noise. This makes these images look very different from FBP reconstructed images.

There were a number of limitations to our study. The first limitation relates to the models proposed, which do not take into account the browsing speed and the forward-backward movements performed by the radiologist in the reading room. In our experiment, we made the implicit assumption that the browsing speed and the browsing pattern do not influence the detection process. If they do, it would be worthwhile to include this in the msCHO strategy. Another limitation with the models is related to the modeling of the human visual system with spatial channels when the task is about detection in spatio-temporal varying noise. In dynamic viewing conditions, the human strategy might be different from those estimated in this study and could integrate the use of the signal's temporal phase by the human observer.⁴⁶ Our study used a signal detection task with spherical targets in homogeneous backgrounds in an SKE paradigm with small ROIs (32×32 pixels side), which could be considered an overly simplistic task compared to the more complex search tasks in clinical images. It has been repeatedly shown that the detection efficiency greatly depends on the complexity of the background and the presence of anatomical structures.^{12,13,25,47} Future studies should investigate how the complexity of background and signal variability influence signal detection in 3-D backgrounds.⁴⁸

5 Conclusions

A multislice model observer with internal noise shows a good correlation with a human observer in a simple 3-D SKE detection task with images taken from a uniform background phantom. Future work should investigate conditions closer to clinical practice with asymmetric signals, images with anatomic-like structures, and model strategies that incorporate temporal features of the human visual system.

Acknowledgments

This work was supported by grant SNF 320030_156032/1. The authors would like to thank Christel Elandoy from Lausanne University Hospital for her help with data acquisition and Ivan Diaz for his valuable insights.

References

1. D. L. Janzen et al., "Acute pulmonary complications in immunocompromised non-AIDS patients: comparison of diagnostic accuracy of CT and chest radiography," *Clin. Radiol.* **47**(3), 159–165 (1993).
2. J. R. Mathieson et al., "Chronic diffuse infiltrative lung disease: comparison of diagnostic accuracy of CT and chest radiography," *Radiology* **171**(1), 111–116 (1989).
3. E. T. Samara et al., "Exposure of the SWISS population by medical x-rays: 2008 review," *Health Phys.* **102**(3), 263–270 (2012).
4. C. H. McCollough et al., "Achieving routine submillisievert CT scanning: report from the summit on management of radiation dose in CT," *Radiology* **264**(2), 567–580 (2012).
5. J.-B. Thibault et al., "A three-dimensional statistical approach to improved image quality for multislice helical CT," *Med. Phys.* **34**(11), 4526 (2007).
6. Z. Deák et al., "Filtered back projection, adaptive statistical iterative reconstruction, and a model-based iterative reconstruction in abdominal CT: an experimental clinical study," *Radiology* **266**(1), 197–206 (2013).
7. M. Beister, D. Kolditz, and W. A. Kalender, "Iterative reconstruction methods in X-ray CT," *Phys. Med.* **28**(2), 94–108 (2012).
8. H. H. Barrett, "Objective assessment of image quality: effects of quantum noise and object variability," *J. Opt. Soc. Am. A* **7**(7), 1266–1278 (1990).
9. H. H. Barrett, *Foundations of Image Science*, Wiley-Interscience, Hoboken, New Jersey (2004).
10. A. E. Burgess, "Statistically defined backgrounds: performance of a modified nonprewhitening observer model," *J. Opt. Soc. Am. A* **11**(4), 1237 (1994).
11. A. E. Burgess, X. Li, and C. K. Abbey, "Visual signal detectability with two noise components: anomalous masking effects," *J. Opt. Soc. Am. A* **14**(9), 2420 (1997).
12. A. E. Burgess, F. L. Jacobson, and P. F. Judy, "Human observer detection experiments with mammograms and power-law noise," *Med. Phys.* **28**(4), 419–437 (2001).
13. J. P. Rolland and H. H. Barrett, "Effect of random background inhomogeneity on observer detection performance," *J. Opt. Soc. Am. A* **9**(5), 649–658 (1992).
14. K. J. Myers and H. H. Barrett, "Addition of a channel mechanism to the ideal-observer model," *J. Opt. Soc. Am. A* **4**(12), 2447 (1987).
15. K. J. Myers et al., "Aperture optimization for emission imaging: effect of a spatially varying background," *J. Opt. Soc. Am. A* **7**(7), 1279 (1990).
16. J. Yao and H. H. Barrett, "Predicting human performance by a channelized Hotelling observer model," *Proc. SPIE* **1768**, 161–168 (1992).
17. 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**(3), 473–488 (2001).
18. 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).
19. M. B. Sachs, J. Nachmias, and J. G. Robson, "Spatial-frequency channels in human vision," *J. Opt. Soc. Am.* **61**(9), 1176 (1971).
20. 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**(10), 1348–1362 (2006).
21. L. Yu et al., "Prediction of human observer performance in a 2-alternative forced choice low-contrast detection task using channelized Hotelling observer: impact of radiation dose and reconstruction algorithms," *Med. Phys.* **40**(4), 041908 (2013).
22. S. Leng et al., "Correlation between model observer and human observer performance in CT imaging when lesion location is uncertain," *Med. Phys.* **40**(8), 081908 (2013).

23. H. W. Tseng et al., "Assessing image quality and dose reduction of a new x-ray computed tomography iterative reconstruction algorithm using model observers," *Med. Phys.* **41**(7), 071910 (2014).
24. Y. Zhang et al., "Correlation between human and model observer performance for discrimination task in CT," *Phys. Med. Biol.* **59**(13), 3389–3404 (2014).
25. L. Platiša et al., "Channelized Hotelling observers for the assessment of volumetric imaging data sets," *J. Opt. Soc. Am. A* **28**(6), 1145–1163 (2011).
26. J. S. Kim et al., "A comparison of planar versus volumetric numerical observers for detection task performance in whole-body PET imaging," *IEEE Trans. Nucl. Sci.* **51**(1), 34–40 (2004).
27. M. Chen et al., "Using the Hotelling observer on multislice and multi-view simulated SPECT myocardial images," *IEEE Trans. Nucl. Sci.* **49**(3), 661–667 (2002).
28. C. Lartzien, P. E. Kinahan, and C. Comtat, "Volumetric model and human observer comparisons of tumor detection for whole-body positron emission tomography1," *Acad. Radiol.* **11**(6), 637–648 (2004).
29. H. C. Gifford et al., "A comparison of human and model observers in multislice LROC studies," *IEEE Trans. Med. Imaging* **24**(2), 160–169 (2005).
30. A. B. Watson, "Detection and recognition of simple spatial forms," in *Physical and Biological Processing of Images*, O. J. Braddick and A. C. Sleight, Eds., Vol. **11**, pp. 100–114, Springer Berlin Heidelberg (1983).
31. L. Platiša et al., "Volumetric detection tasks with varying complexity: human observer performance," *Proc. SPIE* **8318**, 83180S (2012).
32. A. E. Burgess and B. Colborne, "Visual signal detection. IV. Observer inconsistency," *J. Opt. Soc. Am. A* **5**(4), 617–627 (1988).
33. M. P. Eckstein et al., "Optimization of model observer performance for signal known exactly but variable tasks leads to optimized performance in signal known statistically tasks," *Proc. SPIE* **5034**, 123–134 (2003).
34. Y. Zhang, B. T. Pham, and M. P. Eckstein, "Evaluation of internal noise methods for Hotelling observer models," *Med. Phys.* **34**(8), 3312–3322 (2007).
35. A. Burgess and H. Ghandeharian, "Visual signal detection. I. Ability to use phase information," *J. Opt. Soc. Am. A* **1**(8), 900–905 (1984).
36. F. O. Bochud, C. K. Abbey, and M. P. Eckstein, "Search for lesions in mammograms: statistical characterization of observer responses," *Med. Phys.* **31**(1), 24–36 (2004).
37. V. A. Rosslyn, "Digital imaging and communications in medicine (DICOM) part 14: gray scale standard display function," NEMA Standards Publication PS 3.14-2004, National Electric Manufacturers Association, 1–54 (2004).
38. B. D. Gallas, G. A. Pennello, and K. J. Myers, "Multireader multicase variance analysis for binary data," *J. Opt. Soc. Am. A* **24**(12), B70–B80 (2007).
39. R. L. Van Metter, J. Beutel, and H. L. Kundel, Eds., *Handbook of Medical Imaging, Volume 1. Physics and Psychophysics*, SPIE, Bellingham, Washington (2000).
40. M. P. Eckstein and J. S. Whiting, "Lesion detection in structured noise," *Acad. Radiol.* **2**(3), 249–253 (1995).
41. 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**(20), 7159–7182 (2013).
42. R. M. Gagne, B. D. Gallas, and K. J. Myers, "Toward objective and quantitative evaluation of imaging systems using images of phantoms," *Med. Phys.* **33**(1), 83–95 (2005).
43. M. P. Eckstein, A. J. Ahumada, and A. B. Watson, "Visual signal detection in structured backgrounds. II. Effects of contrast gain control, background variations, and white noise," *J. Opt. Soc. Am. A* **14**(9), 2406 (1997).
44. Y. Zhang, C. K. Abbey, and M. P. Eckstein, "Adaptive detection mechanisms in globally statistically nonstationary-oriented noise," *J. Opt. Soc. Am. A* **23**(7), 1549–1558 (2006).
45. A. E. Burgess, F. L. Jacobson, and P. F. Judy, "Human observer detection experiments with mammograms and power-law noise," *Med. Phys.* **28**(4), 419 (2001).
46. M. P. Eckstein, J. S. Whiting, and J. P. Thomas, "Role of knowledge in human visual temporal integration in spatiotemporal noise," *J. Opt. Soc. Am. A* **13**(10), 1960–1968 (1996).
47. F. O. Bochud et al., "Estimation of the noisy component of anatomical backgrounds," *Med. Phys.* **26**(7), 1365 (1999).
48. J. Solomon and E. Samei, "Quantum noise properties of CT images with anatomical textured backgrounds across reconstruction algorithms: FBP and SAFIRE," *Med. Phys.* **41**(9), 091908 (2014).

Francis Verdun is a medical physicist. He has a strong expertise in the physical characterization of mammography units of CT scanners.

François O. Bochud has worked as a medical physicist for about 20 years. He has led several research projects in the field of texture analysis and model observers dedicated to the detection task, especially in mammography and more recently in CT.

Biographies for the other authors are not available.