Evaluation of the channelized Hotelling observer with an internal-noise model in a train-test paradigm for cardiac SPECT defect detection

Jovan G Brankov

Department of Electrical and Computer Engineering, Illinois Institute of Technology, 3300 S Dearborn St, Chicago, IL 60616, USA

E-mail: brankov@iit.edu

Received 6 April 2013, in final form 13 August 2013
Published 19 September 2013
Online at stacks.iop.org/PMB/58/7159

Abstract

The channelized Hotelling observer (CHO) has become a widely used approach for evaluating medical image quality, acting as a surrogate for human observers in early-stage research on assessment and optimization of imaging devices and algorithms. The CHO is typically used to measure lesion detectability. Its popularity stems from experiments showing that the CHO’s detection performance can correlate well with that of human observers. In some cases, CHO performance overestimates human performance; to counteract this effect, an internal-noise model is introduced, which allows the CHO to be tuned to match human-observer performance. Typically, this tuning is achieved using example data obtained from human observers. We argue that this internal-noise tuning step is essentially a model training exercise; therefore, just as in supervised learning, it is essential to test the CHO with an internal-noise model on a set of data that is distinct from that used to tune (train) the model. Furthermore, we argue that, if the CHO is to provide useful insights about new imaging algorithms or devices, the test data should reflect such potential differences from the training data; it is not sufficient simply to use new noise realizations of the same imaging method. Motivated by these considerations, the novelty of this paper is the use of new model selection criteria to evaluate ten established internal-noise models, utilizing four different channel models, in a train-test approach. Though not the focus of the paper, a new internal-noise model is also proposed that outperformed the ten established models in the cases tested. The results, using cardiac perfusion SPECT data, show that the proposed train-test approach is necessary, as judged by the newly proposed model selection criteria, to avoid spurious conclusions. The results also demonstrate that, in some models, the optimal internal-noise parameter is very sensitive to the choice of training data; therefore, these models are prone
to overfitting, and will not likely generalize well to new data. In addition, we present an alternative interpretation of the CHO as a penalized linear regression wherein the penalization term is defined by the internal-noise model.

(Some figures may appear in colour only in the online journal)

1. Introduction

Image quality evaluation is a critical step in optimization of any medical imaging system or image-processing algorithm (ICRU 1996, Barrett and Myers 2004). In diagnostic medical imaging, the human observer (HO) is the principal agent of decision-making. Therefore it is now widely accepted that the diagnostic performance of the HO is the ultimate test of medical image quality. For example, if an image is to be used for cardiac perfusion-defect detection, as in the test data set used in this paper, then image quality should be judged by the ability of a HO to detect perfusion defects within the image. Such an approach has become known as task-based image quality assessment.

Psychophysical studies to assess HO performance are difficult to organize, costly and time-consuming. Therefore, numerical observers (NOs) (also known as model observers)—algorithms capable of predicting HO performance—have gained popularity as a surrogate approach for image quality assessment.

An especially popular anthropomorphic NO, developed by Myers and Barrett (1987) (Yao and Barrett 1992), is the channelized Hotelling observer (CHO), which can be viewed as a linear generalized likelihood ratio test (see Zhang et al (2006) for a nonlinear extension) that introduces channels that are representative of human visual-system response. In some cases, the CHO outperforms human performance (Yao and Barrett 1992, Wollenweber et al 1998, 1999, Gifford et al 1999, Abbey and Barrett 2001, Lartizien et al 2004, Shidahara et al 2006, Park et al 2007) and one must introduce an internal-noise model that can diminish the detection performance of the CHO. Internal noise represents actual phenomena of the human visual system, including variations in neural firing, intrinsic inconsistency in receptor response, and a loss of information during neural transmission (Burgess and Colborne 1988, Lu and Dosher 1999). The use of an internal-noise model within the CHO has been shown in many situations to produce detection performance that correlates well with that of the HO (e.g., (Yao and Barrett 1992, Wollenweber et al 1998, Gifford et al 1999, Abbey and Barrett 2001)).

Thus, the CHO, with or without internal noise, is one of the most popular NOs in the medical imaging community (e.g., (Narayan and Herman 1999, Gifford et al 2000, Abbey and Barrett 2001, Narayanan et al 2002, Oldan et al 2004)), particularly in the field of nuclear medicine.

If one needs to introduce an internal-noise model for a given application, it is necessary to tune the CHO by choosing a specific internal-noise model, and then adjust the parameters of that model empirically based on an available set of HO data (Narayan and Herman 1999, Oldan et al 2004, Zhang et al 2004, 2007, Gilland et al 2006). In machine-learning terminology, the CHO with an internal-noise model is selected and trained to improve predictions of HO performance based on a set of labeled training data. Thus, model training can be viewed as a supervised-learning, system-identification, or machine-learning problem.

Motivated by this viewpoint, we have previously proposed an approach for prediction of HO detection performance for cardiac single-photon emission computed tomography (SPECT) defects, in which the CHO is replaced by a machine-learning algorithm (Brankov et al 2003, 2009), and we have extended this approach to diagnostic tasks other than lesion detection
Evaluation of the CHO with an internal-noise model

In these studies, our so-called learning NO (LNO) has outperformed the CHO. The major advantage of the LNO over the CHO is that the former approach is a nonlinear regression model. However, as the CHO with internal-noise model remains a highly popular approach, in this paper we investigate methods to optimize and test the method.

In a previous comparison (Brankov et al 2009) we considered the CHO with a single internal-noise model and channel type; specifically, we used uniform-variance internal noise (Oldan et al 2004, Gilland et al 2006) and rotationally symmetric bandpass filters (Myers and Barrett 1987, Abbey and Barrett 2001, Barrett and Myers 2004) as channeling operators. In this paper, we expand our investigation to compare ten published internal-noise models utilizing four different channel models in search of the best possible predictor of HO performance within the CHO framework. In addition, we propose the use of four new model-evaluation criteria using the same train-test protocol as in (Brankov et al 2009). Though not the main focus of the paper, a new internal-noise model is proposed that outperformed the ten established models in the cases tested on cardiac SPECT data. In addition we present an alternative interpretation of the CHO as a penalized linear regression in which the penalization term is defined by the internal-noise model.

In this work we investigate the CHO with internal noise through a supervised-learning viewpoint and obtain new findings about selection and tuning of the internal-noise model. As in supervised learning, it is essential to evaluate the CHO with internal-noise model using test data (data used to assess prediction error) that were not used during the tuning of the internal-noise parameters (model fitting). Moreover, in an image quality assessment problem it is important to go a step further, testing the model on images that represent not only new noise realizations, but also new image characteristics. For example, a CHO trained (by tuning the internal-noise model parameters) to predict HO performance for images reconstructed by one algorithm should predict HO performance accurately for a different reconstruction algorithm (or different parameter setting), otherwise one would defeat the very purpose of this approach. In the language of machine learning, the CHO with internal-noise model must be capable of good generalization performance (Vapnik 1998, Brankov et al 2006). The issue of generalization performance has been largely neglected in the CHO literature, where the model is typically tested using the same images, or images obtained by exactly the same algorithm (but different noise realizations), as those used for tuning the internal-noise parameters.

To illustrate the proposed approach to optimization of the internal-noise parameters, this paper considers eleven different internal-noise models (ten from the literature, and one proposed in this paper) within a simple example that typifies image quality assessment studies in a signal-known-exactly/background-known-exactly (SKE/BKE) scenario. The presented formalism could be extended to a SKE/background-known-statistically scenario (SKE/BKS). Note that the concordance of the CHO and HO in the SKE/BKS scenario can sometimes be better than that for SKE/BKE and it may turn out that the influence of an internal-noise model becomes less important. Alternatively in Burgess (1994), Jiang and Wilson (2006), Park et al (2009) the authors propose to incorporate human contrast sensitivity, rather than internal noise, into CHO model observers to match the human performance.

In this paper we specifically consider evaluation of reconstructed images obtained by SPECT with an ordered-subsets expectation-maximization (OSEM) algorithm. The purpose is not to evaluate the OSEM algorithm or to prove conclusively that any one internal-noise model will always perform best. The study is merely an example to illustrate the proposed procedure for selecting the internal-noise model, tuning its parameters, and using the proposed model selection criteria. The example reveals pitfalls that are encountered when the test data are not different from the training data.
In the proposed evaluation, after tuning of the internal-noise models on a broad set of images, the CHO is then tested on a different, but equally broad, set of images. Specifically, we tuned the internal-noise model parameters using images for six values of the full-width at half-maximum (FWHM) of the post-reconstruction filters and one iteration of OSEM, and then tested using images for six filters FWHM values and five iterations of OSEM, which are not included in training, then the roles of one and five iterations were reversed.

2. Methods

In a SKE/BKE defect-detection task the HO is typically asked to provide a score (confidence rating) S as to which of two hypotheses is true: defect present ($H_1$) or defect absent ($H_0$). The images under the two hypotheses are usually modeled as:

$$H_0 : \quad f = f_0 + n$$

$$H_1 : \quad f = f_1 + n$$

where the image is represented as a vector $f$ by using lexicographic ordering of the pixel values, $f_0$ denotes a background image, $f_1 = f_0 + \Delta f$ represents the image with defect present (where $\Delta f$ denotes the defect signature that observer aims to detect), and $n$ is zero-mean noise. In a detection study, image quality is assessed by the degree to which the HO can correctly distinguish the two hypotheses. This performance is typically quantified by using metrics from decision theory, notably the area under the receiver operating characteristic (ROC) curve, abbreviated as AUC (Barrett et al 1998, Abbey and Barrett 2001), which can be calculated using software such as ROCKIT (Metz et al 1998). In short, the goal of a NO is to predict the HO’s AUC.

Next we will continue with a brief introduction to the well-known CHO, a defect-detection NO model in which information is extracted through channels of a simplified representation of the human visual system and a statistical detector. The CHO is a cascade of two linear operators—a channeling operator and an observer—which, in practice, can be combined into one.

2.1. Channeling operator

The first operator, $U$, called the channeling operator, measures numerical features of the image by applying filters that are intended to model the human visual system (Myers and Barrett 1987). The experiments reported in this paper tested the four most commonly used types of channeling operators.

1. CHO-BP: rotationally symmetric, square profile, non-overlapping, bandpass filters, having cutoff frequencies of $[1/32, 1/16, 1/8, 1/4, 1/2]$ cycles/pixel, yielding a total of four channels (see figure 1(a)) (Myers and Barrett 1987, Abbey and Barrett 2001, Barrett and Myers 2004).

2. CHO-GB: Gabor filters with three spatial frequency bands of $[1/8, 1/4, 1/2]$ cycles/pixel and eight orientations from 0 to $\pi$. This results in 24 channels (see figure 1(b)) (Barrett and Myers 2004, Zhang et al 2004, 2006). Like the bandpass filters, the Gabor filters are intended to reflect the response of neurons in the primary visual cortex.

3. CHO-LG: Laguerre–Gauss polynomials. This paper considers polynomials of order 0–6 with three orientations: vertical, horizontal and rotationally invariant, yielding 18 channels (see figure 1(c)) (Barrett and Myers 2004, Zhang et al 2004).
(4) CHO-DOG: rotationally symmetric, overlapping, difference-of-Gaussians (DOG) profiles as defined in (Narayan and Herman 1999, Abbey and Barrett 2001) yielding ten channeling filters (see figure 1(d)).

We also tested the so-called sparse DOG, using profiles as defined in Abbey and Barrett (2001). However, as this channel model is a special case of the DOG and yields similar results, we dropped it from further consideration.

All channels are designed to have non-zero values on a 71 × 71 pixel window centered at the defect location. The window size is chosen such that average defect spectral support FWHM is approximately 1/4 cycles/pixel.

Letting \( u_i, i = 1, 2, \ldots, M \), denote vectors obtained by lexicographic ordering of the channels’ spatial response functions, and letting \( M \) represent the number of channels (in our experiments \( M = 4, 24, 18 \) or 10 corresponding to BP, GP, LG and DOG channeling operator, respectively), the channeling operator \( U \) is defined as:

\[
U = [u_1, u_2, \ldots, u_M]^T.
\]  

Without loss of generality let us assume that the channels are normalized as follows:

\[
u_i^T u_i = 1, \quad i = 1, 2, \ldots, M;
\]

thus the channel outputs are given by:

\[
x = Uf.
\]
Here it is worth noting that only the BP channel model has non-overlapping (independent) channels, i.e., $UU^T = I$; in the other models, the channel outputs are correlated.

As explained later, an internal-noise model is used in the CHO to enhance prediction accuracy for HO performance (Burgess and Colborne 1988, Lu and Dosher 1999). Specifically, a noise vector with normal distribution of zero mean and covariance $K_{int}$, $\epsilon \sim N(0, K_{int})$, is injected into all of the channel outputs (Abbey and Barrett 2001), which become:

$$x_i = u_i^T f + \epsilon_i, \quad i = 1, \ldots, M.$$  \hspace{1cm} (4)

### 2.2. Hotelling observer

The channeling operator is followed by a Hotelling observer, a linear classifier that computes a test statistic for choosing between two hypotheses in a cardiac perfusion-defect-detection task—defect present, $H_1$, and defect absent, $H_0$—based on the observed feature vector $x$ as:

$$t(x) = w^T(x + \epsilon) = \sum_{i=1}^{M} w_i (u_i^T f + \epsilon_i),$$  \hspace{1cm} (5)

where

$$w = K^{-1} U \Delta \bar{f},$$  \hspace{1cm} (6)

$$\Delta \bar{f} = \bar{f}_1 - \bar{f}_0,$$

and $K = K_{ext} + K_{int}$, in which the external-noise covariance matrix $K_{ext}$ describing noise originating in the data rather than in the visual system, is given by

$$K_{ext} = \frac{1}{2} \sum_{j=1}^{2} U[(f - \bar{f}_j)(f - \bar{f}_j)^T]_{H_j} U^T,$$  \hspace{1cm} (7)

where $\langle \cdot \rangle_{H_j}$ denotes conditional expectation under hypothesis $H_j$. Further details about the optimality of the CHO can be found in Barrett and Myers (2004). In a BKS scenario $K_{ext}$ will also incorporate statistical information about background variability.

Note that the test statistic, a cascade of two linear operators—a channeling operator and an observer—can be expressed as:

$$t(x) = t(Uf) = w^T Uf = w_f^T f.$$  \hspace{1cm} (8)

Therefore, the CHO effectively applies to the image $f$ an image-domain spatial template $w_f$, defined as:

$$w_f^T = w^T U = (K_{ext} + K_{int})^{-1} U \Delta \bar{f}^T U.$$  \hspace{1cm} (9)

In experiments presented later the external-noise covariance matrix $K_{ext}$ and $\Delta \bar{f}$ are substituted with sample estimates $\tilde{K}_{ext}$ and $\tilde{\Delta} \bar{f}$, exploiting a subset of the available data.

For a comparison between image-domain spatial templates such as those in equation (9), and HOs’ image-domain spatial templates, see (Abbey and Eckstein 2002, Castella et al 2007).

### 2.3. Decision-variable noise model

Burgess and Colborne (1988) and Lu and Dosher (1999) showed that HOs exhibit inconsistencies that can be described by a so-called decision-variable noise model, which can be described mathematically by injecting noise, with normal distribution of zero mean and variance $\sigma_d^2$, into the test-statistic variable as follows:

$$t(x) = w^T(x + \epsilon) + \gamma.$$  \hspace{1cm} (10)
2.4. Area under the receiver operating curve (AUC)

In a detection study, image quality is assessed by the degree to which the HO (approximated by the NO) can correctly perform the task. This performance is typically quantified by using metrics from decision theory, notably the AUC (Barrett et al. 1998, Abbey and Barrett 2001). In this setting, the AUC can be expressed simply in terms of a signal-to-noise ratio (SNR) as follows:

\[
\text{AUC} = \frac{1}{2} \left[ 1 + \text{erf} \left( \frac{\text{SNR}}{\sqrt{2}} \right) \right],
\]

where

\[
\text{SNR} = \frac{\text{var}(t(x)|H_1) - \text{var}(t(x)|H_0)}{\text{var}(t(x)|H_0, H_1)}.
\]

One can show that:

\[
\text{var}(t(x)|H_1) - \text{var}(t(x)|H_0) = \left( \frac{U}{\Delta \bar{f}} \right)^T K \left( K_{\text{ext}} + K_{\text{int}} \right) w + \sigma^2 \gamma,
\]

and, if the variances under the two hypotheses are equal, then the variance of the decision variable \(t(x)\) is given by:

\[
\sigma^2_{\text{(x)}} = \text{var}(t(x)|H_0, H_1) = w^T \left[ K_{\text{ext}} + K_{\text{int}} \right] w + \sigma^2 \gamma,
\]

where \(\sigma^2_{\text{ext}}\) is the variance of decision variable \(t(x)\) due to the external (data) noise, \(\sigma^2_{\text{int}}\) is the variance due to the channels’ internal noise \(\epsilon\) and \(\sigma^2 \gamma\) is the variance of the decision variable’s internal noise \(\gamma\). Note that, for the linear CHO, injection of the channels’ internal noise is equivalent to addition of decision-variable noise with variance \(\sigma^2_{\text{int}}\).

Using equations (8) and (9) and steps similar to those shown in Abbey and Barrett (2001) or Gallas and Barrett (2003), one can show that:

\[
\text{SNR} = \left[ (U \Delta \bar{f})^T (K_{\text{ext}} + K_{\text{int}})^{-1} (U \Delta \bar{f}) \right]^2 \frac{1}{1 + \frac{\sigma^2_{\gamma}}{\sigma^2_{\text{int}} + \sigma^2_{\text{ext}}} \sqrt{1 + \frac{\sigma^2_{\gamma}}{\sigma^2_{\text{int}} + \sigma^2_{\text{ext}}} \left( \frac{U}{\Delta \bar{f}} \right)^T \left( K_{\text{ext}} + K_{\text{int}} \right) w + \sigma^2 \gamma}}.
\]

Note that in the SNR and AUC calculations, \(\Delta \bar{f}\) and \(K_{\text{ext}}\) will be re-estimated for every reconstruction method separately and only \(K_{\text{int}}, \sigma^2 \gamma\) or \(\sigma^2_{\text{int}}\) (whichever exists in the particular internal-noise model being assessed) will be evaluated across different reconstruction methods in the proposed train-test paradigm.

2.5. Alternative interpretation of internal noise

The CHO internal-noise model can be interpreted as a regularized linear regression, as explained next. Let us assume that we wish to use the following channelized linear regression model:

\[
i_j = w_{\text{reg}}^T U f_j, \quad j = 1, 2, \ldots, N,
\]

where \(f_j\) is the \(j\)th image, represented as a vector and \(N\) is the total number of images and \(i_j\) is regression output.

Now for this linear regression model one can employ penalized least-squares estimation, which minimizes differences between the image label \(I_j\) and regression output \(i_j\), to find the optimal value \(w_{\text{reg}}^*\), i.e.:

\[
w_{\text{reg}}^* = \arg \min_{w_{\text{reg}}} \sum_{j=1}^{N} \left( I_j - w_{\text{reg}}^T U f_j \right)^2 + \lambda w_{\text{reg}}^T A w_{\text{reg}},
\]

where \(A\) is a regularization matrix.
in which $A$ is a matrix defining the regularization term $\lambda w_{\text{reg}}^T A w_{\text{reg}}$ and $\lambda$ is the regularization parameter.

Using this model it is easy to show (see the appendix) that:

$$w_{\text{reg}}^* = \left( U \sum_{j=1}^{N} f_j f_j^T U^T + \lambda A \right)^{-1} U \sum_{j=1}^{N} f_j I_j.$$  \hspace{1cm} (18)

Now substituting for $I_j$

$$I_j = \begin{cases} 1 & f_j \in H_1 \\ -1 & f_j \in H_0, \end{cases}$$ \hspace{1cm} (19)

and assuming that:

$$\sum_j f_j = 0.$$ \hspace{1cm} (20)

yields

$$w_{\text{reg}}^* = \frac{1}{2} \left( \tilde{K}_{\text{ext}} + \frac{\lambda}{N} A \right)^{-1} U \Delta \tilde{f},$$ \hspace{1cm} (21)

which has the same form as (6) with $\tilde{K}_{\text{ext}}$ and $\Delta \tilde{f}$ being sample estimates of $K_{\text{ext}}$ and $\Delta \tilde{f}$. Note that the image pixel values are (theoretically) non-negative since they represent concentrations of radiotracer, therefore the assumption in equation (20) must be forced upon images by image centering (for example). However this centering neither has an effect on the CHO performance as calculated according to the SNR in equations (5)–(12), nor on HO performance (since the images are rescaled before displaying them on a computer screen).

The previous derivation helps us to understand the CHO with internal-noise mode and its interpretation.

1. The CHO is a regularized linear regression that predicts the true hypothesis label $I$ rather than the human confidence rating $S$. By utilizing this consideration, the internal-noise covariance $K_{\text{int}}$ acts to stabilize the matrix inversion in equation (6). $K_{\text{ext}}$ itself is invertible but usually varies wildly between different data subsets. Note that equation (19) can be modified by substituting $I_j$ with $S_j$ so that the model fits the human confidence rating. A somewhat similar consideration is presented in Abbey and Eckstein (2002), Castella et al (2007).

2. The channels’ internal noise modulates the regularization term for the channel templates, which consequently regularizes the image-domain templates. This is also evident when comparing image-domain templates (see figures 5–8 and compare unregularized models 1, 8, 9, 10, models without internal channel noise, and regularized models 2, 3, 4, 5, 6, 7, 11, models with internal channel noise).

3. There may be some benefits of imposing regularization directly on the image-domain templates rather than on the channels. This will be explored in future work.

4. In a BKE task, the CHO does not take into account the image background, only the defect signature $\Delta \tilde{f}$. In a BKS scenario, $K_{\text{ext}}$ will incorporate the statistical properties of the background.

2.6. Internal-noise models

This work does not aim to understand the human visual system; instead the goal is to compare noise models in search of the best possible predictor of HO performance in a defect-detection task, within the CHO framework. For a broader view of the visual system and its modeling,
Evaluating the CH0 with an internal-noise model

Readers can refer to Burgess and Colborne (1988), Peli (1996), (2001), Barten (1999), Barrett and Myers (2004) and Zhang et al (2007).

Next, ten existing models are reviewed and evaluated, and a new model is proposed and tested (model 6).

The following are brief descriptions of each of the models considered, the first being an absence of internal noise (model 1). Models 2–7 have internal noise \((\mathbf{K}_{\text{int}} \neq 0)\) but no decision-variable noise \((\sigma^2_\gamma = 0)\). Models 8–10 have decision-variable noise \((\sigma^2_\gamma \neq 0)\), but no internal noise \((\mathbf{K}_{\text{int}} = 0)\). Model 11 is a ‘combined’ model that incorporates both internal noise and decision-variable noise.

The number of model parameters that must be tuned to optimize performance is specified for each of the models described next.

2.6.1. Model 1: no internal noise. This model, in which no internal noise or decision-variable noise is added, will serve as a baseline for comparison (Myers and Barrett 1987, Yao and Barrett 1992):

\[
\mathbf{K}_{\text{int}} = 0 \quad \text{and} \quad \sigma^2_\gamma = 0.
\]

In this case, \(\sigma^2_{\gamma(x)} = \sigma^2_{\text{ext}}\), and SNR is denoted by \(\text{SNR}_{\text{IO}}\) which is calculated as

\[
\text{SNR}_{\text{IO}} = \sqrt{(U\Delta f)^T \mathbf{K}_{\text{ext}}^{-1} U\Delta f}.
\]

Number of model parameters to tune: 0.

2.6.2. Model 2: quantization noise. In (Burgess 1985) the authors suggested that quantization of image intensity by a display device can be seen as a source of internal noise. In this internal-noise model, which was evaluated in (Narayanan et al 2002), \(\mathbf{K}_{\text{int}}\) is a diagonal matrix, the elements of which are given by:

\[
\mathbf{K}_{\text{int}}^{i,i} = Q^2_{\frac{12}{256}}.
\]

\(Q = \max_{H_0 \cup H_1} \langle f \rangle - \min_{H_0 \cup H_1} \langle f \rangle\).

Note. It can be shown (Grubbs and Weaver 1947, Stark and Brankov 2004) that \(Q\) is proportional to the variance of \(f\).

Number of model parameters to tune: 0.

2.6.3. Model 3: uniform-variance internal noise. In this noise model a constant variance \(\sigma^2\) (Oldan et al 2004, Gilland et al 2006) is added to each channel, so that \(\mathbf{K}_{\text{int}}\) is a diagonal matrix, the elements of which are given by:

\[
\mathbf{K}_{\text{int}}^{i,i} = \sigma^2.
\]

Note that, despite appearances, model 2 is a not special case of this model, because \(Q\) is a function of the data variance.

Number of model parameters to tune: 1.

2.6.4. Model 4: non-uniform internal-noise variance, proportional to external-noise variance. In this model the injected noise has channel variances proportional to the external-noise variances (Abbey and Barrett 2001, Oldan et al 2004), so that \(\mathbf{K}_{\text{int}}\) is a diagonal matrix having elements given by:

\[
\mathbf{K}_{\text{int}}^{i,i} = \rho \mathbf{K}_{\text{ext}}^{i,i}.
\]

Number of model parameters to adjust: 1.
2.6.5. Model 5: uniform internal-noise variance, proportional to the maximum external-noise variance. In this model the injected noise variances are proportional to the maximum variance of the channels’ external noise (Barrett et al 1998) so that $K_{\text{int}}$ has diagonal elements given by:

$$[K_{\text{int}}]_{i,i} = p \max_j [K_{\text{ext}}]_{j,j}. \quad (27)$$

Number of model parameters to tune: 1.

2.6.6. Model 6: non-uniform internal-noise variance, proportional to external-noise standard deviation. In this paper we propose the following new model, entirely motivated by the decision-variable internal-noise model 9, described later. In Model 9 the decision-variable noise variance is proportional to the decision-variable standard deviation due to external noise whereas here, in model 6, the channels’ internal noise has variances proportional to the standard deviations of the channels’ external noise, so $K_{\text{int}}$ is a diagonal matrix having elements given by:

$$[K_{\text{int}}]_{i,i} = p\sqrt{[K_{\text{ext}}]_{i,i}}. \quad (28)$$

Number of model parameters to tune: 1.

2.6.7. Model 7: non-uniform compound noise. In this model $K_{\text{int}}$ is a diagonal matrix having elements given by Kulkarni et al (2007):

$$[K_{\text{int}}]_{i,i} = p_1[K_{\text{ext}}]_{i,i} + p_2 \max_j [K_{\text{ext}}]_{j,j}. \quad (29)$$

Number of model parameters to tune: 2.

2.6.8. Model 8: constant variance decision-variable noise. In this model (Nagaraja 1964, Zhang et al 2007) decision-variable noise has constant variance

$$\sigma^2_{\gamma} = \sigma^2. \quad (30)$$

Number of model parameters to tune: 1.

2.6.9. Model 9: decision-variable variance proportional to the external-noise standard deviation. This model, suggested in Burgess and Colborne (1988) and evaluated in Zhang et al (2007), has the decision-variable noise variance proportional to the decision-variable standard deviation due to external noise.

$$\sigma^2_{\gamma} = p\sigma_{\text{ext}}. \quad (31)$$

Number of model parameters to tune: 1.

2.6.10. Model 10: decision-variable variance proportional to the external-noise variance. This model was suggested in Zhang et al (2004), wherein the internal-noise variance is proportional to the external-noise variance:

$$\sigma^2_{\gamma} = p\sigma^2_{\text{ext}}. \quad (32)$$

Number of model parameters to tune: 1.

This model is equivalent to injecting internal noise with covariance matrix proportional to the external-noise covariance matrix; that is,

$$\sigma^2_{\text{int}} = w' (K_{\text{ext}} + pK_{\text{ext}}) w = \sigma^2_{\text{ext}} + p\sigma^2_{\text{ext}}, \quad (33)$$
It is easy to show that:

\[
\frac{\text{SNR}}{\text{SNR}_\text{IO}} = \frac{1}{\sqrt{1 + p}}
\]  

(34)

where \(\text{SNR}_\text{IO}\) is as defined in section 2.6.1. This ratio is usually defined as a relative observer efficiency that Burgess et al (Burgess 1985, Burgess and Colborne 1988, Park et al 2007) calculated to be in the range 0.4–0.8; therefore, \(p \in [2.5, 7.25]\).

2.6.11. Model 11: combination. In Eckstein et al (2003) the authors proposed a noise model that combines internal and decision-variable noise models so that \(K_{\text{int}}\) has diagonal elements as:

\[
[K_{\text{int}}]_{i,i} = p_1[K_{\text{ext}}]_{i,i},
\]  

(35)

with the decision-variable noise variance given by

\[
\sigma^2_y = p_2\sigma^2_{\text{ext}}.
\]  

(36)

Number of model parameters to tune: 2.

2.7. Model evaluation

We will use the following criteria to evaluate performance of the CHO using the various internal-noise models.

(1) Mean-squared error (MSE). MSE is a widely used metric that reflects model-fitting accuracy measured by the squared distance between estimated values and target values. This measure is a good indication of model accuracy, but is not sufficient to serve by itself as the only model selection criterion, as explained next.

(2) Kendall’s tau rank correlation coefficient (\(K_{\text{tau}}\)). (Kendall 1948) The Kendall tau coefficient measures the degree of correspondence between two sets of rankings, ranging from \(-1\) (anticorrelated) to \(+1\) (perfectly correlated). In our evaluation we use the Kendall tau coefficient to assess the degree of correspondence between the rankings of the reconstruction algorithms as predicted by AUC values calculated using NOs, with that of the HOs. Note that \(K_{\text{tau}}\) measures the extent to which, as one variable increases, the other variable tends to increase, without requiring a linear relationship. In practice, when evaluating image-processing methods, the rank-ordering of performance is usually of great interest. However we were unable to use the Kendall tau to optimize internal-noise models because this metric is highly nonlinear and insensitive to large changes of internal-noise model parameters.

(3) Model parameter stability (MPS). For a specific diagnostic task, it is desirable that the optimal internal-noise parameters not vary significantly between data sets. The presence of such variability for a given model would suggest that the model is unstable and, therefore, not useful for practical applications. We use the ratio of internal model parameters to quantify the stability when changing from one data to another. A value of one for this ratio suggests good repeatability (stability); the value can range from zero to infinity. To our knowledge, this aspect of CHO behavior has not been explored previously, although in Zhang et al (2004) the authors noted that the performance of model 10 can depend significantly on the choice of the data set. In cases where internal-noise model has two parameters a ratio of each parameter is taken and an average is reported.
(4) Pearson correlation coefficient (PCC). Just as the model parameters should not vary significantly between data sets, the image-domain spatial template, $\mathbf{w}_t$, as defined in equation (9) should also remain relatively consistent. This may not be true in general, but for the experimental data used in this paper, the images are not significantly different, as can be seen in figure 2. This is even more evident if one considers $\Delta \mathbf{f}$, the defect images, given in the same figure. If the image-domain spatial template is not consistent it may indicate that the model has suffered from overfitting to a given data set, and has failed to capture intrinsic properties of the HO. We quantify this aspect of model stability by using the PCC to compare the obtained spatial templates. PCC between two spatial templates is defined as the covariance between pixel values among two spatial templates, divided by the product of their standard deviations. PCC ranges from $-1$ (anticorrelated) to $+1$ (perfectly correlated). This coefficient has been used successfully to test brain-image analysis procedures for reproducibility (Strother et al 2002).

In summary, we argue that an ideal internal-noise model should yield the following behavior. It should produce (1) AUC values that are close to HO AUC as judged by MSE, (2) high values for the rank correlation coefficient ($K_{\tau} \sim 1$), (3) stable model parameters ($M_{\text{PS}} \sim 1$) and (4) stable templates ($\text{PCC} \sim 1$).

One may argue that a reasonably useful model might not require all these properties. For example, one may view property 2 ($K_{\tau}$) as more important than property 1 (MSE). To the best of our knowledge there has not been a reported study exploring this issue, and no guidance has been proposed as to which metric is the most appropriate. However, as we argue in the next section, using solely the $K_{\tau}$ metric is not suitable, and we suggest that all presented model selection criteria be used.

2.8. Internal-noise model parameter tuning

The process of model training (parameter tuning) consists of finding values for the internal-noise model parameters that optimize an optimality criterion. In early experiments we used $K_{\tau}$ rank ordering, as well as a combination of MSE and $K_{\tau}$, as the optimality criteria for internal-noise parameter tuning; however, model accuracy and correlation with the HO obtained by this approach were not as good as those achieved by using the MSE criterion alone. As we pointed out earlier, the Kendall tau coefficient is extremely nonlinear and insensitive to large changes of the internal-noise model parameters; as such, it is not appropriate for tuning of model parameters, so we instead completed the studies using MSE as the optimality criterion.

In this work, we optimized the internal-noise parameter first by exhaustive search on a coarse grid ranging over 14 orders of magnitude from $10^{-7}$ to $10^{10}$ to ensure that the best match between the AUC of HO and CHO with internal-noise model is not missed. This search was further refined, on a finer grid, by focusing to a range that spans four orders of magnitude.
where the minimum is located. At each successive iteration the search range was reduced by an order of magnitude for a total of ten iterations. Usually, the method reached a stable solution after five iterations.

3. Simulated data and human-observer study

3.1. Human-observer data set

In our experiments we used a previously published HO study (Narayanan et al 2002), in which the MCAT phantom (Pretorius et al 1999) was used to generate average activity and attenuation maps, including respiratory motion, the wringing motion of the beating heart, and heart chamber contraction. The maps were sampled on a grid of $128 \times 128 \times 128$ with a pixel size of 0.317 cm. Projections ($128 \times 128$ images over 60 angles spanning 360) were generated by Monte Carlo methods using SIMIND (Ljungberg and Strand 1989), simulating the effects of non-uniform attenuation, photon scatter and distance dependent resolution corresponding to a low-energy high-resolution collimator. These projections were then resampled on a $64 \times 64$ grid over 60 angles. The simulated perfusion defect was located in the territory supplied by the left-anterior descending artery and had an angular extent of 45. The uptake levels for the perfusion defects were set at 65% of the normal uptake in the left ventricular walls in order to obtain nontrivial detection results. The image noise level is that of a typical clinical study with 0.5M counts from the heart region.

In our evaluation study, we used images reconstructed using the OSEM (Hudson and Larkin 1994) algorithm, with one or five effective iterations, incorporating attenuation correction and resolution recovery. These images were low-pass filtered with three-dimensional Gaussian filters with different FWHM of 0, 1, 2, 3, 4, or 5 pixels (see example images in figure 2). A single short-axis slice was extracted and interpolated to a $160 \times 160$ pixels image. Note that the combinations of filter FWHM and number of iterations yields 12 distinct reconstruction strategies.

Two medical physicists evaluated the defect visibility in a SKE environment (which also assumes location-known-exactly) for images at every combination of the number of iterations and FWHM of the filter. For each parameter combination of the reconstruction algorithm, a total of 100 noisy image realizations were scored by the observers (50 with defect present and 50 with defect absent) on a six-point scale following a HO training session involving an additional 60 images. The estimated AUC was calculated for each setting by using ROCKIT (Metz et al 1998) (see figure 3(a)).

Note that HO AUC curves have a nonlinear shape as a function of the reconstruction parameter (here spatial smoothing), thus complicating the task of matching HO AUC.

We also report $p$-values, i.e., the probability of obtaining by chance AUC values as extreme as the one that was actually observed, under the null hypothesis that there is no difference between methods (see figure 3(b)). The reported $p$-values indicate that it is sufficient to use two observers and 100 images if the goal is to determine whether one reconstruction methods AUC is statistically different from the other (Obuchowski et al 2004). Therefore, the data set used in this study is sufficient for numerical model tuning and testing.

4. Results

4.1. Evaluation of the numerical observer

For each noise model, we performed model evaluation by means of MSE, Ktau, MPS and PCC using two different comparisons.

Comparison 1: fitting accuracy. Here we tested each noise model using the same type of images as in model optimization, but with different noise realizations. A NO’s ability to
fit to HO data is a necessary, but not sufficient, condition for a NO to be useful. Specifically, there is no need to apply a NO on images reconstructed in the same way as those used in the NO training phase, since HO performance for this reconstruction method is already available from the HO study. Such testing would satisfy a general train-testing paradigm, but it is not sufficient for real-life NO applications.

Further, a NO trained to predict HO performance only on images reconstructed by one algorithm may be not accurate in predicting performance for a different reconstruction algorithm, defeating the very purpose of this approach.

To indicate results obtained as part of comparison 1, the names of the metrics are prefixed by the letter ‘F’ (short for ‘fitting’; i.e., F-MSE and F-Ktau). The reported values are averaged over six FWHM.

We used the following training procedure. For each reconstruction method, AUC was calculated according equations (11) using half of the available noise realizations for every value of the filter FWHM, with one iteration of OSEM. The internal-noise parameters were adjusted by exhaustive search to maximize agreement between the HOs’ AUC and the estimated AUC. The internal-noise parameters thus obtained were then applied to form predictions from the remaining noise realizations, yielding six AUC values. We then repeated the experiment with five iterations of OSEM, yielding an additional six AUC values. These 12 AUC values were compared to that of the HOs, and an average MSE and Ktau is reported.

Comparison 2: generalization accuracy. As we pointed out earlier, an important purpose of a NO is to provide an estimate of lesion-detection performance as a measure of image quality for reconstruction methods not yet evaluated by a HO ROC study. Therefore, to be useful, a NO must accurately predict HO performance over a wide range of image-reconstruction parameter

![Figure 3. Human-observer data analysis: (a) defect-detection performance measured by AUC; error bars represent one standard deviation; (b) p-values for rejecting the null hypothesis that there is no difference between methods.](image.png)
settings for which HO data are not available. Thus, the NO must exhibit good generalization properties.

In this comparison, we studied a kind of train-test generalization that is the most representative of the practical use of a NO. In this experiment, after tuning of the internal-noise model parameters for a broad set of images, the NO was then tested on a different, but equally broad, set of images.

Specifically, we tuned the internal-noise model parameters using data for every value of the filter FWHM, and one iteration of OSEM. In tuning, the parameters were adjusted by exhaustive search so as to minimize the average MSE between the six AUC values of the HOs and the six AUC values of the CHO AUC, thus maximizing agreement between human and model observer. The internal-noise parameters thus obtained were then applied to analyze the remaining data; that is, every value of the filter FWHM and five iterations of OSEM, yielding six AUC values. Next, the roles of one and five iterations were reversed, yielding an additional six AUC values. Next, these 12 values were compared to those of the HOs, and average values of MSE and Ktau are reported.

Recall that $K_{ext}$ and $\Delta \bar{f}$ are replaced by sample estimates of these quantities, $\tilde{K}_{ext}$ and $\tilde{\Delta \bar{f}}$, and re-estimated separately for each of the 12 reconstruction methods, and are not the part of generalization evaluation.

To indicate results obtained as part of comparison 2, the names of the metrics are prefixed by the letter ‘G’ (short for ‘generalization’; i.e., G-MSE and G-Ktau). The reported values are averaged over six FWHM values.

MPS is calculated as a ratio of the model parameters optimized in comparison 2. In cases where the internal-noise model has two parameters, the averaged ratio, over two model parameters, is reported.

Finally, an average PCC value is reported. The first component in this average is the average PCC over six image-domain spatial templates, $w_f$, corresponding to different FWHM values, obtained using images reconstructed by five iteration of OSEM. The second component is the average PCC calculated between image templates in iteration one and five. Finally, these two numbers are averaged and reported.

Tables 1–4 show, for each channel model (CHO-BP, CHO-GB, CHO-LG, and CHO-DOG), the average fitting error and rank correlation (F-MSE and F-Ktau) followed by average generalization error and rank correlation (G-MSE and G-Ktau) and MPS. These numbers are followed average PCC. These evaluations uniformly demonstrate that looking simply at fitting accuracy, measured by F-MSE or F-Ktau, leads to a misleading conclusion as to which model performs the best. In each case, the other criteria (G-MSE and G-Ktau, MPS and PCC), which emphasize generalization performance (which we argue is a necessary attribute of a NO), consistently point to the superiority of a different model than that suggested by the fitting-based performance measures.

4.2. CHO-BP evaluation

The results show consistently that the fitting-based metrics lead to conclusions about which NO is best that are not supported when the NO is tested on new data (to measure generalization performance). Because so many such comparisons are made, and these comparisons consistently demonstrate the same point, we only discuss one example comparison to illustrate the theme. The full results are shown in tables 1–4.

To choose an example, let us consider CHO evaluation when using bandpass filters (as in figure 1(a)) for the channeling operator. In this case, models 7, 8, and 10 show excellent fitting accuracy as measured by F-MSE (table 1), and model 6 shows good performance as measured
### Table 1. CHO-BP fitting mean-squared error (F-MSE), fitting rank correlation (F-Ktau), generalization MSE (G-MSE), generalization rank correlation (G-Ktau), model parameter stability (MPS) and the Pearson correlation coefficient (PCC).a

|             | Model 1   | Model 2   | Model 3   | Model 4   | Model 5   | Model 6   | Model 7   | Model 8   | Model 9   | Model 10  | Model 11  |
|-------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| **F-MSE (10^{-3})** | 12.93     | 12.79     | 3.97      | 3.92      | 2.20      | 2.25      | 1.64      | 1.39      | 2.31      | 1.69      | 4.18      |
| F-Ktau      | 0.47      | 0.33      | 0.60      | 0.60      | 0.40      | 0.73      | 0.47      | 0.60      | 0.40      | 0.60      | 0.47      |
| **G-MSE (10^{-3})** | 10.50     | 10.45     | 7.75      | 4.08      | 2.47      | 2.26      | 2.17      | 4.92      | 2.71      | 2.45      | 3.50      |
| G-Ktau      | 0.53      | 0.53      | 0.40      | 0.47      | 0.47      | 0.73      | 0.53      | 0.53      | 0.53      | 0.53      | 0.53      |
| MPS         | NA        | 1.00      | 2.67      | 2.36      | 1.49      | 1.03      | 0.00      | 3.62      | 1.42      | 0.92      | 1.57      |
| PCC         | -0.03     | 0.06      | 0.98      | 0.35      | 0.98      | 0.92      | 0.36      | -0.03     | -0.03     | -0.03     | 0.00      |

*aA grayed box indicates the model that was best in terms of the specific figure of merit.*

### Table 2. CHO-GB fitting MSE (F-MSE), fitting rank correlation (F-Ktau), generalization MSE (G-MSE), generalization rank correlation (G-Ktau), MPS, and the PCC.

|             | Model 1   | Model 2   | Model 3   | Model 4   | Model 5   | Model 6   | Model 7   | Model 8   | Model 9   | Model 10  | Model 11  |
|-------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| **F-MSE (10^{-3})** | 17.74     | 17.62     | 7.53      | 3.35      | 1.54      | 2.73      | 1.57      | 5.39      | 2.82      | 2.95      | 1.31      |
| F-Ktau      | 0.53      | 0.47      | 0.40      | 0.60      | 0.53      | 0.80      | 0.67      | 0.40      | 0.47      | 0.47      | 0.73      |
| **G-MSE (10^{-3})** | 16.04     | 15.98     | 7.38      | 1.48      | 2.37      | 1.88      | 1.57      | 6.38      | 8.97      | 6.43      | 1.60      |
| G-Ktau      | 0.40      | 0.40      | 0.40      | 0.60      | 0.53      | 0.80      | 0.60      | 0.40      | 0.40      | 0.40      | 0.40      |
| MPS         | NA        | 1.00      | 1.87      | 0.74      | 1.03      | 1.20      | 0.93      | 1.06      | 1.24      | 1.60      | 0.07      |
| PCC         | 0.08      | 0.23      | 0.99      | 0.81      | 0.99      | 0.99      | 0.82      | 0.08      | 0.08      | 0.08      | 0.24      |
Table 3. CHO-LG fitting MSE (F-MSE), fitting rank correlation (F-Ktau), generalization MSE (G-MSE), generalization rank correlation (G-Ktau), MPS and the PCC.

| Model  | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 | Model 7 | Model 8 | Model 9 | Model 10 | Model 11 |
|--------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| F-MSE ($10^{-3}$) | 17.27 | 17.19 | 4.81 | 2.50 | 2.22 | 3.21 | 4.13 | 2.21 | 2.90 | 2.20 | 1.82 |
| F-Ktau | 0.53 | 0.67 | 0.60 | 0.53 | 0.55 | 0.73 | 0.47 | 0.47 | 0.60 | 0.60 | 0.67 |
| G-MSE ($10^{-3}$) | 15.40 | 15.27 | 5.56 | 2.34 | 2.67 | 2.33 | 4.92 | 8.79 | 6.29 | 4.29 | 3.37 |
| G-Ktau | 0.53 | 0.53 | 0.60 | 0.40 | 0.47 | 0.80 | 0.40 | 0.53 | 0.53 | 0.53 | 0.33 |
| MPS    | NA     | 1.00   | 1.79   | 0.63   | 0.72   | 1.13   | 0.00   | 4.09   | 0.89   | 1.27   | 3.58   |
| PCC    | 0.22   | 0.63   | 0.96   | 0.94   | 0.98   | 0.96   | 0.91   | 0.22   | 0.22   | 0.22   | 0.08   |

Table 4. CHO-DOG fitting MSE (F-MSE), fitting rank correlation (F-Ktau), generalization MSE (G-MSE), generalization rank correlation (G-Ktau), MPS and the PCC.

| Model  | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 | Model 7 | Model 8 | Model 9 | Model 10 | Model 11 |
|--------|---------|---------|---------|---------|---------|---------|---------|---------|---------|---------|
| F-MSE ($10^{-3}$) | 15.79 | 12.23 | 5.25 | 1.20 | 3.18 | 1.44 | 2.23 | 6.50 | 1.29 | 3.42 | 2.93 |
| F-Ktau | 0.40 | 0.40 | 0.47 | 0.53 | 0.47 | 0.80 | 0.47 | 0.33 | 0.53 | 0.33 | 0.60 |
| G-MSE ($10^{-3}$) | 14.83 | 14.34 | 5.43 | 1.41 | 2.50 | 1.74 | 1.85 | 3.63 | 2.01 | 3.80 | 2.68 |
| G-Ktau | 0.47 | 0.47 | 0.47 | 0.60 | 0.47 | 0.87 | 0.60 | 0.47 | 0.47 | 0.47 | 0.53 |
| MPS    | NA     | 1.00   | 1.57   | 0.99   | 0.97   | 1.07   | 0.83   | 1.82   | 1.52   | 0.57   | 0.39   |
| PCC    | 0.21   | 0.37   | 0.99   | 0.94   | 0.98   | 0.99   | 0.94   | 0.21   | 0.21   | 0.21   | 0.01   |
Figure 4. CHO-BP; estimated image-domain templates.

by F-Ktau; however, models 5, 6, 7, 9 and 10 prove to have much better generalization performance as measured by G-MSE, while only model 6 has good G-Ktau performance. This example shows clearly that testing on the same type of images (i.e., evaluation by F-MSE), may be misleading, and that model testing should be performed on a distinct data set, using metrics such as G-MSE and G-Ktau. Table 1 also summarizes the stability of the internal-noise model parameter—MPS. Here MPS represents the average ratio of the internal-noise parameters, which ideally should equal 1 if the results are repeatable (which would be desirable). This ratio indicates that models 2, 6 and 10 produce stable parameters.

Finally, let us consider the image-domain templates, \( \omega_f \), for each internal-noise model, shown in figure 4. First, let us examine the noiseless difference images, \( \Delta \bar{f} = \bar{f}_1 - \bar{f}_2 \), which appear on the left in figures 4 and 1(e). Note that \( \Delta \bar{f} \) looks similar for images obtained by different reconstruction methods. This similarity was measured quantitatively by PCC computed in two ways: 1) as the average PCC for all difference images within images reconstructed by five iteration of OSEM (result is 0.892); 2) as the average PCC between corresponding difference images in iteration 1 and iteration 5 (result is 0.956) yielding an total average of 0.9240. Therefore, it is reasonable to expect that image-domain spatial templates \( \omega_f \) should somewhat resemble \( \Delta \bar{f} \) templates or at least have similar PCC of 0.9240. The average PCC values are shown at the bottom of table 1. Here we can conclude that models 3, 5 and 6 have similar PCC to that of \( \Delta \bar{f} \) templates. Moreover these templates resemble the difference image, \( \Delta \bar{f} \).

In conclusion, the best CHO-BP noise model overall is model 6 since it produces accurate results as measured by G-MSE and G-Ktau, and proves to be a stable model, as measured by MPS and PCC.

4.3. Other channeling models

To avoid repetition, rather than fully explain every comparison, we will only summarize the key findings for the other channel models; however, the reasoning in each case proceeds similarly to the preceding discussion for the case of BP filters.

CHO-GB. Fitting measures F-MSE and F-Ktau would misleadingly identify model 11 as a good one; however, after considering G-MSE, G-Ktau, MPS and PCC, model 6 emerges as the best, followed by model 7.
CHO-GD. F-MSE and F-Ktau would misleadingly identify model 11 as a good one; however, G-MSE, G-Ktau, MPS and PCC show that model 6 is best, followed by model 5.

CHO-DOG. F-MSE and F-Ktau would identify correctly model 6, but also suggests that models 4 and 9 as good candidates; however, after considering G-MSE, G-Ktau, MPS and PCC, model 6 is found to be the best model, followed by model 4.

4.4. The best model and different channeling operators

The initial aim of this work was not to propose a new internal-noise model, but simply to present possible criteria for evaluation of models and a principled approach to the train-test scheme. However, this study also led us to a new internal-noise model, which turned out to provide the best results in terms of generalization accuracy and stability. In this model the variances of the injected channel noise are proportional to the standard deviations of the external noise. This model proved consistently to be among the best models for all four channeling operators; therefore, as a final comparison, we present results in figure 8, showing AUC curves for the optimized model 6 and all four channeling operators. Error bars represent
7178 J G Brankov

Figure 7. CHO-DOG; estimated image-domain templates.

Figure 8. AUC curves for the best model, model 6, using different channel filters; error bars represent plus or minus one standard deviation. Generalization performance is shown in both figures; that is, the models are tested on data reconstructed in a different way than the data used for training.

This comparison shows similarly good performance of CHO-BP, CHO-GB and CHO-DOG, which is expected after examining the estimated image-domain templates in figures 4, 5 and 7. The performance of CHO-DOG is slightly better in terms of generalization G-MSE and G-Ktau (see tables). The performance of CHO-LG does not seem to be as good as that of the other three methods, as one would expect by examining the estimated templates in figure 6.

5. Discussion and conclusion

In this work we compared the generalization performance of a channelized Hotelling observer (CHO), a widely used NO, in predicting human observer (HO) performance for 11 different internal-noise models and four channeling filter models covering the majority of models which can be found in the current literature.

The findings of this paper are as follows.
In this application, to avoid spurious conclusions and good model selection, the train-test paradigm must not only involve new noise realizations, but also images that are substantially different from those used to develop the model. This paper proposes that one should first adjust the internal-noise model parameters (training phase) to make the CHO agree with HO data using images reconstructed with a broad set of reconstruction techniques (for example) and then evaluate (test) the CHO not only on different noise realizations, but also on images reconstructed by an equally broad set of different reconstructions, which were not available, nor used, during adjustment of the CHO internal-noise model parameters. This issue has been largely neglected in literature, where the CHO has been typically evaluated in terms of fitting accuracy.

To obtain a stable, robust model, one should avoid using models in which the optimal internal-noise parameter is very sensitive to the choice of training data. These models are prone to overfitting, and will not likely generalize well to new data.

This paper proposes and demonstrates the use of four quality metrics for model selection, in addition to traditional mean square error: Kendall’s tau rank correlation coefficient, model parameter stability (MPS), and Pearson correlation coefficient (PCC). The four proposed metrics can help to identify stable internal-noise models for a given application.

An alternative (and sometimes overlooked) interpretation of the CHO internal-noise model is given, which shows that the CHO is a regularized linear regression that predicts the true hypothesis label $I_j$ rather than the human confidence rating $S_j$. In this interpretation, the channels’ internal noise is seen as introducing a regularization term for either the channel templates or image-domain templates.

This paper proposes a new internal-noise model, which performed best in the application and comparisons considered. In this model, the internal noise has variance proportional to the standard deviations of the external noise. In the future, we aim to evaluate this internal-noise model in other settings.

Finally, this study shows that bandpass, Gabor, and difference-of-Gaussians (DOG) filters perform equally well as channeling operators, whereas the Laguerre–Gauss performs less well.

The presented results are only preliminary; for an unambiguous conclusion about the selection of an internal model, one would need to consider additional studies, readers and image sources, including real patient data, since phantom studies are usually limited in terms of object and background variability. The major hurdle, which is not addressed in this paper, is that there is no clear guideline in the literature as to how to set up an experiment to perform a parameter optimization study for image reconstruction. A second question of interest is how to decide which types of images should be selected to perform a pilot HO study used to tune the internal-noise model. Finally it is not clear if having only two readers, as in this work, allows generalizing conclusions to a study with a large number of readers. These questions are left for future studies. The presented results aim to stress the importance of a proper evaluation methodology, and to demonstrate the use of the new model-selection criteria.

Acknowledgments

We thank the medical physics group at the University of Massachusetts Medical Center, led by Michael A King, for generously providing the human-observer data used in the experiments. This work was supported by the National Institutes of Health under grants HL065425 and HL091017.
Appendix

Here we derive the solution of equation (18) which can be rewritten as:

\[ w_{\text{reg}}^* = \text{arg min}_w J(w), \] (A.1)

where

\[ J(w) = [c - 2b^T w + w^T Q w], \] (A.2)

with

\[ Q = U \sum_{j=1}^{N} f_j f_j^T + \lambda A, \quad b = U \sum_{j=1}^{N} f_j I_j \] and \( c = \sum_{j=1}^{N} I_j^2 \).

The solution of equation (18) can be found by taking the derivative of \( J(w) \) with respect to \( w \) and equating it to zero:

\[ \frac{\partial}{\partial w} J(w) = -2b + 2Qw = 0, \] (A.3)

so the final solution has following form:

\[ w_{\text{reg}}^* = Q^{-1}b, \] (A.4)

or:

\[ w_{\text{reg}}^* = \left( \sum_{j=1}^{N} U f_j f_j^T + \lambda A \right)^{-1} U \sum_{j=1}^{N} f_j I_j. \] (A.5)

References

Abbey C K and Barrett H H 2001 Human- and model-observer performance in ramp-spectrum noise: effects of regularization and object variability J. Opt. Soc. Am. A 18 473–88
Abbey C K and Eckstein M P 2002 Optimal shifted estimates of human-observer templates in two-alternative forced-choice experiments IEEE Trans. Med. Imaging 21 429–40
Barrett H H, Abbey C K and Clarkson E 1998 Objective assessment of image quality: III. ROC metrics, ideal observers, and likelihood-generating functions J. Opt. Soc. Am. A 15 1520–35
Barrett H H and Myers K J 2004 Foundations of Image Science (New York: Wiley)
Barten P G J 1999 Contrast sensitivity of the human eye and its effects on image quality Doctoral Thesis Technische Universiteit Eindhoven, The Netherlands
Brankov J G, El-Naqa I, Yongyi Y and Wernick M N 2003 Learning a nonlinear channelized observer for image quality assessment IEEE Nuclear Science Symp. Conf. Record (19–25 Oct 2003) vol 4 pp 2526–9
Brankov J G, Liyang W, Yongyi Y and Wernick M N 2006 Generalization evaluation of numerical observers for image quality assessment IEEE Nuclear Science Symp. Conf. Record (29 Oct–1 Nov 2006) pp 1696–8
Brankov J G, Yang Y, Wei L, Naqa I E and Wernick M N 2009 Learning a channelized observer for image quality assessment IEEE Trans. Med. Imaging 28 991–9
Burgess A 1985 Effect of quantization noise on visual signal-detection in noisy images J. Opt. Soc. Am. A 2 1424–8
Burgess A E 1994 Statistically defined backgrounds: performance of a modified nonprewhitening observer model J. Opt. Soc. Am. A 11 1237–42
Burgess A E and Colborne B 1988 Visual signal-detection: 4. Observer inconsistency J. Opt. Soc. Am. A 5 617–27
Castella C, Abbey C K, Eckstein M P, Verdun F R, Kinkel K and Bochud F O 2007 Human linear template with mammographic backgrounds estimated with a genetic algorithm J. Opt. Soc. Am. A 24 B1–12
Eckstein M P, Bartroff J L, Abbey C K, Whiting JS and Bochud F O 2003 Automated computer evaluation and optimization of image compression of x-ray coronary angiograms for signal known exactly detection tasks Opt. Express 11 466–75
Gallas B D and Barrett H H 2003 Validating the use of channels to estimate the ideal linear observer J. Opt. Soc. Am. A 20 1725–38
Gifford H C, King M A, De Vries D J and Soares E J 2000 Channelized Hotelling and human observer correlation for lesion detection in hepatic SPECT imaging J. Nucl. Med. 41 514–21
Evaluation of the CHO with an internal-noise model

Gifford H C, Pretorius P H and Brankov J G 2009 Tests of scanning model observers for myocardial SPECT imaging
Proc. SPIE 7263 72630R

Gifford H C, Wells R G and King M A 1999 A comparison of human observer LROC and numerical observer ROC for tumor detection in SPECT images IEEE Trans. Nucl. Sci. 46 1032–7

Gilland K L, Tsui B M W, Qi Y J and Gullberg G T 2006 Comparison of channelized Hotelling and human observers in determining optimum OS-EM reconstruction parameters for myocardial SPECT IEEE Trans. Nucl. Sci. 53 1200–4

Grubbs F E and Weaver C L 1947 The best unbiased estimate of population standard deviation based on group ranges J. Am. Stat. Assoc. 42 224–41

Hudson H M and Larkin R S 1994 Accelerated image reconstruction using ordered subsets of projection data IEEE Trans. Med. Imaging 13 601–9

ICRU 1996 Medical Imaging—The Assessment of Image Quality ICRU Report 54 (Bethesda, MD: ICRU)

Jiang Y and Wilson D L 2006 Optimization of detector pixel size for stent visualization in x-ray fluoroscopy Med. Phys. 33 668–78

Kendall M G 1948 Rank Correlation Methods (London: C Griffin)

Kulkarni S, Khur P, Hsiao L, Zhou L and Gindi G 2007 A channelized Hotelling observer study of lesion detection in SPECT MAP reconstruction using anatomical priors Phys. Med. Biol. 52 3601–17

Lartizien C, Kinahan P E and Comtat C 2004 Volumetric model and human observer comparisons of tumor detection for whole-body positron emission tomography Acad. Radiol. 11 637–48

Ljungberg M and Strand S-E 1989 A Monte Carlo program for the simulation of scintillation camera characteristics Comput. Methods Programs Biomed. 29 257–72

Lu Z L and Dosher B A 1999 Characterizing human perceptual inefficiencies with equivalent internal noise J. Opt. Soc. Am. A 16 764–78

Marin T, Kalayeh M M, Hendrik P, Wernick M N, Yang Y and Brankov J G 2011 Numerical observer for cardiac motion assessment using machine learning Proc. SPIE 7966 79660G

Marin T, Pretorius P H, Yang Y, Wernick M N and Brankov J G 2010 Numerical observer for cardiac-motion Proc. IEEE Nucl. Sci. Symp. (at press)

Metz C E, Herman B A and Roe C A 1998 Statistical comparison of two ROC-curve estimates obtained from partially-paired datasets Med. Decis. Making 18 110–21

Myers K J and Barrett H H 1987 Addition of a channel mechanism to the ideal-observer model J. Opt. Soc. Am. A 4 2447–57

Nagaraja N S 1964 Effect of luminance noise on contrast thresholds J. Opt. Soc. Am. 54 950–5

Narayan T K and Herman G T 1999 Prediction of human observer performance by numerical observers: an experimental study J. Opt. Soc. Am. A 16 679–93

Narayan M V, Gifford H C, King M A, Pretorius P H, Farncombe T H, Bruyant P and Wernick M N 2002 Optimization of iterative reconstructions of Tc-99m cardiac SPECT studies using numerical observers IEEE Trans. Nucl. Sci. 49 2355–60

Obuchowski N A, Beiden S V, Berbaum K S, Hillis S L, Ishwaran H, Song H H and Wagner R F 2004 Multireader, multicase receiver operating characteristic analysis: an empirical comparison of five methods Acad. Radiol. 11 980–95

Oldani J, Kulkarni S, Xing Y, Khur P and Gindi G 2004 Channelized Hotelling and human observer study of optimal smoothing in SPECT MAP reconstruction IEEE Trans. Nucl. Sci. 51 733–41

Park S, Badano A, Gallas B D and Myers K J 2009 Incorporating human contrast sensitivity in model observers for detection tasks IEEE Trans. Med. Imaging 28 339–47

Park S, Gallas B D, Badano A, Petrick N A and Myers K J 2007 Efficiency of the human observer for detecting a Gaussian signal at a known location in non-Gaussian distributed lumpy backgrounds J. Opt. Soc. Am. A 24 911–21

Peli E 1996 Test of a model of foveal vision by using simulations J. Opt. Soc. Am. A 13 1131–8

Peli E 2001 Contrast sensitivity function and image discrimination J. Opt. Soc. Am. A 18 283–93

Pretorius P H, King M A, Tsui B M W, Lacroix K J and Xia W 1999 A mathematical model of motion of the heart for use in generating source and attenuation maps for simulating emission imaging Med. Phys. 26 2323–32

Shidahara M et al 2006 Predicting human performance by channelized Hotelling observer in discriminating between Alzheimer’s dementia and controls using statistically processed brain perfusion SPECT Ann. Nucl. Med. 20 605–13

Stark H and Brankov J G 2004 Estimating the standard deviation from extreme Gaussian values IEEE Signal Process. Lett. 11 320–2
Strother S C, Anderson J, Hansen L K, Kjems U, Kustra R, Siltis J, Frutiger S, Muley S, Laconte S and Rottenberg D 2002 The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework Neuroimage 15 747–71

Vapnik V N 1998 Statistical Learning Theory (New York: Wiley)

Wollenweber S D, Tsui B M W, Lalush D S, Frey E C, Lacroix K J and Gullberg G T 1998 Comparison of radially-symmetric versus oriented channel. Models using channelized Hotelling observers for myocardial defect detection in parallel-hole SPECT IEEE Nucl. Sci. Symp. Conf. Record vol 3 pp 2098–4

Wollenweber S D, Tsui B M W, Lalush D S, Frey E C, Lacroix K J and Gullberg G T 1999 Comparison of Hotelling observer models and human observers in defect detection from myocardial SPECT imaging IEEE Trans. Nucl. Sci. 46 2098–103

Yao J and Barrett H H 1992 Predicting human performance by a channelized Hotelling observer model SPIE Proc. 1768 161–8

Zhang Y I, Pham B T and Eckstein M P 2004 Automated optimization of JPEG 2000 encoder options based on model observer performance for detecting variable signals in x-ray coronary angiograms IEEE Trans. Med. Imaging 23 459–74

Zhang Y, Pham B T and Eckstein M P 2006 The effect of nonlinear human visual system components on performance of a channelized Hotelling observer in structured backgrounds IEEE Trans. Med. Imaging 25 1348–62

Zhang Y, Pham B T and Eckstein M P 2007 Evaluation of internal noise methods for Hotelling observer models Med. Phys. 34 3312–22