

Accelerated time-of-flight (TOF) PET image reconstruction using TOF bin subsetization and TOF weighting matrix pre-computation

Abolfazl Mehranian¹, Fotis Kotasidis¹ and Habib Zaidi^{1,2,3}

¹ Division of Nuclear Medicine and Molecular Imaging, Geneva University Hospital, Geneva, Switzerland

² Geneva Neuroscience Centre, University of Geneva, Geneva, Switzerland

³ Department of Nuclear Medicine and Molecular Imaging, University of Groningen, Groningen, Netherlands

E-mail: abolfazl.mehranian@unige.ch

Received 23 September 2015, revised 30 November 2015

Accepted for publication 23 December 2015

Published 21 January 2016



Abstract

Time-of-flight (TOF) positron emission tomography (PET) technology has recently regained popularity in clinical PET studies for improving image quality and lesion detectability. Using TOF information, the spatial location of annihilation events is confined to a number of image voxels along each line of response, thereby the cross-dependencies of image voxels are reduced, which in turns results in improved signal-to-noise ratio and convergence rate. In this work, we propose a novel approach to further improve the convergence of the expectation maximization (EM)-based TOF PET image reconstruction algorithm through subsetization of emission data over TOF bins as well as azimuthal bins. Given the prevalence of TOF PET, we elaborated the practical and efficient implementation of TOF PET image reconstruction through the pre-computation of TOF weighting coefficients while exploiting the same in-plane and axial symmetries used in pre-computation of geometric system matrix. In the proposed subsetization approach, TOF PET data were partitioned into a number of interleaved TOF subsets, with the aim of reducing the spatial coupling of TOF bins and therefore to improve the convergence of the standard maximum likelihood expectation maximization (MLEM) and ordered subsets EM (OSEM) algorithms. The comparison of on-the-fly and pre-computed TOF projections showed that the pre-computation of the TOF weighting coefficients can considerably reduce the computation time of TOF PET image reconstruction. The convergence rate and bias-variance performance of the proposed TOF subsetization scheme were evaluated using simulated, experimental phantom and clinical studies. Simulations demonstrated that as the number of TOF subsets is increased, the convergence rate of MLEM and OSEM algorithms is improved.

It was also found that for the same computation time, the proposed subsetization gives rise to further convergence. The bias-variance analysis of the experimental NEMA phantom and a clinical FDG-PET study also revealed that for the same noise level, a higher contrast recovery can be obtained by increasing the number of TOF subsets. It can be concluded that the proposed TOF weighting matrix pre-computation and subsetization approaches enable to further accelerate and improve the convergence properties of OSEM and MLEM algorithms, thus opening new avenues for accelerated TOF PET image reconstruction.

Keywords: PET, image reconstruction, time-of-flight PET, ordered subsets, OSEM, acceleration

 Online supplementary data available from stacks.iop.org/PMB/61/1309/mmedia

(Some figures may appear in colour only in the online journal)

1. Introduction

Model-based iterative image reconstruction of positron emission tomography (PET) data has experienced major technical advances over the past decades and has now almost superseded conventional analytical filtered-back projection (FBP) and algebraic reconstruction techniques by providing image of improved quality and quantitative accuracy. PET emission data suffer from (i) high statistical noise, governed by the intrinsic sensitivity of the PET scanner, data acquisition time and injected tracer activity, (ii) photon attenuation and Compton scattering, (iii) detector sensitivity variations and (iv) resolution degrading mechanisms such as crystal penetration and scattering, finite-size crystals, non-collinearity of annihilation photons and positron range (Rahmim *et al* 2013). Model-based image reconstruction techniques can account for all these physical factors and resolution degrading mechanisms in a so-called system matrix, thereby allowing for a rich description of the image formation process. Developed on a Bayesian framework, these techniques can accurately model the statistical variability of the emission data and even the prior probabilistic distribution of the underlying PET images, which consequently gives rise to substantial noise and artefact reduction compared to FBP algorithms. In fact, this capability, which is best highlighted by short time or low-dose scanning, was the main motivation for the translation of these reconstruction algorithms into clinical PET scanners. In addition, the incorporation of the blurring effects (usually represented by space-variant point spread function- PSF) into the image reconstruction process, effectively counteracts the low-pass filtering nature of PET and therefore results in resolution recovery and improved image quality.

The enhancements provided by model-based image reconstruction methods, however, come at the expense of a more complicated inverse problem that can only be solved using iterative optimization algorithms, which progressively refine an initial image estimate until a data fidelity criterion is met. In emission tomography, the maximum likelihood (ML) criterion has been extensively used to seek a solution that provides the maximal statistical consistency with measured data. The statistical variability of PET data is modelled using a Poisson distribution; however, if the data have been corrected for accidental coincidences or acquired with high count rate, a Gaussian distribution is a more appropriate model resulting in a minimum (weighted) least squares (LS) criterion (Fessler 1994). Many optimization algorithms have been proposed for image reconstruction based on ML and LS criteria such as maximum likelihood

expectation maximization (MLEM) (Shepp and Vardi 1982, Lange and Carson 1984), successive over-relaxation, conjugate gradient among others (Qi and Leahy 2006, Reader and Zaidi 2007). Iterative reconstruction methods initially suffered from the slow convergence of the optimization algorithms and, as such, increased computation time was needed for running a large number of iterations. Apart from algorithm-specific acceleration techniques, such as line search optimization and preconditioning, convergence rate can be generally improved by designing specific image update schemes. Block-iterative or row-action schemes aim to geometrically or chronologically partition the emission data into a number of subsets or blocks and to sequentially update the entire image with each block. The ordered subsets EM (OSEM) (Hudson and Larkin 1994), rescaled block-iterative EM (RBI-EM) (Byrne 1998) and row-action maximum likelihood (RAMLA) (Browne and De Pierro 1996) algorithms are the best instances of these approaches. The OSEM algorithm, which is based on the subsetization of angular projection bins into a series of mutually exclusive (balanced) subsets, shows a considerably improved initial convergence and therefore was the first sufficiently fast algorithm paving the way for clinical adoption of iterative reconstruction methods. The RBI-EM and RAMLA algorithms were proposed to improve the performance of the OSEM algorithm in terms of subset balance dependency and limit cycle condition, respectively. Another update scheme consists in sequentially updating a subset or group of image voxels instead of all voxels simultaneously as done in EM methods. The iterated coordinate-ascent (ICA) method (Bouman and Sauer 1996) maximizes the ML objective function with respect to each voxel at each iteration, which results into a simple update step and considerable fast convergence rate, given a good initial estimate. The ICA is, however, computationally intensive using standard projectors calculating the contribution of all image voxels to a given line of response (LOR), instead of the contribution of all LORs to a given image voxel or coordinate. The grouped coordinate-ascent (Fessler *et al* 1997) and space-alternating generalized EM (Fessler and Hero 1995) algorithms were proposed to provide a compromise between convergence and computational complexity by sequentially updating a subset of image voxels. The improved convergence rate of these update schemes can be attributed to the fact that the block-wise update of voxels, reduces the cross-dependencies between them and thus leads to spatially invariant convergence. The subsetization of image voxels into uncoupled groups is therefore of importance for faster convergence of these methods. Recently, the combination of block-iterative and grouped coordinate update schemes has been proposed by (Van Slambrouck and Nuyts 2014). It has been shown that the convergence rate of their algorithm is improved as the number of image blocks or patches is increased.

With the advent of detectors with fast decay time and high light output, time-of-flight (TOF) technology has been revived in PET imaging. Using TOF capability, the difference in arrival times of coincident annihilation photons are measured with an uncertainty governed by the coincidence resolving time of the detectors. Thereby, the origin of annihilation events can be confined to a group of voxels along a LOR instead of all voxels intersected by the LOR as in conventional non-TOF PET. During PET image reconstruction, only a group of voxels is thus updated according to corresponding TOF weighting coefficients, which is reminiscent of grouped coordinate update schemes. Therefore, the cross-dependencies between image voxels are reduced, resulting in fast and space-invariant convergence, which improves the detectability of lesions located near large hot regions (e.g. the liver in ^{18}F -FDG scans) (El Fakhri *et al* 2011, Schaefferkoetter *et al* 2013) and reduces noise propagation and sensitivity to errors in data correction (Conti 2011b, Mehranian and Zaidi 2015). The groups of voxels associated with neighbouring TOF bins are, however, spatially overlapped and coupled. As the TOF timing resolution is improved the degree of voxel coupling is reduced leading to further improvements in convergence rate and error propagation.

Most current model-based reconstruction techniques are based on system matrix factorization, (Qi *et al* 1998) and on-the-fly computation of TOF weighting coefficients (Zhou and Qi 2014). The factorization decomposes the system matrix into its blurring and geometric projection components, which allows exploiting the sparsity and base-symmetries of geometric system matrix to make its pre-computation practical and therefore to reduce reconstruction time. Since current TOF timing resolution of clinical PET scanners evolves toward better time resolutions (van Dam *et al* 2013, Nemallapudi *et al* 2015), finer and thus larger number of TOF bins are required to discretize TOF kernels, leading to increased computational burden of TOF projectors. As a result, there is need for efficient strategies to implement TOF image reconstruction algorithms with improved convergence rates. In this regard, the aim of this study is threefold. First, to propose a new data subsetization scheme over TOF bins in order to reduce the voxel coupling of adjacent TOF bins, thereby further accelerating the convergence of current TOF image reconstruction algorithms. Secondly, to perform the pre-computation of TOF weighting coefficients and demonstrate that the same geometric base-symmetries can be employed for this task, which considerably speeds up the reconstruction time. Finally, given the prevalence of TOF PET, we elaborate the practical algorithmic implementation of TOF PET image reconstruction given the lack of comprehensive technical reports in the literature on this subject.

2. Materials and methods

In histogram-based TOF PET, the measured data, y_{ik} , along LOR i and TOF bin k consist of the collection of the coincident annihilation photons recorded in TOF time interval (t_k, t_{k+1}) within the scanner's coincidence window width 2τ . With negligible detector dead time and no data pre-correction, the statistical variability of the measured data is best modelled using a Poisson distribution:

$$y_{ik} \sim \text{Poisson } \{\bar{y}_{ik} + \bar{r}_{ik} + n_i \bar{s}_{ik}\} \quad (1)$$

where \bar{y}_{ik} , \bar{r}_{ik} and \bar{s}_{ik} are the expected number of prompt, random and scattered coincidences in sinogram TOF bin i, k , respectively. The expected values of random and scatter events are usually estimated using delayed coincidence window measurements followed by randoms smoothing and model-based scatter estimation with the inclusion of detector normalization factors n_i . In model-based reconstruction of emission data, it is often assumed that over the range of count rates encountered in clinical settings (i.e. negligible dead time and pulse pileup), the PET system response to an object's activity distribution $f(\nu)$ is linear and spatially-variant, thereby the expected prompts are given by the following linear integral equation:

$$\bar{y}_{ik} = \int_{t_k}^{t_{k+1}} \int_{\Omega} f(\nu) h_i(\nu) p_i(\nu, t) d\nu dt, \quad i = 1, \dots, I, \quad k = 1, \dots, K \quad (2)$$

where $h_i(\nu)$ and $p_i(\nu, t)$ are point and time spread functions of the system in response to an input point source located at the spatial coordinate ν over the finite spatial domain of Ω and time difference interval of (t_k, t_{k+1}) . I and K denote the total number of LORs and TOF time bins, respectively. The continuous object $f(\nu)$ is usually parameterized using a linear combination of J non-overlapping rectangular basis functions (voxels) $b_j(\nu)$, that is, $f(\nu) \cong \sum_j^J f_j b_j(\nu)$, thereby (2) can be written as:

$$\bar{y}_{ik} = \sum_j^J g_{ij} f_j \int_{t_k}^{t_{k+1}} p_i(\nu_j, t) dt \quad (3)$$

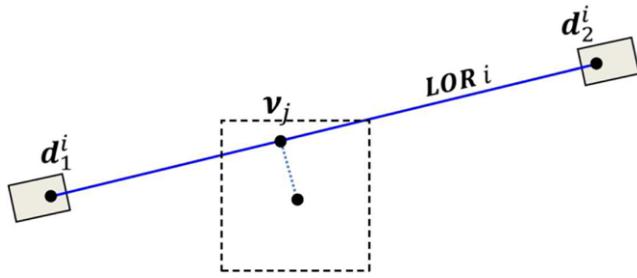


Figure 1. Illustration of the interaction of a line of response (LOR) with an image voxel with projected center of ν_j . Note the endpoints of the LOR, d_2^i and d_1^i , typically take an approximation of the photon's depth of interaction into account.

where $g_{ij} \triangleq \int_{\Omega} b_j(\boldsymbol{\nu}) h_i(\boldsymbol{\nu}) d\boldsymbol{\nu}$ is an element of the system matrix indicating the contribution of annihilation photons emanating from j th image voxel to i th projection bin. The TOF kernel $p_i(\boldsymbol{\nu}_j, t)$ is typically modelled as a Gaussian function with a standard deviation of $\sigma = R_t(8 \ln 2)^{-1/2}$ determined by the nominal TOF timing resolution R_t of the PET scanner, as follows:

$$\begin{aligned} p_i(\boldsymbol{\nu}_j, t) &= \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(t - \mu)^2\right) \\ \mu &= \frac{1}{c} (\|d_2^i - \boldsymbol{\nu}_j\| - \|d_1^i - \boldsymbol{\nu}_j\|) \end{aligned} \quad (4)$$

where c is the speed of light, μ is the signed arrival time difference of annihilation photons along LOR i , $\boldsymbol{\nu}_j$ is the orthogonal projection of the j th voxel center onto the i th LOR connecting the detector points d_2^i and d_1^i as shown in figure 1. $\|\mathbf{x}\| = \sqrt{x_1^2 + \dots + x_n^2}$ is the Euclidean norm of \mathbf{x} .

2.1. System matrix pre-computation

In 3D PET scanners, the geometric component of the system matrix $\mathbf{G} = \{g_{ij}\} \in \mathbb{R}^{I \times J}$, which maps the image space to projection space, is highly sparse, particularly with voxel basis functions that have a minimal spatial support. The sparsity of this matrix can therefore be exploited to pre-compute and store only the non-zero elements of g_{ij} for each LOR.

Let $\mathbf{G}_i = [g_{i1}, \dots, g_{iJ}]^T \in \mathbb{R}^J$ be the i th row of geometric system matrix and $\Theta \subseteq \{1, \dots, J\}$ a subset of image voxels for which $g_{ij} \neq 0$, i.e. $\Theta = \{\Theta_1, \dots, \Theta_Z\} = \{j \mid j, g_{ij} \neq 0\}$. By eliminating the zero components of the \mathbf{G}_i , the sub-vector $\widehat{\mathbf{G}}_i = [g_{i\Theta_1}, \dots, g_{i\Theta_Z}]^T \in \mathbb{R}^Z$ can be defined as the compressed form of \mathbf{G}_i . Similarly the sub-vector $\widehat{\mathbf{f}} = [f_{\Theta_1}, \dots, f_{\Theta_Z}]^T \in \mathbb{R}^Z$ is defined by squeezing \mathbf{f} over Θ . In forward and backprojections during image reconstruction only the $\widehat{\mathbf{G}}_i$'s are used, thereby the computational burden of the reconstruction process is considerably reduced. Let $\bar{\mathbf{y}}_i = [\bar{y}_{i1}, \dots, \bar{y}_{iK}]^T \in \mathbb{R}^K$ be the vector of TOF measurements in the i th LOR, the forward model in (3) can therefore be defined as:

$$\bar{\mathbf{y}}_i = n_i a_i (\widehat{\mathbf{G}}_i \odot \widehat{\mathbf{f}}) \mathbf{W}_i \quad (5)$$

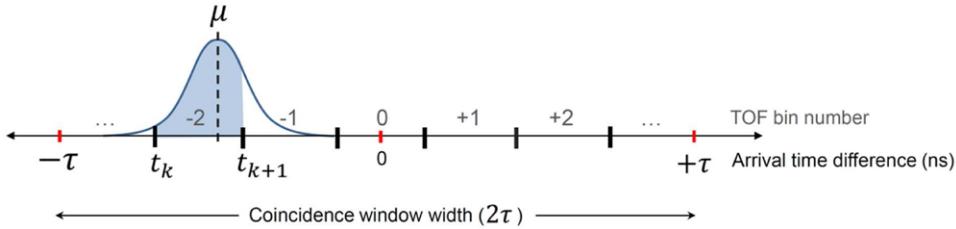


Figure 2. Contribution of the annihilation photons arriving with signed time difference of μ to the k th TOF bin, calculated as the definite integral of the TOF Gaussian kernel centred at μ over the TOF bin boundaries. The bin boundaries are typically equally spaced on the coincidence window width forming K odd TOF bins indexed by $0, \pm 1, \pm 2$ and so forth.

where the scalars n_i and a_i are normalization and attenuation factors, \odot is Hadamard or element-wise matrix multiplication and \mathbf{W}_i is the TOF weighting coefficient matrix associated with the i th LOR, defined as:

$$\mathbf{W}_i = \begin{bmatrix} w_i(\boldsymbol{\nu}_{\Theta_1}, t_1) & \dots & w_i(\boldsymbol{\nu}_{\Theta_1}, t_K) \\ \vdots & \ddots & \vdots \\ w_i(\boldsymbol{\nu}_{\Theta_Z}, t_1) & \dots & w_i(\boldsymbol{\nu}_{\Theta_Z}, t_K) \end{bmatrix} \in \mathbb{R}^{Z \times K} \quad (6)$$

For brevity, without loss of generality, we ignored the sinogram and image space blurring components of the system matrix in (5). The entry $w_i(\boldsymbol{\nu}_j, t_k)$ indicates the contribution of the j th voxels along i th LOR to k th TOF bin. As shown in figure 2, it is calculated as the area under the TOF kernel curve centered at μ with t_k and t_{k+1} as lower and upper limits of the integration. The integration can be obtained using Gauss error function (erf):

$$w_i(\boldsymbol{\nu}_j, t_k) = \int_{t_k}^{t_{k+1}} p_i(\boldsymbol{\nu}_j, t) dt = \frac{1}{2} \operatorname{erf}\left(\frac{t_{k+1} - \mu}{\sigma\sqrt{2}}\right) - \frac{1}{2} \operatorname{erf}\left(\frac{t_k - \mu}{\sigma\sqrt{2}}\right) \quad (7)$$

where $\operatorname{erf}(x)$ can be numerically approximated for positive and negative values of x by (Abramowitz and Stegun 1964):

$$\operatorname{erf}(x) \approx \frac{x}{|x|} \sqrt{1 - \exp\left(-x^2 \frac{4/\pi + ax^2}{1 + ax^2}\right)}, \quad a = 0.140\,012 \quad (8)$$

The TOF weighting coefficient matrix \mathbf{W}_i acts as a convolution matrix that convolves the impulse of photons arriving by the time difference of μ with the time spread function of the scanner. Figure 3(A) shows the circular structure of this matrix for a given LOR in the mCT scanner (Siemens Healthcare, Erlangen, Germany) with 13 TOF bins. As shown, the voxels located at the centre of the field-of-view (FOV) (with indices of around 100 in this case) mainly contribute to the central TOF bins, while those located at the edges of the FOV, contribute to the peripheral TOF bins. Note that the columns of this matrix are not necessarily the cyclic permutations of the first column with an equal offset, because (i) the TOF bins might be chosen to have different widths, particularly at the edges of FOV and (ii) the arrival time difference of photons emitted from the successive voxel centers are not necessarily equally incremented, particularly for oblique LORs. For practical reasons, the coincidence window width is typically divided into K number of TOF bins with widths equal to $\Delta = 2\tau/K$, whereby

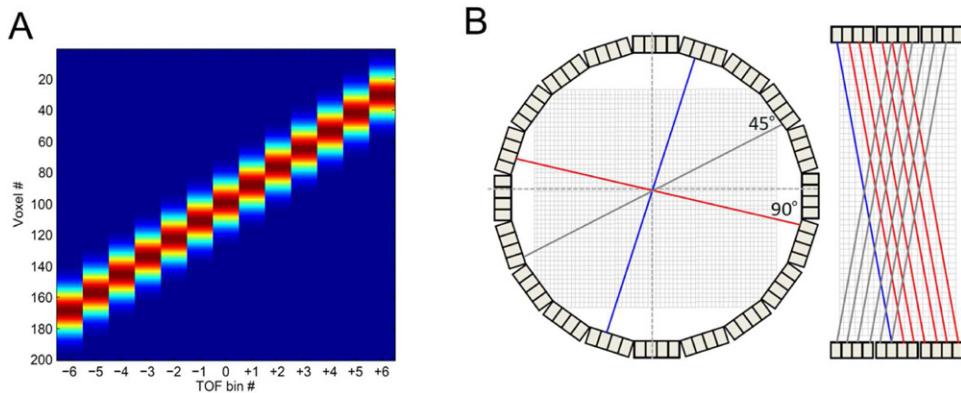


Figure 3. (A) TOF weighting coefficient matrix of a given LOR of the Siemens mCT scanner with 13 TOF bins, indicating the contribution of the intersected image voxels to each TOF bin. (B) The in-plane 90° and 45° rotation symmetries and axial translation and mirror symmetries used in the pre-computation of the geometric system matrix can also be used in the pre-computation of the TOF weighting matrix associated to each LOR. The blue lines in (B) indicate the base LOR or axial planes for which the system matrix is pre-computed.

the TOF bin boundaries $\{-\tau, \Delta - \tau, 2\Delta - \tau, \dots, \tau - \Delta, \tau\}$ will span the coincidence window width. The minimum number of TOF bins can be heuristically determined by:

$$K_{\min} = \text{Odd}\left(\left\lfloor \frac{4\tau}{R_t} \right\rfloor\right) \quad (9)$$

where $\text{Odd}(\lfloor \cdot \rfloor)$ returns the greatest lower odd integer. Based on Nyquist criterion, this rule ensures at least two TOF bin samples for the proper sampling of the TOF Gaussian kernel. Note that K must be an odd integer to centralize the pair of photons with almost zero arrival time difference on the center of the FOV corresponding to the central or zeroth TOF bin (see figure 2).

The pre-computation of \mathbf{W}_i is particularly efficient for backprojection where the counts in a given TOF bin should be distributed among the contributing voxels. In addition, it is noticeable that the pre-computation \mathbf{W}_i can be limited to a certain number of LORs based on the same base geometrical symmetries that are commonly used for the pre-computation of geometric system matrix. Therefore, substantial reduction of storage space and computation time can be achieved in 3D TOF PET reconstruction. As shown in figure 3(B), the base-symmetries include the in-plane $\pi/2$ and $\pi/4$ rotation symmetries and axial translation and mirror symmetries (Qi *et al* 1998). In 3D PET, the axial symmetries are of particular importance for storage reduction. In fact, for a scanner with span S and number of segments (or polar angles) N_{seg} , the system matrix should be pre-computed for only N_{sym} axial base planes as:

$$N_{\text{sym}} = \frac{S+1}{2} \left(\frac{N_{\text{seg}}+1}{2} - 1 \right) + 1 \quad (10)$$

For example, the mCT 128-Edge scanner accommodates emission data into 621 plans over 9 segments with span 11. Therefore, the system matrix can be stored for only 25 base axial planes, resulting in 96% reduction of storage requirement.

2.2. TOF bin subsetization

Based on the factored forward model in (5), the conventional TOF OSEM image reconstruction algorithm can be defined as:

$$\hat{\mathbf{f}}^{m+1,l} = \frac{\hat{\mathbf{f}}^{m,l}}{n_i a_i (\hat{\mathbf{G}}_i \odot \mathbf{W}_i^T \mathbf{1})} \odot \left(n_i a_i \hat{\mathbf{G}}_i \odot \mathbf{W}_i^T \frac{\mathbf{y}_i}{n_i a_i (\hat{\mathbf{G}}_i \odot \hat{\mathbf{f}}^{m,l}) \mathbf{W}_i + \bar{\mathbf{r}}_i + n_i \bar{\mathbf{s}}_i} \right), i \in Q_m \quad (11)$$

where $\mathbf{1} \in \mathbb{R}^K$, $\mathbf{y}_i = [y_{i1}, \dots, y_{iK}]^T$ and $\bar{\mathbf{s}}_i = [\bar{s}_{i1}, \dots, \bar{s}_{iK}]^T$ are the TOF concatenated vectors of prompts and expected scattered coincidences, $\{Q_m\}$ are M mutually exclusive subsets of projection bins and the division and addition operations are element-wise. Due to their random nature, the random coincidences are TOF independent and almost uniformly distributed over TOF bins. For storage reasons, in clinical scanners the random sinograms are first TOF integrated and then stored in 3D matrices. In (11), the expected randoms in TOF bins of the i th LOR can be approximated by $\bar{\mathbf{r}}_i = \frac{\bar{n}_i}{K} \mathbf{1}$.

We hypothesize that for a given TOF timing resolution, further convergence rate improvements of the MLEM and OSEM (for a given number of angular subsetization) algorithms can be achieved by subsetization of TOF data over the TOF bins. In connection to grouped coordinate ascent optimization algorithms, where a group of image voxels are sequentially updated, one can notice that in TOF MLEM/OSEM reconstruction only a group of image voxels corresponding to each TOF bin are updated. However, for a given TOF resolution, the voxel groups of adjacent TOF bins are spatially overlapped, which can impede the convergence rate. In this work, we propose the subsetization of TOF bins by sequential interleaving of the TOF bin data, thus reducing the coupling between groups of voxels.

Figure 4(A) shows the backprojection of an individual TOF bin of a simulated scanner with a TOF resolution of 350 ps and 23 TOF bins, for a given projection angle and all radial bins. The spatial uncertainty associated with this TOF resolution, $\Delta x = cR_t/2$ (Conti 2011a), is 5.25 cm. Figures 4(B) and (C) show the backprojections of all TOF bins and a subset of every three TOF bins. As can be seen, the backprojection of a subset of TOF bins leads to reduced coupling and cross-dependencies of image voxels. Figure 4(D) shows the spatial intensity profiles of odd and even TOF bins of the simulated scanner along a given LOR. As can be seen, the adjacent TOF bins are spatially overlapped, which will be more pronounced for better TOF timing resolution.

In TOF MLEM algorithm, all angular projections with all TOF bins are simultaneously employed to update the image, while in TOF OSEM a subset of those projections are used in each round of image update. Let $C = \{1, \dots, K\}$ be the set of TOF bins, where K is the total number of TOF bins. In the proposed TOF bin subsetization, the TOF bins are organized into L subsets with interleaved bin members, as follows:

$$\begin{aligned} S_1 &= \{1, 1 + L, 1 + 2L, \dots, K - 1\} \\ S_2 &= \{2, 2 + L, 2 + 2L, \dots, K - 2\} \\ &\vdots \\ S_L &= \{L, 2L, 3L, \dots, K - L\} \end{aligned} \quad (12)$$

such that the union of the subsets S_i spans the set C . For image reconstruction, one approach is to employ a given TOF subset for all projection angles (or subset of angles in the case of OSEM) and update the image sequentially by iterating over L TOF subsets (see supplemental material figure S1(A) (stacks.iop.org/PMB/61/1309/mmedia)). However, in each iteration of this approach only some regions of image space covered by the TOF member bins, will be

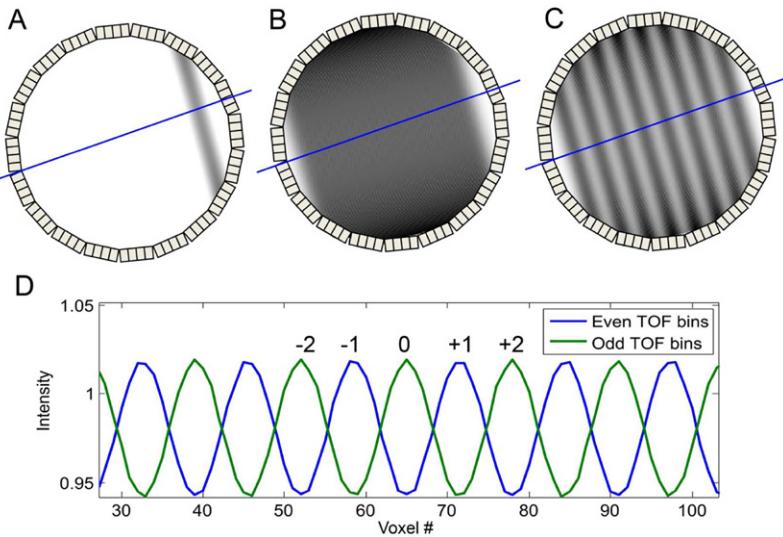


Figure 4. (A)–(C). Backprojection images of a single TOF bin, all TOF bins and every three TOF bins of a simulated scanner with 23 TOF bins along a given projection angle. (D) The intensity profiles of the backprojection of odd and even TOF bins, showing the coupling of the groups of voxels associated with adjacent TOF bins.

updated, which might render the sensitivity images (obtained by the backprojection of attenuation and normalization sinograms) to get zero values in non-updated regions, particularly with improved TOF resolutions. To address this possible issue and to provide a more balanced subsetization of TOF bins, we propose to employ different TOF subsets for adjacent projection angles with cyclic permutation, which ensures that all regions of image space are being updated in each round of image update. Figure 5(A) illustrates this approach for TOF subsetization of the MLEM algorithm with $L = 3$ subsets, at the first round of image update (out of three in this case). As shown, for projection angles 1 through 3, the subsets S_1, S_2, S_3 are used and this order is repeated for the angles 4 through 6 and so on. In the next rounds, for each projection angle, a TOF subset different from previous rounds should be used. Hence, in the second round, the cyclic permutation of the subsets used in the first round, i.e. S_2, S_3, S_1 is used for the projection angles 1 through 3 and so on. This cyclic pattern is repeated for the third round of update by assigning the subsets S_3, S_1, S_2 to projection angles 1 through 3 and so on (see supplemental material figure S1(B) (stacks.iop.org/PMB/61/1309/mmedia)). As shown in figure 5(B), the same approach is applicable to OSEM algorithm in which a subset of projection angles is used in each image update. This approach, however, necessitates to organize the TOF bins into a *valid* number of subsets based on the total number of projection angles, N_a , in MLEM or subset size, N_a/M , in OSEM reconstruction, that is, L must be chosen such that N_a or N_a/M are dividable into groups of angles to which subsets S_1, \dots, S_L are assigned. For instance, for the mCT scanner with $N_a = 168$ and $K = 13$, the valid number of TOF subsets in TOF subsetized MLEM are $\{1, 2, 3, 4, 6, 7, 8, 12\}$, while those of OSEM with 14 azimuthal subsets are $\{1, 2, 3, 4, 6, 12\}$. In the case of MLEM with $L = 3$ TOF subsets, the 168 angular projections are divided into 56 groups, each one containing 3 projection angles. The subsets S_1, S_2, S_3 are then assigned to each angle within the groups. In the case of OSEM with $L = 3$ TOF and $M = 14$ angular subsets, the 168/14 angular projections of each subset are divided into 4 groups, each one containing 3 projection angles. Similarly, the subsets S_1, S_2, S_3 are assigned to each angle of the groups.

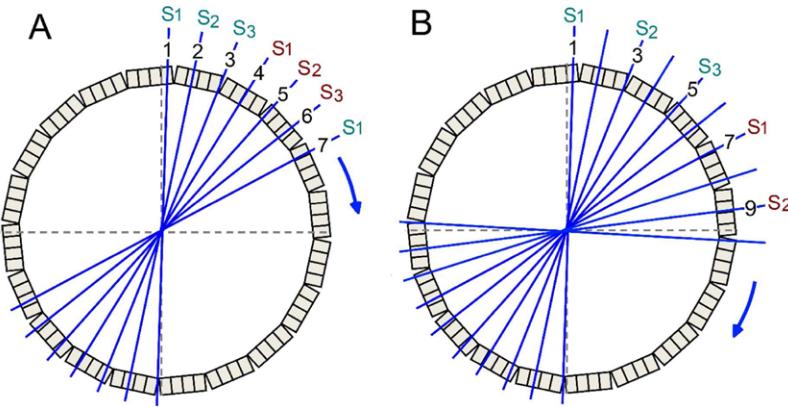


Figure 5. Illustration of the proposed TOF subsetization scheme with $L = 3$ TOF subsets, S_1, S_2 and S_3 . In this approach, the image is updated L times per iteration, each time with different TOF subsets for adjacent projection angles. Panel (A) shows the first round of update for TOF subsetized MLEM reconstruction. The subset S_1 is used during TOF forward/backprojection of projection angles $\{1, 1+L, 1+2L, \dots\}$. The subsets S_2 and S_3 are employed for angles $\{2, 2+L, 2+2L, \dots\}$ and $\{L, 2L, 3L, \dots\}$, respectively. In the next rounds of update, a subset different from previous updates is used for each projection angle (see supplemental material figure S1(B) (stacks.iop.org/PMB/61/1309/mmedia)). (B) The same idea is applicable for OSEM reconstruction, in this illustration, with $M = 2$ angular and $L = 3$ TOF subsets.

2.3. Implementation and evaluation

The proposed TOF system matrix pre-computation and TOF bin subsetization schemes were implemented and evaluated for image reconstruction on the Biograph mCT Flow Edge scanner. The PET subsystem of this TOF PET/CT scanner consists of 4 block rings accommodating 32448 physical LSO crystals with dimensions of $4 \times 4 \times 20$ mm 3 . The total number of physical and virtual crystals per ring is 672, which with an angular mashing factor of 2 gives rise to 168 angular projections. The effective detector radius with LOR depth of interaction is 43.72 cm. The PET data are acquired with a coincidence window width of 4.06 ns with a nominal TOF timing resolution of 580 ps and recorded in 13 TOF bins of 312 ps width. The 4D TOF sinograms are of dimensions: 400 radial bins, 168 azimuthal angles, 621 planes (in 9 segments with span 11) and 13 TOF bins. The CT subsystem is a 128-slice CT scanner (SOMATOM Definition AS+) with an extended transaxial FOV of 780 mm.

An in-house TOF PET reconstruction software was developed for the native geometry of this scanner using MATLAB R2013a, with single program multiple data (SPMD) parallel processing (MathWorks Inc., Natick, MA), running on a 12-core workstation with 2.4 GHz Intel Xeon processors and 32 GB random-access memory. In our SPMD implementation, the task of forward (or backward) projection of 641 sinogram planes was split into 25 tasks based on the axial base symmetries of the mCT scanner (see equation (9)), which were parallelized over multiple processors with different input data. Depending on the number of available and recruited processors, the parallelization was performed in multiple sessions to complete the 25 tasks. A task equalization technique was performed to reduce the idle time of recruited processors in each session. The SPMD requires the transfer of different parts of data (i.e. all sinogram planes covered by each base axial plane and the corresponding geometric and TOF weighting system matrices) to each processor. For increased efficiency, the data should be pre-loaded in memory, which obviously increases memory demand. To address this issue,

the radial and TOF bin trimming of both sinogram data and system matrices were performed prior to image reconstruction. Depending on the user-defined trimming values (which can be estimated from attenuation correction sinograms), this approach enables to discard non-informative sinogram data (i.e. random and scattered coincidences outside of subject's body contour) and their corresponding system matrix elements and therefore not only to reduce memory demand but also the reconstruction time.

The computation time reduction associated with TOF weighting matrix pre-computation was evaluated against on-the-fly TOF projector. For this purpose, we conducted 2D TOF forward projections using different TOF timing resolutions and image matrix sizes for the geometry of the mCT scanner. In these comparisons, the only difference in the implementation of the pre-computed and on-the-fly forward projectors was that in the former, the TOF weighting coefficients in equation (7) were pre-computed and in the latter they were calculated during run-time. In both cases, a pre-computed geometric system matrix and in-plane rotational symmetries were used.

For the evaluation of TOF subsetization, we performed simulation, experimental phantom and clinical studies. A 2D simulated NEMA phantom was used for convergence rate evaluation of the proposed technique using MLEM and OSEM reconstruction algorithms with TOF resolutions of 580 ps and 350 ps with 23 TOF bins, respectively, and with a range of different TOF subsets. Since the convergence to the underlying true activity image was of interest, only noise-free simulations were performed. The convergence rate was monitored by the logarithm of the normalized Euclidean distance between the image estimate at i th iteration \mathbf{f}^l and the ground truth image \mathbf{f}^{True} :

$$\eta(l) = \log\left(\frac{\|\mathbf{f}^l - \mathbf{f}^{\text{True}}\|}{\|\mathbf{f}^{\text{True}}\|}\right) \quad (13)$$

For the bias-variance evaluations, the NEMA IEC body phantom was filled with ~ 100 MBq ^{18}F -FDG and scanned on the mCT scanner for 15 min. PET images were reconstructed using a matrix size of $400 \times 400 \times 109$ with $4 \times 4 \times 2 \text{ mm}^3$ voxels. The bias-variance performance was assessed in terms of contrast recovery coefficients (CRCs) as a metric of bias and image roughness (IR) as a metric of variance, according to the international atomic energy agency protocol (IAEA 2009).

On each sphere a circular same-diameter region of interest (ROI) was drawn on the central plane to measure the mean activity concentration for each hot sphere (C_i^h). Similarly, for each sphere, 60 same-diameter ROIs were drawn on the background of the central plane and planes $\pm 1 \text{ cm}$ and $\pm 2 \text{ cm}$ away from that plane (12 ROIs per plane) to measure the mean estimated activity concentration for the i th size background sphere (C_i^b). For each sphere ($i = 1, \dots, 4$), the CRC and IR metrics were calculated by:

$$\begin{aligned} \text{CRC}_i &= \frac{(C_i^h/C_i^b) - 1}{(A_i^h/A_i^b) - 1} \times 100 \\ \text{IR}_i &= \frac{1}{N} \sum_{n=1}^N \text{IR}_{i,n}, \quad \text{IR}_{i,n} = \frac{1}{C_{i,n}^b} \sqrt{\sum_{j=1}^{N_i^b} \frac{(C_{i,n,j}^b - C_{i,n}^b)^2}{N_i^b - 1}} \end{aligned} \quad (14)$$

where A_i^h and A_i^b are measured activity concentration in the hot spheres and background, $N = 60$ is the number of background ROIs, $C_{i,n}^b$ is the mean activity concentration for the n th background ROI of the i th sphere, N_i^b is the number of pixels in the i th background sphere,

Table 1. The elapsed computation time (s) for the 2D TOF forward projections using pre-computed (Pr-c) and on-the-fly (O-t-f) TOF weighting coefficients for different TOF timing resolutions and number of TOF bins and image matrix sizes.

Image matrix size	580ps /13		350ps/23		100ps/81	
	O-t-f	Pr-c	O-t-f	Pr-c	O-t-f	Pr-c
200 × 200	494.8	2.2	585.7	6.2	674.2	31.8
400 × 400	1007.2	8.3	1039.5	41.6	1281.5	68.4

$C_{i,n,j}^b$ is the reconstructed activity concentration of the j th pixel of the n th background ROI of the i th sphere. In order to assess the overall quantitative performance, the IR and CRC values of the hot spheres were averaged based on the size of the spheres (number of voxels in ROIs corresponding to each sphere).

In addition, combination of different angular and TOF subsetization setups was evaluated to investigate if the convergence gain obtained by TOF subsetization can be equated by increasing the number of angular subsets used in conventional OSEM. Finally, the proposed subsetization approach was evaluated using a clinical ^{18}F -FDG study. The tumour-to-background ratio versus noise in each iteration was evaluated for different tumour sizes and TOF subset schedules. In this work, all data correction sinograms were obtained using the Siemens e7 VG-50 reconstruction software.

3. Results

3.1. Computation time

The computational speed-up gained from the pre-computation of the TOF weighting coefficients was evaluated against on-the-fly calculation using 2D TOF forward projections with TOF resolutions of 580, 350 and 100 ps and with 13, 23 and 81 TOF bins respectively, for image matrix sizes of 200×200 and 400×400 for the geometry of the mCT scanner. Table 1 compares the time elapsed for these forward projections. No radial and TOF bin trimmings were utilized. The results clearly demonstrate that the pre-computation of TOF weighting coefficients significantly reduces the computation time of the TOF forward projection and therefore the overall reconstruction time. In fact, on-the-fly calculation of TOF coefficients for a given LOR includes (i) the identification of the centre coordinate of the voxels intersected by the LOR, (ii) the projection of voxel centres onto the LOR, (iii) the calculation of arrival time difference μ of the photons emitted from the voxel centers according to equation (4) and (vi) the calculation of TOF weighting coefficients in equation (7). These mostly time-consuming steps are repeated for all LORs, therefore their pre-calculations are substantially cost-effective in 3D TOF reconstruction. Note that in our comparisons, we used code vectorization in the on-the-fly implementation of steps i) through iii), which effectively reduced their computation time.

3.2. Simulation and experimental studies

The convergence rate and bias-variance performances of the proposed TOF subsetization scheme were evaluated using 2D simulated and experimental NEMA IEC phantoms shown in figures 6(A) and (B). Since the range of valid TOF subsets is limited by the number of TOF bins, in our simulations we employed a TOF resolution of 350 ps with 23 TOF bins to

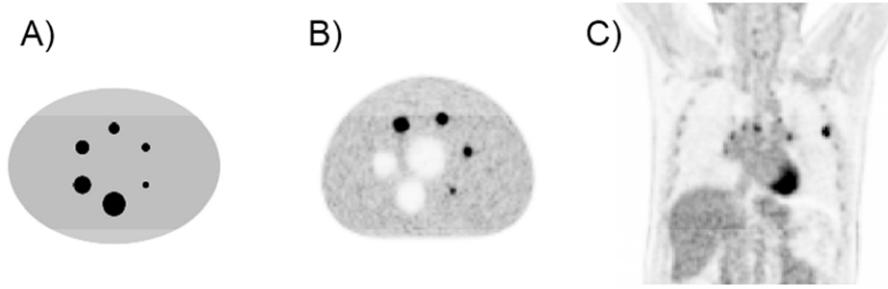


Figure 6. (A) Simulated 2D NEMA phantom. (B) The experimental NEMA IEC body phantom. (C) An ^{18}F -FDG PET clinical study with mediastinal lymphoma and chest wall invasion.

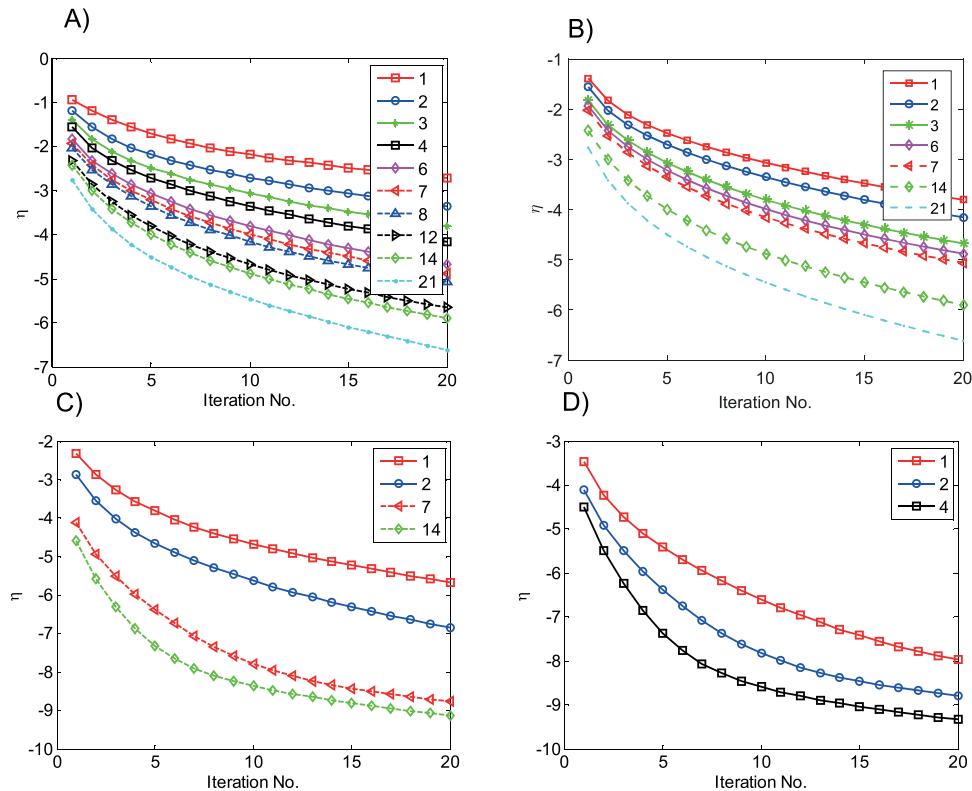


Figure 7. The convergence rate of (A) MLEM and (B)–(D) OSEM reconstruction algorithms in the simulated 2D NEMA phantom accelerated with different TOF subsets for a TOF timing resolution of 350 ps and 23 TOF bins. The legends indicate the number of TOF subsets. (B)–(D) show the TOF subsetized OSEM with 4, 12 and 42 angular subsets, respectively.

demonstrate the potential of TOF subsetization with a larger number of valid TOF subsets compared to the typical TOF resolution of 580 ps and 13 TOF bins.

Figure 7 shows the convergence rate of the TOF subsetized MLEM and OSEM algorithms in the simulated phantom based on the normalized Euclidean distance between image estimates and the ground truth image in equation (13). In figure 7(A), the results of the MLEM

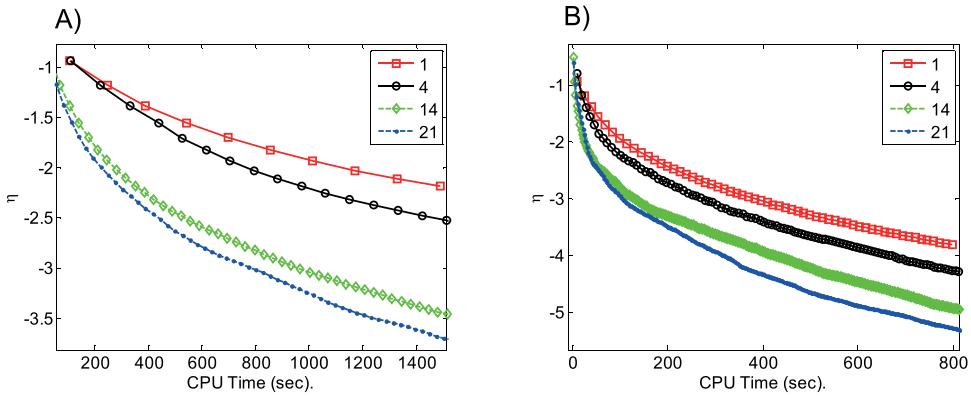


Figure 8. The computation time elapsed per image update of the TOF subsetized (A) MLEM and (B) OSEM (12 angular subsets) algorithms using the simulated NEMA phantom.

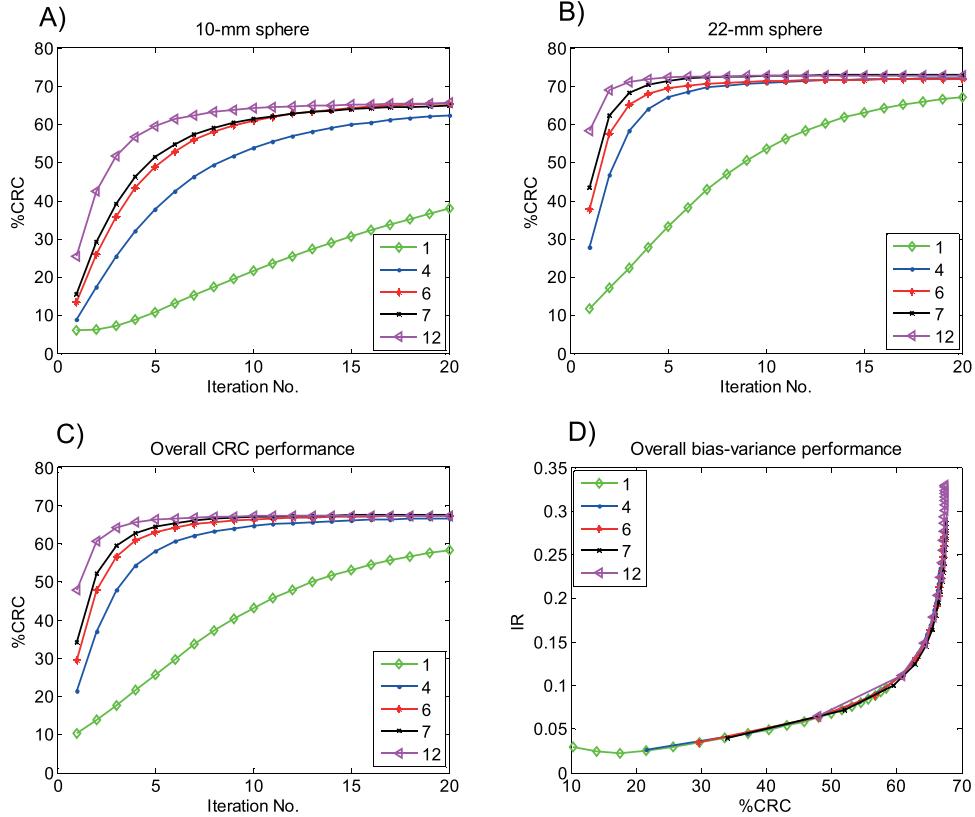


Figure 9. The contrast recovery coefficient (CRC) and image roughness (IR) performance of the TOF subsetized MLEM estimated using the experimental NEMA phantom. (A) and (B) CRC% versus Iteration number for the 10 and 22 mm spheres. (C) Overall CRC performance averaged over all hot spheres. (D) Overall bias-variance trade off TOF subsetization averaged over all hot spheres. The legends indicate the number of employed TOF subsets.

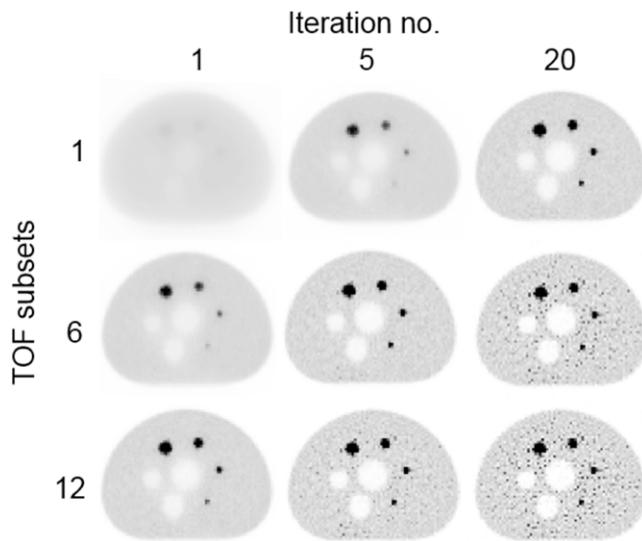


Figure 10. Reconstructed images of the experimental NEMA phantom at different iterations of the MLEM algorithm accelerated using different TOF subsets.

algorithm are shown, while those of the OSEM algorithm with 4, 12 and 42 angular subsets are shown in figures 7(B)–(D), respectively. As the results clearly show, the initial and overall convergence of both MLEM and OSEM algorithms increases consistently when increasing the number of TOF subsets. That is, for a given level of η , the algorithm with higher L meets that level after a fewer number of iterations. This improved performance should be attributed to the reduced cross-dependencies between image voxels associated with each TOF bin in addition to the fact that the TOF subsetization scheme makes use of part of the data to reconstruct the image similar to angular subsetization concept. Moreover, the results show that the convergence of OSEM with any angular subset can be further increased using additional subsetization over TOF bins.

The computational cost of the proposed method was also evaluated. Figures 8(A) and (B) show the convergence rate versus computation (CPU) time for MLEM and OSEM algorithms. Note that each data point corresponds to an image update instead of iteration number. As can be seen, for a given CPU time, a higher convergence gain is achieved by increasing the number of TOF subsets both in MLEM and OSEM reconstructions, showing the cost-effectiveness of the proposed approach. This gain is due to the fact that in each image update, a subset of TOF bins or, in other words, a subset of the columns of TOF weighting coefficient matrix shown in figure 3(A), is used along each LOR, leading to reduced arithmetic operations and computational burden.

The bias-variance trade-off of the proposed subsetization method was evaluated using the experimental NEMA phantom based on contrast recovery coefficient (CRC) and image roughness (IR) criteria. Figures 9(A)–(C) show the CRC versus iteration number for the 10 mm and 22 mm spheres and for all spheres averaged based on the size of the spheres (i.e. number of voxels in ROIs corresponding to each sphere), respectively. The results show that the convergence rate of the MLEM algorithm in terms of CRC can be significantly increased by TOF subsetization. There is a substantial convergence gain even starting with 4 TOF subsets. Figure 9(D) compares the overall bias-variance of the TOF subsetized MLEM algorithms

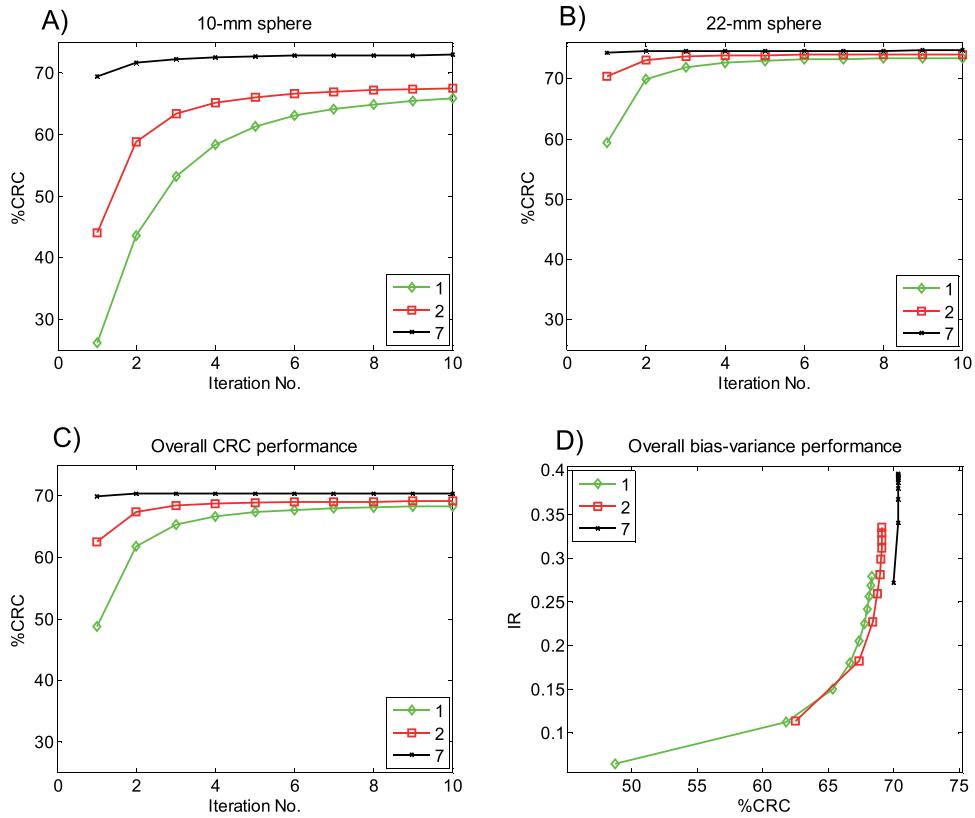


Figure 11. Same as figure 9, but for TOF subsetized OSEM with 12 angular subsets and 1, 2 and 7 valid TOF subsets.

averaged over all hot spheres. As can be seen, the accelerated MLEM methods have a relatively matched performance with the conventional MLEM, however, for higher numbers of TOF bins, such as $L = 12$, as the iteration number is increased, the contrast recovery gets saturated and the noise is increased, as expected from the conventional OSEM algorithms.

Figure 10 shows reconstructed images of the experimental NEMA phantom at different iterations of the MLEM algorithm accelerated using different TOF subsets. Consistent with the above-mentioned results, as the number of TOF subsets is increased, the hot spheres are recovered at a higher rate. However, with a higher number of iterations, the ML algorithm results in the maximum statistical consistency with the measured data and the images become noisier.

The proposed technique was also assessed for the acceleration of the OSEM algorithm using the NEMA phantom. Figures 11 and 12 present the results of contract recovery performance and bias-variance trades off for TOF subsetization of the OSEM algorithm with 12 and 21 angular subsets, respectively. Similar to figure 9, the contrast recovery coefficients for 10mm and 22 mm spheres and for all spheres together with overall CRC versus IR are shown. As expected, the results show that TOF subsetization improves the convergence of the OSEM algorithm for a given angular subset.

As can be seen, the CRCs of the 10mm sphere are noticeably higher for higher TOF subsets, while those of the 20mm sphere (the largest hot sphere in our experiment) are almost

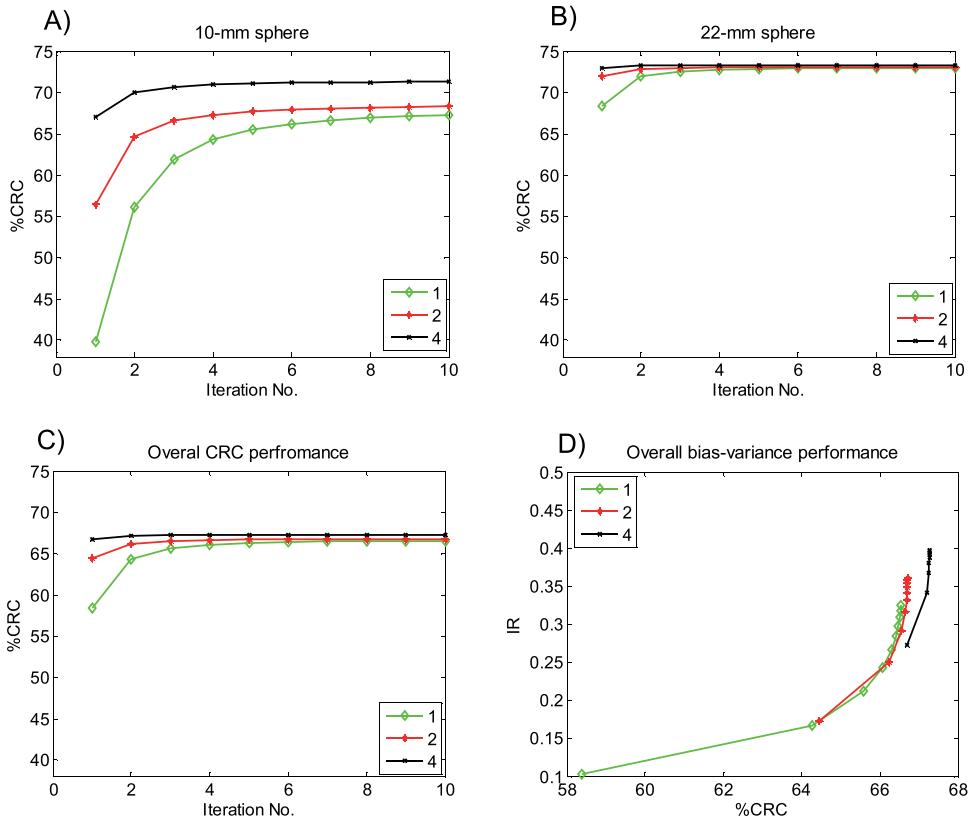


Figure 12. Same as figure 9, but for TOF subsetized OSEM with 21 angular subsets and 1, 2 and 4 valid TOF subsets.

comparable when using a high number of iterations. In figure 11(D), it is noticeable that 1 iteration of OSEM with 7 TOF subsets achieves as better overall bias-variance trade-off than 10 iterations of the conventional OSEM without TOF subsetization. A similar trend is observed in figure 12(D) for OSEM reconstructions with 4 TOF subsets. The higher CRC of the 10mm sphere in these figures should be ascribed to the fact that the TOF subsetization indeed reduces the voxel interference and cross-dependencies which eventually led to better recovery of small lesions.

Figures 13(A) and (B) compare the bias-variance performance of OSEM image reconstruction of the NEMA phantom for the 10 mm sphere and all hot spheres using a combination of different angular and TOF subsets with all resulting in 12 image updates per iteration. The results show that the studied reconstruction setups possess almost equal bias-variance performance. Therefore, as for angular subsetization, one might expect that the acceleration gain obtained by the proposed TOF subsetization is increased by the number of TOF subsets. Figure 13(B) shows the zoomed-in plots for the last 5 iterations. As can be seen, there are, however, subtle differences (<1%) in the performance of the reconstructions. Indeed, it turns out that based on the number of image updates, the convergence gain obtained from TOF subsetization can be roughly equated by increasing the number of angular subsets.

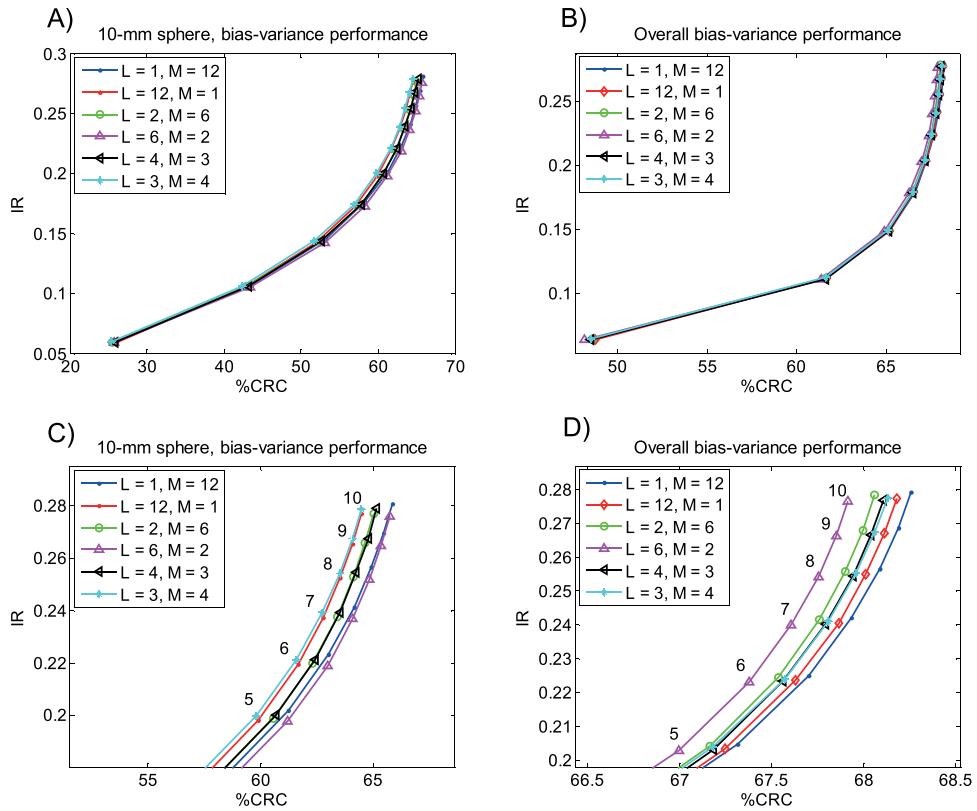


Figure 13. The bias-variance performance of the OSEM algorithm in the experimental NEMA phantom for different combinations of angular (M) and TOF subsets (L) for a total number of 10 iterations. (A) and (B) show the results for the 10 mm sphere and all hot spheres, respectively. (C) and (D) The zoomed-in plot of figures (A) and (B) over the last 5 iterations showing a slight difference in the performance of the studied reconstruction setups.

3.3. Clinical evaluation

The performance of the proposed TOF subsetization was evaluated for a clinical FDG-PET study with multiple lesions in the mediastinal and chest wall regions (figure 6(C)). Four ROIs were drawn on 4 lesions, plus a large circular ROI in the liver served as background uptake. Figure 14(A) indicates the selected lesions. The ratio of the average lesion uptake to average uptake in the liver (tumour to background ratio) versus the noise in the liver was evaluated for two different reconstruction set-ups. Figures 14(B) and (C) show the results of OSEM reconstruction with 4 angular subsets and 1 and 6 TOF subsets, and those of OSEM reconstruction with 14 angular subsets and 1 and 3 TOF subsets. These results were obtained for a total number of 10 iterations. The results show that a higher initial convergence is achieved by TOF subsetization, however in expense of higher noise at the same iteration number, which appears similar to the angular subsetization. In some lesions such as number 3 and 4, even a higher contrast is obtained by TOF subsetization which is in consistent with the results of 10 mm sphere in the NEMA phantom, where a higher CRC was observed too. The improved performance of the TOF-subsetized OSEM is more noticeable for the largest lesion with higher uptake. Overall, the results show that using TOF subsetization the lesion's contrast

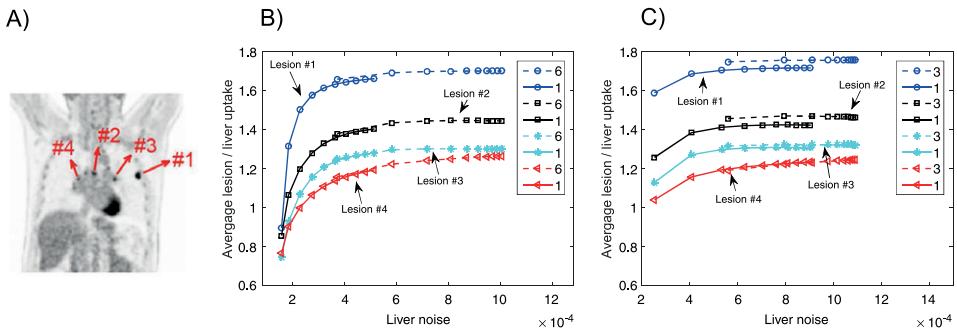


Figure 14. Lesion to background (liver) ratio versus noise in the liver per iteration. (A) The selected lesions in a clinical FDG-PET/CT scan. (B) and (C) The results of TOF subsetized OSEM reconstruction with 4 and 14 angular subsets, respectively. The legends indicate the number of TOF subsets used in each reconstruction.

almost reaches a plateau at the first iteration, leading to improved convergence of the OSEM algorithm. However, to avoid noise amplification early stopping or regularization is required.

4. Discussion

Model-based iterative image reconstruction techniques, such as OSEM, are the most widely used reconstruction algorithms in clinical PET imaging. The time-of-flight technology has recently dominated the design concept of commercial PET/CT scanners by providing higher image quality and improved lesion detectability. As the prevalence of TOF technology especially in PET/MRI scanners is growing (Zaidi *et al* 2011, Deller *et al* 2014), the importance of efficient TOF implementation and development of convergent model-based image reconstruction algorithms is increasing. In this work, we elaborated the practical aspects of TOF PET image reconstruction implementation using the pre-computation of TOF weighting coefficients matrix in conjunction with the base-geometric symmetries. In addition, we proposed a novel data subsetization scheme by partitioning the TOF data into a number of balanced TOF subsets to further improve the convergence of current TOF PET reconstruction algorithms.

Our results showed that the pre-computation of TOF weighting coefficients, instead of their on-the-fly calculation, can significantly reduce the reconstruction time, since the most time-consuming steps for indicating the contribution of emission points (i.e. voxel centres from which the annihilation photons are emitted) to each TOF bin, are stored in memory and instantly used during forward and backprojections. The pre-computation of the area under the TOF kernel as a function of path length (or arrival time) difference and TOF bin has been previously reported with the aim of accelerating TOF single-scatter simulation (Jin *et al* 2013). However, as described in the present work, all steps involved in the calculation of TOF coefficients (see section 3.1) can be precomputed in conjunction with TOF symmetries, which can lead to further acceleration of TOF reconstruction or scatter simulation.

For on-the-fly calculations, we utilized all common in-plane rotational symmetries, however, as noted in (Hong and Burbar 2013), other symmetries such as radial and TOF bin symmetries can be exploited to speed up on-the-fly TOF projection. The latter symmetries are also applicable for the pre-computation of TOF coefficients matrix in order to reduce the memory and storage demands. For clinical 3D TOF reconstruction implementation, we utilized the SPMD parallelization to speed up the reconstruction time through splitting

forward/back-projection of sinogram planes into a number of tasks, according to equation (9). This axial-plane parallelization scheme was chosen to direct SPMD resources toward utilizing the axial translation and mirror symmetries, which have the highest contribution on reducing the computation time, particularly in PET scanners with lower axial compression (span) and higher maximum ring difference. However, the SPMD methodology can also be used to parallelize all in-plane and axial symmetries (Hong and Burbar 2013), though it necessitates precisely-designed task equalization and job distribution between recruited processors in order to improve its efficiency due to dissimilarity of the tasks.

Motivated by grouped coordinate ascent algorithms, where a group of image voxels is sequentially updated, we proposed the partitioning of TOF data in each LOR into a number of TOF subsets. Thereby, the group of voxels associated with TOF bins are updated with reduced coupling. Therefore, one can expect an improved convergence rate using an interleaved TOF bin subsetization. Our simulation and experimental results show that the convergence rate of the MLEM and OSEM algorithms can be substantially increased using the proposed scheme. The results shown in figure 8 exemplified the computation cost-effectiveness of the proposed method, as at the same computation time an improved convergence can be obtained. Similar to angular subsetization, as the number of TOF subsets is increased, the computation time per image update is decreased and the overall convergence is increased. The bias-variance analysis of the experimental NEMA phantom using different reconstruction set-ups showed that the performance of the MLEM and OSEM algorithms can be improved. Specifically, the results showed that the overall convergence rate of the MLEM algorithm and the initial convergence of the OSEM algorithm can be noticeably improved using the proposed subsetization approach.

In (Van Slambrouck and Nuyts 2014), a patchwork image update scheme was proposed to further improve the convergence of the OSEM algorithm. It was shown that as the number of patches is increased, the convergence of the resulting reconstruction algorithm is markedly improved. However, the question that remains to be answered is whether the same convergence gain, obtained by new data or image subsetization techniques, can be achieved by the conventional OSEM through increasing the number of angular subsets. And if so, what are the potential advantages and promises of the new subsetization schemes, which can justify their application for accelerated PET image reconstruction. To explore these key questions, we compared the bias-variance performance of the MLEM algorithm by combining different angular and TOF subsets using the experimental NEMA phantom. As shown in figure 13, these combinations were chosen in such a way that the images are updated 12 times per iteration ($L \times M$). Interestingly, the results showed that the studied reconstruction set-ups, present similar bias-variance trade-off. In other words, the convergence gain obtained from TOF subsetization can indeed be equated by increasing the number of angular subsets. It is noticeable that the performance of the MLEM algorithm with 12 TOF subsets is to a large extent similar to the OSEM algorithm with 12 angular subsets. Although the results show equal performance between these two data partitioning schemes based on the number of image updates, we believe that the proposed TOF subsetization method has some potential and present many advantages that can open new avenues in TOF PET image reconstruction.

One of the promises of the proposed subsetization scheme is its exploitation in projection-based PSF PET image reconstruction using EM algorithms. In this reconstruction technique, the sinograms obtained after forward projection, are blurred in the radial, axial and angular directions and compared with the measured data to obtain a correction sinogram. During back-projection, the resulting correction sinogram is also blurred in the above-mentioned directions. For MLEM reconstruction, where all projection data are simultaneously used during forward/backward projection, such PSF-based blurring approach is feasible. However, for OSEM

reconstruction where only a subset of angular projections is used during the forward/backprojection, the sinogram blurring in the angular direction is not feasible. Therefore, in projection-based PSF OSEM reconstruction method, the angular component of PSFs is ignored (Panin *et al* 2006), which can undermine the benefits of the PSF reconstruction, especially in PET scanners with polygonal geometries (such as the Philips Ingenuity PET/MRI scanner (Zaidi *et al* 2011)) where the angular dependency of the scanner's PSF is considerable. In this regard, the acceleration of the MLEM algorithm using TOF subsetization is preferable since it would be possible to utilize all projection angles during forward/backprojections and therefore to include the angular component of projection-based PSF reconstruction.

In PET, the object is uniformly sampled across projection angles; thereby each projection view contains equally important emission data. In conventional angular subsetization, it is therefore important to choose the projection angles of each subset in such a way that the data are uniformly partitioned among subsets, which can be viewed as a balancing of the subsets. In TOF PET, the emission data of an object positioned at the centre of the FOV are preferentially recorded in central TOF bins than peripheral bins. This is more pronounced for the relatively poor TOF resolution of current TOF PET scanners (~600 ps). In this study, we proposed and evaluated an interleaved TOF subsetization to reduce the spatial cross-dependencies of neighbouring TOF bins. However, another possible approach, and promise of the proposed method, is the TOF subsetization through importance weighting of TOF bins to favour the fact that the content of data in central TOF bins is more than in peripheral ones. For this purpose, cumulative subsets (Hudson and Larkin 1994) can be exploited starting with a subset that initially contains the central TOF bins and gradually encompasses the peripheral bins. In this approach, the regions of the image where most of data come from are thus updated more than those regions with less important data. Therefore, one can expect an optimized way of image update using this cumulative TOF subsetization scheme.

One of the limitations of the OSEM algorithm is that this algorithm is not generally convergent and can enter a limit cycle. Although this should be proved by thorough investigation, one can expect that the TOF subsetized MLEM algorithm can also enter a similar limit cycle, depending on the subsetization scheme (interleaved or cumulative). One approach to avoid this possible limit cycle of the proposed subsetization is to gradually decrease the number of TOF subsets per iteration toward the conventional MLEM algorithm. This approach is consistent with the fact that the proposed approach especially improves the initial convergence of the employed EM algorithms.

To this end, it is worth to highlight the limitations of this study and the future directions for improving the proposed methodologies in the implementation and acceleration of TOF PET image reconstruction. To reduce the computational burden of on-the-fly TOF forward projection, the pre-computation of TOF weighting coefficients matrix was proposed. Although the results showed a substantial reduction in reconstruction time, this pre-computation technique imposes an additional demand on the memory, particularly as the number of TOF bins and therefore columns of the TOF weighting matrix are increased. However, there are a number of ways to meet this increased memory demand. As mentioned earlier, one approach is the application of radial and TOF bin trimming on both data and geometric and TOF system matrices. Another approach is to exploit additional radial and TOF bin symmetries as described in (Hong and Burbar 2013). Finally, the TOF weighting matrices are mainly sparse; therefore they can be effectively stored using sparse matrices. In our bias-variance analysis, a single noise level corresponding to 15 min scan of NEMA phantom was considered. The evaluation of the studied reconstruction set-ups under different noise levels can further characterize the performance of the algorithms. Future work falls into two directions, further evaluation of the proposed interleaved TOF subsetization scheme for different noise levels and for future

generation TOF PET scanners with improved TOF resolution and detailed assessment of new TOF subsetization schemes such as importance-weighted TOF subsets.

5. Conclusion

In this work, the practical aspects of TOF PET image reconstruction implementation in clinical TOF PET scanners were elaborated. In addition, the pre-computation of TOF weighting coefficients of LORs in conjunction with geometrical symmetries were proposed and evaluated against on-the-fly TOF calculation. Our results showed that this approach leads to a substantial reduction of reconstruction time. Moreover, we proposed the subsetization of EM algorithms using interleaved partitioning of TOF PET data into a number of TOF subsets, thereby reducing the spatial coupling of TOF bin data in image space. As a result, our simulation, experimental and clinical evaluations showed that the proposed TOF subsetization approach can significantly improve the convergence of the MLEM and OSEM algorithms.

Acknowledgments

This work was supported by the Swiss National Science Foundation under Grant SNSF 31003A-149957 and the Indo-Swiss Joint Research Programme ISJRP-138866.

References

- Abramowitz M and Stegun I A 1964 *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables* vol 55 (New York: Dover)
- Bouman C and Sauer K 1996 A unified approach to statistical tomography using coordinate descent optimization *IEEE Trans. Image Process.* **5** 480–92
- Browne J and De Pierro A R 1996 A row-action alternative to the EM algorithm for maximizing likelihoods in emission tomography *IEEE Trans. Med. Imaging* **15** 687–99
- Byrne C L 1998 Accelerating the EMML algorithm and related iterative algorithms by rescaled block-iterative (RBI) methods *IEEE Trans. Image Process.* **7** 100–9
- Conti M 2011a Focus on time-of-flight PET: the benefits of improved time resolution *Eur. J. Nucl. Med. Mol. Imaging* **38** 1147–57
- Conti M 2011b Why is TOF PET reconstruction a more robust method in the presence of inconsistent data? *Phys. Med. Biol.* **56** 155–68
- Deller T W, Grant A M, Khalighi M M, Maramraju S H, Delso G and Levin C S 2014 PET NEMA performance measurements for a SiPM-based time-of-flight PET/MR system *IEEE Nuclear Science Symp. and Medical Imaging Conf.* (Seattle, WA, 8–15 November 2014)
- El Fakheri G, Surti S, Trott C M, Scheuermann J and Karp J S 2011 Improvement in lesion detection with whole-body oncologic time-of-flight PET *J. Nucl. Med.* **52** 347–53
- Fessler J A 1994 Penalized weighted least-squares image reconstruction for positron emission tomography *IEEE Trans. Med. Imaging* **13** 290–300
- Fessler J A, Ficaro E P, Clinthorne N H and Lange K 1997 Grouped-coordinate ascent algorithms for penalized likelihood transmission image reconstruction *IEEE Trans. Med. Imaging* **16** 166–75
- Fessler J A and Hero A O 1995 Penalized maximum-likelihood image reconstruction using space-alternating generalized EM algorithms *IEEE Trans. Image Process.* **4** 1417–29
- Hong I and Burbar Z 2013 Ultra fast TOF 3D reconstruction using SIMD and symmetry superior to GPU implementation *IEEE Nuclear Science Symp. and Medical Imaging Conf.* pp 1–2
- Hudson H M and Larkin R S 1994 Accelerated image reconstruction using ordered subsets of projection data *IEEE Trans. Med. Imaging* **13** 601–9
- IAEA 2009 *IAEA Human Health Series 1: Quality assurance for PET and PET/CT Systems* (Vienna: International Atomic Energy Agency)

- Jin X, Chan C, Mulnix T, Panin V, Casey M E, Liu C and Carson R E 2013 List-mode reconstruction for the Biograph mCT with physics modeling and event-by-event motion correction *Phys. Med. Biol.* **58** 5567–91
- Lange K and Carson R 1984 EM reconstruction algorithms for emission and transmission tomography *J. Comput. Assist. Tomogr.* **8** 306–16 (PMID: 6608535)
- Mehranian A and Zaidi H 2015 Impact of time-of-flight PET on quantification errors in MRI-based attenuation correction *J. Nucl. Med.* **56** 635–41
- Nemallapudi M V, Gundacker S, Lecoq P, Auffray E, Ferri A, Gola A and Piemonte C 2015 Sub-100 ps coincidence time resolution for positron emission tomography with LSO:Ce codoped with Ca *Phys. Med. Biol.* **60** 4635–49
- Panin V Y, Kehren F, Michel C and Casey M 2006 Fully 3D PET reconstruction with system matrix derived from point source measurements *IEEE Trans. Med. Imaging* **25** 907–21
- Qi J and Leahy R M 2006 Iterative reconstruction techniques in emission computed tomography *Phys. Med. Biol.* **51** R541–78
- Qi J, Leahy R M, Cherry S R, Chatzioannou A and Farquhar T H 1998 High-resolution 3D Bayesian image reconstruction using the microPET small-animal scanner *Phys. Med. Biol.* **43** 1001–13
- Rahmim A, Qi J and Sossi V 2013 Resolution modeling in PET imaging: theory, practice, benefits, and pitfalls *Med. Phys.* **40** 064301
- Reader A J and Zaidi H 2007 Advances in PET image reconstruction *PET Clin.* **2** 173–90
- Schaefferkoetter J, Casey M, Townsend D and El Fakhri G 2013 Clinical impact of time-of-flight and point response modeling in PET reconstructions: a lesion detection study *Phys. Med. Biol.* **58** 1465–78
- Shepp L A and Vardi Y 1982 Maximum likelihood reconstruction for emission tomography *IEEE Trans. Med. Imaging* **1** 113–22
- van Dam H T, Borghi G, Seifert S and Schaart D R 2013 Sub-200 ps CRT in monolithic scintillator PET detectors using digital SiPM arrays and maximum likelihood interaction time estimation *Phys. Med. Biol.* **58** 3243–57
- Van Slambrouck K and Nuyts J 2014 Reconstruction scheme for accelerated maximum likelihood reconstruction: the patchwork structure *IEEE Trans. Med. Imaging* **61** 173–81
- Zaidi H, Ojha N, Morich M, Griesmer J, Hu Z, Maniawski P, Ratib O, Izquierdo-Garcia D, Fayad Z A and Shao L 2011 Design and performance evaluation of a whole-body Ingenuity TF PET-MRI system *Phys. Med. Biol.* **56** 3091–106
- Zhou J and Qi J 2014 Efficient fully 3D list-mode TOF PET image reconstruction using a factorized system matrix with an image domain resolution model *Phys. Med. Biol.* **59** 541–59