

Massively parallelizable list-mode reconstruction using a Monte Carlo-based elliptical Gaussian model

G. Sportelli^{a)}

Biomedical Image Technologies Group, ETSI de Telecomunicación, Universidad Politécnica de Madrid, E-28040, Madrid, Spain; CIBER de Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), E-28040, Madrid, Spain; and Istituto Nazionale di Fisica Nucleare, Sezione di Pisa I-56127, Pisa, Italy

J. E. Ortuño

Biomedical Image Technologies Group, ETSI de Telecomunicación, Universidad Politécnica de Madrid, E-28040, Madrid, Spain and CIBER de Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), E-28040, Madrid, Spain

J. J. Vaquero and M. Desco

Departamento de Bioingeniería e Ingeniería Aeroespacial, Universidad Carlos III de Madrid, E-28911, Madrid and Instituto de Investigación Sanitaria Gregorio Marañón E-28007, Madrid, Spain

A. Santos

Biomedical Image Technologies Group, ETSI de Telecomunicación, Universidad Politécnica de Madrid, E-28040, Madrid, Spain and CIBER de Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN), E-28040, Madrid, Spain

(Received 19 March 2012; revised 22 November 2012; accepted for publication 27 November 2012; published 27 December 2012)

Purpose: A fully three-dimensional (3D) massively parallelizable list-mode ordered-subsets expectation-maximization (LM-OSEM) reconstruction algorithm has been developed for high-resolution PET cameras. System response probabilities are calculated online from a set of parameters derived from Monte Carlo simulations. The shape of a system response for a given line of response (LOR) has been shown to be asymmetrical around the LOR. This work has been focused on the development of efficient region-search techniques to sample the system response probabilities, which are suitable for asymmetric kernel models, including elliptical Gaussian models that allow for high accuracy and high parallelization efficiency. The novel region-search scheme using variable kernel models is applied in the proposed PET reconstruction algorithm.

Methods: A novel region-search technique has been used to sample the probability density function in correspondence with a small dynamic subset of the field of view that constitutes the region of response (ROR). The ROR is identified around the LOR by searching for any voxel within a dynamically calculated contour. The contour condition is currently defined as a fixed threshold over the posterior probability, and arbitrary kernel models can be applied using a numerical approach. The processing of the LORs is distributed in batches among the available computing devices, then, individual LORs are processed within different processing units. In this way, both multicore and multiple many-core processing units can be efficiently exploited. Tests have been conducted with probability models that take into account the noncolinearity, positron range, and crystal penetration effects, that produced tubes of response with varying elliptical sections whose axes were a function of the crystal's thickness and angle of incidence of the given LOR. The algorithm treats the probability model as a 3D scalar field defined within a reference system aligned with the ideal LOR.

Results: This new technique provides superior image quality in terms of signal-to-noise ratio as compared with the histogram-mode method based on precomputed system matrices available for a commercial small animal scanner. Reconstruction times can be kept low with the use of multicore, many-core architectures, including multiple graphic processing units.

Conclusions: A highly parallelizable LM reconstruction method has been proposed based on Monte Carlo simulations and new parallelization techniques aimed at improving the reconstruction speed and the image signal-to-noise of a given OSEM algorithm. The method has been validated using simulated and real phantoms. A special advantage of the new method is the possibility of defining dynamically the cut-off threshold over the calculated probabilities thus allowing for a direct control on the trade-off between speed and quality during the reconstruction. © 2013 American Association of Physicists in Medicine. [<http://dx.doi.org/10.1118/1.4771936>]

Key words: list-mode reconstruction, PET, kernel model, GPU reconstruction

I. INTRODUCTION

List-mode (LM) reconstruction has long been a promising method for PET imaging.^{1–3} In high-resolution tomographs, especially in dynamic PET studies with low-statistics three-dimensional (3D) scans, the number of events acquired in a histogram set may be less than the number of response bins used in histogram mode (HM) reconstruction and, therefore, LM reconstruction can, in principle, be performed more quickly and efficiently. Another advantage of LM reconstruction comes from the ease of including additional information acquired by the PET scanner (i.e., photon energy, arrival time, or full detector readout), which increases the accuracy of the system model used in the reconstruction process. Thus, the image signal-to-noise ratio can be improved significantly if the reconstruction algorithm uses a per-event energy⁴ or time-of-flight (TOF) information.^{5–7}

A PET system response is described using a system matrix (SM) that maps the relationships between the 3D radionuclide distribution and the acquired data.⁸ An enormous effort has been expended to provide accurate estimations of statistical and physical effects involved in the system response. Many effects involved are object dependent, such as attenuation and scatter effects, positron range, and gamma-rays non-collinearity, while other effects are system dependent, and can be characterized based on the PET scanner geometry, detection physics, and associated electronics. The methods used to obtain accurate estimations of the system response include Monte Carlo (MC) simulations of the acquisition process,^{9–12} point source measurements,^{13,14} theoretical models,^{15–18} or hybrid solutions.¹⁹

When the SM is calculated using MC methods, real measurements, or complex numerical approximations, it must be precomputed and stored off line to keep the reconstruction time low, thus requiring enormous storage space for 3D PET imaging. Size reduction is provided by histogram compression, axial and rotational symmetries,^{20,21} polar-voxel symmetries,^{22,23} quasi-symmetries,⁸ axial mashing,²⁴ and factorization as a product of sparse matrices.²⁵

It has been shown that MC calculated SM can be stored in programmable graphic processing units (GPUs), which have typically very limited memory resources. One possibility is to use approximated symmetries,²⁶ so as to reduce the size of the SM by two orders of magnitudes. A different approach has been also proposed,²⁷ that factorizes the system response into geometrical, voxel blurring, and detector blurring components. However, compression can still be insufficient for LM reconstruction in GPUs, where part of the memory must be dedicated to the input LOR dataset, and when the SM must take into account both timing and energy information.

In addition to this, the histogram compression technique reduces the SM dimensionality. This drawback implies that the SM factorization scheme can only be performed in the image space and not in the projection space. Thus, SM factorization suffers from accuracy limits in modeling projection-dependent blurring effects when applied to LM data.¹⁵

For system matrices calculated on the fly, several techniques have been adopted to reduce the computational com-

plexity, ranging from Siddon's ray-tracing model²⁸ to tube-shaped kernels²⁹ and volumes of response.¹⁸ Recently, it has been shown that, using GPUs, it is possible to achieve good results with the Gaussian blurring approximation for times compatible with practical environments,³⁰ by using the concept of the symmetric blurring kernel.^{2,31,32} Kernels represent the functions that associate a projection weight to a voxel for a given line of response (LOR), based on the relative position of the voxel with respect to the LOR, the crystal's efficiency, the photon's energy, the depth of interaction, and TOF information.⁷ A symmetric Gaussian kernel was used in the work cited above.⁷ Although more complex kernel models have been proposed, the inherent reconstruction architecture has been specially optimized for circular tubes of response (TORs). However, the shape of a system response for a given LOR has been shown to be asymmetrical around the LOR. Its transversal profile can be better approximated as a two-dimensional (2D) Gaussian function, with the two standard deviation variables of the crystal's thickness and angle of incidence.^{16,17} More accurate image-based approximations can be achieved using nonstationary and non-Gaussian blurring functions in the image domain.^{15,33}

In this work, we introduce a LM reconstruction method that is especially optimized for PET scanners composed of parallel planar detectors^{34–36} that is able to use kernel models with an elliptical section based on MC simulations.

Processing parallelization was the main area of study with this work. This is because replacing circular kernels with elliptical kernels hampers the efficient utilization of the techniques previously used to accelerate the calculation of the system probabilities. Therefore, a new scheme had to be implemented to increase the parallelization efficiency.

II. MATERIALS AND METHODS

II.A. The reconstruction algorithm

The LM 3D-ordered-subsets expectation-maximization (OSEM) algorithm used in this work is based on that described in Ref. 37. It consists of the following iterative process:

$$\lambda_j^{k+1} = \frac{\lambda_j^k}{s_j} \sum_{i \in L_n} \frac{a_{ij}}{r_i + t_i + \sum_{j'} a_{ij'} \lambda_{j'}^k}. \quad (1)$$

All the observed coincidence events, or LORs, are divided into N subsets L_n ; λ_j^k is the estimated intensity of voxel j at subiteration $k = hN + n$, where n is the subset number and h is the iteration number; s is the sensitivity of the image, and a_{ij} is the likelihood that an emission from voxel j is detected by LOR i .

In the current implementation, random coincidences estimates (r_i) and scatter estimates (t_i) are neglected. To prevent reconstruction biases, LM subsets were composed by sampling uniformly the entire acquired dataset with offsets $n \in [0, N - 1]$. The sensitivity image s was obtained by projecting all possible LORs. In the case of rotational geometries, the computation can be relatively time consuming, but nonuniform sampling techniques,³⁸ rotational

symmetries, and parallel computation of different angular directions are planned to be used to accelerate the process.

The concept of *kernel*^{2,30,32} was used to model the system response. The blurring effect was obtained by sampling the scalar field defined by the kernel model in the voxel positions close to each LOR. The volume of the sampled ROR was regulated by using a constant cut-off peripheral probability. Lower cut-off values produced larger volumes, thus reducing the noise and slowing the reconstruction.

In Secs. II.B–II.F, different kernel models will be defined and the sampling method will be described. Both the accuracy and the reconstruction times could be regulated by selecting the level of approximation of the kernel and the sampling cut-off.

II.B. The kernel space frame

The projector kernel, i.e., the scalar field that describes the TOR, was derived with respect to a Cartesian reference system, in which the ideal LOR is aligned with one of the axes. We will call this reference system the *kernel space frame* and we will refer to it using coordinates x' , y' , and z' , in contrast with the *image space frame*, whose coordinates x , y , and z are aligned with the axes of the image.

For any LOR, a transformation matrix is used to convert the coordinates of each voxel center $\mathbf{p} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$ to the new reference frame $\mathbf{p}' = x'\mathbf{i}' + y'\mathbf{j}' + z'\mathbf{k}'$, where \mathbf{i}' is aligned with the projection on the detector surface of the vector joining the two crystal centers, \mathbf{k}' is aligned with the LOR, and \mathbf{j}' is orthogonal to \mathbf{i}' and \mathbf{k}' .

II.C. Kernel models

Our simplest PDF approximation was a TOR whose section is a 2D circular Gaussian function with constant standard deviation, σ . We refer to this as a circular constant Gaussian kernel or shortly *0D kernel*,

$$a_{ij} = A \exp\left(-\frac{d_{ij}^2}{2\sigma^2}\right), \quad (2)$$

where a_{ij} is the same as in (1), d_{ij} is the distance between the center of voxel j (x', y', z') and the LOR i (i.e., axis z'), and A is the scale factor. The standard deviation σ was derived in our experiments as the weighted mean of all standard deviations observed in the MC simulations, being weighted proportional to the LOR probability.

A better PDF approximation consists of TORs with circular Gaussian sections of variable standard deviation σ_i , a function of the angle φ between the LOR i and the normal to the planar detectors (Fig. 1),

$$a_{ij} = A \exp\left(-\frac{d_{ij}^2}{2\sigma_i^2}\right), \quad (3)$$

where a_{ij} is the same as in (1) and d_{ij} is the same as in (2). We refer to this as a circular varying kernel or *1D kernel*. A more accurate approximation of the PDF is a TOR with an elliptical Gaussian section parameterized with two separated standard

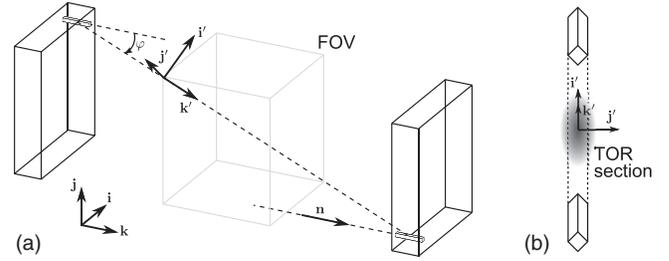


FIG. 1. Illustration of the global (x, y, z) reference system and the one local to the LOR (x', y', z'). A point $\mathbf{p} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$ in the global reference system corresponds to $\mathbf{p}' = x'\mathbf{i}' + y'\mathbf{j}' + z'\mathbf{k}'$ in the local one. In (a) φ is the angle between the LOR and the normal \mathbf{n} to the detectors. The schematic in (b) illustrates the alignment between the new reference system and the crystals. The TOR section represents the probability distribution projected on the plane orthogonal to the LOR.

deviations, $\sigma_{i,x'}$ and $\sigma_{i,y'}$, which are functions of the angle φ between the LOR i and the normal to the planar detectors. The expression for the elliptical varying kernel (*2D kernel*) is

$$a_{ij} = A \exp\left(-\frac{x'^2}{2\sigma_{i,x'}^2} - \frac{y'^2}{2\sigma_{i,y'}^2}\right), \quad (4)$$

where (x', y') are the coordinates in the kernel space frame and a_{ij} is the same as in (1). It is important to note that the proposed reference system simplifies the general 2D Gaussian function to a separable elliptical Gaussian function. Moreover, although there is a one-to-one correspondence between the value of the Gaussian function and the distance from the center in the 1D kernel, this is not true for 2D kernels. This means that to neglect uniformly low-likelihood voxels, the cut-off condition must be expressed and tested directly against a_{ij} .

II.D. MC simulations

Several representative LORs were simulated using the MC method and the obtained probability distributions were fitted to our proposed numerical kernels. LORs were simulated with incident angles φ (shown in Fig. 1) regularly distributed between the minimum and maximum possible values. In the adopted geometry, φ was directly related to the discrete distances expressed as the coordinate differences between coincident crystals $c_{x,1} - c_{x,2}$ and $c_{y,1} - c_{y,2}$, where $c_{x,1}$, $c_{y,1}$, $c_{x,2}$, and $c_{y,2}$ are the x and y discrete crystal coordinates in the first and second coincident detectors, respectively. These differences were used as the indices of a look-up table of standard deviations accessed during the forward and back projections.

To improve the simulation speed, a preliminary simulation was launched to obtain the minimum cylinder that contained all the voxels whose detection probability was non-negligible. As the generation density is constant in the cylinder and the solid angle of emission is constant, there was no bias in the MC simulations. Both the positron generation region and the direction of annihilated gamma rays were then constrained to the obtained cylinder.

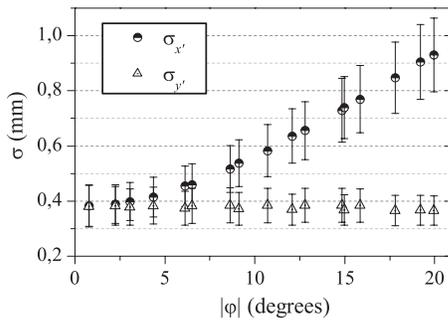


FIG. 2. $\sigma_{i,x'}$ and $\sigma_{i,y'}$ fitted values of representative LORs from the MC simulations as a function of the absolute value of angle φ between the LOR and the vector normal to the detector surface.

The custom MC method incorporates the effects of intercrystal scatter, crystal penetration, positron range, and photon-pair noncollinearity in the same way it has been already done for SM-based reconstructions.¹¹

The positron range was modeled as the sum of two exponentials and was simulated using an accept-reject algorithm.³⁹ Noncollinearity was modeled using a Gaussian distribution⁴⁰ and the intercrystal scatter followed the Klein-Nishina distribution sampled using a double-rejection technique.⁴¹

The coordinates of the positron sources associated with the detected coincidences between the selected pair of scintillating crystals were stored on a disk for further evaluation. The kernel space frame transformation was applied to obtain a source distribution aligned with the z' axis. The histogram of the projections onto the $x' - y'$ plane were fitted to our proposed 2D kernel model and averaged along z' to derive the parameters $\sigma_{i,x'}$ and $\sigma_{i,y'}$ of representative LORs Fig. 2. The distribution variability along axis z' was not considered during the reconstruction. Isotropic generation of gamma rays included in the MC method guarantees the adjustment for TOR solid angle.

The value of $\sigma_{i,x'}$ showed to depend on the absolute value of the angle φ associated with the LOR, while $\sigma_{i,y'}$ remained almost constant. The standard deviation of the 1D was calculated as $\sigma_i = \sqrt{\sigma_{i,x'}\sigma_{i,y'}}$ for all LOR i . The standard deviation of the 0D kernel was chosen as the mean value of σ_i over all the possible LORs.

II.E. Identification of the region of response

For a given model, forward and backprojections were performed by sampling the kernel at the centers of the voxels of the ROR. The kernels can then be seen as being the body of the innermost loop over such voxels. Therefore, their optimization is of paramount importance for the overall reconstruction performance, as well as the optimization of the looping technique.

To compute kernels efficiently, we focused on using the least number of operations possible. In addition, to parallelize the computation to a large degree, it is important to use few memory resources. To accomplish these requirements, a specific FOV sampling system was developed that was able to

process only the surroundings of each LOR within a region, defined using a dynamically calculated peripheral threshold.

As was carried out in Ref. 30, in our method, we divided the volume in slices and identified the voxel closest to the LOR in each slice. In addition, we identified the region by processing recursively the neighbors of each voxel, instead of looping around the center within a fixed, predefined range. In this way, it was possible to tailor efficiently the target region for any peripheral condition, calculated on line, and to process regions that were different from circular TORs, as in presence of TOF-capable scanners or scatter-corrected algorithms.

The voxel identification algorithm was based on a hierarchical search approach, in which each identified voxel corresponds to a node of a tree that represents the ROR. In the search algorithm, the voxels are divided into three partitioning levels, each referring to one of the three dimensions of the image. The first level, referred to as the primary level, is obtained by selecting the closest voxel to the LOR for each plane of the FOV orthogonal to the main direction of the LOR. The main direction is characterized by the greatest difference in coordinates between the two vertices of the LOR. We refer to this as the primary direction (Fig. 3, up-left). The selected (primary) voxels represent the center of propagation for voxels of inferior levels. The next (secondary) level is obtained by choosing a different axis of propagation. In Fig. 3 (up-right side), the secondary direction corresponds to axis j . In a tree-like scheme, each secondary voxel is obtained using a straight propagation in the positive and negative directions from a primary voxel, and it is a child of that voxel. By analogy with the secondary voxels, the tertiary voxels are obtained from a straight propagation of the secondary voxels. By propagating each primary voxel in the secondary and tertiary directions, up to the contour calculated online for a given kernel and cut-off condition, it is possible to achieve a ROR that contains the most significant part of the PDF relative to a given LOR (Fig. 3, down).

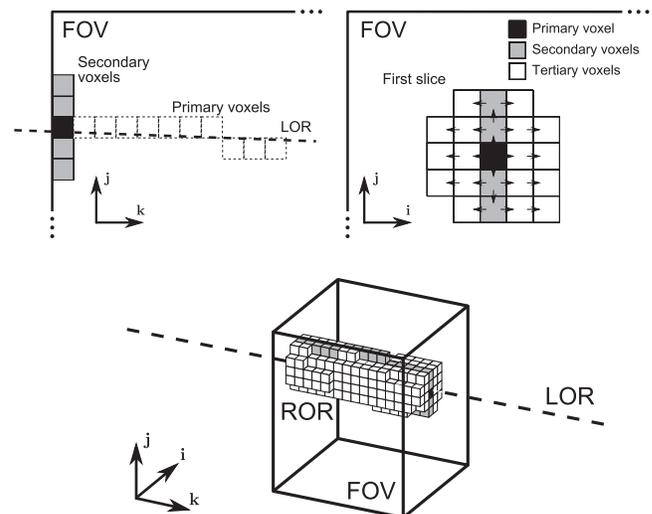


FIG. 3. (Top) Geometric diagrams of the ROR propagation technique. The primary direction is along axis k , the secondary direction is along axis j , and the tertiary direction is along axis i . (Bottom) 3D diagram of the full identified ROR. A FOV of $15 \times 15 \times 18$ voxels has been used as an example.

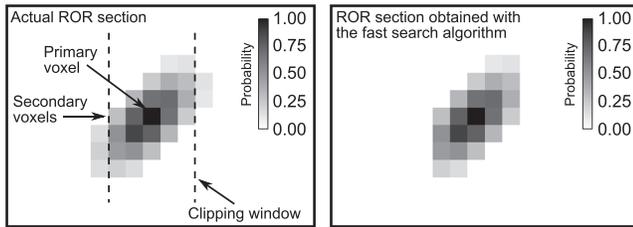


FIG. 4. Clipping effect for the most oblique LOR using an elliptical Gaussian projector model and a cutoff threshold of 10% of the Gaussian peak.

In the current implementation, axis \mathbf{k} was always taken as the primary direction, axis \mathbf{j} as secondary direction, and axis \mathbf{i} as the tertiary direction. The importance of choosing the best secondary and tertiary axes for each independent LOR increases for highly eccentric tubes and with high (less accurate) cutoff thresholds. This is because the search algorithm neglects tertiary voxels whose probability is over threshold, but which are not reachable because the parent secondary voxel has been discarded. This behavior can be imagined as if a clipping window were applied to each ROR section, centered on the primary voxel, and as large as the total number of secondary voxels over the threshold. The resulting clipping effect, illustrated in Fig. 4, is discussed in Sec. IV.D.

In the hierarchical structure, identifying the entire ROR corresponds to performing a depth-first traversal. Our proposed ROR identification scheme has the important characteristic of processing each voxel in the surroundings of the LOR only once, without requiring any memory write operation, except the recursive update of current voxel coordinates during straight propagations. Voxel coordinates can be either cached and used for both forward and back projections, or used directly during the discovery process to access the FOV. In the latter case, ROR identification must be performed twice per LOR in OSEM reconstruction.

Using this new technique, no voxels were processed unless they contributed significantly to the projections, or belonged to the region contour that was cutoff. The propagation scheme requires only integer operations and a few scalars to be stored in the memory to control the peripheral conditions.

II.F. Parallelization strategies

Processing parallelization is a necessary step to make LM reconstruction feasible using modern computing resources. The pioneering approach of Ref. 7 has been shown to be effective, and was the first applicable in practice for LM reconstruction in massive parallel processing units. The above work resolved the problem of write synchronization and data dependence by dividing the FOV into slices and processing each in a different GPU core. Because all the parallel processors are busy processing the same LOR, we refer to this method as *intra-LOR* parallelization. With the increasing number of processing cores available in modern graphics cards and the decreasing number of target slices in TOF-enabled reconstruction, the question arises of what happens when the number of cores exceeds the number of slices. In fact, no strategies have

TABLE I. Spatial resolution FWHM measured on a point source at the center and at 20 mm from the center in the transaxial plane.

Position	Radial (mm)	Tangential (mm)	Axial (mm)
Center	1.1	1.1	0.9
20 mm from the center	1.2	1.1	1.0

been described to employ such exceeding cores. Moreover, the *intra-LOR* parallelization technique imposes an overhead on the number of voxels to be processed that may limit its efficiency for more eccentric elliptical TORs.

To respond to this situation, we propose using two new parallelization techniques that operate at the *inter-LOR* level, i.e., allow us to process multiple lines of response in parallel. The first technique allows for flexible work parallelization, under the conditions of medium/high granularity, the latter being the ratio between the computational size of the parallelized blocks, and the size of the full process itself. This coarse-grained technique consists of grouping several LORs into different processing blocks, each with a private memory space. The high granularity is obtained by using sufficiently large blocks. Parallelization is then achieved using a parallel reduction pattern,⁴² applied using the (1) in the following form:

$$\begin{aligned} \lambda_j^{k+1} &= \sum_{i \in L_n} \frac{\lambda_j^k}{s_j} \frac{a_{ij}}{r_i + t_i + \sum_{j'} a_{ij'} \lambda_{j'}^k} \\ &= \sum_{p=0}^{\Pi-1} \sum_{i \in L_{n,p}} \frac{\lambda_j^k}{s_j} \frac{a_{ij}}{r_i + t_i + \sum_{j'} a_{ij'} \lambda_{j'}^k} = \sum_{p=0}^{\Pi-1} \lambda_{j,p}^{k+1}, \end{aligned} \quad (5)$$

where Π is the number of available processors and $\lambda_{j,p}^{k+1}$ is the partial result obtained by reconstructing only the p th block $L_{n,p}$. All the partials $\lambda_{j,p}^{k+1}$ for $p \in (0, \dots, \Pi - 1)$ are independent of each other and, therefore, they can be computed in parallel. The additional summatory function required in Eq. (5), referred to as a reduction operation, introduces an overhead on the entire reconstruction process that is negligible under the assumption of medium/high granularity, which is true in the case of dividing the input LM file into a number of blocks of the same order, but higher than the number of cores present in a multicore CPU. In the current version, software parallelization was implemented using the Intel Threading Building Blocks library.⁴³ The process of splitting the LM subsets and reducing the results can be carried out dynamically to equalize workloads. As a result, the acceleration factor is practically equal to the number of processors available, Π . In principle, the same technique can be used in CPU clusters. The difference would be that the reduction operation, i.e., the sum of all the voxels in the image, has to take place through Ethernet connections, which are typically slower than a system main memory, thus resulting in a time overhead per processing unit greater than in the case of multicore parallelization.

Although the above parallelization technique is efficient and flexible, it requires additional memory allocation for each vector $\lambda_{j,p}^{k+1}$. This is not a problem for computing clusters and

multicore CPUs, but it is not practically feasible if we wish to distribute the computation among GPU cores where the available memory per core is generally limited to hundreds of kilobytes. To overcome this situation, we used a second, fine-grained parallelization approach to accelerate the computation of each single block $L_{n,p}$. The work subdivision is performed in the same way as in Eq. (5), but using a shared memory space. In this case, the problem is that multiple cores can write to the same memory location at the same time, thus incurring write-after-write race conditions. This problem can be resolved by using thread-safe memory structures, at the cost of a severely degraded write speed or to use atomic operations, today available in most parallel computing devices and GPUs.

The second parallelization level was implemented on an NVIDIA graphics adapter able to provide atomic operations using OpenCL 1.1.⁴⁴ At the present state, dynamic workload control has not been implemented, so the work distribution has been decided previously based on the size of the reconstructed list-mode data set.

III. EXPERIMENTAL VALIDATION

III.A. Scanner geometry

Our method is generalizable for all scanner geometries, although a specific implementation has been developed for scanners with parallel planar detectors with pixelated scintillating crystals. The rPET scanner^{35,45} used in our validations is a small animal model consisting of two pairs of planar detector modules in the coincidence mode, with 160 mm separation distance between opposing modules. Each detector is composed of a 30×35 array of $1.5 \times 1.5 \times 12.0$ mm³ LSO

pixelated crystals, assembled on a 100 μ m thick matrix of a plastic reflector. The detectors are mounted on a gantry rotating in the continuous mode with a 180° span.

III.B. Reconstructed phantoms

The image quality phantom defined in the NEMA standard for small-animal PET scanners⁴⁶ was modeled using the MC simulation toolkit GATE,⁴⁷ and up to 5.1×10^7 coincidence events were stored in a LM file. This experiment was designed to evaluate the noise characteristics and recovery coefficients of the proposed algorithms. In accordance with,⁴⁶ the recovery coefficients (RCs) are evaluated for the image slices covering the central 10 mm length of the rods, which were averaged to obtain a single image slice with lower noise. Circular ROIs with diameters of 2 and 4 mm were drawn around the smaller rods with diameters of 1 and 2 mm, respectively. In both ROIs, the maximum value was measured and divided by the mean activity concentration to determine the RCs for both rods. The mean activity concentration was evaluated for a 22.5 mm diameter using a 10 mm long cylindrical ROI, centered on the uniform region. The noise-to-signal ratio was calculated as the standard deviation of a circular ROI divided by its mean value.

In order to assess the actual performance of the method, real NEMA Image Quality and Derenzo phantom were acquired using the rPET small animal scanner. A ²²Na point source was also acquired and reconstructed using the 2D model for 12 iterations with 10 subsets. The diameter of the source was 0.25 mm. The size of the epoxy capsule containing the source was 1 cm³. The point source was positioned at the center of the field of view and at 20 mm from the center in the transaxial plane.

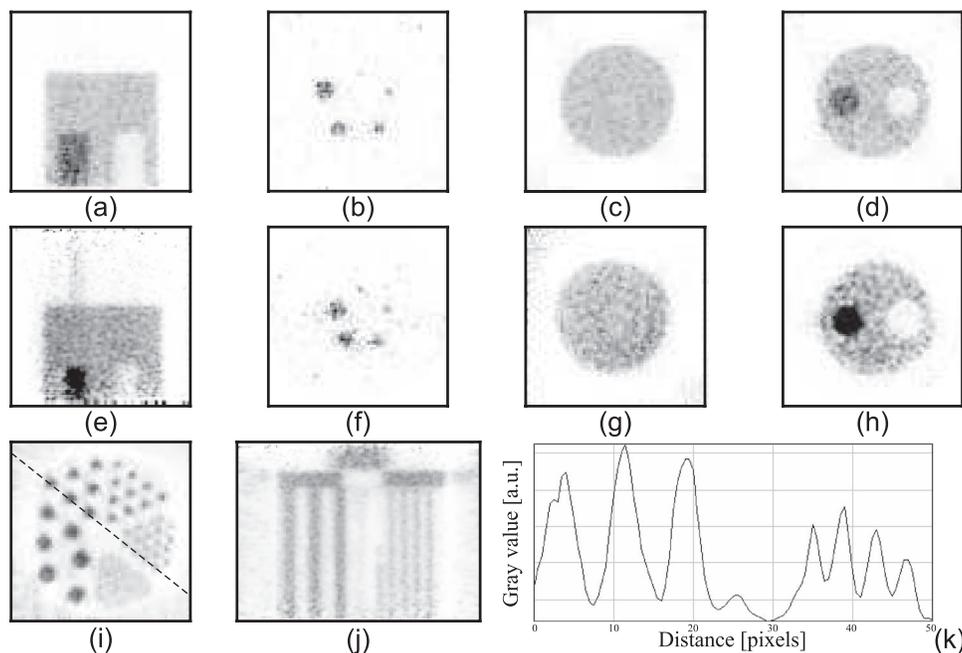


FIG. 5. Axial and transaxial slices of the simulated (a)–(d) and real (e)–(h) image quality phantoms reconstructed with the 2D kernel. Transaxial (i) and axial (j) slices of a reconstructed real Derenzo phantom and profile of its 3.2 and 1.6 mm rods (k).

TABLE II. Improvement in RC and S/N ratio obtained by using the 2D kernel with respect to the previous histogram mode, SM-based algorithm.

Rod diameter (mm)	RC (iteration 4) (%)	RC (iteration 12) (%)	S/N (iteration 4) (%)	S/N (iteration 12) (%)
1	4.9	6.3	8.7	8.9
2	2.2	3.4	8.7	8.9

III.C. Hardware description

Reconstruction experiments were conducted using two different systems to assess the different parallelization strategies. The first coarse parallelization level was tested using a Dual Xeon Quad Core E5506 processor operating at 2.13 GHz. With this system, eight cores were running in parallel, each with a private memory space. The second, finer parallelization level was tested using an Intel i5 750-based machine with four cores operating at 2.67 GHz, equipped with two GeForce GTX 580 with 512 cores operating at 770 MHz.

IV. RESULTS

IV.A. Point source

The resolution of the acquired Na^{22} point source has been measured as the FWHM of the Gaussian fit in the radial, tangential, and axial planes for both positions. The obtained measurements are reported in Table I.

IV.B. Image quality

Reconstruction experiments were conducted to compare the maximum image quality achievable with an OSEM algorithm based on a precomputed SM,¹¹ previously used with the rPET commercial scanner for small animals.

The slices of the image quality phantom reconstructed from both simulated and real data using the LM-OSEM algorithm with a 2D kernel are shown in Figs. 5(a)–5(h). The transaxial (i), axial (j), slices and the profile (k) of real 3.2 and 1.6 mm rods in the Derenzo phantom are also shown in Fig. 5.

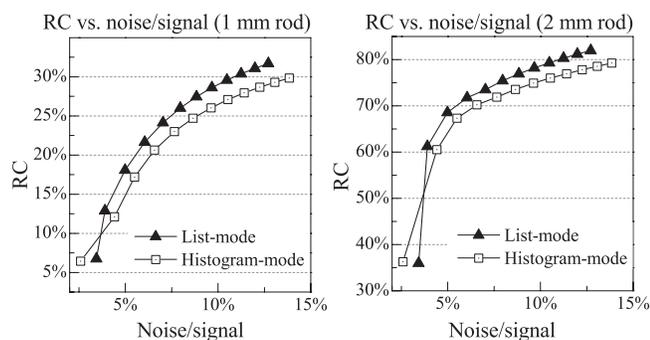


FIG. 6. RC vs noise-to-signal measured using a 1 mm diameter rod (left-hand) and a 2 mm diameter rod (right-hand) of the simulated image quality phantom for SM-based reconstructions and LM reconstructions. Results are shown for up to 12 iterations with 10 subsets (i.e., 120 subiterations). Each point in the figure represents a full iteration.

The recovery coefficients and noise-to-signal ratios, measured on the simulated data, are reported in Fig. 6 for up to 12 iterations, with 10 ordered subsets per iteration. The obtained improvements in RC and signal-to-noise ratio are summarized in Table II.

IV.C. Kernel accuracy

We have compared the modeled image quality obtained with the kernel models described in Sec. II.C. The recovery coefficients versus the noise-to-signal ratios are reported in Fig. 7 for the 1 and 2 mm diameter rods of the image quality phantom. The values after 4 and 12 iterations are also summarized in Table III. The data were reconstructed using 12 iterations and 10 ordered subsets. As shown in this figure, any improvement achieved in using the 2D model instead of the 1D model becomes negligible after several iterations. The limited improvement introduced by the more accurate model is suspected to be caused by the rotational nature of the scanner.

IV.D. Cut-off variation

Experiments were conducted to assess the trade-off between the reconstruction speed and the image quality, decided in terms of cut-off thresholds in the kernel sampling algorithm. The output of the fast ROR search technique was also studied to assess the number of voxels lost owing to the clipping effect. The voxels lost have been determined using another implementation of the algorithm that uses a slower ROR search technique unaffected by the clipping effect.

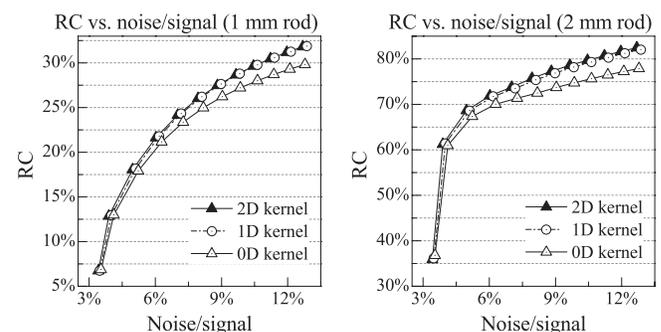


FIG. 7. RC vs noise-to-signal measured using the 1 mm diameter rod (left-hand) and 2 mm diameter rod (right-hand) of the simulated image quality phantom for three different models of the LM reconstruction: 2D kernel, 1D kernel, and 0D kernel. The LM-OSEM reconstructions used ten subsets.

TABLE III. RC and noise-to-signal measured using the 1 mm diameter rod (left-hand) and 2 mm diameter rod (right-hand) after 4 and 12 OSEM iterations with 10 subsets for the 0D, 1D, and 2D kernel models.

Rod (mm)	Kernel	RC (iteration 4) (%)	RC (iteration 12) (%)	N/S (iteration 4) (%)	N/S (iteration 12) (%)
1	0D	21.2	29.8	6.3	12.8
1	1D	21.8	31.8	6.2	12.9
1	2D	21.6	31.8	6.0	12.7
2	0D	70.0	77.9	6.3	12.8
2	1D	71.8	82.0	6.2	12.9
2	2D	71.9	82.5	6.0	12.7

Figure 8 shows the percentage of voxels that were lost for a LOR that connects the top left-hand crystals of a detector to any of the others on the facing detector when using a 1D kernel with probability threshold of 10% (a), a 2D kernel with thresholds 10% (b), and 1% (c). Figure 8(d) shows the corresponding percentage probability lost for the case of the 2D model with a threshold of 1% of the Gaussian peak. The clipping effect is almost absent in the 0D and 1D kernels, while it mainly affects the most oblique LORs modeled with the 2D kernel. Histograms for the probability losses owing to the clipping effect are shown in Figs. 8(e)–8(h): the loss per LOR was approximately 2% in the worst case, and less than 1% in 97% of all the possible LORs when using a cutoff threshold equal to 1% of the Gaussian peak. The images reconstructed using the search technique unaffected by the clipping effect showed no appreciable difference with respect to the proposed implementation for probability thresholds lower than 10% of the Gaussian peak.

The results in Fig. 9 show the recovery coefficient vs the noise-to-signal ratio trade-off when reducing the cutoff from

10% to 1%. Data were measured for the 1 and 2 mm diameter rods of the simulated image quality phantom for LM-OSEM reconstructions using the 0D and 2D kernels.

IV.E. Acceleration by parallelization

Figure 10 summarizes the number of LORs reconstructed per second using the kernels described above. Data on the CPU implementation correspond to the reconstruction times for the image quality phantom for a FOV of $120 \times 140 \times 120$ voxels.

Experiments show that the first level of the inter-LOR parallelization technique accelerates the process by a factor almost proportional to the number of processing units used, with a mean proportionality factor, i.e., *parallelization efficiency*, of about 80%. Thus, using eight cores, it is possible to reconstruct up to 2 5000 LORs per second with a cut-off threshold of 10%, or 15 000 LORs per second with a cut-off threshold of 1%. It is evident that a single CPU is insufficient for fast reconstructions.

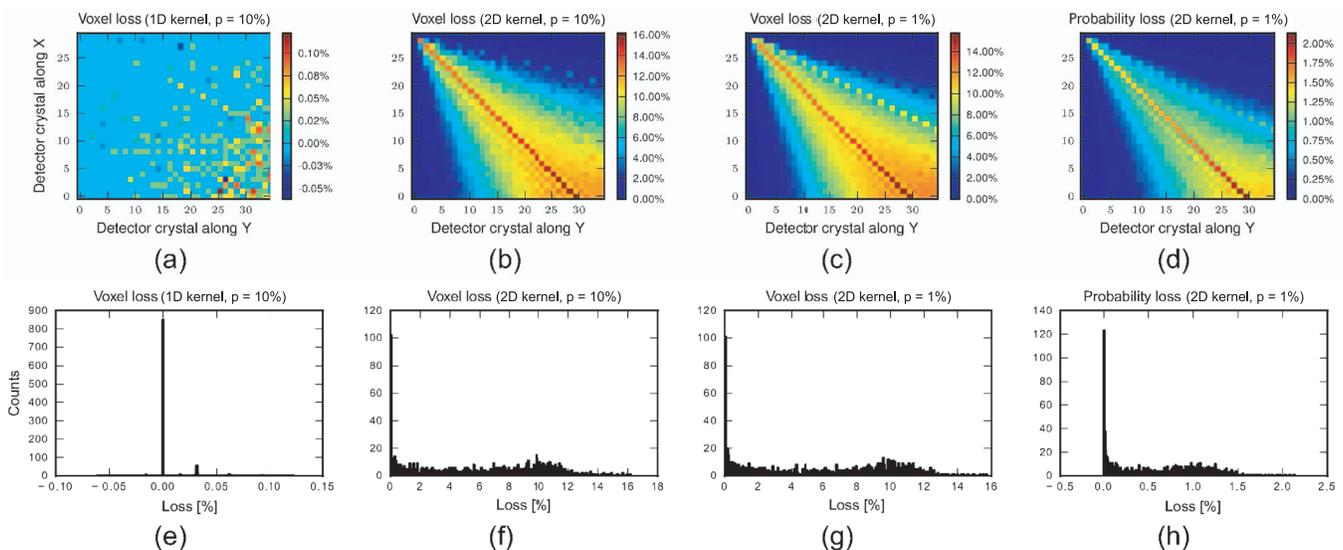


FIG. 8. Percentage of voxels and corresponding probability that were lost owing to the clipping effect. In the first row, the loss is represented in color-scale for every crystal from one detector in coincidence with the top left-hand crystal of the other detector. In the second row, the histograms of the same losses are reported. The first three columns represent the loss in terms of number of voxels missing over the total for the 1D kernel with 10% threshold (a)–(e) and the 2D kernel with 10% (b)–(f) and 1% (c)–(g) thresholds. The fourth column (d)–(h) represents the probability losses, i.e., the tails of the Gaussian distribution that have been lost due to the clipping effect, for the 2D kernel with 1% threshold.

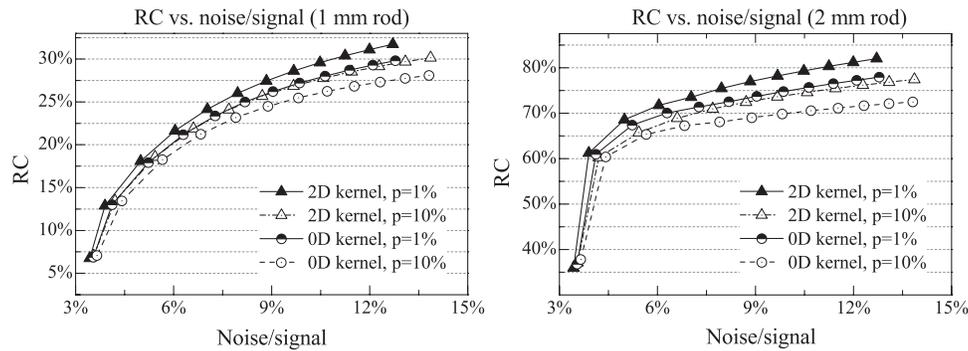


FIG. 9. RC vs noise-to-signal measured for the 1 mm diameter rod (left-hand) and 2 mm diameter rod (right-hand) of the simulated image quality phantom for LM-OSEM reconstructions using the 2D kernel and the 0D kernel. Cut-off values of $p = 1\%$ and $p = 10\%$ are compared for both cases.

The OpenCL implementation has been tested on both CPU and GPU. This implementation lacks any voxel caching, thus requiring double the computation per LOR, however, its computation speed is higher because of a better utilization of CPU parallel resources, so that the final reconstruction speed is almost the same for the two CPU implementations.

The GPU cores run at a frequency that is about one-third that of the CPU counterpart. The GPU reconstruction speed is about four times (or eight times for 2 GPUs) that of the speed obtained using the eight-core CPU. Considering these factors, the mean CPU to GPU parallelization efficiency, was estimated to be approximately 40%, while it was observed that doubling the number of GPUs the reconstruction speed doubles (i.e., 100% GPU parallelization efficiency). For the three kernel models, the maximum observed reconstruction speed on 2 GPUs is roughly 206 000 LORs per second with a probability threshold of 10% and 103 000 LORs per second with a threshold of 1%.

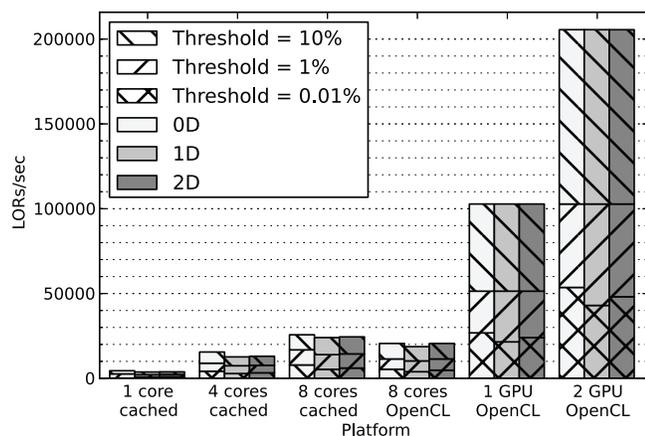


FIG. 10. A comparison of the reconstruction speed in terms of the LORs forward and back projections per second for a FOV of $120 \times 140 \times 120$ voxels. In the above figure, the speeds for all the kernel models described in the previous paragraphs are compared when executed using the same machine with one, four, and eight cores running at 2.1 GHz versus the speed of a GPU-based version running on 512 cores at 770 MHz.

V. DISCUSSION AND CONCLUSIONS

A LM reconstruction method has been proposed based on new techniques aimed at improving the reconstruction speed and the image signal-to-noise ratio of the OSEM algorithm used previously for the commercial rPET small animal scanner.

We have described a methodology for the efficient derivation of elliptical kernel models based on a LOR-local reference system a MC simulations. The new reference system was necessary because the kernel models used were not spatially invariant like the circular models are. A new ROR identification technique was also described, based on integer operations and with minimal memory usage, that is suitable for low-memory processing units. A special advantage of the new identification technique is the possibility of defining regions of response based on a constant contour threshold over the calculated probabilities.

The new technique allows for fast, accurate reconstructions with minimal memory usage per processing core. The performances in terms of image quality are comparable to SM based approaches,^{8,27} because the used model include per-LOR MC simulations, but the minimal memory requirements allow for LM storage and reconstruction in the GPU as in analytical approaches.⁷ Moreover, the proposed implementation does not present the mismatch between “ray-driven” forward projection and “voxel-driven” back projection pointed out by Ref. 27 on other previous works.

An undesirable effect of the voxel search technique proposed here may limit the maximum threshold applicable: the ROR clipping effect, which introduces a mismatch between the actual ROR and the one obtained by voxel propagation. The magnitude of the voxel and probability losses caused by this mismatch have been characterized and shown to be negligible (2% in the worst case, and less than 1% in 97% for all the possible cases) when using cut-off thresholds equal to, or less than, 1% of the Gaussian peak. More advanced versions of the ROR search technique that can make use of LOR-driven secondary axis selection and oblique propagation are planned as future work. In any case, the clipping effect only implies a lower limit on image quality and it has been shown

that does not produce any sensible effect with the above threshold.

The method was validated using simulated and real NEMA 2008 image quality phantoms, Derenzo phantoms and a real Na^{22} point source. The results were compared with a specific commercial histogram-mode OSEM algorithm based on a precalculated system matrix. It was shown how the LM reconstruction improves considerably the image quality with respect to the previous SM-based algorithm (Fig. 6). The main reason for this difference is attributed to the full-angle coverage adopted by the LM version, while the SM version is limited by the adopted symmetries. Thus, the number of coincidence events used in the reconstruction process for the same acquisition is larger in the LM method, which results in a lower noise level. Another reason for the relatively low noise is that the MC simulation of the SM scheme adds statistical noise, which is not present in the numerical fitted model used in the LM method. Finally, the binning and interpolation process involved in the histogram-mode reconstruction is subject to approximation errors.

Figure 7 shows how the 2D provides a better image quality than do 0D and 1D kernels, although the difference with respect to the 1D version is limited. The 2D kernel model performs only slightly better than the 1D model, even if it is more accurate in principle. The rotational symmetry of the scanner used in the experiments, as well as the symmetries in the NEMA phantom, could be the cause of the small difference between the outcomes of 1D and 2D models. More research would be needed in this field. However, given that the 2D kernel model is also slightly faster than the 1D version, it represents in any case the best option for both quality and speed.

The cut-off value has a clear impact on the quality of the results. As can be seen in Fig. 10, all models reconstruct at almost half the speed if the cut-off value is changed from 10% of the Gaussian peak to 1% of the Gaussian peak. The effect of changing the threshold was evaluated quantitatively and is summarized in Fig. 9.

Multicore and many-core parallelized implementations have been realized and characterized. For a given model, the images obtained with the different implementations are completely equivalent. Reconstruction speed was evaluated for both CPU- and GPU-based implementations, demonstrating that the reconstruction could be accelerated by about 80% of the number of processors used on the same CPU architecture. Conversely, GPU cores perform at 40% the speed of CPU cores, but then the computation time decreases identically with the number of GPUs.

ACKNOWLEDGMENTS

This work was partially supported by Spain's Ministry of Science and Innovation through CDTI's CENIT program (AMIT project) and INNPACTO (PRECISION project), Instituto de Salud Carlos III (PI09/91058 and PI09/91065), and Project Nos. TEC2010-21619-C04-03 and TEC2011-28972-C02-02, Comunidad de Madrid (ARTEMIS S2009/DPI-1802), and the European Regional Development Funds

(FEDER). CIBER-BBN is an initiative funded by the VI National R&D&I Plan 2008–2011, Iniciativa Ingenio 2010, Consolider Program, CIBER Actions, and financed by the Instituto de Salud Carlos III with assistance from the European Regional Development Fund.

^{a)}Electronic mail: gsportelli@die.upm.es

- ¹H. H. Barrett *et al.*, "List-mode likelihood," *J. Opt. Sc. Am. A* **14**, 2914–2923 (1997).
- ²L. Parra and H. H. Barrett, "List-mode likelihood: EM algorithm and image quality estimation demonstrated on 2-D PET," *IEEE Trans. Med. Imaging* **17**, 228–235 (1998).
- ³A. J. Reader, "The promise of new PET image reconstruction," *Phys. Medica* **24**, 49–56 (2008).
- ⁴B. Guérin and G. E. Fakhri, "Novel scatter compensation of list-mode PET data using spatial and energy dependent corrections," *IEEE Trans. Med. Imaging* **30**, 759–773 (2011).
- ⁵S. Surti *et al.*, "Investigation of time-of-flight benefit for fully 3-DPET," *IEEE Trans. Med. Imaging* **25**, 529–538 (2006).
- ⁶M. Conti, "State of the art and challenges of time-of-flight PET," *Phys. Medica* **25**, 1–11 (2009).
- ⁷G. Pratz *et al.*, "Fast list-mode reconstruction for time-of-flight PET using graphics hardware," *IEEE Trans. Nucl. Sci.* **58**, 105–109 (2011).
- ⁸J. L. Herraiz *et al.*, "FIRST: Fast iterative reconstruction software for PET tomography," *Phys. Med. Biol.* **51**, 4547–4565 (2006).
- ⁹S. Stute *et al.*, "A method for accurate modelling of the crystal response function at a crystal sub-level applied to PET reconstruction," *Phys. Med. Biol.* **56**, 793–809 (2011).
- ¹⁰L. Zhang *et al.*, "Fast and memory-efficient Monte Carlo-based image reconstruction for whole-body PET," *Med. Phys.* **37**, 3667–3676 (2010).
- ¹¹J. E. Ortuño *et al.*, "Efficient methodologies for system matrix modelling in iterative image reconstruction for rotating high-resolution PET," *Phys. Med. Biol.* **55**, 1833–1861 (2010).
- ¹²M. Rafecas *et al.*, "A Monte Carlo study of high-resolution PET with granulated dual-layer detectors," *IEEE Trans. Nucl. Sci.* **48**, 1490–1495 (2001).
- ¹³M. S. Tohme and J. Y. Qi, "Iterative image reconstruction for positron emission tomography based on a detector response function estimated from point source measurements," *Phys. Med. Biol.* **54**, 3709–3725 (2009).
- ¹⁴V. Y. Panin *et al.*, "PET reconstruction with system matrix derived from point source measurements," *IEEE Trans. Nucl. Sci.* **53**, 152–159 (2006).
- ¹⁵C. Cloquet *et al.*, "Non-Gaussian space-variant resolution modelling for list-mode reconstruction," *Phys. Med. Biol.* **55**, 5045–5066 (2010).
- ¹⁶A. Iriarte *et al.*, "A theoretical model for EM-ML reconstruction algorithms applied to rotating PET scanners," *Phys. Med. Biol.* **54**, 1909–1934 (2009).
- ¹⁷S. Staelens *et al.*, "A three-dimensional theoretical model incorporating spatial detection uncertainty in continuous detector PET," *Phys. Med. Biol.* **49**, 2337–2350 (2004).
- ¹⁸J. J. Scheins *et al.*, "Analytical calculation of volumes-of-intersection for iterative, fully 3-d PET reconstruction," *IEEE Trans. Med. Imaging* **25**, 1363–1369 (2006).
- ¹⁹S. Moehrs *et al.*, "Multi-ray-based system matrix generation for 3D PET reconstruction," *Phys. Med. Biol.* **53**, 6925–6945 (2008).
- ²⁰C. Johnson *et al.*, "A system for the 3D reconstruction of retracted-septa PET data using the EM algorithm," *IEEE Trans. Nucl. Sci.* **42**, 1223–1227 (1995).
- ²¹L. Zhang *et al.*, "Monte Carlo system modeling for PET reconstruction: A rotator approach," in *IEEE Nuclear Science Symposium Conference Record, Dresden, Germany* (IEEE, NY, 2008), pp. 5101–5106.
- ²²J. Cabello *et al.*, "High performance 3D PET reconstruction using spherical basis functions on a polar grid," *Int. J. Biomed. Imaging* **2012**, 1–11.
- ²³J. Cabello and M. Rafecas, "Comparison of basis functions for 3D PET reconstruction using a Monte Carlo system matrix," *Phys. Med. Biol.* **57**, 1759–1777 (2012).
- ²⁴S. Vandenberghe *et al.*, "Fast reconstruction of 3D time-of-flight PET data by axial rebinning and transverse mashing," *Phys. Med. Biol.* **51**, 1603–1621 (2006).
- ²⁵J. Y. Qi *et al.*, "High-resolution 3D bayesian image reconstruction using the microPET small-animal scanner," *Phys. Med. Biol.* **43**, 1001–1013 (1998).

- ²⁶J. L. Herraiz *et al.*, "GPU acceleration of a fully 3D iterative reconstruction software for PET using CUDA," in *IEEE Nuclear Science Symposium Conference Record, Orlando, FL* (IEEE, NY, 2009), pp. 4064–4067.
- ²⁷J. Zhou and J. Qi, "Fast and efficient fully 3D PET image reconstruction using sparse system matrix factorization with GPU acceleration," *Phys. Med. Biol.* **56**, 6739–6757 (2011).
- ²⁸R. L. Siddon, "Fast calculation of the exact radiological path for a three-dimensional CT array," *Med. Phys.* **12**, 252–255 (1985).
- ²⁹L. M. Popescu and R. M. Lewitt, "Ray tracing through a grid of blobs," in *IEEE Nuclear Science Symposium Conference Record, Rome, Italy* (IEEE, NY, 2004), pp. 3983–3986.
- ³⁰G. Prax *et al.*, "Fast, accurate and shift-varying line projections for iterative reconstruction using the GPU," *IEEE Trans. Med. Imaging* **28**, 435–445 (2009).
- ³¹A. J. Reader *et al.*, "One-pass list-mode EM algorithm for high-resolution 3-D PET image reconstruction into large arrays," *IEEE Trans. Nucl. Sci.* **49**, 693–699 (2002).
- ³²A. J. Reader *et al.*, "EM algorithm system modeling by image-space techniques for PET reconstruction," *IEEE Trans. Nucl. Sci.* **50**, 1392–1397 (2003).
- ³³E. Rapisarda *et al.*, "Image-based point spread function implementation in a fully 3D OSEM reconstruction algorithm for PET," *Phys. Med. Biol.* **55**, 4131–4151 (2010).
- ³⁴A. Del Guerra *et al.*, "YAP-PET: First results of a small animal positron emission tomograph based on YAP:Ce finger crystals," *IEEE Trans. Nucl. Sci.* **45**, 3105–3108 (1998).
- ³⁵J. Vaquero *et al.*, "rPET detectors design and data processing," in *IEEE Nuclear Science Symposium Conference Record, Wyndham El Conquistador, Puerto Rico* (IEEE, NY, 2005), pp. 2885–2889.
- ³⁶G. Sportelli *et al.*, "Reprogrammable acquisition architecture for dedicated positron emission tomography," *IEEE Trans. Nucl. Sci.* **58**, 695–702 (2011).
- ³⁷A. J. Reader *et al.*, "Accelerated list-mode EM algorithm," *IEEE Trans. Nucl. Sci.* **49**, 42–49 (2002).
- ³⁸J. Y. Qi, "Calculation of the sensitivity image in List-Mode reconstruction for PET," *IEEE Trans. Nucl. Sci.* **53**, 2746–2751 (2006).
- ³⁹C. S. Levin and E. J. Hoffman, "Calculation of positron range and its effect on the fundamental limit of positron emission tomography system spatial resolution," *Phys. Med. Biol.* **44**, 781–799 (1999).
- ⁴⁰A. Sanchez-Crespo and S. A. Larsson, "The influence of photon depth of interaction and non-collinear spread of annihilation photons on PET image spatial resolution," *Eur. J. Nucl. Med. Mol. Imaging* **33**, 940–947 (2006).
- ⁴¹X. M. Hua, "Monte Carlo simulation of comptonization in inhomogeneous media," *Comput. Phys.* **11**, 660–668 (1997).
- ⁴²G. Blelloch, *Synthesis of Parallel Algorithms* (Morgan Kaufmann, San Francisco, CA, 1993), pp. 35–60.
- ⁴³J. Reinders, *Intel Threading Building Blocks*, 1st ed. (O'Reilly & Associates, Sebastopol, CA, 2007).
- ⁴⁴J. E. Stone *et al.*, "OpenCL: A parallel programming standard for heterogeneous computing systems," *Comput. Sci. Eng.* **12**, 66–73 (2010).
- ⁴⁵E. Lage *et al.*, "Design and performance evaluation of a coplanar multi-modality scanner for rodent imaging," *Phys. Med. Biol.* **54**, 5427–5441 (2009).
- ⁴⁶NEMA, *Performance Measurements of Small Animal Positron Emission Tomographs*, Standards publication NU 4-2008 (NEMA, Rosslyn, VA, 2008).
- ⁴⁷S. Jan *et al.*, "GATE: A simulation toolkit for PET and SPECT," *Phys. Med. Biol.* **49**, 4543–4561 (2004).