Full 3D-OSEM reconstruction with compressed response of the system

J. López Herraiz[1], S. España[1], J. J. Vaquero[2], Member, IEEE, M. Desco[2] ,J. M. Udias[1]
[1] Dpto. Física Atómica, Molecular y Nuclear, Complutense University, Madrid, Spain
[2]Unidad de Medicina y Cirugía Experimental, Hospital GU "Gregorio Marañón", Madrid, Spain

***Abstract*—Small animal PET scanners require high spatial resolution (< 1 mm) and good sensitivity. To obtain high resolution images, iterative reconstruction methods, like OSEM, applied to image reconstruction in three-dimensional (3D) positron emission tomography (PET), have superior performance over analytical reconstruction algorithms like FBP. However, the high computational cost of iterative methods remains a serious drawback to their development and clinical routine use. The increase in performance of current computers should make iterative image reconstruction fast enough to attain clinical viability. However, dealing with the large number of probability coefficients for the response of the system in high-resolution PET scanners becomes a difficult task that prevents the algorithms from reaching peak performance. Taking into account all possible axial, in-plane and other symmetries, we have reduced the storage needs what allows us to store the whole response of the system in dynamic memory of ordinary industry standard computers, so that the reconstruction algorithm can achieve near peak performance.**

I. INTRODUCTION

THERE is a need for fast and accurate reconstruction software for high resolution and high sensitivity PET scanners. This kind of detectors, typically employed in small animal PET studies, are designed with the goal of optimizing spatial resolution while keeping good detection sensitivity. They consist of several opposite scintillation detectors, each with an array of small crystals, arranged in a static ring or in a rotating device. The detector ring diameter and the size of their field of view (FOV) is in general less than 20 cm of diameter. Because spatial resolutions of the order of 1 mm are required, many small crystals are required in these detectors and the number of lines of response (LORs) defined by a pair of crystals is very large. 3D acquisitions (and reconstruction) are mandatory due to the sensitivity requirements.

Statistical reconstruction methods [1,9] have shown superior image quality to any conventional analytic reconstruction techniques. Moreover, EM has some desirable properties as number of counts conservation, non-negativity, good linearity and dynamical range.

One of their key advantages lies on the ability to incorporate accurate models of the PET acquisition process through use of the system response matrix (SRM). However, SRM for 3D systems are of the order of several thousands of Megabytes in size, and this impose serious demands for statistical iterative methods in terms of the time required to complete the reconstruction procedure and the computer memory needed for the storage of the SRM.

The image process can be described as $y(i) = A(i,j).x(j)$, where $A(i,j)$ is the SRM, the vector x corresponds to the voxelized image and y to the measured data. Each element $\{A(i,j)\}$ is defined as the probability of detecting an annihilation event emitted from image pixel j by a detector pair i. This depends on various factors such as the solid angle subtended by this voxel to the detector element, the attenuation and scatter in the source volume and detector response characteristics. In the way we have written it, the reconstruction method is based upon a linear model. However, some dependence on x(j) can be included in A(i,j), for instance attenuation or scatter corrections depending on x(j), and then the model would be highly nonlinear.

Forward projection is the operation that estimates the projection data that corresponds to a given source activity distribution. It is given by the former expression: $y(i) = A(i,j).x(j)$.

Backward projection is the complementary operation to forward projection. It estimates a source volume distribution of activity from a given projection data. Let b(j) denote the element of the backward projection image. This operation corresponds to the expression: $b(j) = A(i,j).y(i)$. Both the forward and backward projection operations require the knowledge of the SRM.

Some implementations trade accuracy for speed by making approximations that do neglect some physical processes. All algorithms repeatedly use the forward and backward projection operations, which are the most time consuming parts of the iterative reconstruction programs. Many implementations simplify these operations to gain increments in speed, but the trade off is usually getting worse images.

Manuscript received October 28, 2004.
J.L. Herraiz acknowledges support from UCM grant.
e-mail: joaquin@nuc1.fis.ucm.es
S. España acknowledges support from "Fundacion para la investigación biomédica del Hospital Gregorio Marañon" grant.
e-mail: samuel@nuc1.fis.ucm.es
J.J.Vaquero e-mail: juanjo@dns.hggm.es
M. Desco e-mail: desco@mce.hggm.es
J.M. Udias e-mail: jose@nuc2.fis.ucm.es

Evaluation and storage of the SRM elements has been the main battlefield of large number of researchers in this area. Ideally, the SRM may be calculated somehow (e.g. using MC methods, or empirical data) and stored once and for all before the start of the reconstruction algorithm. In practice, time and memory requirements for doing this is prohibitive. A number of different methods have been proposed to create and handle a huge but sparse matrix like SRM.

Some implementations compute the elements {A(i,j)} on the fly, only as and when they are required [5]. This avoids the need to store the SRM, but the computational simplicity required by on-the-fly calculations often overlooks important effects. The SRM is very sparse and by using clever storage schemes and symmetries of the system [6] it can be kept on disk. This slows down the reconstructions considerably, as disk access is very slow and the SRM elements are used at every step. Other groups factorized the SRM as a product of independent contributions: geometry, attenuation, and detector sensitivity [7]. The factorization assumption however is not accurate and thus the reconstructed images are not the best possible ones.

We have employed a method to compress the SRM to allocate it in dynamical memory. This has proved to be a good choice as the reconstruction algorithm has achieved a sustained performance of around 50% of the theoretical peak computing capability of the processors.

II. .SYSTEM RESPONSE MATRIX [SRM]

The SRM is composed of all the nV x nL probability elements C_{VL} representing the probability to detect an event coming from voxel V at detector LOR (Line of Response) L. Forward and backward projection require the knowledge of all of them. This matrix depends on factors such as the solid angle subtended from voxel to detector element, the physics of beta decay, the attenuation and scatter in the source volume and detector response characteristics. For a reconstruction method to be accurate, these effects must be taken into account.

Storing all the elements of the SRM would require more than 10^4 Gbytes for a typical high resolution scanner. This exceeds the resources of any ordinary PC, so we must to get rid of all redundant elements and makes some approximations in order of being able to store the SRM in memory.

A. Sparseness of the matrix

As most of matrix elements are null, there is no need to save them. This reduces considerably the storage requirements. Using resolutions of 177x177x60, the average number of voxels connected with a LOR (or CHORD [7]) is about 4000 for a typical chord size of 150 x 7 x 4 voxels. In this case, storing just the 0.2% nonzero elements requires over 100 Gbytes, still too high.

B. System Symmetries

When the exact axial (translation and reflection) and in-plane symmetries are taken into account, a reduction factor of approximately 40 in the number of non null SRM elements that need to be stored can be obtained[7]. Storage needs can then be reduced to a few (say 5 for typical cases) GBytes, small enough to fit in hard-disks, yet too much for the RAM amount of ordinary industry standard PCs. We aim to reduce the problem size so that it would fit in less than 200 Mbytes, thus we have to be more aggressive in the symmetries taken into account.

C. Compressed SRM

The method that we propose here is to use additional non exact symmetries, or quasi-symmetries in order to allow for additional compression of the SRM. If we allow for relatively small differences, we can group *a priori* different LOR's within groups of the same quasi-symmetry class, where the differences in the elements of the SRM among LOR's beloging to each class are much smaller than among LOR's from different classes. Quasi-equivalent classes can be obtained, for instance, grouping together LOR's from crystals with different, but close, LOR-crystal orientations.

Depending on the geometry of the system, using quasi-exact symmetries, the number of non quasi-equivalent LOR's classes can be 5 to 20 times smaller than the number of exact symmetries classes, if tolerances of the order of 5 to 10% between corresponding elements of quasi-equivalent LOR's are accepted. Further reduction can be achieved if the relatively small differences inside the same quasi-symmetry class are taken into account by a factorized method inside the class.

D. Resolution Independence

We encode the SRM elements as transverse and longitudinal profile functions, in an almost resolution (voxel size) independent way, so that the same quasi-equivalent non zero elements can be used to build the SRM for a reasonable range of voxel sizes. We choose parameters so that SRM with resolutions varying in a factor 2 above or below the central resolution can be recovered.

Eventually, we end up with a compressed SRM that fits in less than 150 Mbytes. Depending of chosen resolution and system size, recovery of the SRM "on–the-fly" from the compressed one and matching the profile functions to the actual voxel size employed accounts for 10% to 30% of the total reconstruction time.

The compressed SRM can be computed with any means and stored once and for all. MC methods are in principle the best theoretical way of giving realistic estimates of SRM elements. In our case, we use our own MC model that includes scatter, positron range and non-colinearity effects. Enough simulated events are accumulated so that an statistical uncertainty of less than 5% is achieved at the center of typical LORs.

III. ITERATIVE IMAGE RECONSTRUCTION ALGORITHMS

Probably the most widely applied algorithm for finding the maximum-likelihood (ML) estimation of activity x given the projections y, is expectation-maximization (EM), first applied to the emission tomographic problem by Shepp and Vardi [3], though ML is a general statistical method, formulated as a general solution to any optimization problem.

Usually, iterative algorithms obtained from the ML statistical model assume that the data being reconstructed retain Poisson statistics. Preserving the Poisson statistical nature of data requires that any pre-corrections on the data made by the acquisition system are removed. The corrections for randoms, scatter and other effects should be incorporated into the reconstruction procedure itself, rather than being applied as pre-corrections. Uncorrected data in raw 3D-LOR histogram mode preserve Poisson statistics.

The slow convergence of EM is its greatest disadvantage [3]. This is due to the fact that the image is updated only after one full iteration, that is, after having projected and back-projected passing by all the LORs at least once. In the Ordered Subset EM (OSEM) algorithm, proposed by Hudson and Larkin [3], the image is updated more often, which has been proved to reduce the number of necessary iterations to achieve equivalent convergence to EM.

EM methods have also another important difficulty: noisy images are obtained from over-iterated reconstructions. This is due to: A) There is no stopping rule in this kind of iterative reconstructions. B) The statistical (noisy) nature of the detection process and reconstruction method.

Several techniques have been proposed to remedy for this, like filtering the image either after the whole reconstruction or along or in between iterations. Removal of noise from the data using wavelets methods has been proposed [4]. Maximum A Priori (MAP) algorithms are also widely used [2]. MAP adds *a priori* information during the reconstruction process. Typical assumption is that, due to the inherent finite resolution of the system, "the reconstructed image doesn't have any abrupt edge". Thus MAP methods apply a penalty function to those voxels that differs too much and/or too abruptly from its neighbors. Whether the maximum effective resolution achievable is reduced by the use these methods is still an open issue.

We define the following parameters and functions:

X_J - Value of voxel J (J=1..number of voxels)
X_J^n – Expected value of voxel J at iteration n
a_{IJ} – Probability that a photon emitted from voxel J were detected at detector I
P_I – Projection from the object measured at detector I (Experimental Data)
R_I^n – Projection estimated for the image reconstructed at iteration n

$$R_I^n = \sum_{J=1}^{Num.Voxels} a_{IJ} \cdot X_J^n$$

EM algorithm is defined as:

$$X_J^{n+1} = X_J^n \cdot \left[\frac{\sum_{I=1}^{Num.Detect} a_{IJ} \cdot \frac{P_I}{R_I^n}}{\sum_{I=1}^{Num.Detect} a_{IJ}} \right]$$

The MAP-OSEM algorithm is a modified version of EM:

$$X_J^{n,s+1} = X_J^{n,s} \cdot \left[\frac{\sum_{I \in SubSet\ S} a_{IJ} \cdot \frac{P_I}{R_I^n}}{\sum_{I \in SubSet\ S} a_{IJ} \cdot \left(1 + Penalty(J,n)\right)} \right]$$

In order to take into account scatter, it can be incorporated into the MAP-OSEM algorithm as follows [8]:

$$X_J^{n,s+1} = X_J^{n,s} \cdot \left[\frac{\sum_{I \in SubSet\ S} a_{IJ} \cdot \frac{P_I}{\left(R_I^n + S_I\right)}}{\sum_{I \in SubSet\ S} a_{IJ} \cdot \left(1 + Penalty(J,n)\right)} \right]$$

Simulation of scatter inside the object reveals this to be important for small animal scanners: about 30% of the data comes from scatter. An accurate modeling of scatter on the object during reconstruction will improve the image quality. In this work scatter in the object is evaluated assuming and isotropic and homogeneous model of the object.

IV. OPTIMIZATION TECHNIQUES FOR FULL 3D RECONSTRUCTION

We have designed an EM based reconstruction software (FIRST®) looking for the maximum flexibility compatible with the best performance. We implemented the accelerated version of EM (OS-EM). In our code, the number of subsets in each iteration can be chosen freely, not being limited by any system symmetry. Moreover, this number of subsets can be changed between iterations, even inside them.

Parallel computing on multiple processors is an attractive option to reduce computational time. New software protocols like the Message Passing Interface (MPI) and LAM-MPI enable a cluster of networked, independent industry standard PC's (Beowulf clusters) to be used together like a multi-processor unit.

FIRST® can be run in parallel in a Beowulf clusters of several CPUs in a master/slave implementation. There exists a Master-process and several (usually as many as CPU available) Slave-processes. The Master distributes the job between the Slaves ones and continuously balances the workload looking for the best performance taking into account differences in individual speed or workload for each CPU.

Always bearing in mind flexibility as a goal of design, FIRST® can work with variable image resolution that can even be changed during reconstruction, between or inside any iteration.

V. SIMULATION RESULTS

A. Test sets for evaluating the method

In order to test our implementation, before applying it on acquisitions from real scanners cases, we have reconstructed images from simulated projections using different phantoms (from uniform cylinders to more sophisticated test sets as the "Spiral-phantom") as original activity images. Events were generated from these tests sets using a MC method. Positron range and non-colinearity was taken into account for the emission. Nor attenuation neither scatter within the object was included. The response of the detector was also simulated considering the main physical effects. For each study 10 billion events were simulated and stored as projection data.

The scanner parameters chosen for these simulations were the ones of the ARGUS-drT small animal PET scanner [10]. It is a ring-type scanner with an 11.8 cm ring diameter, a transverse FOV of 6.8 cm and an axial FOV of 4.6 cm. It is based on a phoswhich scintillator depth-of-interaction technology with detector modules arranged in single/double rings 12. cm in diameter. The detector modules are comprised of a 13 x 13 array of crystals with 1.50 mm pitch size. The number of LORs in this scanner is over 2.8×10^7.

The images reconstructed from these simulations have a resolution of 175 x 175 x 60 voxels. The size of the phantoms and the images were chosen to be the same as the FOV of ARGUS-drT.

B. Evaluation of the method

As a first test, we have verified that using the compressed SRM and the uncompressed one produces images of the same quality and introduces no artifacts.

Secondly, an estimate of the PSF is obtained by using a phantom consisting of an array of small sources, located at different radial and axial positions (10 mm between them) were simulated. FWHM resolutions of 0.8 mm (at center of the scanner) to 1.0 mm (2.0 cm off axis) were obtained.

In order to study the linearity of the reconstruction method as well as, conservation of the number of counts and noise properties, the "Spiral phantom" was designed (Figure 1). It is comprised of three cylinders (background) each one with two spirals inside: a hot one (activity 4 times greater than the background) and a cold one (activity 4 times smaller).

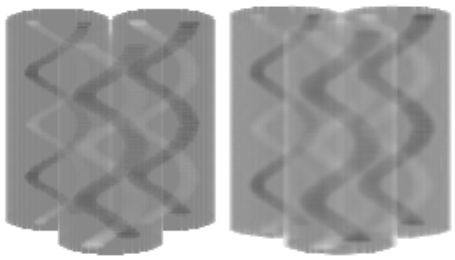

Fig. 1. Spiral-Phantom (left) and reconstructed image: MAP-OS-EM [3 Iterations, 50+50+50 Subsets] (right).

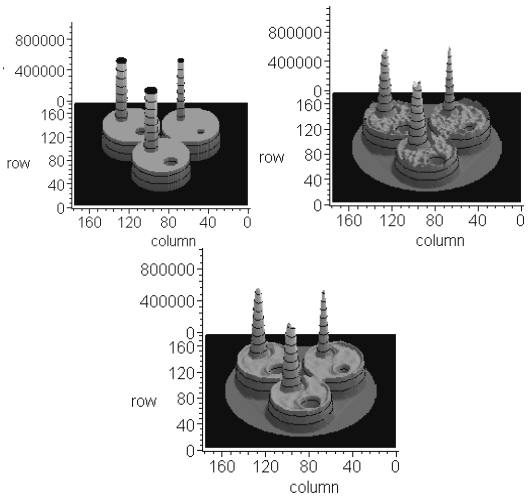

Fig. 2. 3D representation of a transverse section of the "Spiral- Phantom" and its reconstructed images. Z-Axis represents emitter activity. From left to right: True Spiral phantom, OS-EM 3 iterations (100 + 100 + 50 subsets) and MAP-OS-EM 3 iterations [100 + 100 + 50 subsets].

Figures 1-3 show the "Spiral-Phantom" and the reconstructed (OS-EM) and (MAP-OS-EM) images. MAP-OS-EM reconstructions have less noise and show no resolution degradation. In these figures the very linear response of the reconstruction method is noticed.

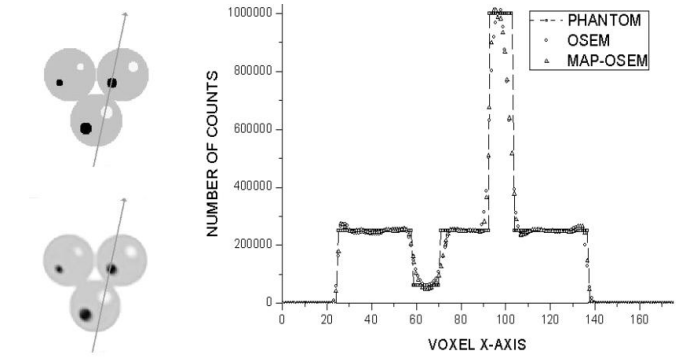

Fig. 3. Traces through the same "Spiral-phantom" study. Shown is emitter activity: True Phantom (solid line); OS-EM (circles) and MAP-OS-EM (triangles).

VI. RESULTS ON REAL DATA

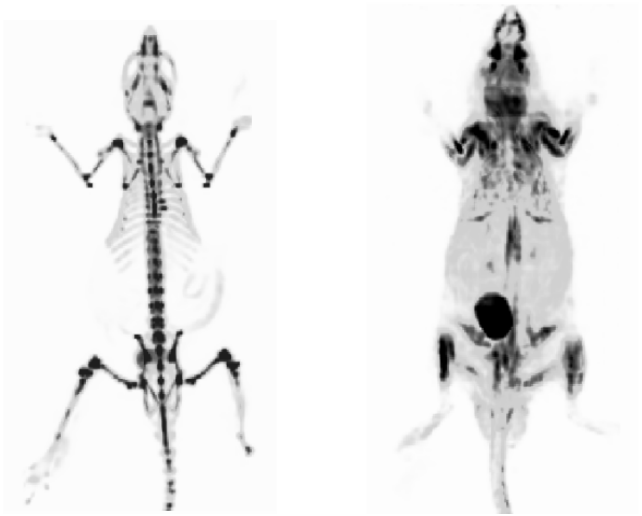

Fig. 5. Reconstructed images of 18-F (left) and FDG (right) injected mice body data acquired with an ARGUS [10] PET scanner.

Our reconstruction software was applied to real data. 18-F and FDG injected mice whole-body data were acquired with an ARGUS [10] PET scanner. Figure 5 shows the reconstructed images obtained using the OS-EM algorithm: 3 Iterations [50+50+50 SubSets]. The number of voxels is 175x175x168. Resolution obtained from real acquisition is around 1 mm what can be clearly appreciated in the images shown.

Typical reconstruction time from real data is 1 hour for a full iteration using 1 CPU (Opteron Dual 244, 1800 GHz, 2Gb RAM). Reconstruction time nearly scales with the product of the number of LORs [2.8×10^7] times the number of voxels in a LOR [4000]. The parallel version of FIRST® reduces this time by the number of CPUs available.

VII. DISCUSSION & CONCLUSIONS

We have implemented FIRST, a fully 3D-OS-EM and 3D-OS-EM-MAP with a compressed SRM with all the resolution recovery properties of EM. Its flexibility, reconstruction time and the accuracy and resolution of reconstructed images can make FIRST to become an important tool in real clinical studies of high resolution small animal PET scanners.

VIII. ACKNOWLEDGMENT

The authors thank Dr. Martin G. Pomper and James Fox from the Johns Hopkins University School of Medicine for providing access to the ARGUS data.

IX. REFERENCES

[1] L. A. Shepp and Y. Vardi, "Maximum-likelihood reconstruction for emission tomography," *IEEE Trans. Med. Imag.*, vol. MI-1, pp. 113–121, 1982.

[2] P. J. Green, "Bayesian reconstructions from emission tomography data using a modified EM algorithm," *IEEE Trans. Med. Imag.*, vol. 9, pp. 84–93, Feb 1990.

[3] H. M. Hudson and R. S. Larkin, "Accelerated image reconstruction using ordered subsets of projection data," *IEEE Trans. Med. Imag.*, vol. 13, pp. 601–609, Aug. 1994.

[4] B. A. Mair, R. B. Carroll and J. M. M. Anderson, "Filter Banks and the EM algorithm," Nuclear Science Symposium, 1996, Conference Record, 1996 IEEE, vol. 3, 2-9 Nov. 1996.

[5] H. Kudrolli, W. Worstell and V. Zavarzin – "SS3D-Fast fully 3D PET iterative reconstruction using stochastic sampling." IEEE Trans. Nuc. Sci., vol. 49, issue 1, pp. 124-130, Feb 2002

[6] Jinyi Qi, Richard M Leathy, Simon R Cherry, Arion Chatziioannou and Thomas H Farquhar, "High-resolution 3D Bayesian image reconstruction using the microPET small-animal scanner" Phys. Med. Biol. 43 (1998) 1001-1013

[7] Calvin A. Johnson, Ariela Sofer, "A Data-Parallel Algorithm for Tomographic Image Recontruction" *Proc.7th Symposium on the Frontiers of Massively Parallel Computation* IEEE Computer Society Press, 126-137 Washington (1999).

[8] Robert M. Lewitt and Samuel Matej, "Overview of methods for image reconstruction from projections in emission computed tomography" Proc. IEEE, vol 91,nº 10, Oct. 2003

[9] J. Browne and Alvaro R. De Pierro, "A row-action alternative to the EM algorithm for maximizing likelihoods in emission tomography" IEEE Trans. Med. Imag., vol 15, no. 5, Oct 1996

[10] J. J. Vaquero, J. Pascau, A. Molins, J. M. Arco, M. Desco, "Performance characteristics of the ARGUS-drT small animal PET scanner: preliminary results." Medical Imaging Conference, Rome 2004, Conference Record, preprint.