CHANNELIZED HOTELLING OBSERVERS FOR THE DETECTION OF 2D SIGNALS IN 3D SIMULATED IMAGES

Ljiljana Platiša[a], Bart Goossens[a], Ewout Vansteenkiste[a], Aldo Badano[b], Wilfried Philips[a]

[a]Ghent University, TELIN-IPI-IBBT, Ghent, Belgium
[b]CDRH, FDA, Silver Spring MD, US

ABSTRACT

Current clinical practice is increasingly moving in the direction of volumetric imaging. However, model observers for 3D images have been little explored so far. This study is investigating the task of detecting 2D signals in multi-slice simulated image data. We propose a novel design of a multi-slice model observer. To evaluate it, we compare three different model designs of the channelized Hotelling observer (CHO), two multi-slice and one single-slice model. The multi-slice models are built as a sequence of a 2D CHO and 1D HO, where the CHO is used to calculate a vector of metrics for each slice in the planar view and the HO is used to calculate the final scalar statistic of the model. The single-slice model is a 2D CHO applied on the location of the lesion. Our results show that the multi-slice models outperform the single-slice one, and here the new model surpasses the existing one.

Index Terms— Image classification, Medical decision-making, Observers, Signal detection

1. INTRODUCTION

The primary goal of medical images is to assist physicians in the diagnostic process. Often, this diagnostic image reading corresponds to the task of detecting a lesion (signal) of interest, such as lung nodule detection in chest CT scans, in order to make a classification decision: normal case (signal-absent) or abnormal case (signal-present). In this respect, medical image quality may be assessed concerning how well the physicians, *human observers*, perform the task of signal detection. Still, such psychovisual studies are complex and time-consuming. Therefore, *model observers* have been developed to assist or even substitute humans in the detection task.

Current clinical practice is increasingly moving in the direction of volumetric imaging. This tendency is widely observed in various anatomical as well as functional 3D imaging modalities including ultrasound, MRI, CT, SPECT. In this study, we focus on multi-slice images where the signal of interest is two-dimensional as it is typically the case in lung nodule detection in chest CT scans or microcalcifications detection in breast tomosynthesis.

In past years, the study case of model observers for signal detection in 2D images has been widely explored. Specifically, a model known as the channelized Hotelling observer (CHO) [1] is shown to be a good estimator of the ideal signal detector [2]. Moreover, the CHO models are recognized to approximate human observer performance reasonably well [3], [4].

However, the domain of *3D image data* has been poorly explored in the sense of model observers and only few solutions have been proposed so far. In Ref. [5], for example, it has been shown that the performance of a human observer as well as one of the model observers is higher in case of multi-slice images compared to the single-slice ones. The 3D CHO model observers described in literature are usually designed to satisfy conditions of a particular clinical application, often in the domain of PET [6] or SPECT imaging [7], rather than investigate the problem in a more fundamental sense. Also, these studies all refer to detection of a 3D signal rather than a 2D one. In clinical practice, however, the greater distance between slices or greater thickness of those often results in single slice (2D) signals. Usually, detection of a 2D signal is a far more difficult task and exactly the focus of this study.

In this study, we propose a novel multi-slice CHO design motivated by the assumption that human observers may be more likely to examine (a certain number of) multiple consecutive slices of a stack with a *unique* signal matched filter in mind. This is in contrast to the literature, which assumes humans tuning this filter to the varying background information thus using a *separate* filter for each individual slice in the stack. We evaluate the design on a set of simulated image data and compare it to one of the state-of-the-art models proposed by Chen *et al* [5]. In addition, we compare the two multi-slice models to a single-slice CHO.

The paper is organized as follows: the next section summarizes the essential background information about the model observers. In section 3, we describe the three models used in this study and explain the experiment setup. The results are presented in section 4 and discussed in section 5. Finally, concluding remarks are given in section 6.

978-1-4244-5654-3/09/$26.00 ©2009 IEEE

2. MODEL OBSERVERS

We consider a binary classification task in which signal detection theory is determined by the two hypotheses: signal is present (H_1) or signal is absent (H_0). The observer decides which of the two hypotheses is true. An observer is defined by its discriminant function, t(**g**), which maps an image **g** to its test statistic, t_0. The decision is made by comparing t_0 to a certain threshold.

The ideal observer is defined as one that has full knowledge of the problem in terms of the conditional probability density functions of image data **g** under each hypothesis, pr(**g**|H_i), i={0,1}. The test statistic of the ideal observer is defined as the likelihood ratio, Λ = pr(**g**|H_1) / pr(**g**|H_0). In practice, it is often complicated or impossible to know the probabilities required to calculate Λ. Therefore, a linear approximation of the ideal observer has been defined, with linearity referring to its discriminant function

$$\mathrm{t}(\mathbf{g}) = \sum_{m=1}^{\mathrm{M}} \mathrm{w}_m \mathrm{g}_m \,, \qquad (1)$$

where M is the number of pixels in the image **g**. The weights, w_m, m={1,…,M}, form an image **w** called the *template* of the observer. Thus, (1) may be written as

$$\mathrm{t}(\mathbf{g}) = \mathbf{w}^{\mathrm{t}}\mathbf{g}\,. \qquad (2)$$

The ideal *linear* observer is known as the Hotelling observer (HO), and its template is defined as

$$\mathbf{w}_{\mathrm{HO}} = \mathbf{K}_{\mathbf{g}}^{-1}\Delta\mathbf{g}\,. \qquad (3)$$

Here, $\Delta\mathbf{g} = \langle \mathbf{g}|\mathrm{H}_1\rangle - \langle \mathbf{g}|\mathrm{H}_0\rangle$ where $\langle\,.\,\rangle$ denotes ensemble average, and $\mathbf{K_g}$ is the average of the ensemble covariance matrices of the signal-absent and signal-present data:

$$\mathbf{K_g} = \frac{1}{2}\left(\mathbf{K}_{\mathbf{g},0} + \mathbf{K}_{\mathbf{g},1}\right), \qquad (4)$$

$$\mathbf{K}_{\mathbf{g},\mathrm{i}} = \frac{1}{2}\langle(\mathbf{g}-\overline{\mathbf{g}}_{\mathrm{i}})(\mathbf{g}-\overline{\mathbf{g}}_{\mathrm{i}})^{\mathrm{t}}|\mathrm{H}_{\mathrm{i}}\rangle, \quad \mathrm{i}=\{0,1\},$$

$$\overline{\mathbf{g}}_{\mathrm{i}} = \langle \mathbf{g}|\mathrm{H}_{\mathrm{i}}\rangle.$$

When the images are Gaussian random vectors, the HO equals the ideal observer. However, the HO requires many image samples to properly estimate $\mathbf{K_g}$ resulting in high dimensionality problems.

To overcome this difficulty, the channelized Hotelling observer (CHO) was defined [1]. This is a linear observer. It may be seen as a specialization of the HO model which makes use of selective channels to model the human visual system while reducing the dimensionality of the problem. The channels can be seen as M-dimensional images, $\mathbf{u}_j$, j={1,…,J} where J is the number of channels. In contrast to the HO where all image data is used to build the template $\mathbf{w}_{\mathrm{HO}}$, the CHO model only makes use of the channel outputs, $\mathbf{v}=\mathbf{U}^{\mathrm{t}}\mathbf{g}$, where **U** denotes the channel matrix, **U**=[$\mathbf{u}_1$, $\mathbf{u}_2$, … , $\mathbf{u}_{\mathrm{J}}$]. If we denote the ensemble covariance matrix of the channelized data as $\mathbf{K}_{\mathrm{v}}$, the template of a CHO model is

$$\mathbf{w}_{\mathrm{CHO}} = \mathbf{K}_{\mathbf{v}}^{-1}\Delta\mathbf{v}. \qquad (5)$$

Finally, the test statistic is calculated as a linear combination of all channel responses, $\mathrm{t}_{\mathrm{CHO}}(\mathbf{v}) = \mathbf{w}_{\mathrm{CHO}}^{\mathrm{t}}\,\mathbf{v}$.

Commonly, to select the type of the channel and the channel parameters we refer to prior knowledge of the signal and consider the purpose of the model. In this study, we use two types of channels: Laguerre-Gauss (LG) channels [2] which are known to be efficient channels in case of rotationally symmetrical signals, and dense difference-of-Gaussian (DDOG) channels [3] which in the two-dimensional domain are recognized as anthropomorphic channels. Figure 1 illustrates the first 5 LG channels and the first 5 DDOG channels used in the study.

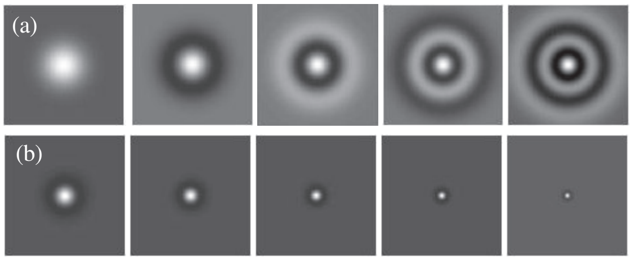

Figure 1. Images of the channels used in the study. (a) First 5 LG channels, the channel width is 75. (b) First 5 DDOG channels.

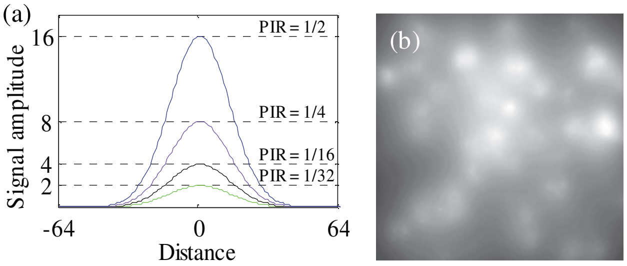

Figure 2. (a) Contrast profiles of the simulated designer nodules: PIR={1/4, 1/8, 1/16, 1/32}. (b) An example of images used in the study: signal-absent image slice. 3D CLB parameters: mean number of clusters K=160, mean number of blobs per cluster N=20, Lx=3, Ly=2, Lz=3. The image size is 256x256x64 pixels.

3. METHODS

3.1. Simulated data

We use a total of 2200 simulated images synthesized as 3D clustered-lumpy backgrounds (CLB) described in [8]. The images are of 256x256x64 pixel size, where the number of slices is N=64, each slice with M=256^2 pixels (Figure 2b).

Half of the backgrounds are used as signal-absent images. To generate signal-present images, we insert a 2D designer nodule signal [4] of four different intensities (Figure 2a) in the central slice of the remaining half of 1100 backgrounds. We define the peak intensity ratio (PIR) as the relative value of maximum gray level of the designer nodule compared to maximum gray level of the 3D CLB.

The image data set is used as follows: 1000 pairs (hereafter called *trainers*) of signal-present and signal-absent images are used as training data and 100 image pairs (hereafter called *testers*) are used as test data.

3.2. Design of the observer models

We compare three different designs: a single-slice model and two multi-slice models. The single-slice [2] observer model (hereafter called M1) is a 2D CHO run on the central slice in the stack in which the signal is inserted. The two multi-slice observer models are built as a sequence of 2D CHO and 1D HO. The CHO is used to calculate a vector of metrics for each slice in the planar view, $\mathbf{t}_{\mathrm{CHO}}$,

$$t_{\mathrm{CHO}n} = \mathbf{w}_{\mathrm{CHO}n}{}^{\mathrm{t}}\, \mathbf{v}_n, \quad n = 1,\dots,\mathrm{N}, \tag{6}$$

where N denotes the number of slices in the multi-slice image. The $\mathbf{t}_{\mathrm{CHO}}$ is then used by the HO to calculate the final scalar statistic of the model, namely

$$t_{\mathrm{M}}(\mathbf{t}_{\mathrm{CHO}}) = \mathbf{w}_{\mathrm{HO}}{}^{\mathrm{t}}\, \mathbf{t}_{\mathrm{CHO}} = \left(\mathbf{K}_{\mathbf{t}_{\mathrm{CHO}}}{}^{-\mathbf{1}} \Delta\mathbf{t}_{\mathrm{CHO}}\right)^{\mathrm{t}} \mathbf{t}_{\mathrm{CHO}}\ , \tag{7}$$

Nevertheless, the two models differ in how they build the template for the CHO. Let us denote the template matrix as

$$\mathbf{w}_{\mathrm{CHO}n} = \mathbf{K}_{\mathbf{v}n}{}^{-1} \Delta\mathbf{v}_n, \quad n = 1,\dots,\mathrm{N}. \tag{8}$$

In one case (hereafter called M2), a separate 2D CHO template is built for each position of the slice. For example, to build a template for the first slice in the stack we use only the first slices of the trainers. Hence, in (6) there are N different templates, $\mathbf{w}_{\mathrm{CHO}n}$. This approach corresponds to the work of Chen *et al.* [5] and Gifford *et al.* [7].

Alternatively, we assume that humans examine multiple consecutive slices of a stack with a unique signal template in mind. To model this, we build one 2D CHO template only and apply it on any slice in the tester stack, independent of the slice position within the stack. In view of (6) this translates into $\mathbf{w}_{\mathrm{CHO}n} = \mathbf{w}_{\mathrm{CHO}}$, $n = 1,\dots,\mathrm{N}$. To estimate this template, only the central slices from the trainer stacks are used. Hereafter, we refer to this design as M3.

3.3. Study design

The study is designed to evaluate the performance of three model observers: M1, M2 and M3. All experiments are performed for both LG and DDOG type of the channels, while PIR={1/4, 1/8, 1/16, 1/32}.

We employ the design of fully-crossed multiple-reader multiple-case (MRMC) study where each reader reads each case. In particular, the number of readers is Nrd=5, each trained with an independent set of Ntr_1=200 trainers, and each reading the same set of testers, Nts=100.

In addition, to understand how the number of trainers affects the performance of the model observers, we repeat the experiments for Ntr_2=500 and Ntr_3=1000 where the number of readers is 2 and 1, respectively.

As is common practice in objective image quality assessment [9], we use the area under the ROC curve (AUC) and the signal to noise ratio (SNR) as figures of merit. To evaluate the variability associated with the results, we use the one-shot variance analysis proposed in [10]. Finally, to evaluate the influence of the number of trainers on the models, we calculate the efficiency of the model observers relative to their performance for the highest considered number of trainers: $\eta_i = \mathrm{SNR}_{\mathrm{Ntr}i}{}^2 / \mathrm{SNR}_{\mathrm{Ntr}3}{}^2$, where $\mathrm{SNR}_{\mathrm{Ntr}i}$ denotes the SNR in case of Ntr_i trainers, i ={1,2}.

4. RESULTS

For all three model designs: M1, M2, and M3, the results of the MRMC studies and summarized in Figure 3. The results are arranged based on the type of 2D CHO channels, LG or DDOG. We note that for M3 the AUC was either very close to one or one, thus the variability of AUC could not be calculated directly using the one-shot method. Therefore, we restrict the analysis of M3 to AUC and SNR metrics. In case of DDOG channels, at the peak signal intensity level corresponding to PIR=1/32, all three models demonstrate low performance and high variability in terms of AUC, so we decide not to include them in further analysis. The SNR values, for PIR=1/16, are in the range of 0.5 to 2 for M1 and M2, while for M3 they increase up to 10 or even 25 in case of DDOG and LG channels, respectively.

Finally, we analyze the influence of the number of trainers on the observer performance. For all three designs and both channel types, the efficiency of models built with fewer trainers relative to the case with the greatest number of trainers used in the study is presented in Table 1.

Table 1. Efficiencies of the model observers trained with Ntr_1=200 and Ntr_2=500 relative to their performance for Ntr_3=1000. For both channel types used in the experiments (LG, DDOG) and each of the three model designs (M1, M2, M3), efficiency of the model is calculated using the SNR measurement at PIR=1/16 (the weakest signal detected by both LG and DDOG based models).

$\mathrm{N_{tr}}$	Channel type: LG			Channel type: DDOG		
	η_{M1}	η_{M2}	η_{M3}	η_{M1}	η_{M2}	η_{M3}
200	87%	78%	81%	95%	39%	41%
500	96%	>100%	99%	>100%	64%	55%

5. DISCUSSION

In order to compare different CHO designs, we examine model performance on the basis of three factors: signal intensity, type of 2D CHO channels, and number of trainers.

To evaluate the aspect of signal intensity, we refer to Figure 3. Overall, we notice that M1 and M2 follow approximately similar trends while M3 is nearly not affected by the selected levels of peak signal intensities. With current statistical properties of the backgrounds and signals, the detection performance of all three models degrades significantly as the signal intensity is further decreased. In terms of the variability, we observe lower performance of M1 compared to M2. Comparing single-slice model (M1) to the multi-slice ones (M2, M3), we notice that M1 performs below the other two designs in terms of AUC averaged over readers. These observations are expected since the single-slice model is only using the limited amount of available data comprised in the central slices of the stacks.

More interestingly, we make a comparison between the two multi-slice models, M2 and M3. In terms of AUC values, and even more notably the SNR values, the initial results from this study demonstrate considerably higher performance of M3 compared to M2.

Next, we compare the model designs based on the type of 2D CHO channels. The results suggest that the relative observed rankings between M1, M2 and M3 are well preserved for both LG and DDOG channels. In view of the absolute performance measurements, the LG channels slightly outperform the DDOG ones. This corresponds well to the theory and earlier reported results in literature [3].

Finally, the third aspect of interest in this study is the degree of influence of the number of trainers on the model observer performance. This criterion is of exceptional importance for real data studies where the number of relevant clinical samples is usually very limited. Table 1 compares the efficiencies of the models relative to the case with the greatest number of trainers considered in this study, Ntr=1000. Here we point out that for some models the performance for Ntr=500 seems to surpass the one for Ntr=1000 (η >100%) which indicates that in these cases the number Ntr=1000 of trainers is probably still too small for high confidence statistical analysis of the results. We notice that the influence is less significant for the single-slice model compared to the multi-slice ones, and here DDOG channels exhibit slightly lower dependency. In case of the multi-slice observer, the dependency on Ntr is much stronger in the case of DDOG channels for both M2 and M3 designs.

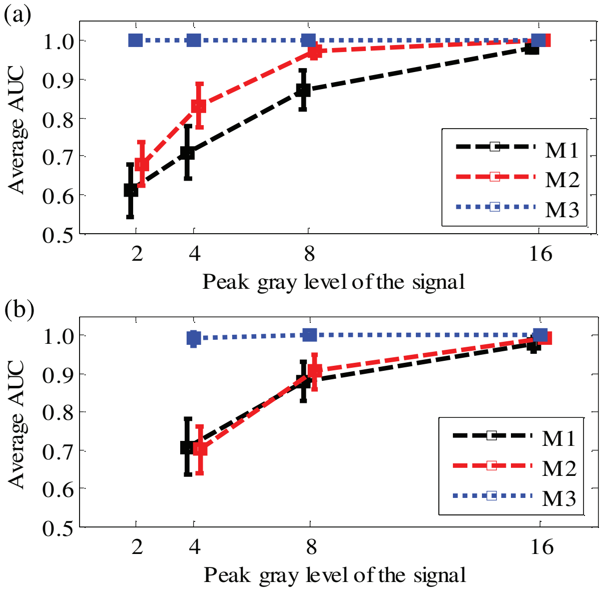

Figure 3. Average AUC of the three model observer designs: M1, M2 and M3. Number of readers in MRMC study Nrd=5, number of trainer image pairs per reader Ntr=200, and number of tester pairs Nts=100. Error bars are ± 2 standard deviations estimated by the "one-shot" method. (a) 2D channels are 10 LG channels with channel parameter a=75; PIR={1/4,1/8,1/16,1/32}. (b) 2D channels are 10 DDOG channels; PIR={1/4,1/8,1/16}.

6. CONCLUSIONS

To summarize, our results for all three criteria suggest that in the case of 2D signal detection in multi-slice images where the signal location is known, the new M3 model clearly outperforms M2. As explained above, the design of model M3 is motivated by our assumption that humans may be more likely to observe multiple consecutive slices in a stack with the same 'template of a signal' instead of changing this 'template' from one slice to another. This would particularly apply for high speed of stack-mode image browsing. However, assuming that our initial assumption is true (based on our experimental data), it may be of interest for future research to determine the number of consecutive slices for which the *same* 'template' can be used and how this is affected by the image content and the signal.

ACKNOWLEDGEMENTS

This work is supported by a grant from IBBT in the context of the MEVIC project.

REFERENCES

[1] K. J. Myers, and H. H. Barrett, "Addition of a channel mechanism to the ideal-observer model," J. Opt. Soc. Am. A, vol. 4-12, pp. 2447-2457, 1987.

[2] B. D. Gallas, and H. H. Barrett, "Validating the use of channels to estimate the ideal linear observer," J. Opt. Soc. Am. A 20, pp. 1725-1738, 2003.

[3] 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, vol. 18, pp. 473–488, 2001.

[4] E. A. Burgess, F. L. Jacobson, and P. F. Judy, "Human observer detection experiments with mammograms and power-law noise," Med Phys 28, pp. 419–437, 2001.

[5] M. Chen, J. Bowsher, A. Baydush, K. Gilland, D. DeLong, and R. Jaszczak, "Using the Hotelling observer on multi-slice and multi-view simulated SPECT myocardial images," Conf. Rec. IEEE Nuc. Sc. Symp., vol. 4, pp. 2258-2262, 2001.

[6] J. S. Kim, P. E. Kinahan, C. Lartizien, C. Comtat, and T. K. Lewellen, “A comparison of planar versus volumetric numerical observers for detection task performance in whole-body PET imaging,” IEEE Trans. Nucl. Sci., vol. 51, no. 1, pp. 34–40, 2004.

[7] H.C. Gifford, M.A. King, P.H. Pretorius, and R.G. Wells, "A comparison of human and model observers in multislice LROC studies," IEEE Trans. Med. img., vol. 24, issue 2, pp. 160-169, 2005.

[8] H. Liang, S. Park, B. D. Gallas, K. J. Myers, A. Badano, "Image Browsing in Slow Medical Liquid Crystal Displays," Acad. Radiol. 15, issue 3, pp. 370-382, 2008.

[9] H. H. Barrett, C. K. Abbey, and E. Clarkson, “Objective assessment of image quality. III. ROC metrics, ideal observers, and likelihood-generating functions,” J. Opt. Soc. Am. A, vol. 15, pp. 1520-1535, 1998.

[10] B. D. Gallas, “One-shot estimate of MRMC variance: AUC,” Acad Radiol. 13, issue 3, pp. 353-62, 2006.