Comparison of channelized hotelling and human observers in determining optimum OS-EM reconstruction parameters for myocardial SPECT

Karen L. Gilland  
*Johns Hopkins Medical Institutes*

Benjamin M. W Tsui  
*Johns Hopkins Medical Institutes*, btsui@jhmi.edu

Yujin Qi  
*Johns Hopkins Medical Institutes*, yujin@uow.edu.au

Grant T. Gullberg  
*E. O. Lawrence Berkeley National Laboratory*

Publication Details  
Comparison of channelized hotelling and human observers in determining optimum OS-EM reconstruction parameters for myocardial SPECT

Abstract
The performance of the Channelized Hotelling Observer (CHO) was compared to that of human observers for determining optimum parameters for the iterative OS-EM image reconstruction method for the task of defect detection in myocardial SPECT images. The optimum parameters were those that maximized defect detectability in the SPECT images. Low noise, parallel SPECT projection data, with and without an anterior, inferior or lateral LV wall defect, were simulated using the Monte Carlo method. Poisson noise was added to generate noisy realizations. Data were reconstructed using OS-EM at 1 & 4 subsets/iteration and at 1, 3, 5, 7 & 9 iterations. Images were converted to 2D short-axis slices with integer pixel values. The CHO used 3 radially-symmetric, 2D channels, with varying levels of internal observer noise. For each parameter setting, 600 defect-present and 600 defect-absent image vectors were used to calculate the detectability index (dA). The human observers rated the likelihood that a defect was present in a specified location. For each parameter setting, the AUC was estimated from 48 defect-present and 48 defect-absent images. The combined human observer results showed the optimum parameter setting could be in the range 5-36 updates ([number of subsets] × number of iterations). The CHO results showed the optimum parameter setting to be 4-5 updates. The performance of the CHO was much more sensitive to the reconstruction parameter setting than that of the human observers. The rankings of the CHO detectability values did not change with varying levels of internal noise.

Keywords
determining, channelized, optimum, os, em, reconstruction, parameters, myocardial, spect, hotelling, human, observers, comparison

Disciplines
Engineering | Science and Technology Studies

Publication Details

This journal article is available at Research Online: http://ro.uow.edu.au/eispapers/952
Comparison of Channelized Hotelling and Human Observers in Determining Optimum OS-EM Reconstruction Parameters for Myocardial SPECT

Karen L. Gilland, Member, IEEE, Benjamin M. W. Tsui, Fellow, IEEE, Yujin Qi, Member, IEEE, and Grant T. Gullberg, Fellow, IEEE

Abstract—The performance of the Channelized Hotelling Observer (CHO) was compared to that of human observers for determining optimum parameters for the iterative OS-EM image reconstruction method for the task of defect detection in myocardial SPECT images. The optimum parameters were those that maximized defect detectability in the SPECT images. Low noise, parallel SPECT projection data, with and without an anterior, inferior or lateral LV wall defect, were simulated using the Monte Carlo method. Poisson noise was added to generate noisy realizations. Data were reconstructed using OS-EM at 1 & 4 subsets/iteration and at 1, 3, 5, 7 & 9 iterations. Images were converted to 2D short-axis slices with integer pixel values. The CHO used 3 radially-symmetric, 2D channels, with varying levels of internal observer noise. For each parameter setting, 600 defect-present and 600 defect-absent image vectors were used to calculate the detectability index \( d' \). The human observers rated the likelihood that a defect was present in a specified location. For each parameter setting, the AUC was estimated from 48 defect-present and 48 defect-absent images. The combined human observer results showed the optimum parameter setting to be 4–5 updates. The performance of the CHO was much more sensitive to the reconstruction parameter setting than was that of the human observers. The rankings of the CHO detectability values did not change with varying levels of internal noise.

I. INTRODUCTION

In SPECT imaging, to make a fair comparison among different imaging system geometries and collimator geometries, the reconstruction parameters must be separately optimized for each system being compared. The optimum reconstruction parameters for a detection task are generally assumed to be those which maximize an index of image quality such as the area under the ROC curve (AUC) [1] or the detectability index \( d' \) [2],[3]. Because there are so many parameter values to test in determining the optimum parameter settings, numerical observers, such as the Channelized Hotelling observer (CHO) [2]–[5], provide a more practical method for parameter optimization than human observer studies, provided the numerical observer can predict human observer performance. Previous studies have compared the performance of the CHO to human observers for tumor detection in \(^{67}\)Ga SPECT imaging [6]–[8] and cardiac defect detection in \(^{99m}\)Tc-sestamibi SPECT imaging [9]–[13]. The \(^{67}\)Ga SPECT studies showed good agreement between the CHO and human observer performance; however, for the \(^{99m}\)Tc-sestamibi cardiac SPECT studies, the results were mixed. Wollenweber [9] and Chen [10] reported good correlation between CHO and human observer performance, but the CHO was not used to optimize reconstruction parameters in either of these studies. Narayanan [11] showed the good correlation between CHO and human observer performance for parameter optimization, but only if the images were grayscale instead of real-valued or if an observer internal noise model [14] was included in the CHO. Sankaran and Frey [12], [13] showed good correlation between the CHO and human observer performance for optimization of the filter cutoff frequency in post-reconstruction filtering, but did not compare the CHO and human performance in optimizing the number of iterations in their iterative reconstruction. The Sankaran and Frey studies did use grayscale images, but did not use internal noise in the CHO and used a very low lesion contrast (12.5%).

The study presented in this paper uses many of the same methods as the Sankaran and Frey studies, but explores some issues not addressed in those studies, including comparing the CHO and human observers for optimizing the number of iterations in reconstruction of the images. In this study, we increase the lesion contrast to a more clinically relevant level of 20%, but use the internal observer noise model to adjust the AUC values to a level that allows meaningful statistical analysis and comparison with human observer results. In addition, in this study we will use a more anatomically realistic phantom and simulate SPECT projection data with a more realistic Monte Carlo simulation as opposed to the analytical projection data simulation used by Frey [13]. Specifically, in this study, the CHO with the internal noise model will be compared to human observer performance for the optimization of the (number of subsets)/iteration and the number of iterations in the iterative ordered subset-expectation maximization (OS-EM) image reconstruction algorithm [15] used in obtaining the SPECT images.
The SPECT images were reconstructed using the iterative OS-EM algorithm [15] with 10 different subset and iteration settings. The 10 settings were 1 and 4 (number of subsets)/iteration at 1, 3, 5, 7, and 9 iterations. The 1 subset/iteration resulted in 64 views/subset and is equivalent to the ML-EM algorithm. The 4 subsets/iteration resulted in 16 views/subset. The OS-EM reconstructions were performed with nonuniform attenuation correction only. After reconstruction, the images were reoriented to the short-axis (SA) view and a single slice, containing the defect centroid, was extracted. The SA images were then low-pass filtered with a Butterworth filter of order 5 and cutoff of 0.15 pixels⁻¹. These filter parameters were chosen based on previous work which found optimal filter parameters for similar simulated SPECT images [12], [13].

The SA images were then processed for the observer studies. First, a 32 × 32 pixel region containing the myocardium was extracted from each SA image. Next, the images were interpolated to 256 × 256 pixels using bilinear interpolation. Finally, the image pixel values were converted from real values to integer values ranging from 0 to 255. This processing is necessary for the human observer study display. For consistency, the CHO study used more images than the human observer study, but all the images used in the human observer study were also used in the CHO study.

D. Human Observer Study

In the human observer study, the observers were shown a series of images and asked to rate, on a discrete scale from 1 to 6, the likelihood that a defect was present in a given location in the LV wall. The location was indicated by a cross-hair, which the observers could toggle on and off. The displayed image size was approximately 10 cm × 10 cm for a single SA slice. The images were shown against a black background.

There were 5 human observers, all are researchers in SPECT imaging. For each observer, the study was split into two sessions, each session about 1 hour in length. Each session consisted of 5 blocks of images. Each block consisted of images reconstructed with a given parameter setting ([number of subsets]/iteration and number of iterations). Each block consisted of a training set of 48 images followed by a test set of 96 images. Within both the training and test sets, there was an equal number of defect-present and defect-absent images as well as an equal number of images for each of the three defect locations. The image order within each block was completely random. The order of the blocks was different for each observer and was designed to minimize reading order effects.

An ROC analysis and subsequent analysis of variance (ANOVA) of the human observer data were done using the jackknife method for multiple readers and multiple cases (MRMC) [18]. This analysis was performed with the LABMRMC [19] program which estimated the mean and standard deviation of the AUC for each parameter setting (i.e., treatment) and each observer and then performed the subsequent ANOVA. The

so that the human observer study task would be of medium difficulty given the defect contrast of 20%.

C. Image Reconstruction and Processing

The SPECT images were reconstructed using the iterative OS-EM algorithm [15] with 10 different subset and iteration settings. The 10 settings were 1 and 4 (number of subsets)/iteration at 1, 3, 5, 7, and 9 iterations. The 1 subset/iteration resulted in 64 views/subset and is equivalent to the ML-EM algorithm. The 4 subsets/iteration resulted in 16 views/subset. The OS-EM reconstructions were performed with nonuniform attenuation correction only. After reconstruction, the images were reoriented to the short-axis (SA) view and a single slice, containing the defect centroid, was extracted. The SA images were then low-pass filtered with a Butterworth filter of order 5 and cutoff of 0.15 pixels⁻¹. These filter parameters were chosen based on previous work which found optimal filter parameters for similar simulated SPECT images [12], [13].

The SA images were then processed for the observer studies. First, a 32 × 32 pixel region containing the myocardium was extracted from each SA image. Next, the images were interpolated to 256 × 256 pixels using bilinear interpolation. Finally, the image pixel values were converted from real values to integer values ranging from 0 to 255. This processing is necessary for the human observer study display. For consistency, the CHO study was applied to images processed in exactly the same way as the human observer study images. The CHO study used more images than the human observer study, but all the images used in the human observer study were also used in the CHO study.

D. Human Observer Study

In the human observer study, the observers were shown a series of images and asked to rate, on a discrete scale from 1 to 6, the likelihood that a defect was present in a given location in the LV wall. The location was indicated by a cross-hair, which the observers could toggle on and off. The displayed image size was approximately 10 cm × 10 cm for a single SA slice. The images were shown against a black background.

There were 5 human observers, all are researchers in SPECT imaging. For each observer, the study was split into two sessions, each session about 1 hour in length. Each session consisted of 5 blocks of images. Each block consisted of images reconstructed with a given parameter setting ([number of subsets]/iteration and number of iterations). Each block consisted of a training set of 48 images followed by a test set of 96 images. Within both the training and test sets, there was an equal number of defect-present and defect-absent images as well as an equal number of images for each of the three defect locations. The image order within each block was completely random. The order of the blocks was different for each observer and was designed to minimize reading order effects.

An ROC analysis and subsequent analysis of variance (ANOVA) of the human observer data were done using the jackknife method for multiple readers and multiple cases (MRMC) [18]. This analysis was performed with the LABMRMC [19] program which estimated the mean and standard deviation of the AUC for each parameter setting (i.e., treatment) and each observer and then performed the subsequent ANOVA. The

so that the human observer study task would be of medium difficulty given the defect contrast of 20%.
ANOVA tested the null hypothesis that the AUCs for the 10 parameter settings were equal. The results are generalizable to both similar cases and similar observers.

E. Channelized Hotelling Observer Study

The basic CHO paradigm used in this study is very similar to that used in the previous study by Frey [13]. The reader is referred to that study for a detailed description of the paradigm. This section details the specific parameters used in our CHO study.

The CHO model consisted of 3 radially-symmetric, square-profiled, non-overlapping channels with octave-based center frequencies and channel widths. The starting channel width and frequency were 1/64 pixels⁻¹.

For each parameter setting, a separate CHO was calculated using a training set of 600 defect-present (3 defect locations × 200 noise realizations) and 600 corresponding defect-absent images. Then, for each parameter setting, a collection of scalar rating values was obtained by taking the inner product of the 3-element CHO vector and each image vector in a “test” set. The test set was a similar but statistically-independent set of 600 defect-present and 600 defect-absent image vectors.

Since the collections of rating values were approximately Gaussian distributed, an estimated mean detectability, \( \hat{d}_A \), and corresponding variance, \( \hat{\sigma}_{A, \text{int}}^2 \), were calculated as given by Abbey et al. [3], using the following equations:

\[
\hat{d}_A = \frac{\hat{\mu}_{A,+} - \hat{\mu}_{A,-}}{\sqrt{\frac{1}{2}(\hat{\sigma}_{A,+}^2 + \hat{\sigma}_{A,-}^2) + \hat{\sigma}_{\text{int}}^2}}
\]

(1)

and

\[
\hat{\sigma}_{A, \text{int}}^2 = \frac{2}{(\hat{\sigma}_{A,+}^2 + \hat{\sigma}_{A,-}^2)} \left[ \frac{\hat{\sigma}_{A,+}^2 + \hat{\sigma}_{A,-}^2}{N_+} + \frac{\hat{\sigma}_{\text{int}}^2}{N_-} \right] + \frac{(\hat{d}_A/2)^2}{(\hat{\sigma}_{A,+}^2 + \hat{\sigma}_{A,-}^2)} \left[ \frac{\hat{\sigma}_{A,+}^2 + \hat{\sigma}_{A,-}^2}{N_+ - 1} + \frac{\hat{\sigma}_{\text{int}}^2}{N_- - 1} \right]
\]

(2)

where \( \hat{\mu}_{A,+} \) and \( \hat{\sigma}_{A,+}^2 \) are the mean and variance of the rating values obtained from the defect-present images, \( \hat{\mu}_{A,-} \) and \( \hat{\sigma}_{A,-}^2 \) are the mean and variance of the rating values obtained from the defect-absent images, \( \hat{\sigma}_{\text{int}}^2 \) is the internal noise factor and \( N_+ \) and \( N_- \) are the number of defect-present and defect-absent images, respectively, and finally

\[
\hat{\sigma}_{A, \text{int}}^2 = (\hat{\sigma}_{A,+}^2 + \hat{\sigma}_{\text{int}}^2).
\]

(3)

The internal observer noise model, as given in equation (1), has been suggested and applied for model observers such as the CHO [3], [20], because it is known that human observers exhibit some randomness in rating images [14], [21]. There is no guide for how much internal noise to add in the absence of human observer results, so we tried many different levels of internal noise ranging from no noise to a level that was high enough to produce detectability values close to zero (AUCs close to 0.5) for all parameter settings. In this paper, the results will be presented at 3 values of internal noise: 0, 10, and 100.

For comparison with the human observer results, the \( d_A \) values were converted to AUC values using the following well-known relationship:

\[
\text{AUC} = 0.5 + 0.5 \times \text{erf}(d_A/2),
\]

(4)

III. RESULTS

A. Human Observer Study

The estimated mean AUC values for the individual human observers are shown in Fig. 2 along with their corresponding 95% confidence intervals. Note the variation in the AUC values and trends among observers. At 1 and 3 updates, especially, there is a wide range in the absolute AUC values among observers. However, for all 5 observers the trend in AUC with an increasing number of updates has some similarities. The AUC tends to increase from 1 to 5 updates, reaches a maximum at 5 to 9 updates and then tends to decrease slightly from 9 to 36 updates. For all 5 human observers, the maximum mean AUC occurred at 1 subset and 5 to 9 iterations.

The MRMC ANOVA analysis tested the null hypothesis that the AUC was equal among the 10 parameter settings. The results of this analysis strongly rejected the null hypothesis with \( p < 0.0001 \). Thus, the ANOVA indicates that at least one of the parameter settings has a different AUC than the others, but the ANOVA does not tell us which one (or ones). Therefore, we performed a second MRMC ANOVA using only 8 of the parameter settings, specifically from 4 to 36 updates. For this analysis, the null hypothesis could not be rejected (\( p = 0.1928 \)). Thus, for the range of 4 to 36 updates, the difference among the AUC values are not statistically significant. We can only say, with confidence, that the optimum parameter setting is neither 1 nor 3 updates.

B. Channelized Hotelling Observer Study

The estimated detectability (\( d_A \)) values obtained from the CHO are shown in Fig. 3, for 3 levels of internal noise. The errors bars on the mean \( d_A \) values show the corresponding 95% confidence intervals. For any two mean \( d_A \) values, if their 95% confidence intervals do not overlap, then difference between the two mean \( d_A \) values is statistically significant at the \( p=0.05 \) level. For all internal noise levels, the maximum mean \( d_A \) occurred at 5 updates. The difference between the mean \( d_A \) at 5 updates and the mean \( d_A \) at 1, 3, 7, 9, 12, 20, 28, 36 updates is statistically significant at the \( p=0.05 \) level. The difference between the mean \( d_A \) at 5 updates and the mean \( d_A \) at 4 updates is not statistically significant at the \( p=0.05 \) level. Therefore, we can say with confidence, that the optimum OS-EM parameter setting, as determined by the CHO, is 4 or 5 updates.

The level of internal noise affected the absolute \( d_A \) values and the relative differences among the \( d_A \) values, but the not rankings of the \( d_A \) values. It is the rankings that determine the choice of an optimum parameter setting. Thus, regardless of the level of internal noise, the CHO predicted the optimal parameter setting to be either 4 updates (4 subsets and 1 iteration) or 5 updates (1 subset and 5 iterations).
The AUC values for the CHO (obtained from the $d_A$ values shown in Fig. 3) are shown in Fig. 4 along with the mean AUC values averaged across all 5 human observers. The errors bars on the mean AUC values for the CHO and human studies show the 95% confidence intervals. Since the relationship between the $d_A$ and the AUC is nonlinear, $d_A$ values greater than approximately 3.6 translate to AUC values very close to 1; thus, the differences among the those AUC values and their corresponding 95% confidence intervals cannot be appreciated on the graph shown.

Fig. 4 shows that without the addition of internal noise, the AUCs for the CHO are much higher than for the human observers, indicating the task was much easier for the CHO than for the human observers. It is common in CHO studies for the CHO to give much higher absolute AUC values than the human observers [7]–[13].

IV. CONCLUSION AND DISCUSSION

The purpose of this study was to determine how well the CHO would predict human observer performance in determining op-
timum OS-EM reconstruction parameters for a $^{99m}$Tc-sestamibi myocardial SPECT defect detection study. The CHO study predicted the optimum setting to be 4-5 updates, while the best we can conclude from the human study is that the optimum setting lies in a range of 4–36 updates. The CHO’s optimum setting of 4–5 updates is within that range of optimum settings given by the human study. Furthermore, for any of the individual human observers, the maximum estimated mean AUC value occurred at either 5, 7, or 9 updates, which is very close that predicted by the CHO.

The inclusion of internal noise did not change the rankings of the CHO AUC values nor did it change the statistical significance of the differences among the AUC values. Thus, for parameter optimization studies, where one is only interested in the rankings of the detectability values, the level of internal noise does not appear to be important.

The problem the CHO has in predicting human observer performance in this study is that the CHO does not predict the broad range of OS-EM reconstruction parameter values over which the mean AUC doesn’t change much for the human observers. The graph in Fig. 3 shows a pronounced peak at 5 updates and shows the defect detection for the CHO to be fairly sensitive to the number of updates. However, the graph in Fig. 4 shows defect detection for the human observers to be much less sensitive to the number of updates. This result may be due to inadequate statistical power in the human observer study. The human observer study used 96 images per AUC estimate as compared to the human study which used 1,200 images per AUC estimate. This result may be due to inadequate statistical power in the human observer study. The human observer study used 96 images per AUC estimate as compared to the CHO study which used 1,200 images per AUC estimate. However, it may also be that the CHO is too simple a model of human observers in some respects. Perhaps the CHO is too sensitive to changes in noise and resolution in the images as compared to individual observers, and perhaps the CHO needs to model or account for the variations among human observers at the task of defect detection. Further study of the CHO and its predictability of human observers is needed.

ACKNOWLEDGMENT

The authors would like to acknowledge Dr. Eric Frey and Dr. Howard Gifford for their insights and comments. The content of this work is solely the responsibility of the authors and does not necessarily represent the official view of the PHS or its various institutes.

REFERENCES