GPU-based Particle Transport
for PET Reconstruction
Thesis points of PhD dissertation

Milán Magdics

Supervisor:
LÁSZLÓ SZIRMAY-KALOS, Doctor of HAS

Budapest University of Technology and Economics
Faculty of Electrical Engineering and Informatics
Department of Control Engineering and Information Technology

Budapest

10th january 2014.
Synopsis

Research area and research objectives

Tomography reconstruction methods allow the examination of the inner structure of objects or living beings and thus have become important tools of modern medical diagnostics and research. With the rapid evolution of hardware engineering, modern scanners can provide high accuracy and spatial resolution, while also generating a huge amount of data. Processing (or the so-called reconstruction) of this big measurement data requires an enormous computational power. In order to fully utilize the capabilities of tomography scanners and provide as high image quality as possible — e.g. high resolution, high contrast etc. — in reasonable time the performance of a super-computer is required, even for a single measurement.

In the past few years, the graphics hardware or Graphics Processing Unit (GPU), which offers the computational throughput of a super-computer at low cost, has proven to be the most effective option for such computation intensive problems. The graphics hardware has, however, a massively parallel architecture very different from the traditional Neumann type processors, which has to be taken into account not only for algorithm design, but even on the level of mathematical modeling. For example, we need to aim for coherent execution and memory operations of the parallel threads, or to minimize their communication.

The objective of my research was to develop GPU-conform reconstruction methods that accurately model the physical phenomena that occur during the examination, for a specific type of tomography, Positron Emission Tomography (PET).

The goal of PET reconstruction is to find out the spatial density of the radioactive tracer that was injected into the subject before the examination. The tracer typically consists of materials essential for metabolism (e.g. oxygen or glucose) and thus, it is transported by the blood flow to regions with high cell activity enabling in vivo examination of organic functions. As radioisotopes of PET undergo positron emission decay, the tracer density is given by the number of emitted positrons. Positrons travel through the tissue following a chaotic path, terminated by positron-electron annihilation. As a result of annihilation two nearly oppositely directed gamma-photons are born, each with 511 keV energy. The PET scanner detects nearly coincident gamma-photon hits in detector crystals placed in a uniform grid, in a cylindric shape, around the subject. Gamma-photons, on the other hand, may interact with electrons of the surrounding medium both in the measured object and inside the detector crystals, which results in either the loss (due to photoelectric absorption) or direction change (due to scattering) of the photons.

Recent reconstruction methods are iterative. Among these, the most commonly used algorithm is an iterative maximum likelihood estimation, the ML-EM (Maximum Likelihood, Expectation Maximization), consisting of subsequent forward projection and back projection steps. Starting from (an initially arbitrary positive) estimation of the unknown tracer density, the forward projection computes the expected number of coincident photon-pair hits inside
the detectors by simulating the different physical phenomena. The back projection applies a correction on the tracer density estimation based on the ratio of the estimated and measured hits; this corrected estimation then becomes the input of the iteration step.

The relation between the finite element representation of the activity density and the detector pairs, i.e. the probability that a positron born in a particular volume element (a voxel) causes a coincident photon-pair hit in a particular pair of detectors, is modeled by a system matrix. Since these probabilities depend on the physical phenomena, the computation of the elements of the system matrix and thus the accurate reconstruction requires the modeling of the entire physical process starting from the positron decay and translation, through the interaction of the gamma-photons with the measured object and the detector crystals up to the photon detection. The difficulty of this problem arises from the facts that the system matrix is huge — usually consists of $10^7 \times 10^8$ elements —, depends on the measured object — e.g. the particle-medium interaction depends on the material — and the matrix elements are — due to the characteristics of multiple scattering — high dimensional integrals.

Methods

The reconstruction process therefore requires the estimation of a large number of high dimensional integrals, which vary from measurement to measurement. I used Monte Carlo methods to estimate high dimensional integrals of PET. Additionally, I built on the factoring of the system matrix, which decomposes the matrix into a product, where the terms correspond to the physical phenomena of a PET measurement. The main advantage of factorization is that every physical phenomenon can be modeled according to its impact on image quality and with different strategy, thus we can apply sampling strategies and numerical methods that best fit the target hardware, reuse samples and avoid computations that have negligible effect on the final result. Consequently, we can achieve a more accurate result in a given time budget compared to the non-factorized approach. On the other hand, factorization often uses approximations — an element of the system matrix which is described as an integral of a product is approximated by a product of integrals which may not be equivalent —, thus even with dense sampling of the system matrix the iteration does not converge to the true solution.

During my work, I studied the convergence and the provided image quality of the proposed methods both on simulated and real data. Simulations were generated by a widely used and validated tool (GATE), which allows the inclusion or exclusion of each physical phenomena into or from the simulation, allowing to verify and validate the proposed models of the physical phenomena independently. Using simulated data the tracer density is assumed to be known, which allowed me to study the distance between the reconstructed and the true values using standard distance metrics in addition to the regular image quality metrics, such as contrast or homogeneity. Real data was taken from two scanners where different physical phenomena have major influence on image quality, due to the differing geometries of the scanners. For small animal PET, positron range and inter-crystal scattering are the dominant image degrading effects making it an ideal platform for testing these phenomena. Contrarily, photon absorption and scattering in the measured object play a major role in human PET and the effect of positron range and the detector model are less significant.

In order to fully utilize the capabilities of the target hardware, its characteristics had to be considered. Efficiency of GPU implementations is mostly determined by the coherence of thread execution and memory operations and the thread independence, which is best achieved by so-called gathering type algorithms. When computing a number of high dimensional integrals using Monte Carlo quadrature, an efficient gather type algorithm assigns a set of
samples to threads that lead to coherent memory reads, and write operations are organized coherently and independently of the input data. The design of gather type algorithms often requires to transform the problem — a good example is the forward projector of thesis 2.1, which transforms the volumetric integral model to surface integrals of line integrals and thus solves the adjoint problem, and the relation between the two models is given by the Jacobi determinant. Contrarily to gather type algorithms, in scatter (or shooting) type algorithms the address of write operations depend on the input data which may lead to random memory access and cause write conflicts that require atomic operations. Thus, during my research I tried to avoid scatter type methods.
New scientific results

Theses can be categorized into two main classes. Thesis groups 1–4 deal with the factorized models of the physical effects that have the highest impact on the image quality of PET reconstruction: positron range, direct component, photon–material interaction inside the measured object, and inter-crystal scattering in the detectors. Thesis group 5 discusses the application of mostly existing sampling techniques to PET reconstruction in order to increase the accuracy of the physical models with negligible additional computational cost.

Thesis group 1: Positron range simulation

Thesis 1.1: Positron range simulation for heterogeneous materials in the frequency domain

Between their decay and annihilation, positrons may travel up to a few millimeters which, if not taken into account, causes a significant blur to the reconstructed image. Positron range can be approximated as a material dependent blurring on the estimated positron emission density. In high-resolution small animal PET systems, the average free path length of positrons may be many times longer than the linear size of voxels, which means that the blurring kernel should have a very large support so its voxel space calculation would take prohibitively long.

I proposed a fast GPU-based solution to compensate positron range effects which executes filtering in the frequency domain, thus provides a performance that is independent of the size of the blurring kernel. To handle heterogeneous media, we execute Fast Fourier Transforms for each material type and for appropriately modulated tracer densities and merge these partial results into a density that describes the composed, heterogeneous medium. As Fast Fourier Transform requires the filter kernels on the same resolution as the tracer density is defined, I also presented efficient methods for re-sampling the probability densities of positron range for different resolutions and basis functions. [C11].

Thesis group 2: Geometric projections

When gamma-photon scattering and acollinearity are ignored, the photon pair travels a linear path. Coincident photon hits may result only from an annihilation event that happened on a line connecting the surface of the two detectors (Line of Response or LOR). Thus, the system matrix is sparse and can be modeled by geometric projection, without the computation of zero elements of the system matrix. Efficient parallel implementation requires the geometric projection to be LOR driven in the forward projector, that is, a single computational thread accumulates the contribution of every voxel that may affect a specific LOR, and voxel driven
Figure 1: Effect of positron range modeling on a ring phantom. The images of the upper and lower rows show the reconstructions without and with positron range compensation, respectively. The first three columns correspond to different isotopes, while the last column shows line profiles.

in the back projector, i.e. parallel threads apply correction on voxels corresponding to the intersecting LORs.

**Thesis 2.1: LOR driven estimator for the GPU**

Existing LOR driven forward projector methods assume piece-wise constant basis functions and use analytic approximations for the five-dimensional integral of the geometric projection. The deterministic error made by the analytic approximations results in a biased estimator and thus modifies the fixed point of the iteration. Additionally, existing methods use varying sample number to evaluate line integrals and thus would assign different computational load to parallel threads causing their divergence if these methods are ported to the GPU.

I proposed an unbiased sampling scheme that offers efficient parallel implementation using the same set of samples for each thread and derived the sample density formulae based on integral transformations. The surfaces of the detectors are re-sampled uniformly in every iteration step, and a random offset is added for the line samples along the line to guarantee that every point that may correspond to a LOR is sampled with a positive probability. [J8, C3, C5].

**Thesis 2.2: Voxel driven estimator with importance sampling**

LOR driven approaches may be wasting in the sense that they do not consider the annihilation density during sampling, thus are poor for importance sampling. Existing voxel driven methods assume a cylindrical detector ring and thus for scanners built of planar detector panels they sample intersecting LORs of a voxel unevenly.

I proposed a voxel driven geometric projection scheme that computes the contribution of
a voxel to LORs and derived the sample density formulae based on integral transformations. First sample points in the volume of interest are generated mimicking the annihilation density, then for each sample point detector surface points are sampled fairly uniformly. This allows the activity distribution to be taken into account in the forward projection, using importance sampling of the voxels. Furthermore, being a voxel centric approach, it provides an efficient parallel implementation of the back projector [J8, C5].

Thesis group 3: Scattering in the measured object

Scattering in the measured is more relevant in human PET, where about 30–50% of the measured coincident photons are scattered. Ignoring the scattered component causes a shift-variant underestimation of the coincident hits which results in a shift-variant overestimation of the reconstructed intensity. Additionally, scattered hits cause a blur on the reconstructed image, therefore, accurate scatter models can greatly increase image contrast. The scattered component can be modeled as a Neumann-series, where the terms correspond to the number of scattering events, i.e. the length of the polylines corresponding to scattered photon paths.

Thesis 3.1: Scatter simulation with photoelectric absorption

Watson’s method is a popular choice of single scatter simulation and its implementation becomes very efficient with the reuse of line segments. However, it simulates only single photon–material interaction and is not feasible for dense materials since it ignores photoelectric absorption and downsamples the set of detectors.

I proposed several GPU-based improvements for Watson’s algorithm. First, in order to make the method suitable for dense materials, I showed how to include photoelectric absorption into the model, without loosing the ability to precompute paths. Second, I proposed the application of importance sampling for the selection of scattering samples. Third, by giving an efficient GPU implementation that includes path reuse, I showed that the method can work in 3D without needing to downsample the detector space. [J1, J2, J4, B1, C4, C5, C7, D2].

Thesis 3.2: Multiple forward scattering for free

In practice, scattering simulation can only be evaluated up to a limited number of scattering events. Due to truncation of the Neumann series where terms represent higher order bounces, particle transport results are underestimated, thus the radiotracer density in the reconstruction becomes overestimated. This negative bias can be eliminated by Russian roulette which is inefficient on the GPU and it trades bias for noise. The contribution of the terms above truncation can also be approximately re-introduced by blurring and scaling the calculated contribution. However, these methods cannot accurately consider patient specific data and have the added computational cost of filtering.

I presented a simple approximate method to improve the accuracy of scatter computation in PET without increasing the computation time. The proposed method exploits the facts that higher order scattering is a low frequency phenomenon and the Compton effect is strongly forward scattering in the energy window of PET. I showed that the directly not evaluated terms of the Neumann series can approximately be incorporated by an appropriate modification of the scattering cross section while the highest considered term is calculated. The correction factor depends just on the geometry of the detector and is robust to the variation of patient specific data. [C9, D9].
Thesis group 4: Detector model

Thesis 4.1: Detector model with Monte Carlo LOR filtering

Gamma-photons may scatter inside the detectors too, they can even reach and become detected in neighboring crystals instead of the incident crystal. This phenomenon is mainly important in small animal PET systems where, due to the small sizes of the crystals, photons may scatter even to 10 crystals away. When ignored, this causes a significant blur on the reconstructed image.

When modeling inter-crystal scattering to increase the accuracy of PET, we can take advantage of the fact that the structure of the detector is fixed, and most of the corresponding scattering calculation can be ported to a pre-processing phase. Pre-computing the scattering probabilities inside the crystals, the final system response is the convolution of the geometric response obtained with the assumption that crystals are ideal absorbers and the crystal transport probability matrix. This convolution depends on the incident direction and is four-dimensional which poses complexity problems as the complexity of the naive convolution evaluation grows exponentially with the dimension of the domain.

I proposed a Monte Carlo method to attack the curse of dimension in higher dimensional spatial varying convolution. The method replaces the summation of the signal values weighted with the filter kernel by a random sum of signal values at points sampled with the density of the filter kernel. Decoupling the geometric phase from the detector model, i.e. pre-computing the direct contribution before the convolution is evaluated, I demonstrated that these techniques have just negligible overhead on the GPU. [C5, C6, C8, D5].

Thesis group 5: Sampling techniques

Thesis 5.1: Filtered sampling for PET

On-the-fly system matrix generation, i.e. approximation of high dimensional integrals is usually attacked by Monte Carlo quadrature and importance sampling. Determining the number of samples used by the estimators belongs to the classical tradeoff problem between accuracy and computational time. However, the approximation error mainly comes from the measurement noise and high frequency components of the measured object that cannot be captured by the given sample density. Filtered sampling applies low-pass filter on the integrand before sampling in order to suppress both noise and high frequency details.

I proposed the application of filtered sampling for the forward projection step of iterative ML-EM based PET reconstruction to decrease the variance of the integrand and thus to reduce the error of integral estimation for a given set of samples. The input of the forward projection is filtered using a low-pass filter, which reduces noise and increases the probability that samples do not miss high frequency peaks — e.g. a point source — and requires only negligible overhead on the GPU. I showed that the iteration converges to a modified fixed point, from which the original function can be extracted by applying the same filter. [C5, C10].

Thesis 5.2: Multiple importance sampling for PET

Voxel driven methods can focus on point like features while LOR driven approaches are good in reconstructing large, homogeneous regions. Existing methods use voxel and LOR driven approaches exclusively which means that they cannot achieve good performance for every types of input.
Figure 2: Mouse $^{18}$F bone PET study taken by NanoPET/CT reconstructed without accurate detector model (left) and with the proposed LOR filtering scheme (right). Data courtesy of P. Blower, G. Mullen, and P. Mardsden, Rayne Institute, King’s College, London.
I proposed 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. To make the combined estimator unbiased and of low variance, the densities of all individual methods are determined and the integrand values are compensated by their sum. [J8].

Thesis 5.3: Averaging and Metropolis iteration for PET

High dimensional integrals of PET are estimated by Monte Carlo quadrature. If the sample locations are the same in every iteration step of the ML-EM scheme, then the approximation error will lead to a modified reconstruction result. However, when random estimates are statistically independent in different iteration steps, then the iteration may either diverge or fluctuate around the solution. Our goal is thus to increase the accuracy and the stability of the iterative solution while keeping the number of random samples and therefore the reconstruction time low. One way to achieve this is to exploit additional samples from previous iteration steps.

I proposed two modifications of the Maximum Likelihood, Expectation Maximization (ML-EM) iteration scheme to reduce the reconstruction error due to the on-the-fly Monte Carlo approximations of forward and back projections with negligible additional cost: Averaging iteration and Metropolis iteration. Averaging iteration averages forward projection estimates during the iteration sequence. Metropolis iteration rejects those forward projection estimates that would compromise the reconstruction and also guarantees the unbiasedness of the tracer density estimate. I demonstrated that these techniques make the estimation consistent and significantly increase the stability of the iteration sequence. As a result, we can obtain accurate reconstructions with less samples, decreasing the reconstruction time. [J5, J6, D10].
Application of the new scientific results

This research was carried out in close collaboration with the Mediso company and was sponsored by the OTKA K-719922 and the TeraTomo: Tomography reconstruction for PET on the GPU projects. The proposed methods were continually tested during their development both on GATE simulated data and real measurements. The algorithms were built into Mediso’s nanoPET\textsuperscript{TM}/CT small animal PET system, their clinical validation is still in progress.

Collecting several methods, together with Mediso we submitted a US patent application, which is currently under the review process.
Own publications


