Performance of MAP Reconstruction for Hot Lesion Detection in Whole-body PET/CT: an Evaluation with Human and Numerical Observers

Johan Nuyts *Member, IEEE*, Christian Michel *Member, IEEE*, Lieselot Brepoels, Liesbet De Ceuninck, Christophe Deroose, Karolien Goffin, Felix M. Mottaghy, Sigrid Stroobants, Jelle Van Riet, Raf Verscuren

***Abstract*—For PET (positron emission tomography) imaging, different reconstruction methods can be applied, including ML (maximum likelihood) and MAP (maximum a-posteriori) reconstruction. Post-smoothed ML images have approximately position and object independent spatial resolution, which is advantageous for (semi-) quantitative analysis. However, the complex object dependent smoothing obtained with MAP might yield improved noise characteristics, beneficial for lesion detection. In this contribution, MAP and post-smoothed ML are compared for hot spot detection by human observers and by the channelized Hotelling observer (CHO). The study design was based on the "multiple alternative forced choice" approach. For the MAP reconstruction, the relative difference prior was used. For post-smoothed ML, a Gaussian smoothing kernel was used.**
Both the human observers and the CHO performed slightly better on MAP images than on post-smoothed ML images. The average CHO performance was similar to the best human performance. The CHO was then applied to evaluate the performance of priors with reduced penalty for large differences. For these priors, a poorer detection performance was obtained.

I. Introduction

IT is generally accepted that statistical reconstruction of PET images using maximum-likelihood (ML) or maximum-a-posteriori (MAP) algorithms results in improved image quality when compared to filtered backprojection [1]. When iterated to convergence, ML yields very noisy images. However, when these images are filtered with a well-chosen low pass filter, they have very attractive features: they have nearly position and object independent resolution and noise characteristics superior to those of filtered backprojection (FBP) images. The quantitative accuracy is similar or slightly better than that of FBP in high count regions, although positive bias is seen for low count regions [2], [3].

In contrast, if the prior is well tuned, MAP can be iterated to convergence, yielding excellent images that need no further processing. If a shift-invariant prior is used, however, the resolution is object and position dependent, which is usually considered a disadvantage for clinical use. When a prior is tuned to impose uniform resolution, the images produced by MAP and post-processed ML are equivalent [4]–[6].

This work was supported by F.W.O. grant G.0174.03 and by Siemens

All authors except CM are with Nuclear Medicine, K.U.Leuven, Belgium (e-mail: Johan.Nuyts@uz.kuleuven.be), C. Michel is with Siemens Medical Solutions, MI, Knoxville, TN, USA, (e-mail: Christian.J.Michel@Siemens.com).

The non-isotropic and object dependent resolution of MAP is caused by the non-uniformity of the variance of the projections, and by the non-isotropic nature of the tomographic problem (attenuation, sensitivity, spatial resolution etc.). Shift invariant priors smooth more when the data are less reliable. This may be beneficial for non-ideal observers who are not able to optimally process correlated noise: they may need more smoothing when the data are noisier.

To avoid the aforementioned problems with spatial resolution, we currently use post-smoothed ML with a Gaussian kernel for our clinical images. This study investigates if the use of MAP would be recommended for clinical PET applications where hot spots must be detected in a noisy background, as is typically the case in oncological PET examinations. For the human observer study we have used the relative difference (RD) prior [7], which is similar to the quadratic prior, except that it penalizes relative rather than absolute differences. The channelized Hotelling observer was applied to the same set of images. In addition, it was applied to predict the effect of using priors with increased tolerance for large edges.

The next section describes the relative difference prior and priors with improved edge tolerance. It also discusses the design of the MAFC (multiple alternative forced choice) human observer study, and explains how the CHO was applied to execute that same detection task. The third section then describes the two experiments: the first detection experiment compares post-smoothed ML with MAP based on the relative difference prior, by quantifying the performance of human observers and the CHO. The second experiment uses only the CHO to predict the effect of edge-preserving priors on detection performance. The results are presented in section IV and discussed in section V.

II. Methods

A. RD, quadratic and Huber prior

The ML reconstruction is obtained by the iterative maximisation of (the logarithm of) the likelihood, assuming that the PET data are a realization of a Poisson distribution [8]. In MAP, the logarithm of a prior distribution is added to the log-likelihood term. The prior is used here to suppress noise by favoring smooth reconstructions. The objective function, the posterior, then becomes:

$$\text{posterior}(\Lambda, Y) = L(\Lambda, Y) + M(\Lambda), \tag{1}$$

where L is the logarithm of the likelihood, M is the logarithm of the prior, Y is the PET measurement and Λ is the image to be reconstructed by maximizing the posterior. Using concave priors, a concave posterior is obtained, because the log-likelihood L is concave as well (except for the Huber prior, the priors discussed here are strictly concave, yielding a posterior with unique maximum).

Most MAP implementations model the prior as a Gibbs distribution, which is a function of the absolute difference between neighboring pixel values [1]. The one mostly used is the quadratic prior, defining the logarithm of the prior as

$$M_{\mathrm{Q}}(\Lambda) = -\frac{\beta}{2}\sum_j \sum_{k\in N_j} w_{jk}(\lambda_j - \lambda_k)^2, \qquad (2)$$

where λ_j and λ_k are the values of pixels j and k of image Λ, N_j is the set of neighbor pixels of j, w_{jk} are position dependent weights and β is the strength of the prior. The weights w_{jk} are usually chosen to be shift-invariant, typically set to 1 for direct and $1/\sqrt{2}$ for diagonal neighbors.

In [7], a prior was proposed that penalizes relative differences instead of absolute differences. This prior can be written as

$$M_{\mathrm{RD}}(\Lambda) = -\beta\sum_j \sum_{k\in N_j} w_{jk}\frac{(\lambda_j - \lambda_k)^2}{(\lambda_j + \lambda_k) + \gamma|\lambda_j - \lambda_k|}. \qquad (3)$$

The parameter γ determines the edge tolerance. With $\gamma = 0$, the prior is very similar to the quadratic prior, except that its strength is not constant but inversely proportional to the local pixel intensities [7]. As a result, the penalty depends on the relative difference between two pixels, not on the absolute difference. Just like the quadratic prior, the RD-prior produces non-uniform resolution. However, due to its position dependent weights, the non-uniformities will be different from those of the quadratic prior. Some simulation experiments illustrating this different behaviour are presented in [7].

Mumcuoğlu et al [9] proposed to use the Huber prior instead of the quadratic prior:

$$M_{\mathrm{H}} = \begin{cases} -\beta \displaystyle\sum_{j,k\in N_j} \frac{w_{jk}}{2\delta}(\lambda_j - \lambda_k)^2 & \text{if } |\lambda_j - \lambda_k| < \delta \\ -\beta \displaystyle\sum_{j,k\in N_j} w_{jk}(|\lambda_j - \lambda_k| - \frac{\delta}{2}) & \text{if } |\lambda_j - \lambda_k| \geq \delta \end{cases} \qquad (4)$$

This prior is identical to the quadratic prior for small differences, but becomes linear rather than quadratic for large differences. The decreased penalty for large differences results in better tolerance for strong edges. A similar effect is obtained by setting γ to a non-zero value in (3): the prior becomes linear when the relative difference $|\lambda_j - \lambda_k|/(\lambda_j + \lambda_k)$ is large compared to $1/\gamma$. Priors with edge tolerance produce even more non-uniform resolution than the quadratic prior, because their resolution characteristics in the linear mode are different from those in the quadratic mode.

In this study, we evaluated MAP with the RD-prior and $\gamma = 0$, by comparing it to ML with Gaussian post-smoothing, for a hot lesion detection task. The comparison was done based on the performance of human observers and the performance of the CHO. In addition, a CHO detection study was carried out to evaluate the performance of the two edge-preserving priors (Huber and RD with $\gamma > 0$) mentioned above.

B. Channelized Hotelling observer

Currently, the CHO is often used as a model for the human observer in lesion detection tasks [10]. In our study, we used a "constant-Q" CHO as described in [11], [12]: it used three adjacent rectangular bandpass channels, defined by the four frequencies $[q^{-3}B, q^{-2}B, q^{-1}B, qB]$, with B = 0.4 per pixel and q = 2.3. The corresponding impulse responses in the spatial domain were computed and centered at the lesion position. The CHO applies the three channels to that location and obtains a column vector A with three values, one for each channel. It then forms the optimal linear combination of the channel outputs for detecting the presence or absence of the lesion. If the channel outputs are Gaussian random variables and if the covariance matrix is independent of the hypothesis, then the CHO is equivalent to the channelized ideal observer [10]. As a result, the CHO statistic can be considered as the logarithm of the likelihood ratio under the assumption of Gaussian noise, estimating the probability for the observation A as

$$p(A|\bar{A}_i) = N_i \exp\left(-\frac{1}{2}(A - \bar{A}_i)' C_{A_i}^{-1}(A - \bar{A}_i)\right) \qquad (5)$$

where $\bar{A}_i$ is the mean under the hypothesis $i = 0, 1$ (lesion present or absent), C_{A_i} is the covariance matrix, prime denotes matrix transpose and N_i is a normalization constant. To select the most likely hypothesis, the CHO computes the logarithm of $p(A|\bar{A}_1)/p(A|\bar{A}_0)$. Assuming that $C_{A_0} = C_{A_1} = C_A$, one obtains:

$$q(A) = (\bar{A}_1 - \bar{A}_0)' C_A^{-1} A - \frac{1}{2}(\bar{A}_1' C_A^{-1} \bar{A}_1 - \bar{A}_0' C_A^{-1} \bar{A}_0). \qquad (6)$$

The first term is the standard expression for the CHO-statistic in a single lesion experiment, but (6) is equivalent to it because the second term is independent of A. However, this second term depends on the local system response, and must be taken into account when the CHO is used to choose between different lesion positions (this term is equivalent to the expression (7) in [12], where it was introduced for the same reason).

The performance of the CHO for the detection of an individual lesion can be estimated by computing the signal-to-noise ratio of q, which is given by

$$\mathrm{SNR} = \sqrt{(\bar{A}_1 - \bar{A}_0)' C_A^{-1}(\bar{A}_1 - \bar{A}_0)}. \qquad (7)$$

The signal-to-noise ratio can be used as a measure of image quality [10], [11].

C. Multiple alternative forced choice study

Many researchers have used ROC or LROC approaches to quantify human observer performance in detections tasks. This study design requires that the observer provides a quantitative estimate of his or her confidence. Instead, we have chosen to use a multiple alternative forced choice (MAFC) approach

[13]. Images with exactly one lesion are presented, and the observer has to select the lesion position from a finite set of possible positions. This approach does not yield an ROC or LROC curve. However, it is less demanding for the observers, mainly because they don't have to rate their certainty. The output of the MAFC experiment is the fraction of correct responses, which we use as a measure of detection performance. The detectability index is a monotonic function of the fraction of correct responses [13], so both numbers are equivalent for ranking algorithms.

To compare human and CHO performance, the CHO was applied to the MAFC data as well. For that purpose, (6) was computed for every lesion position, and the position with maximum q was selected.

III. Experiments

A. Human observer study

The human observer study was done to compare the value of post-smoothed MLEM and MAP for lesion detection in PET/CT, and this for different degrees of regularisation: i.e. different widths of the Gaussian smoothing kernel for post-smoothed MLEM, and different values of β for MAP. For the MAP algorithm, the RD-prior with $\gamma = 0$ was used (eq (3)). For simplicity, the study was restricted to a single slice. The phantom was derived from a CT image (pixel size 0.825 mm), shown in figure 1. The image was segmented with thresholding, and PET activities and attenuation values were assigned to the different tissue types. Attenuated projections were computed for 144 angles. These projections were smoothed to simulate a detector resolution of 5 mm FWHM (full width at half maximum) and rebinned to obtain a PET sinogram of 150 detectors and 144 angles with a pixel size of 3.3 mm, and a total count of 245000. For ML and MAP reconstruction, the equivalent of 103 iterations were used, using a scheme with a gradually decreasing number of subsets. At the 16 possible lesion positions, 15 "benign" and 1 "malignant" lesion were inserted. The lesions are disks with a diameter of 1 cm and a uniform tracer distribution. The tracer uptake in the malignant lesion was set to 2.3 times the value in normal tissue, the tracer uptake in the benign lesions was the same as that in normal tissue. This tumor intensity was determined in a preliminary calibration study, to obtain a performance of between 70 and 90% correct detections.

Together with the simulated PET image, the CT image was shown to the observer. Figure 1 shows a snapshot of the interactive tool. The observer could select the color lookup table, adjust its minimum and maximum, change the image size on the screen etc., similar as with the viewers used for clinical purposes. Optionally the possible lesion positions were indicated with a circle of adjustable size and color, as shown in the figure. Scoring was done with a mouse click in one of the lesion positions. The PET image was shown alternatively at the left and the right, in an attempt to prevent the observer from rapidly browsing through the images, using changes between subsequent images to detect tumors.

The width of the 2D Gaussian post-smoothing kernels were 2, 3, 4 and 5 pixels FWHM (the pixel size is 3.3 mm $\times$ 3.3 mm). The values of β for the RD-prior were 3, 5, 7 and 10, which produced approximately the same range of spatial resolutions in the final images. These values were chosen based on an earlier study [14], predicting that the optimal performance would be captured with these values.

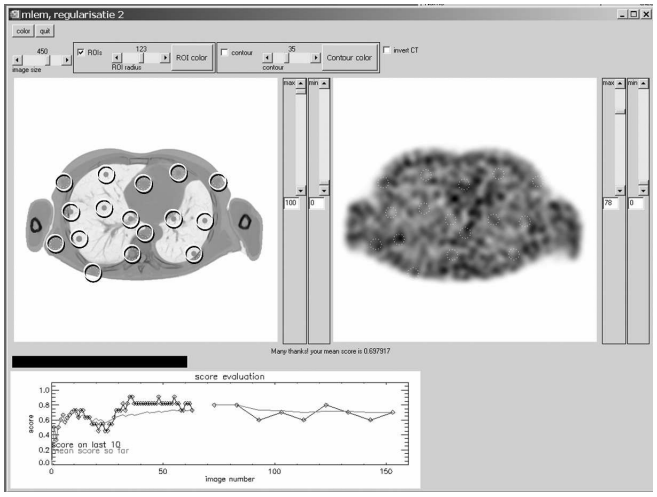

Figure 1. The interactive tool for the human observer study. The 16 lesion positions are indicated.

Nine observers have participated in the observer study. For each observer, 160 noisy sinograms were generated: there were 16 possible lesion positions, and for each position 10 noise realisations were produced. The 10 noise realisations were divided in two groups: 4 for training and 6 for scoring. Each sinogram was reconstructed with both algorithms and for each of the 4 regularizations, resulting in 8 sets of 160 reconstructions. Consequently, each observer had to complete 8 sessions, each consisting of 64 training images and 96 scoring images. Per observer, the 8 sessions were obtained by reconstructing the same 160 noisy sinograms with a different algorithm. However, a different set of 160 noisy sinograms was generated for each observer. The order of the training and scoring images was changed randomly for each session, to minimize the chance that an observer would recognize a particular noise realisation. Reusing the same noise realisations for each algorithm should increase the statistical power (enabling a paired comparison between algorithms).

During the training session, the observer received immediate feedback after every mouse click: the program reported if the detection was correct and also showed the noise free image for a few seconds. Two curves were updated after every detection, showing the overall mean score and the mean score on the last ten lesions (fig 1). For the scoring session, no feedback was given, except that the curves were updated after 10 detections, which helped to keep the observers motivated (they usually scored much better than they thought). Only the scoring session was used for the performance analysis. The primary output of each session is the fraction of correct lesion detections, called the score in the rest of the paper. The response time was also recorded during both training and scoring sessions. Outliers in response time were excluded to allow for unwanted interruptions in the session.

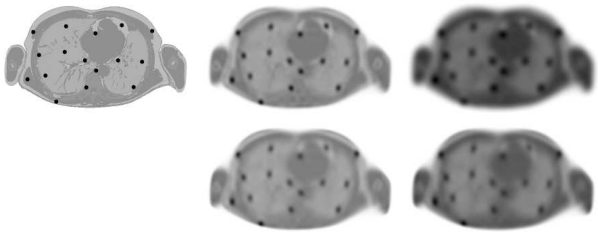

Figure 2. The true image (top left) and noise free reconstructions with post-smoothed MLEM (top row) and MAP with RD-prior (bottom row). The center column is with weak regularisation, the right column with strong regularisation. All images are scaled to their own maximum.

B. First CHO study

The channelized Hotelling observer was applied to score all the images from the human observer study. That means 9 $\times$ 8 $\times$ 160 images, because there were 8 independent sessions for each of the 9 observers.

The CHO needs an estimate of the covariance matrix C_A for every lesion position. This was estimated using the training data only, assuming that the covariance was the same for the lesion present and lesion absent case. Consequently, there was one C_A per lesion position, and it was estimated from all training data (both lesion present and lesion absent).

After that, the detection performance of the CHO was computed from its score on the scoring images. This produced 9 independent CHO-scores for each algorithm.

C. Second CHO study

Finally, the same 9 $\times$ 8 $\times$ 160 sinograms were reconstructed with other MAP algorithms as well, and scored with the CHO in the same way as explained above. These other MAP algorithms used the following priors: the quadratic prior, the RD-priors with γ = 10 and 20, the Huber prior with $\delta = 0.1$ (see eq. 4, the uptake in normal tissue was 1) and finally the RD-prior with γ = 200. The quadratic prior is similar to the RD- prior, except that its strength is object independent. The RD-priors with moderate γ and the Huber prior operate in two modes: a quadratic mode for small differences and an (almost) linear mode for large differences. The RD-prior with very high γ operates always in linear mode, and is therefore identical to the Huber prior with very small δ. Related edge preserving priors have been studied by others [15]–[18].

IV. Results

A. Human observer study

Nine observers completed all 8 sessions, a single session took on the average 32 min, the longest session was 70 min (including possible interruptions during the session) and the shortest was 14 min.

Figure 3 plots the fraction of correct responses for each of the nine observers and for each reconstruction algorithm, together with the mean over all observers. For each algorithm (MLEM and MAP) there were 4 different levels of regularisation. Assuming that in an ideal case, one could optimize the regularisation for each observer, we compared the scores

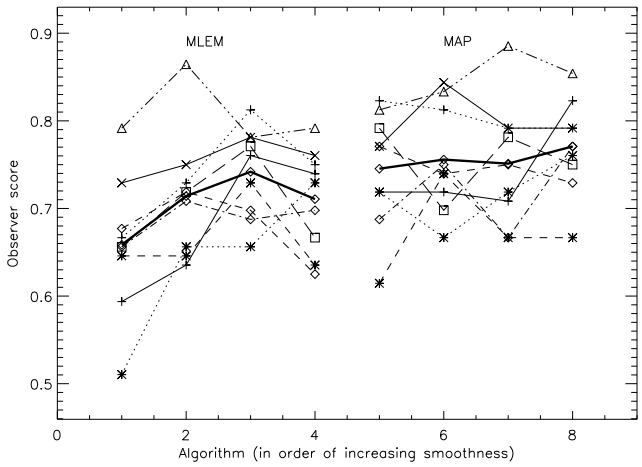

Figure 3. Human observer scores, for both algorithms, plotted for the four different regularisation levels. Each individual observer result is plotted with different symbols and/or line styles, the thick lines represent the mean over all observers.

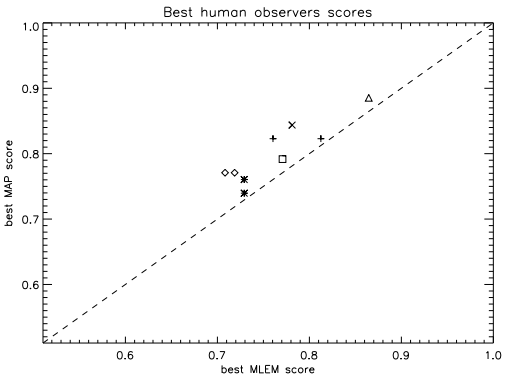

Figure 4. Paired comparison of the best detection score on post-smoothed MLEM and the best score on MAP for each observer.

obtained at the best regularisation (i.e. the highest score of the four points on each curve in fig 3). This produces 9 pairs of scores, which are plotted in fig 4. For all observers, the highest score was obtained for the same algorithm. The probability this would happen by chance if MLEM and MAP were equivalent, is p = $1/2^8 \simeq 0.004$, which indicates that MAP reconstructions are slightly but significantly better for this detection task than post-smoothed MLEM images. A two-sided paired t-test gave p = 0.0012. Fig 3 shows that also the mean performance was slightly better for MAP.

Figure 5 shows the total response times for each observer, and also the mean over all observers. The response time was measured as the time interval between the previous and the current lesion-selection mouse clicks. The total response time is the sum over all response times during the scoring session. Response times longer than the mean + 3 standard deviations were discarded, because they may have been caused by voluntary or undesired interruptions. The response times tend to decrease with increasing detection performance. The mean response time for a correct answer was 7.75 s, and for a wrong answer 14.6 s. For all 9 observers and all 8 algorithms,

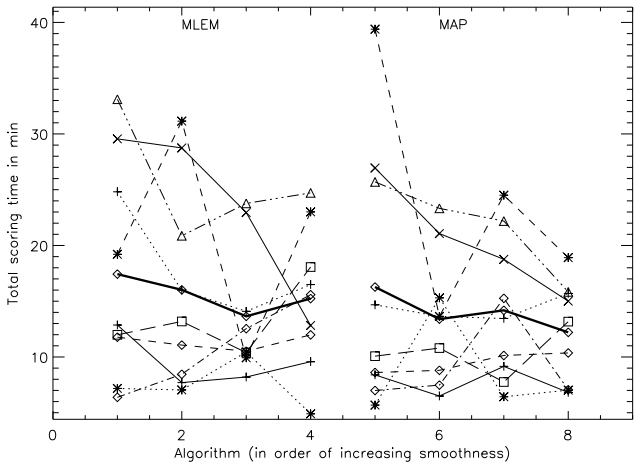

Figure 5. Human observer response times. Each individual observer result is plotted with different symbols and/or line styles, the thick lines represent the mean over all observers.

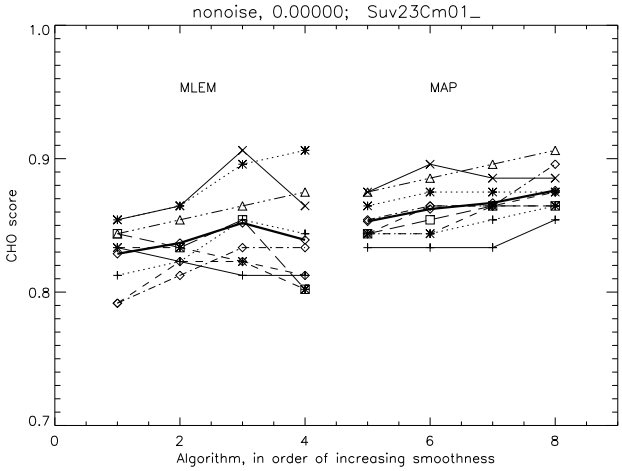

Figure 6. The score of the CHO observer obtained on the same nine data sets that were used in the human observer study, together with the mean performance (thick plain curve). The symbols and lines correspond to those in figures 3 and 5, so one can compare the CHO performance to that of the observer who scored the same data set.

the mean response time was longer for wrong answers than for correct ones. This indicates that shorter response times correspond to higher observer confidence.

B. First CHO study

The channelized Hotelling observer has been applied to the same datasets that were used in the human observer study. The resulting scores are shown in figure 6. The CHO detects between 80 and 90% of the lesions, which is higher than the average human observer score. However, the best human observer produced a score within the same range. Except for the shift to a better performance, the CHO curves are very similar to those of the human observer study, producing the same ranking. We therefore assumed that in this study, the CHO can be considered as a model for a very accurate human observer. Consequently, there was no need to improve the agreement with human observers by the addition of internal noise.

C. Second CHO study

Assuming that the CHO indeed predicts human performance, it was applied to an additional set of 5 reconstruction methods. For these reconstructions, the same 9 sets of noisy sinograms have been used, allowing paired comparison of all CHO-scores. Figure 7 shows the fraction of correct detections for all 7 reconstruction algorithms, as a function of the estimated spatial resolution of the reconstructed images.

This resolution estimate was obtained by fitting the ideal lesion, smoothed with a 2D Gaussian, to the "lesion response". The "lesion response" denotes the difference between the noise-free reconstructions with and without the lesion. The FWHM of the Gaussian that produced the best fit was used as a measure of resolution. For each reconstruction algorithm, the resolution was estimated for the 16 lesions, and the resulting mean, minimum and maximum FWHM-values are shown in fig 8. The impulse responses of the MAP-images are often asymmetrical, position dependent, and they have a shape that

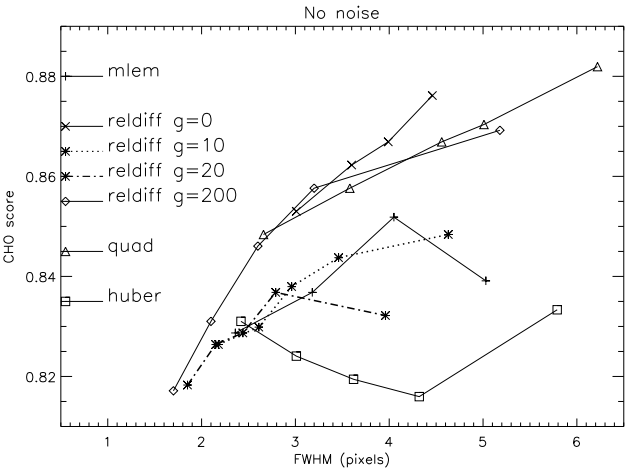

Figure 7. The CHO detection score as a function of the average resolution for the 7 different reconstruction algorithms. The curves for MLEM and MAP are those obtained in the first CHO study.

differs from the more compact Gaussian response of post-smoothed ML (see fig 2). This produced a large range of FWHM-values for the MAP algorithms. With post-smoothed ML, the resolution is nearly uniform, producing a much narrower range, and a mean value close to the FWHM of the Gaussian smoothing kernel, as can be seen in fig 8.

Fig 7 reveals similar CHO performance for MAP with the RD-prior and $\gamma = 0$ or $\gamma = 200$, and for MAP with the quadratic prior. Poorer performance is obtained for post-smoothed MLEM and for MAP with the edge-preserving priors (Huber prior and RD-prior with $\gamma = 10$ and $\gamma = 20$).

An alternative evaluation is the computation of the signal-to-noise ratio in every lesion position with eq (7). Averaging the signal-to-noise ratio over all lesion positions yielded a similar ranking as in fig 7 (results not shown).

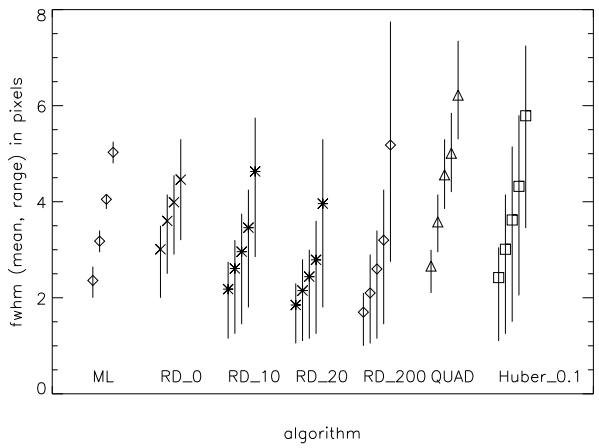

Figure 8. The average FWHM in mm for the seven algorithms: post-smoothed ML, MAP with RD-prior and γ = 0, 10, 20 and 200, MAP with quadratic prior and MAP with Huber prior. The symbols and line segments indicate the mean and the range of the FWHM for the 16 lesions in the image.

V. Discussion

We have done a similar comparison between post-smoothed MLEM and MAP earlier [14]. However, in that study, the task of the observers was made easier by avoiding the explicit learning phase. Instead, the lesion contrast was gradually decreased, which allowed the observers to learn about the image characteristics before detection became really hard. With that approach, no significant difference in observer performance was found. The current study design includes an explicit learning phase, which optimizes the observer performance for both methods. This was expected to improve the sensitivity of the experiment, and indeed revealed a (relatively small) advantage of MAP with the relative difference prior for lesion detection. With MAP, the observer response was also slightly shorter, indicating that the observer not only performed better, but perceived the detection task as slightly easier as well.

Apparently, with the new study design, the best average human performance shifted to a stronger regularisation for MAP, but not for post-smoothed MLEM. Because of this, the new experiment probably missed the maximum performance for MAP. Consequently, the true performance difference may be higher than what is suggested by this study.

Considering that the relative difference prior is similar to the more common quadratic prior, the superior performance with MAP was already predicted by the CHO experiments of Qi et al [19], [20]. Our own CHO experiment agreed with the human observer study as well. Several groups have proposed to add internal noise to the CHO to improve its agreement with human observers [12], [21]–[24]. In the current study, the CHO was "trained" using the 64 training images of a single session only. This yields inexact estimates for the mean and the covariance matrix of the channel outputs. This handicap was sufficient to reduce the CHO-performance to the range of the most performant human observers (we verified that a more accurate covariance matrix indeed yielded a higher CHO-performance, results not shown). When averaged over all 9 sessions, this CHO yielded very similar curves as the human observer studies, albeit shifted to a higher performance. For that reason, we have not added internal noise, avoiding the need for selecting a particular noise model. An alternative approach would be to model the internal noise as being proportional to the channel covariance matrix, similar to a model proposed in [24]. This would shift the CHO-detectability curves to a lower performance without changing the shape. Tuning of the constant of proportionality would then result in good agreement with the average human observer, rather than with the most performant ones.

Qi et al have [25] reported that CHO performance did not increase when MAP used an edge preserving prior rather than a quadratic prior. A similar result is obtained here. The Huber prior has a quadratic response for small differences and a linear response for large differences. The relative difference prior has a similar feature, which is tuned with its parameter γ. With intermediate values for γ it operates in two modes, as does the Huber prior. Similar to the Huber prior, it then yielded reduced CHO performance. With $\gamma = 0$ it operates always in quadratic mode, and yielded high CHO-performance, similar to that of the quadratic prior. With $\gamma = 200$, it operates (virtually) always in linear mode, and produced high CHO-performance as well. This suggests that it is the combination of different modes, rather than the actual modes themselves, that reduces the CHO performance. It is not clear to what extent this observation can be extrapolated to human observers.

With this experiment, only a limited range of regularisation has been studied. The CHO-scores of the quadratic prior, the RD-prior with $\gamma = 0$ and the RD-prior with $\gamma = 200$ were similar in figure 7. However, because none of these curves has reached its maximum, this experiment does not provide information about the best achievable detection performance of these methods. It is possible that these three curves will diverge for higher values of β. With very strong noise suppression, the edge preserving priors will operate in the quadratic mode nearly everywhere, so the curves of the edge preserving priors will converge to those of the corresponding priors without edge tolerance at high β values.

Based on this observer study, optimal reconstruction in clinical routine is not easily implemented. Indeed, the non-uniform point spread of the MAP reconstructions is undesirable for (semi-)quantitative analysis, e.g. based on maximum standardized uptake values or total lesion glycolysis [26]. Optimal processing would, at least for some applications, require two reconstructions per scan: one for optimal detection and another for optimal quantification. Unless a dedicated user interface would be created to deal with it, most physicians will probably prefer to avoid this solution. For that reason, it may be useful to find a good compromise between the two.

This study has several limitations. Only a single slice was studied, it cannot be excluded that for slices with different anatomy and for a volumetric analysis [27], [28] different results would be obtained. We studied lesions of a particular size and with a fixed intensity, while the relative performance of reconstruction algorithms may depend to some extent on the lesion size and intensity [20]. A Gaussian post-filter has been used because that is our current clinical practice. However, the local impulse response of the MAP algorithm has a different

shape [5]. Post-smoothing with a different smoothing kernel might affect the detection performance as well.

VI. Conclusion

The human observer study indicates that for lesion detection, MAP reconstruction with the relative difference prior outperforms MLEM with Gaussian post-smoothing, which is our current clinical method. Applying the channelized Hotelling observer to the same detection task yielded scores that agree well with those of the human observers.

The second channelized Hotelling observer study indicates that the use of edge-preserving priors may have an adverse effect on lesion detection.

References

[1] J Qi, RM Leahy. "Iterative reconstruction techniques in emission computed tomography" (topical review), *Phys Med Biol* vol 51, pp R541-R578, 2006.

[2] R Boellaard, A van Lingen, AA Lammertsma. "Experimental and clinical evaluation of iterative reconstruction (OSEM) in dynamic PET: Quantitative characteristics and effects on kinetic modeling", *J Nucl Med* vol 42, pp 808-817, 2001.

[3] L. Brèzes-Besset, J. Nuyts, R. Boellaard, I. Buvat, C. Michel, C. Pierre, N. Costes, A. Reilhac M13-217. "Simulation Based-Evaluation of NEG-ML Iterative Reconstruction of Low Count PET Data", *IEEE Nuclear Science Symposium*, M13-217, 2007.

[4] JW Stayman, JA Fessler, "Compensation for nonuniform resolution using penalized-likelihood reconstruction in space-variant imaging systems," *IEEE Trans Med Imaging*, vol 23, pp 269-84, 2004.

[5] J Nuyts, JA Fessler, "A penalized-likelihood image reconstruction method for emission tomography, compared to post-smoothed maximum-likelihood with matched spatial resolution." *IEEE Trans Med Imaging*, vol 22, pp 1042-52, 2003.

[6] J A Fessler. "Analytical approach to regularization design for isotropic spatial resolution." *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, vol. 3, pp. 2022-6, 2003.

[7] J Nuyts, D Bequé, P Dupont, L Mortelmans. "A concave prior penalizing relative differences for maximum-a-posteriori reconstruction in emission tomography." *IEEE Trans Nucl Sci*, vol 49, pp 56-60, 2002.

[8] LS Shepp, Y Vardi, "Maximum likelihood reconstruction for emission tomography," *IEEE Trans Med Imaging*, vol MI-1, pp. 113-122, 1982.

[9] EU Mumcuoğlu, RM Leahy, SR Cherry. "Bayesian reconstruction of PET images: methodology and performance analysis." *Phys Med Biol*, vol 41, pp 1777-1807, 1996.

[10] HH Barret, KJ Myers. "Foundations of Image Science", pp. 851-852 and 936-940, John Wiley & Sons, 2004.

[11] HC Gifford, MA King, DJ de Vries, EJ Soares. "Channelized Hotelling and human observer correlation for lesion detection in hepatic SPECT imaging.". *J Nucl Med* vol 41, pp. 514-521, 2000.

[12] HC Gifford, MA King, PH Pretorius, RG Wells. "A comparison of human and model observers in multislice LROC studies." *IEEE Trans Med Imaging*, vol 24, pp 160-169, 2005.

[13] AE Burgess. "Comparison of receiver operating characteristics and forced choice observer performance measurement methods", *Med Phys* 1995; 22: 643-655.

[14] J Nuyts, C Michel. "Performance of the relative difference prior for hot lesion detection in whole-body PET/CT: an evaluation with numerical and real observers". *IEEE Nuclear Science Symposium Conference Record*, vol 4, pp 2155 - 2159, 2005.

[15] SJ Lee, "Penalized likelihood image reconstruction for emission tomography using higher order convex-nonquadratic priors", *Proc. SPIE: Mathematical Modeling, Bayesian Estimation, and Inverse Problems*, vol. 3816, pp. 172-182, Jul. 1999.

[16] JA Fessler, H Erdoğan, WB Wu. "Exact distribution of edge-preserving MAP estimators for linear signal models with Gaussian measurement noise". *IEEE Trans Image Processing*, vol 9, pp 1049-1055, 2000.

[17] I-T Hsiao, A Rangarajan, G Gindi. "A new convex edge-preserving median prior with applications to tomography", *IEEE Trans Med Imaging*, vol 5, pp 580-585, 2003.

[18] S Alenius, U Ruotsalainen, "Bayesian image reconstruction for emission tomography based on median root prior", *Eur J Nucl Med*, vol 24, pp. 258-265, 1997.

[19] J Qi. "Comparison of statistical reconstructions with isotropic and anisotropic resolution in PET". *IEEE Trans Nucl Sci*, vol 53, pp 147-151, 2006.

[20] J Qi, RH Huesman, "Penalized maximum-likelihood image reconstruction for lesion detection". *Phys Med Biol*, vol 51, pp 4017-4029, 2006.

[21] KL Gilland, Y Qi et al. "Comparison of channelized Hotelling and human observers in determining the optimum OSEM reconstruction parameters for myocardial SPECT", *IEEE-MIC*, 2003, M11-287.

[22] J Oldan, S Kularni et al. "Channelized Hotelling and human observer study of optimal smoothing in SPECT MAP reconstruction", *IEEE Trans Nucl Sci* vol 51, pp.733 - 741 , 2004.

[23] PP Bruyant, HC Gifford, G Gindi, PH Pretorius, MA King. "Human and numerical observer studies of lesion detection in Ga-67 images obtained with MAP-EM reconstructions and anatomical priors." proceedings of *IEEE NSS-MIC, Rome 2004*, M10-320.

[24] CK Abbey, HH Barrett. "Human- and model-observer performance in ramp-spectrum noise: effects of regularization and object variability.". *J. Opt. Soc. Am. A*, vol 18, pp 472-488, 2001.

[25] J Qi. "Comparison of lesion detection and quantification in MAP reconstruction with Gaussian and non-Gaussian priors". *Int J Biomed Imaging*, vol 2006, Article ID 87567, 10 pages, 2006.

[26] YE Erdi, H Macapinlac, KE Rosenzweig, JL Humm, SM Larson, AK Erdi, ED Yorke. "Use of PET to monitor the response of lung cancer to radiation treatment." *Eur J Nucl Med* vol 27, pp 861-866, 2000.

[27] C Lartizien, P Kinahan, C Comtat. "Volumetric model and human observer comparisons of tumor detection for whole-body positron emission tomography", *Acad Radiol*, vol 11, pp 637-548, 2004.

[28] JS Kim, PE Kinahan, C Lartizien, C Comtat, TK Lewellen. "A comparison of planar versus volumetric numerical observers for detection task performance in whole-body PET imaging". *IEEE Trans Nucl Sci*, vol 51, pp 34-40, 2004.