Multiple Importance Sampling for PET

László Szirmay-Kalos, Milán Magdics, and Balázs Tóth
Budapest University of Technology and Economics

***Abstract*—This paper proposes the application of Multiple Importance Sampling (MIS) in fully 3D PET to speed up the iterative reconstruction process. The proposed method combines the results of LOR driven and voxel driven projections keeping their advantages, like importance sampling, performance and parallel execution on GPUs. Voxel driven methods can focus on point like features while LOR driven approaches are efficient in reconstructing homogeneous regions. The theoretical basis of the combination is the application of the mixture of the samples generated by the individual importance sampling methods, emphasizing a particular method where it is better than others. The proposed algorithms are built into the Tera-tomo™ system.**

I. Introduction

In Positron Emission Tomography (PET) we need to find the spatial positron emission density. At a positron–electron annihilation, two oppositely directed photons are generated. A PET/CT system [TB02] collects the numbers $\mathbf{y} = (y_1, y_2, \ldots, y_{N_{\text{LOR}}})$ of simultaneous photon incidents in detector pairs, also called *Lines Of Responses* or *LORs*, while the material map of the examined object is obtained by a CT scan. The output of the reconstruction method is the *tracer density* function $x(\vec{v})$, which is approximated in a *finite function series* form:

$$x(\vec{v}) = \sum_{V=1}^{N_{\text{voxel}}} x_V b_V(\vec{v}), \tag{1}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_{N_{\text{voxel}}})$ are the coefficients to be computed, and $b_V(\vec{v})$ ($V = 1, \ldots, N_{\text{voxel}}$) are *basis functions*, which are typically defined on a *voxel grid*.

The correspondence between positron density $x(\vec{v})$ and the expected number of hits $\tilde{y}_L$ in LOR L is described by *scanner sensitivity* $\mathcal{T}(\vec{v} \to L)$ that expresses the probability of generating an event in LOR L given that a positron is emitted in point $\vec{v}$ of volume $\mathcal{V}$:

$$\tilde{y}_L = \int_{\vec{v} \in \mathcal{V}} x(\vec{v}) \mathcal{T}(\vec{v} \to L) \mathrm{d}v. \tag{2}$$

The scanner sensitivity is a high-dimensional integral of variables unambiguously defining the path of particles from positron emission point $\vec{v}$ to the detector electronics.

During iterative *Expectation Maximization* (ML-EM) reconstruction [SV82], the expected number of hits is computed for each LOR from the current positron density estimate, then the voxel coefficients are updated according to the ratios of the measured and expected number of hits [RZ07]. In every iteration cycle a large number of integrals need to be evaluated simultaneously, for which numerical quadrature rules can be applied [SKMT13].

The integrand of Equ. 2 is a product of source intensity $x(\vec{v})$ and scanner sensitivity $\mathcal{T}(\vec{v} \to L)$. In order to improve performance, instead of handling each emission point $\vec{v}$ or LOR L independently, it is worth simultaneously computing the scanner sensitivity for many emission points or for many LORs since this way we can *reuse* partial results obtained with other emission points and LORs. If we simultaneously obtain the scanner sensitivity for all (or many) LORs and a single emission point, then the method is called *voxel driven*. On the other hand, if we simultaneously compute the scanner sensitivity for many volume points and a single LOR, then the approach is *LOR driven*.

Whether a voxel driven or a LOR driven approach should be preferred depends on the following factors:

- *Importance sampling*: The error of the quadrature depends on the distribution of the samples, which should be concentrated where the integrand is large. In a LOR driven method voxels are identified at the end of the sampling process, thus their positron density cannot be mimicked by the sample density. A voxel driven approach, on the other hand, can start with selecting important voxels, thus is more accurate when the activity is concentrated in a few voxels.
- *Parallel processing*: The computational power required by PET reconstruction is available in GPUs [XM07], [GMDH08] favoring *gathering type computational threads* where each thread computes its own result from the shared input data without communication and synchronization. As ML-EM executes forward projections obtaining LOR values and back projections updating voxels, in forward projection a thread is preferred to be LOR driven while in back projection voxel driven.
- *Performance*: The performance difference of LOR driven and voxel driven approaches comes from the efficiency of selecting relevant LORs for a voxel or vice versa. In a LOR driven approach, the selection of voxels for a LOR is well supported by the *3D texturing hardware* of the GPU. In a voxel driven method, however, we should efficiently identify the LORs that are affected by this voxel in the 4D LOR (aka sinogram) space, for which no optimized texture format is available in GPUs.

Summarizing, both voxel driven and LOR driven approaches have certain, often complementary advantages, which also depend on the volume to be reconstructed. In this paper, we propose their combination in a way that their advantages can be preserved. From mathematical point of view, both approaches are sampling methods and integral quadratures.

The proposed method takes advantage of the mixture of samples of the two or more techniques in the sense of *Multiple Importance Sampling* (MIS) [VG95].

In Section II, we first review the theory of MIS. In Section III a LOR driven method and a voxel driven method for geometric projection with attenuation are analyzed and combined based on their sample densities. Section IV extends the idea to scatter compensation. Finally, we present results and evaluate the performance of the new method.

II. Multiple importance sampling

A Monte Carlo quadrature generates N sample points $\mathbf{z}_i$ randomly with probability density $p(\mathbf{z}_i)$ in the integration domain and divides integrand $f(\mathbf{z}_i)$ evaluated at the sample points by *sample density* $d(\mathbf{z}_i) = Np(\mathbf{z}_i)$:

$$\int f(\mathbf{z})\mathrm{d}\mathbf{z} \approx \frac{1}{N}\sum_{i=1}^{N}\frac{f(\mathbf{z}_i)}{p(\mathbf{z}_i)} = \sum_{i=1}^{N}\frac{f(\mathbf{z}_i)}{d(\mathbf{z}_i)}. \tag{3}$$

Suppose that we have M different quadrature schemes defined by densities $d_1, \ldots, d_M$ and using $N_1, \ldots, N_M$ number of samples for the same integral. The mixture of samples from all methods is characterized by the following density:

$$\hat{d}(\mathbf{z}) = \sum_{k=1}^{M} d_k(\mathbf{z}). \tag{4}$$

Thus the integral quadrature using the mixture of individual samples is

$$\int f(\mathbf{z})\mathrm{d}\mathbf{z} \approx \sum_{m=1}^{M}\sum_{i=1}^{N_m}\frac{f(\mathbf{z}_{m,i})}{\hat{d}(\mathbf{z}_{m,i})} = \sum_{m=1}^{M}\sum_{i=1}^{N_m}\frac{f(\mathbf{z}_{m,i})}{\sum_{k=1}^{M} d_k(\mathbf{z}_{m,i})}, \tag{5}$$

where $\mathbf{z}_{m,i}$ is the ith sample of the mth sampling method. Note that the application of the sample mixture means the addition of the estimators of different quadrature schemes and also the modification of their densities to a single common density that is the sum of individual sampling densities.

MIS can also be interpreted as an additional weighting of samples of the techniques to be combined:

$$\int f(\mathbf{z})\mathrm{d}\mathbf{z} \approx \sum_{m=1}^{M}\sum_{i=1}^{N_m}\lambda_m(\mathbf{z}_{m,i})\frac{f(\mathbf{z}_{m,i})}{d_m(\mathbf{z}_{m,i})}, \tag{6}$$

where the weighting scheme corresponding to Equ. 5 is

$$\lambda_m(\mathbf{z}) = \frac{d_m(\mathbf{z})}{\sum_{k=1}^{M} d_k(\mathbf{z})}. \tag{7}$$

This weighting is called *balance heuristics*. Combining unbiased estimators, the combined estimator will also be unbiased if the sum of weights $\sum_{m=1}^{M}\lambda_m(\mathbf{z})$ is 1 for any sample, which is true for balance heuristics.

Why this combination is worth doing can be understood if we consider the modification of a density where it was small or large in a particular method. If the density around sample $\mathbf{z}_{m,i}$ in a particular method m was great with respect to other combined methods, then this particular method puts samples in this neighborhood densely, and a sample represents a small domain fairly accurately. Thus, this particular method provides a more reliable estimate here than other methods, which should be preserved despite the estimates of other techniques. Indeed, the combination formula of Equ. 4 and the weight of Equ. 7 result in $\hat{d}(\mathbf{z}_{m,i}) \approx d_m(\mathbf{z}_{m,i})$ and $\lambda_m(\mathbf{z}_{m,i}) \approx 1$ if $d_m(\mathbf{z}_{m,i})$ is significantly larger than the densities of other methods, thus contribution $f(\mathbf{z}_{m,i})/d_m(\mathbf{z}_{m,i})$ of the samples of method m does not decrease in this region despite the addition of other estimators. On the other hand, if the density of method m is much smaller than other densities for particular sample $\mathbf{z}_{m,i}$, then the samples of this method in this neighborhood is sparse and the estimate is unreliable. Thus, this contribution should be suppressed, which is achieved by Equ. 4 or Equ. 7 resulting in $d_m(\mathbf{z}_{m,i}) \ll \hat{d}(\mathbf{z}_{m,i})$ and $\lambda_m(\mathbf{z}_{m,i}) \approx 0$.

The limits of the method can be understood by considering the objective of importance sampling. The variance of the estimator is small if density $\hat{d}(\mathbf{z})$ mimics integrand $f(\mathbf{z})$ and is as large as possible. If we include more estimators, density $\hat{d}(\mathbf{z})$ will increase, which is a positive effect. However, if an included estimator is so bad that it makes the combined density less proportional to the integrand than other estimators would do, then the variance of the combined estimator may be higher than the original one. When it happens, we can solve the problem with preferring good sampling methods even more than suggested by their relative density. For example, in Equ. 7 the weights can be defined as

$$\lambda_m(\mathbf{z}) = \frac{d_m^{\alpha}(\mathbf{z})}{\sum_{k=1}^{M} d_k^{\alpha}(\mathbf{z})}, \tag{8}$$

which still guarantees that the sum of weights is equal to 1, but suppresses methods more in regions where they have small density if power α is greater than 1. This weighting is called *power heuristics* [VG95]. When $\alpha = 1$ we get balance heuristics back. The other extreme case, called *maximum heuristics*, corresponds to $\alpha = \infty$, when $\lambda_m(\mathbf{z}) = 1$ if $d_m(\mathbf{z})$ is greater than all other densities $d_k(\mathbf{z})$, and zero otherwise.

III. Application to the computation of unscattered contribution

For the sake of simplicity, we first assume that detectors are ideally black, ignore positron range and scattering, and consider only phantom attenuation in this section. However, even this simplified approach can be applied in systems of physically plausible simulation if the system sensitivity is factored. On the other hand, the basic idea is extended to incorporate scattering in Section IV.

If positron range and acollinearity are ignored, annihilation photons born in point $\vec{v}$ have two opposite directions $\vec{\omega}$ and $-\vec{\omega}$ of uniform distribution and this pair contributes to a LOR if the line of position $\vec{v}$ and direction $\vec{\omega}$ crosses the surfaces of the LOR's two detectors and none of the photons gets scattered or absorbed (Fig. 1). Let us denote the intersections of this line with the detector surfaces by $\vec{u}$ and $\vec{w}$, which are unambiguously determined by line point $\vec{v}$ and direction $\vec{\omega}$. As the photon direction has uniform distribution on the half sphere

Ω_H of solid angle 2π, the scanner sensitivity assuming zero number of scattering is an integral over the set of directions:

$$\mathcal{T}_0(\vec{v} \to L) = \int\limits_{\vec{\omega} \in \Omega_H} \frac{1}{2\pi} A(\vec{u}, \vec{w}) \xi_L(\vec{u}, \vec{w}) \mathrm{d}\omega, \tag{9}$$

where ξ_L is the indicator function that is 1 if intersection points $\vec{u}$ and $\vec{w}$ belong to the crystals of LOR L, and

$$A(\vec{u}, \vec{w}) = \exp\left(- \int\limits_{\vec{l}=\vec{u}}^{\vec{w}} \sigma_a(\vec{l}) + \sigma_s(\vec{l}) \mathrm{d}l \right) \tag{10}$$

is the *attenuation factor* defined as the line integral of absorption cross section σ_a and scattering cross section σ_s. The attenuation factor expresses the probability that photons get neither absorbed nor scattered in the measured object.

Including scanner sensitivity $\mathcal{T}_0(\vec{v} \to L)$ defined in Equ. 9 into Equ. 2, the formula of expected hits from unscattered photons is

$$\tilde{y}_L^{(0)} = \int\limits_{\mathcal{V}} \int\limits_{\Omega_H} \frac{x(\vec{v})}{2\pi} A(\vec{u}, \vec{w}) \xi_L(\vec{u}, \vec{w}) \mathrm{d}\omega \mathrm{d}v. \tag{11}$$

In the following subsections we discuss a LOR driven method and a voxel driven method to compute these LOR integrals. When different quadrature schemes are studied, we find their corresponding densities $d_m(\vec{v}, \vec{\omega}) = N_m p_m(\vec{v}, \vec{\omega})$.

A. LOR driven sampling

If photon paths are linear, annihilation point $\vec{v}$ and direction $\vec{\omega}$ unambiguously identify detector hit points $\vec{u}$ and $\vec{w}$, or alternatively, from detector hit points $\vec{u}$ and $\vec{w}$, we can determine those annihilation points $\vec{v}$ and directions $\vec{\omega}$, which can contribute: contributing annihilation points are on the line segment between $\vec{u}$ and $\vec{w}$ and direction $\vec{\omega} = \vec{\omega}_{\vec{u} \to \vec{w}}$ points from $\vec{u}$ to $\vec{w}$. We modify our view point from the annihilation points and directions to detector points, and using the correspondence between them, the detector response is expressed as an integral over the detector surfaces.

The Jacobian of the change of integration variables is

$$J_L(\vec{u}, \vec{w}) = \frac{\mathrm{d}\omega \mathrm{d}v}{\mathrm{d}l \mathrm{d}w \mathrm{d}u} = \frac{\cos\theta_{\vec{u}} \cos\theta_{\vec{w}}}{|\vec{u} - \vec{w}|^2}, \tag{12}$$

where $\theta_{\vec{u}}$ and $\theta_{\vec{w}}$ are angles between the surface normals and the line connecting points $\vec{u}$ and $\vec{w}$ (Fig. 1).

Including the Jacobian of the change of integration variables into Equ. 11, the expected number of hits can be expressed as a triple integral over the two detector surfaces D_1 and D_2 of the given LOR and over the line segment connecting points $\vec{u}$ and $\vec{w}$ belonging to the two detector surfaces:

$$\tilde{y}_L^{(0)} = \int\limits_{D_1} \int\limits_{D_2} \int\limits_{\vec{l}=\vec{u}}^{\vec{w}} \frac{x(\vec{l})}{2\pi} A(\vec{u}, \vec{w}) J_L(\vec{u}, \vec{w}) \mathrm{d}l \mathrm{d}w \mathrm{d}u. \tag{13}$$

As the integration domain is the surface points of LOR L, indicator function $\xi_L(\vec{u}, \vec{w})$ has value 1 for all samples.

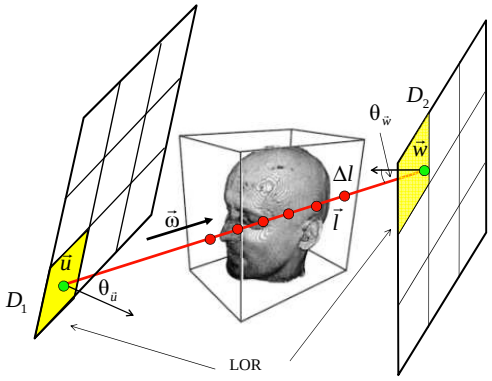

Fig. 1. A single computational thread of the LOR driven projection marches on rays between sample points $\vec{u}, \vec{w}$ of the detector surfaces.

The three integrals can be estimated analytically or with MC quadrature, or we can even mix the two approaches and some integrals are estimated with simple analytical formula while others are computed from random samples. For example, the integrals over the pair of detector surfaces can be estimated taking random point pairs on the LOR's two detectors, and for the evaluation of the line integrals Siddon's algorithm [Sid85] or Joseph's method [Jos82] can be used. The *distance-driven approach* [MB04], on the other hand, would sample only one endpoint and would simultaneously approximate the surface integral of the other endpoint and the line integral. *Solid angle based methods* [QLC+98] approximate surface integrals. Approximating some integrals deterministically we can increase the accuracy when low number of MC samples are used. However, approximations have a deterministic error which makes the method biased, i.e. the error will not converge to zero when the number of MC samples goes to infinity. The proposed MIS scheme can be plugged in all these approaches to make the MC quadrature more accurate while the error of deterministic approximations is not affected.

Here we demonstrate MIS with a simple algorithm where all three integrals are computed with unbiased MC quadrature. The integrals over the pair of detector surfaces are estimated by N_{ray} discrete line samples, i.e. point pairs $(\vec{u}_i, \vec{w}_i)$ on the LOR's two detectors. A point is sampled from uniform distribution on the crystal surface of size $D = |D_1| = |D_2|$. We take N_{march} equidistant points $\vec{l}_{ij}$ started at a random offset on each line segment $(\vec{u}_i, \vec{w}_i)$ and approximate the line integral with *ray marching*. Step size Δl_i is the length of the line segment divided by N_{march}. The random offset, i.e. the distance of the first point of ray marching from the start of the line segment is uniformly distributed in $[0, \Delta l_i]$. This Monte Carlo strategy is called *systematic sampling*.

Note that the implementation of this algorithm does not need conditional instructions, computational threads execute the same instruction sequence and thus do not diverge, and the method is very fast if $x(\vec{l}_{ij})$ is fetched from a 3D texture of the GPU since the probability that neighboring threads

need neighboring voxels is high, thus the texture cache works efficiently. As this approach takes only discrete point samples in the volume of interest, it can be used for arbitrary finite element basis. Tri-linear interpolation is directly supported by the texturing hardware and higher order splines can also be traced back to tri-linear interpolation [SH05].

With these, the integral estimator is:

$$\tilde{y}_L^{\mathrm{A1}} = \frac{D^2}{N_{\mathrm{ray}}} \sum_{i=1}^{N_{\mathrm{ray}}} \sum_{j=1}^{N_{\mathrm{march}}} \frac{x(\vec{l}_{ij})}{2\pi} A(\vec{u}_i, \vec{w}_i) J_L(\vec{u}_i, \vec{w}_i) \Delta l_i. \quad (14)$$

Comparing this estimator to the integrand of Equ. 11, we can conclude that the density of this LOR driven approach is:

$$\tilde{y}_L^{\mathrm{A1}} = \sum_{i=1}^{N_{\mathrm{ray}}} \sum_{j=1}^{N_{\mathrm{march}}} \frac{x(\vec{l}_{ij}) A(\vec{u}_i, \vec{w}_i)/(2\pi)}{d^{\mathrm{A1}}(\vec{l}_{ij}, \vec{\omega}_{\vec{u}_i \to \vec{w}_i})} \quad \Longrightarrow$$

$$d^{\mathrm{A1}}(\vec{l}, \vec{\omega}) = \frac{N_{\mathrm{ray}}}{D^2 J_L(\vec{u}, \vec{w}) \Delta l}, \quad (15)$$

where $\vec{u}$, $\vec{w}$ and Δl can be unambiguously determined from the line of $\vec{l}$ and $\vec{\omega}$, and from the geometry of the detector and the volume to be reconstructed.

The LOR centric method has several advantages in forward projection. As it is a gathering algorithm, it requires no atomic operations on the GPU. On the other hand, it samples annihilation points occupying the 3D space, which can be well supported by the 3D texture hardware.

B. Voxel driven sampling

In voxel driven sampling, first annihilation point $\vec{v}$ is selected, then directions are found for lines crossing in this annihilation point. In order not to sample directions that would not intersect detector modules and thus would not result in LORs, we change the variables in the integral. Direction $\vec{\omega}$ is expressed as the difference of detector point $\vec{u}$ and emission point $\vec{v}$. The Jacobian of this substitution is

$$J_V(\vec{u}, \vec{v}) = \frac{\mathrm{d}\omega}{\mathrm{d}u} = \frac{\cos\theta_{\vec{u}}}{|\vec{u} - \vec{v}|^2}, \quad (16)$$

where $\theta_{\vec{u}}$ is the angle between direction $\vec{v} - \vec{u}$ and the surface normal of the detector. Expected value $\tilde{y}_L^{(0)}$ in LOR L is

$$\tilde{y}_L^{(0)} = \int_{\mathcal{V}} \int_{D_1} \frac{x(\vec{v})}{2\pi} A(\vec{u}, \vec{w}) \xi_L(\vec{u}, \vec{w}) J_V(\vec{u}, \vec{v}) \mathrm{d}u \mathrm{d}v, \quad (17)$$

where $\vec{w}$ is the intersection point of the detector surface and the line connecting points $\vec{u}$ and $\vec{v}$.

This integral is estimated with MC quadrature. First N_v annihilation points $\vec{v}_i$ are sampled with a probability density that is proportional to the activity according to the principles of importance sampling:

$$p_{\vec{v}}(\vec{v}) = \frac{x(\vec{v})}{\mathcal{X}}, \qquad \mathcal{X} = \int_{\mathcal{V}} x(\vec{v}) \mathrm{d}v, \quad (18)$$

where $\mathcal{X}$ is the total activity. Based on the finite function series form of $x(\vec{v})$ (Equ. 1), the sampling process has two steps. First, a voxel coefficient x_V is found with discrete probability $x_V / \sum x_V$. The second step is to sample with basis function $b_V(\vec{v})$. If piece-wise constant basis functions are used, then $\vec{v}_i$ is uniformly distributed inside the voxel. In case of tri-linear interpolation or higher order spline basis functions, we could exploit that these basis functions can be obtained as the self convolution of box functions. As the probability density of the sum of two independent random variables is the convolution of the individual densities, samples can be generated with tri-linear density by adding two uniform random samples obtained with constant density, with quadratic splines by adding three uniform random samples, etc.

For each annihilation point $\vec{v}_i$, we put just a single sample $\vec{u}_i$ randomly generated with a uniform probability on each crystal. As the crystal area is D, the density of $\vec{u}_i$ is $1/D$ (Fig. 2), and the estimated number of hits in LOR L is

$$\tilde{y}_L^{\mathrm{A2}} = \frac{D}{N_v} \sum_{i=1}^{N_v} \frac{x(\vec{v}_i)}{2\pi p_{\vec{v}}(\vec{v}_i)} A(\vec{u}_i, \vec{w}_i) \xi_L(\vec{u}_i, \vec{w}_i) J_V(\vec{u}_i, \vec{v}_i). \quad (19)$$

Note that this sampling method may generate line samples that do not intersect crystal surface D_2 of LOR L. However, in these cases the integrand in Equ. 11 is zero since indicator function ξ_L is also zero.

Comparing this estimator to the integrand of Equ. 11, we obtain the following density for this voxel driven approach:

$$\tilde{y}_L^{\mathrm{A2}} = \sum_{i=1}^{N_v} \frac{x(\vec{v}_i) A(\vec{u}_i, \vec{w}_i) \xi_L(\vec{u}_i, \vec{w}_i)/(2\pi)}{d^{\mathrm{A2}}(\vec{v}_i, \vec{\omega}_{\vec{u}_i \to \vec{v}_i})} \quad \Longrightarrow$$

$$d^{\mathrm{A2}}(\vec{v}, \vec{\omega}) = \frac{N_v p_{\vec{v}}(\vec{v})}{D J_V(\vec{u}, \vec{v})} = \frac{N_v x(\vec{v})}{D \mathcal{X} J_V(\vec{u}, \vec{v})}. \quad (20)$$

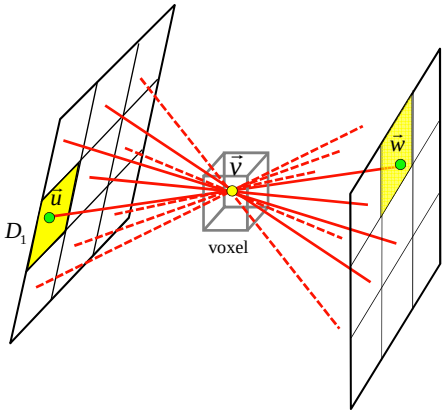

Fig. 2. A single computational thread of the voxel driven projection samples $\vec{v}$ in proportion to positron density $x(\vec{v})$ and processes a line crossing this point for each LOR.

The voxel driven method can focus on high activity regions. However, it requires atomic operations in forward projection and a single thread accesses many LORs stored in a 4D data structure, which are slow on the GPU.

C. Combined sampling

According to the theory of MIS, when two methods are combined, the sampling algorithms are left unchanged, only the sample weights are modified to include the density of all combined methods. Then, the estimators of different techniques are simply added. The combined density is

$$\hat{d}(\vec{v},\vec{\omega}) = \frac{N_{\text{ray}}}{D^2 J_L(\vec{u},\vec{w})\Delta l} + \frac{N_v x(\vec{v})}{D\mathcal{X} J_V(\vec{u},\vec{v})}. \tag{21}$$

With balance heuristics, the modified LOR driven and voxel driven projections compute the following estimates:

$$\hat{y}_L^{\text{A1}} = \sum_{i=1}^{N_{\text{ray}}} \sum_{j=1}^{N_{\text{march}}} \frac{x(\vec{l}_{ij})A(\vec{u}_i,\vec{w}_i)/(2\pi)}{\hat{d}(\vec{l}_{ij},\vec{\omega}_i)},$$

$$\hat{y}_L^{\text{A2}} = \sum_{i=1}^{N_v} \frac{x(\vec{v}_i)A(\vec{u}_i,\vec{w}_i)\xi_L(\vec{u}_i,\vec{w}_i)/(2\pi)}{\hat{d}(\vec{v}_i,\vec{\omega}_i)}. \tag{22}$$

The final estimator is the sum of the two estimates:

$$\tilde{y}_L^{(0)} \approx \hat{y}_L^{\text{A1}} + \hat{y}_L^{\text{A2}}. \tag{23}$$

The implementation of the combined sampling scheme has two phases. First, based on the current activity distribution a LOR centric projection is executed, which initializes every LOR value $\hat{y}_L^{\text{A1}}$. In this phase a computation thread is responsible for a LOR. Then, a voxel centric projection is run in parallel, where each thread adds its contribution $\hat{y}_L^{\text{A2}}$ to the affected LOR values. Sample points $\vec{v}_i$ of the voxel centric method are generated on the CPU, and a separate thread is started for every sample point to compute the contribution of this point to all LORs meeting here. While the LOR driven method is of gathering type in forward projection, it is of scattering type in back projection. On the other hand, voxel driven methods are of scattering type in forward projection and of gathering type in back projection. Thus, in the combined method it is worth preferring LOR sampling or voxel sampling depending on whether we execute forward or back projection.

IV. Application to scattering materials

MIS can be used also for physically more plausible projection models. Here we consider scattering in the measured object. In case of scattering, the system sensitivity and the expected hits are high dimensional integrals, which can be expressed as summing the contributions of paths representing increasing number of scattering events

$$\mathcal{T}(\vec{v}\to L) = \sum_{S=0}^{\infty} \mathcal{T}_S(\vec{v}\to L), \tag{24}$$

$$\tilde{y}_L = \sum_{S=0}^{\infty} \tilde{y}_L^{(S)} = \sum_{S=0}^{\infty} \int_{\vec{v}\in\mathcal{V}} x(\vec{v})\mathcal{T}_S(\vec{v}\to L)\mathrm{d}v, \tag{25}$$

where $\mathcal{T}_S(\vec{v}\to L)$ is the probability that a photon pair born in $\vec{v}$ undergoes exactly S scattering events in total and contributes to LOR L.

We consider two samplers, the first is the LOR driven projector with attenuation of Section III-A, which can compute only the unscattered contribution $\tilde{y}_L^{(0)}$. The second sampler is a Monte Carlo Photon Tracer that can handle direct contribution $\tilde{y}_L^{(0)}$ as well as single and multiple scattering $\sum_{S=1}^{\infty}\tilde{y}_L^{(S)}$.

A. Voxel driven Monte Carlo Photon Tracer

In scattering media, the contribution to a LOR, i.e. Equ. 2 is an infinite dimensional integral over the photon path space. In *Photon Tracing* (*PT*), first annihilation point $\vec{v}$ is sampled with a density that is proportional to the activity, then the paths of the two annihilation photons are obtained with scanner sensitivity $\mathcal{T}(\vec{v}\to L)$. To do this, initial direction $\vec{\omega}$ is drawn from uniform distribution. Two photons are started from the annihilation point and their free paths are sampled to find the photon–material interaction points. At scattering, a new direction is generated mimicking the Klein–Nishina formula, and the photon energy is adjusted according to the Compton law. When one of the photons leaves the detector or its energy drops below the discrimination threshold, the photon pair is lost and no LOR is contributed. If photons hit the detector surface, the simulation of this path is terminated and the affected LOR is given contribution $\mathcal{X}/N_{\text{PT}}$ where $\mathcal{X}$ is the total activity and N_{PT} is the number of simulated paths. The estimator is

$$\tilde{y}_L^{\text{PT}} = \frac{\mathcal{X}}{N_{\text{PT}}}\#(\text{hits in } L) \implies d^{\text{PT}} = N_{\text{PT}}\frac{x(\vec{v})\mathcal{T}(\vec{v}\to L)}{\mathcal{X}}. \tag{26}$$

This general formula has a simpler form for the case when the number of scattering events is zero:

$$d_0^{\text{PT}}(\vec{v},\vec{\omega}) = N_{\text{PT}}\frac{x(\vec{v})A(\vec{u},\vec{w})\xi_L(\vec{u},\vec{w})}{2\pi\mathcal{X}}. \tag{27}$$

B. Combined method

We run the two projections compensating with the combined density and add up their contributions. The first method is a LOR driven estimator of the unscattered component with density $d^{\text{A1}}(\vec{l},\vec{\omega})$, which can be controlled by the number of rays per LOR, N_{ray}, and the number of ray marching steps per LOR, N_{march}. The second method is the Particle Tracer that estimates paths of arbitrary lengths and has density d^{PT}.

The unscattered contribution is estimated by both methods, so their densities should be added when an unscattered path is obtained. With balance heuristics, the LOR driven unscattered estimator becomes

$$\hat{y}_L^{\text{A1}} = \sum_{i=1}^{N_{\text{ray}}} \sum_{j=1}^{N_{\text{march}}} \frac{x(\vec{l}_{ij})A(\vec{u}_i,\vec{w}_i)/(2\pi)}{d^{\text{A1}}(\vec{l}_{ij},\vec{\omega}_i) + d_0^{\text{PT}}(\vec{l}_{ij},\vec{\omega}_i)}, \tag{28}$$

where $\vec{u}_i$ and $\vec{w}_i$ are the intersection points of the ray and the detector surfaces. The PT sampler should separate unscattered paths and add the following contribution to the affected LOR:

$$\hat{y}_L^{\text{PT},(0)} = \sum_{i=1}^{N_{\text{PT}}} \frac{x(\vec{v}_i)A(\vec{u}_i,\vec{w}_i)\xi_L(\vec{u}_i,\vec{w}_i)/(2\pi)}{d^{\text{A1}}(\vec{v}_i,\vec{\omega}_i) + d_0^{\text{PT}}(\vec{v}_i,\vec{\omega}_i)}, \tag{29}$$

where $\vec{\omega}_i$ is the direction between hit points $\vec{u}_i$ and $\vec{w}_i$.

The scattered contribution is computed only by PT, so its estimator is unchanged:

$$\hat{y}_L^{\mathrm{PT},(1+)} = \frac{\mathcal{X}}{N_{\mathrm{PT}}} \#(\text{scattered hits in } L). \tag{30}$$

The combined estimator is the sum of the estimators of the elementary methods:

$$\tilde{y}_L \approx \hat{y}_L^{\mathrm{A1}} + \hat{y}_L^{\mathrm{PT},(0)} + \hat{y}_L^{\mathrm{PT},(1+)}. \tag{31}$$

V. Results

The presented algorithm has been implemented in CUDA and tested on NVIDIA GeForce 690 GPUs. We use the discussed MIS scheme in the forward projector of the reconstruction algorithm. The back projector is the voxel based method of Subsection III-B for maximum efficiency, which computes geometric effects and attenuation but does not involve scatter simulation. The reason of using a simplified back projector is that it increases the initial convergence speed and reduces the time needed for a single iteration cycle.

A. Performance in geometric projection

The accuracy of geometric projection is crucial in high resolution small animal PET where the voxel edge length can be significantly smaller than the edge length of the detector crystals. We modeled Mediso's *nanoScan PET/CT* [Med10b], which consists of twelve detector modules of 81×39 crystals of surface size 1.12×1.12 mm^2, thus the total number of LORs is 180 million when crystals of a module are connected by LORs to crystals of three opposite modules.

For validation in geometric projection, we took three mathematical phantoms, an off-axis *Point source*, the *Derenzo* with rod diameters $1.0, 1.1, \ldots, 1.5$ mm in different segments, and the *Homogeneity* that is built of 8 constant activity cubes (Fig. 3). We used GATE [Jea04] to generate a "ground truth" reference projection $\tilde{y}_L^{\mathrm{ref}}$ with 10^{12} samples and calculated the LOR space L_2 error of the proposed projectors obtained with different number of samples. To compare techniques working with different sample types, the LOR space L_2 error is depicted in Fig. 3 with respect to the computation time of a single projection. We consider the discussed LOR driven and voxel driven methods and three MIS versions, including balance heuristics (MIS-Balance), power heuristics with $\alpha = 2$ (MIS-Power), and maximum heuristics (MIS-Max).

We observe that the error decreases with increasing computation times, i.e. number of samples in all cases, which is compatible with the expected unbiased character of MC estimators. We note that voxel driven sampling is particularly efficient for the Point, while LOR driven sampling is good for the Homogeneity. The combined method is significantly better for the Derenzo and is similar to the best of the LOR driven and voxel driven approaches when the Homogeneity and the Point are projected. When the performances of LOR driven and voxel driven sampling are similar, then balance heuristics is optimal, but when data strongly favors either voxel driven or LOR driven sampling, maximum or power heuristics has minor advantages.

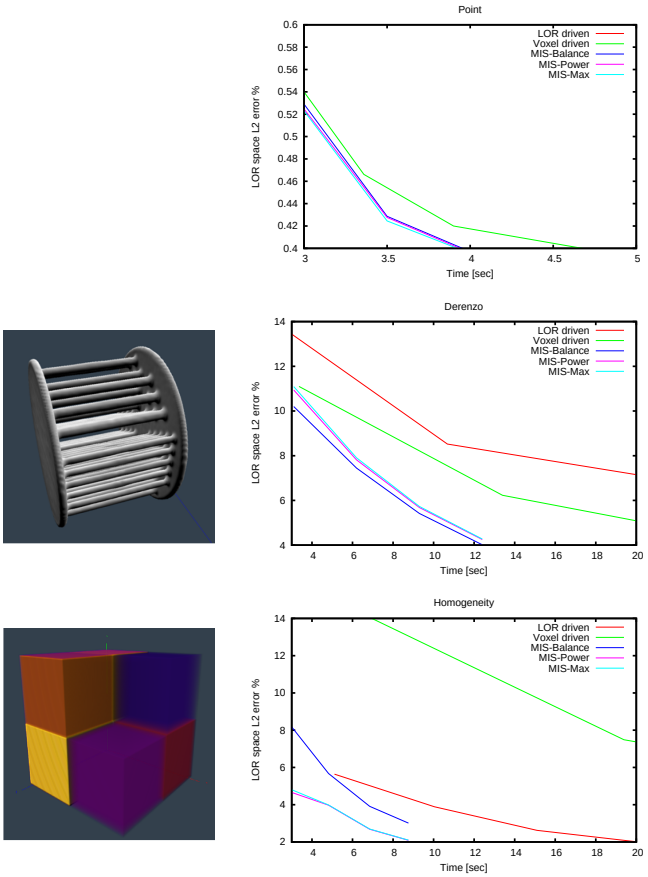

Fig. 3. LOR space L_2 error of different projectors with respect to the computation time of the projection for the Point (top), Derenzo (middle), and the Homogeneity (bottom) phantoms. Note that the top figure does not include the curve of the LOR driven sampling because its error is an order of magnitude higher than those of the voxel driven and the combined methods.

In the second evaluation phase, the projectors are included in ML-EM reconstruction. To obtain measured value y_L, we assigned 0.1 MBq activity to the *Point source*, 5 MBq to the *Derenzo*, 1.2 MBq to the *Homogeneity*, and simulated a 1000 sec long measurement for each with GATE, which mimics physical phenomena and thus obtains the measured data with realistic Poisson noise.

The *Signal to Noise Ratios*

$$\mathrm{SNR} = \frac{\sum_{L=1}^{N_{\mathrm{LOR}}} \tilde{y}_L^{\mathrm{ref}}}{\sum_{L=1}^{N_{\mathrm{LOR}}} |y_L - \tilde{y}_L^{\mathrm{ref}}|}$$

of the Point, Derenzo and Homogeneity measurements are 22.99, 10.22 and 4.04, respectively.

The phantoms are reconstructed on a grid of $144^2 \times 128$ voxels of edge length 0.23 mm. Fig. 4 shows the voxel space *Cross Correlation* (*CC*) error curves and the results of the reconstruction for the three phantoms using different N_{ray}, N_{march} and N_v parameters. When N_{ray} is zero, the method is voxel driven. When N_v is zero, we run a LOR driven algorithm. The combined approach is characterized by non zero N_{ray}, N_{march} and N_v parameters.

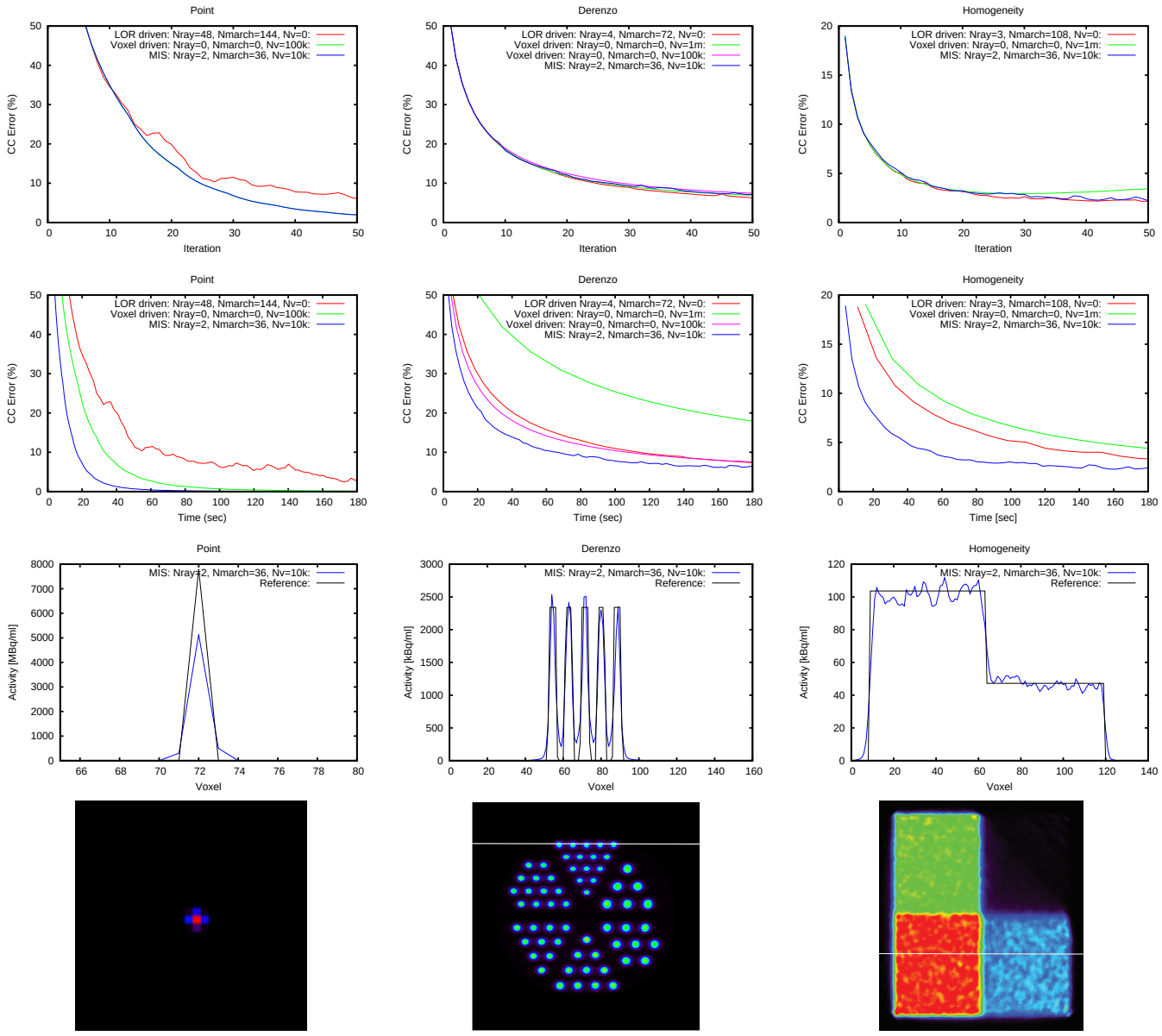

Fig. 4. Voxel space CC error curves with respect to the iteration number (first row) and to the total forward projection time (second row), profiles (third row) and slice images (fourth row) of the reconstructed Point (left), Derenzo (middle) and Homogeneity phantoms (right). The error and profile curves were made with different N_{ray}, N_{march} and N_v samples. The method is LOR driven when the number of voxel samples N_v is zero. The method is voxel driven when the number of LOR samples N_{ray} is zero. Finally, in MIS-combined reconstructions both the number of voxel samples and the number of LOR samples are non zero. The MIS error curves using different heuristics are very similar, so we depicted the error curves of the power heuristics only. We executed full EM iterations in all cases. The last row shows the reconstruction results of the MIS-combined method after 50 iterations.

If the number of samples is sufficiently high, then the error curves of different projectors run together when they are drawn with respect to the iteration number (first row of Fig. 4).

However, different methods are associated with significantly different computation times. To show this, in the second row of Fig. 4 we also include the errors as functions of the time in seconds devoted to execute forward projections. As expected, the Point phantom can be efficiently reconstructed with the voxel driven method, while the LOR centric approach is good for the Homogeneity phantom. Both the voxel driven and the LOR driven methods are outperformed by MIS for all three phantoms, which allows the reduction of the number of line samples N_{ray} and ray marching steps N_{march}, and adds relatively few N_v volume points to compensate the missing samples at important regions. Note that for about 10^6 voxels, only 10^4 added volume samples are sufficient. The random selection and projection of 10^4 volume samples onto 180 million LORs need just 0.3 seconds on the GPU, which is negligible with respect to the times of other processing steps.

B. Performance in scatter compensation

Scattering in the measured object is significant in human PET, so for the purpose of examining the proposed approach in scatter compensation, we examined the projection and reconstruction of the NEMA NU2-2007 *Human IQ phantom* in Mediso *AnyScan human PET/CT* [Med10a]. AnyScan has 24 detector modules consisting of 37×38 crystals detectors of $3.9 \times 3.9 \times 20$ mm^3. The voxel grid has $166^2 \times 75$ resolution and the voxel edge length is 2 mm. We set the energy discrimination window to [100, 750] keV. With such a wide window 35% of the measured events are direct hits, 27% are single scatters and 38% are multiple scatters.

As AnyScan detector modules cover just a small solid angle of the directional sphere, only about 2% of the photons have chance to hit the detectors. To attack this problem, in the Particle Tracer we sample annihilation photon directions non-uniformly, while the density is weighted accordingly.

For comparison, we also implemented a Watson type [Oll96], [Wat00] *Single Scatter Simulation* (SSS) algorithm on the GPU with the following modifications. Instead of computing polylines connecting the two detector crystals and the scattering point separately for each LOR, the process is decomposed to three phases. In the first phase, N_{scatter} scattering points are sampled globally from a probability density that mimics the scattering cross section, which will be used in the quadratures of all LORs. In the second phase, each detector crystal is connected to each of the scattering points, and along these line segments the line integrals of the activity, absorption and out-scattering are computed. Absorption and scattering cross sections depend on the energy of gamma photons, which is not available yet, thus these line integrals are temporarily computed assuming 511 keV photons. In the third phase, the line segments sharing a scattering point are paired, resulting in N_{scatter} polylines in each LOR. When a polyline is formed, the scattering angle and the Compton formula are evaluated, and the line integrals are corrected according to the ratios of the real photon energy and 511 keV. We note that it would be possible to extend Watson's method to approximately account for multiple scatters by scaling its results to fit the tails of the sinogram [Wat00] or by modifying the cross sections while the Watson method is executed [MST12], but we compare our method to the basic approach simulating only single scatter.

First, the projectors are validated computing the LOR space L_2 error with respect to a reference projection generated by GATE with 10^{13} annihilation photons. We considered different samples that are multiples of $N_{\text{ray}} = 1, N_{\text{march}} = 21$, $N_{\text{scatter}} = 50$, and $N_{\text{PT}} = 10^6$, and the error curves are depicted with respect to the computation time (Fig. 5). The LOR driven and the Watson type methods compute only the unscattered term and at most single scattering, respectively, thus they are fast converging at the beginning but for higher number of samples the projection error stops decreasing, i.e. these methods are biased. MIS combines the unbiasedness of the Particle Tracer and the speed of LOR driven methods.

The performance of the projectors is also evaluated in ML-EM reconstruction. We get GATE to simulate a 500 sec long measurement of the Human IQ phantom of 40 MBq activity, which resulted in LOR data of 3 SNR. Fig. 6 shows the error curves with respect to the number of iterations and the total forward projection time, and also a transaxial slice obtained with the combined method. The profile curves on the centerline are depicted by Fig. 7. Note that using a LOR driven method ($N_{\text{ray}} = 4$, $N_{\text{march}} = 84$, $N_{\text{PT}} = 0$) alone, we cannot expect fully accurate reconstruction since this method computes only the direct contribution and ignores scattered photon hits. As a consequence, false activity is added that can be observed in the profile curve, which is just partially compensated by the Watson algorithm as it ignores multiple scattering. The Particle Tracer ($N_{\text{ray}} = 0$, $N_{\text{march}} = 0$, $N_{\text{PT}} = 40 \cdot 10^6$) and the MIS-combined ($N_{\text{ray}} = 2$, $N_{\text{march}} = 42$, $N_{\text{PT}} = 4 \cdot 10^6$) methods involve unbiased multiple scattering estimators, thus they can theoretically lead to accurate reconstructions. The Particle Tracer requires at least 40 million photon pairs of non-uniform initial directional distribution per iteration to prevent the process from diverging, but the reconstruction result remains noisy. The combined method is not only more accurate but also much faster since it needs just 4 million photon pairs per iteration to add scattering and to help the LOR centric approach in the computation of the direct contribution. We note that the performance of single and multiple scattering compensation could be improved by turning scatter calculation on just in later phases of the iteration, and also by re-computing the scattered contribution less frequently and not in every iteration cycle.

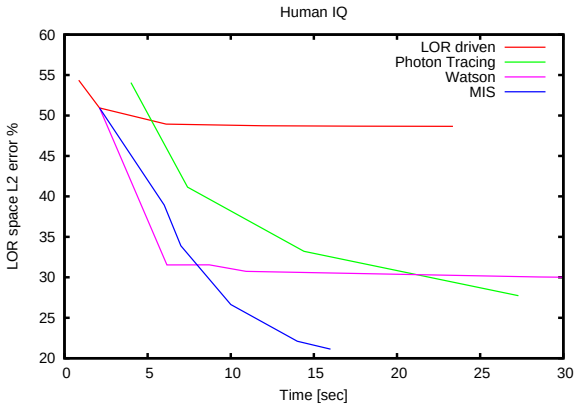

Fig. 5. LOR space L_2 error of different projectors with respect to the computation time of the projection for the Human IQ phantom.

VI. Conclusions

This paper proposed the MIS-based combination of different MC projection methods, including LOR centric and voxel centric approaches. The individual methods have different advantages and drawbacks concerning numerical accuracy and GPU execution performance. The proposed combination automatically finds an optimal weighting, which keeps the advantages of all techniques. The combined sampling can

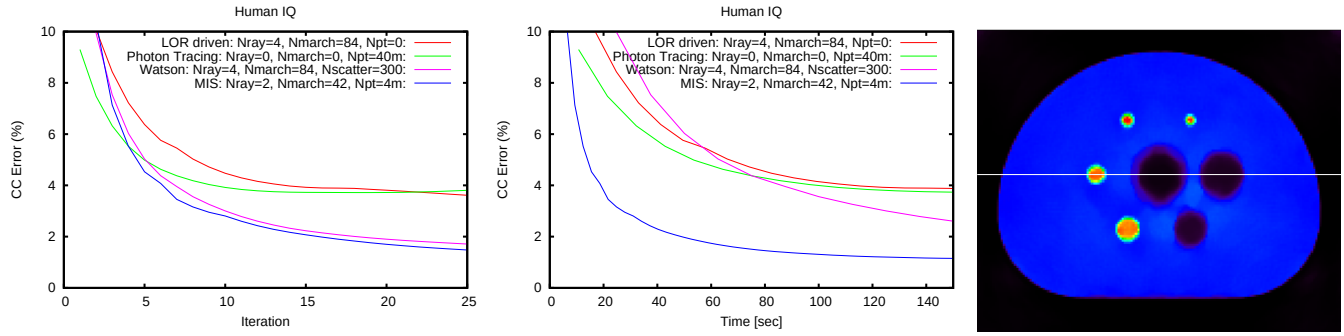

Fig. 6. CC error curves reconstructing the Human IQ phantom with different N_{ray}, N_{march}, N_{scatter} and N_{PT} sample numbers, depicted as functions of the iteration number (left) and the time devoted to forward projections (middle).

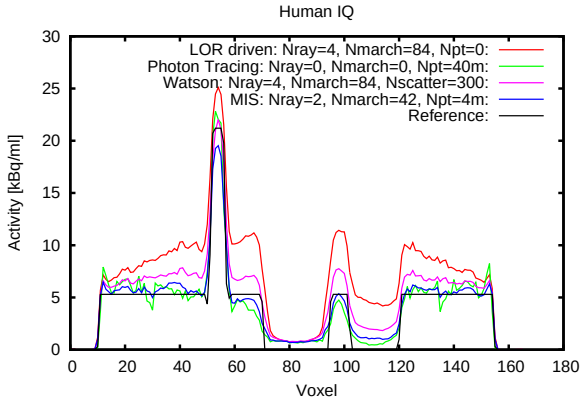

Fig. 7. Profile curves of the Human IQ phantom along the centerline crossing a hot sphere, the lung, and a cold sphere.

result in accurate projections using fewer samples and thus can reduce the time of reconstruction.

We have applied the concept for the computation of geometric projection with attenuation and also for multiple scattering compensation. These methods are built into the forward projector of the Tera-tomo$^{\mathrm{TM}}$ system [Mea10]. MIS can also be applied in other MC estimators developed for the same or other physical phenomena. For example, we can consider the combination of more efficient geometric projectors, like the distance driven method. Furthermore, scattering in the detector crystals can also be simulated with input crystal driven or output crystal driven approaches, whose advantages can be combined with MIS.

Acknowledgement

This work is supported by OTKA K-104476 (Hungary). The authors are grateful to Gergely Patay (Mediso) for the GATE simulations and NVIDIA for donating the GPU cards.

References

[GMDH08] N. Gac, S. Mancini, M. Desvignes, and D. Houzet. High speed 3D tomography on CPU, GPU, and FPGA. *EURASIP Journal on Embedded Systems*, 2008. ID 930250.

[Jea04] S. Jan and et al. GATE: A simulation toolkit for PET and SPECT. *Phys. Med. Biol.*, 49(19):4543–4561, 2004.

[Jos82] P. M. Joseph. An improved algorithm for reprojecting rays through pixel images. *IEEE Trans. Med. Imaging*, 1(3):192–196, 1982.

[MB04] B. de Man and S. Basu. Distance-driven projection and backprojection in three dimensions. *Phys. Med. Biol.*, 49:2463–2475, 2004.

[Mea10] M. Magdics et al. Tera-Tomo project: a fully 3D GPU based reconstruction code for exploiting the imaging capability of the NanoPET/CT system. In *World Molecular Imaging Congress*, 2010.

[MST12] M. Magdics, L. Szirmay-Kalos, B. Tóth, and T. Bükki. Higher order scattering estimation for PET. In *IEEE NSS/MIC Conf. Rec.*, pages 2288–2294, 2012.

[Med10a] Mediso Anyscan-PET/CT, www.mediso.com/products.php?fid=1,9&pid=73.

[Med10b] Mediso Nanoscan-PET/CT, www.mediso.com/products.php?fid=2,11&pid=86.

[Oll96] J. M. Ollinger. Model-based scatter correction for fully 3D PET. *Phys. Med. Biol.*, 41:153–176, 1996.

[QLC+98] J. Qi, R. M. Leahy, S. R. Cherry, A. Chatziioannou, and T. H. Farquhar. High-resolution 3D Bayesian image reconstruction using the microPET small-animal scanner. *Phys. Med. Biol.*, 43(4):1001, 1998.

[RZ07] A. J. Reader and H. Zaidi. Advances in PET image reconstruction. *PET Clinics*, 2(2):173–190, 2007.

[SH05] C. Sigg and M. Hadwiger. Fast third-order texture filtering. In *GPU Gems 2*, pages 313–329. Addison-Wesley, 2005.

[Sid85] R. L. Siddon. Fast calculation of the exact radiological path for a three-dimensional CT array. *Medical Physics*, 12(2):252–257, 1985.

[SV82] L. Shepp and Y. Vardi. Maximum likelihood reconstruction for emission tomography. *IEEE Trans. Med. Imaging*, 1:113–122, 1982.

[SKMT13] L. Szirmay-Kalos, M. Magdics, B. Tóth, and T. Bükki. Averaging and Metropolis iterations for positron emission tomography. *IEEE Trans. Med. Imaging*, 32(3):589–600, 2013.

[TB02] D. W. Townsend and T. Beyer. A combined PET/CT scanner: the path to true image fusion. *British Journal of Radiology*, 75:24–30, 2002.

[VG95] E. Veach and L. Guibas. Optimally combining sampling techniques for Monte Carlo rendering. In *ACM SIGGRAPH '95 Proceedings*, pages 419–428, 1995.

[Wat00] C.C. Watson. New, faster, image-based scatter correction for 3D pet. *IEEE Trans. Nucl. Science, ,* 47(4):1587–1594, 2000.

[XM07] F. Xu and K. Mueller. GPU-acceleration of attenuation and scattering compensation in emission computed tomography. In *9th Fully Three-Dimensional Image Reconstruction in Radiology and Nuclear Medicine*, 2007.