Physics Simulation and Inverse Problem Solution on Massively Parallel Systems

Final report

1. Objectives of the project

Particle transport as many real life phenomena are mathematically described by differential or integral equations, thus the prediction of the behavior of such systems requires the numerical solutions of these equations. In such systems the *inverse problem* means the determination of system parameters that would lead to a prescribed behavior. Inverse problems are typically attacked by being formulated as an optimization problem with a goal function of minimum difference between the expected and obtained behavior. To handle the measurement noise, maximum likelihood estimation is possible. The optimization problem can then be solved iteratively, repeating a simulation step, called forward projection, which computes the behavior from the actual guess of the parameters, followed by a back projection that corrects the parameters based on the distance of the behavior of the current system and the target behavior. Thus, the solution of an inverse problem needs the execution of forward projection many times, once in every simulation cycle.

Inverse problems are usually *ill-posed*, i.e. there is no exact, unique solution, and approximate solution methods are *ill-conditioned*, i.e. the error or noise in the measured or target data may be upscaled in the result, invalidating it completely. Such problems require *regularization*, i.e. the inclusion of an a-priori information based penalty term that can exclude unacceptable solutions, or constrained optimization that solves the problem of ill-conditioning without compromising the minimum error requirement. Another question relates to how sensitive the iteration process is to numerical errors made in the forward projection. To tackle complexity, we can use massively parallel computers like the graphics card (*GPU*). However, parallel hardware imposes special requirements on the design of solution algorithms.

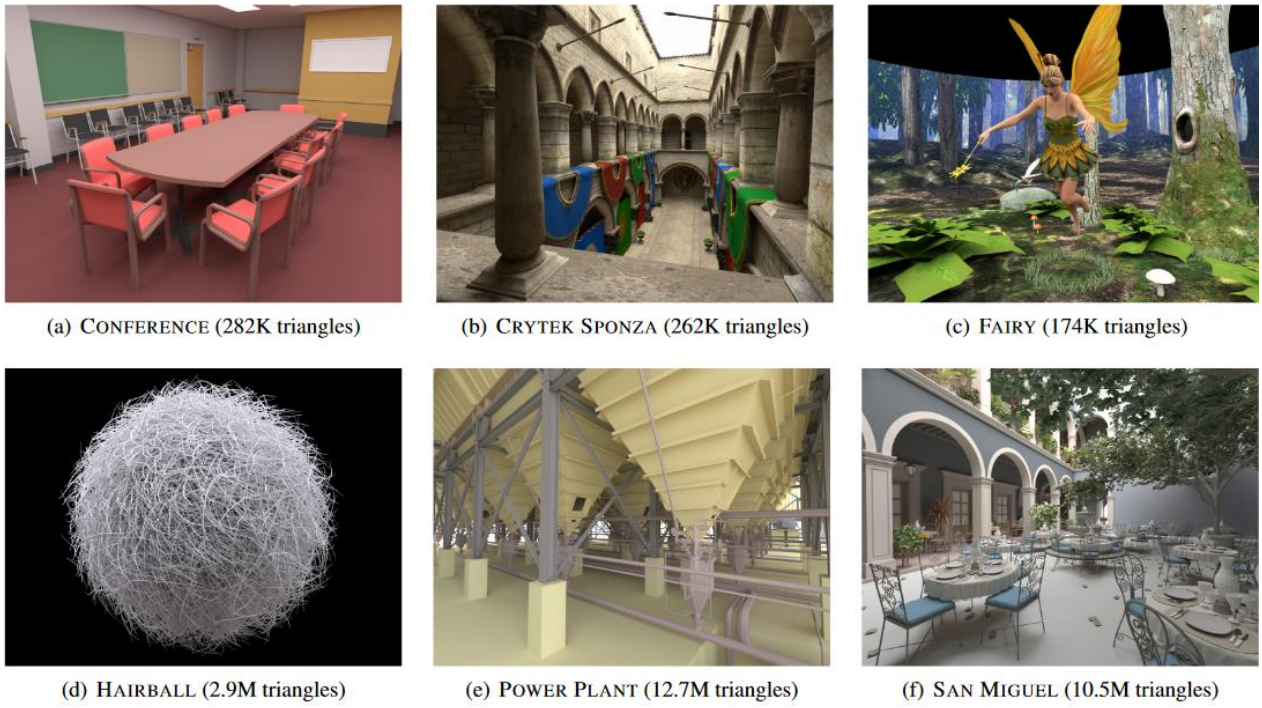
This project aimed at GPU-based parallel algorithms that attack complex system simulation and inverse problem solution primarily in the fields of particle or energy transport and tomography.

2. Research results

For simulation, we developed gathering type methods for partial differential equations and integral equations for efficient GPU implementation of particle/energy transport computation. We aimed at gathering type algorithms that can be implemented on parallel computers without causing write collisions. Gathering type approaches often correspond to adjoint transport operators, for which the governing equations must be established. In particular, we focused on the light transfer in surface models and participating media, and positron and gamma photon transport in volumetric models describing living organs or the body.

2.1. GPU based ray tracing without requiring stack[1]

Ray tracing is a fundamental recursive algorithm that finds the object that is first intersected by a ray. Recursion requires a stack, which is not available on GPUs. Stackless traversal algorithms for ray tracing acceleration structures require significantly less storage per ray than ordinary stack-based ones. This advantage is important for massively parallel rendering methods, where there are many rays in flight. On SIMD architectures, a commonly used acceleration structure is the multi bounding volume hierarchy (MBVH), which has multiple bounding boxes per node for improved parallelism. It scales to branching factors higher than two, for which, however, only stack-based traversal methods have been proposed so far. We have introduced a novel stackless traversal algorithm for MBVHs with up to 4-way branching. Our approach replaces the stack with a small bitmask, supports dynamic ordered traversal, and has a low computation overhead. We also presented efficient implementation techniques for recent CPU, MIC (Intel Xeon Phi), and GPU (NVIDIA Kepler) architectures.



(a) CONFERENCE (282K triangles) (b) CRYTEK SPONZA (262K triangles) (c) FAIRY (174K triangles)

(d) HAIRBALL (2.9M triangles) (e) POWER PLANT (12.7M triangles) (f) SAN MIGUEL (10.5M triangles)

Figure 1. Scenes rendered with our stack-less algorithm at interactive rates

2.2. Approximation of the infinite Neumann series[2]

The solution of the particle transport problem, which is the core part of tomography reconstruction, is mathematically equivalent to the evaluation of a Neumann series of increasing dimensional integrals, where the first term represents the direct contribution, the second the single scatter contribution, the third the double scattering etc. High dimensional integrals are computationally very expensive, thus this infinite series is truncated after a few (typically after the first or second) terms, which underestimates particle transport results. We have presented a simple approximate method to improve the accuracy of the scatter computation in positron emission tomography without increasing the computation time. We exploited the facts that higher order scattering is a low frequency phenomenon and the Compton effect is strongly forward scattering in 100–511 keV range. Analyzing the integrals of the particle transfer, we came to the conclusion that the directly not evaluated terms of the Neumann series can approximately be incorporated by the modification of the scattering cross section while the highest considered term is calculated.

[1] Attila T Áfra, László Szirmay-Kalos: ***Stackless Multi-BVH Traversal for CPU, MIC and GPU Ray Tracing***, COMPUTER GRAPHICS FORUM 33:(1) pp. 129-140., 2014

[2] Milán Magdics, László Szirmay-Kalos, Balázs Tóth, Balázs Csébfalvi, Tamás Bükki. Higher Order Scattering Estimation for PET. In: IEEE Nuclear Science Symposium and Medical Imaging Conference. Anaheim, USA, 2012.

2.3. Positron range simulation for Positron Emission Tomography[3]

We have developed a fast GPU-based solution to compensate positron range effects in heterogeneous media for iterative PET reconstruction. In a factorized approach, positron range is the first effect, which causes a spatially varying blurring according to local material properties. 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. This means that positron range significantly compromises the reconstruction quality if it is not compensated, and also that the material dependent blurring should have a very large support so its voxel space calculation would take prohibitively long. Frequency domain filtering does not have such computational complexity problems, but its direct form is ruled out by the fact that we need a spatially variant filtering in heterogeneous media. 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. Fast Fourier Transform requires the filter kernels on the same resolution as the tracer density is defined, so we also presented efficient methods for re-sampling the probability densities of positron range for different resolutions and basis functions.

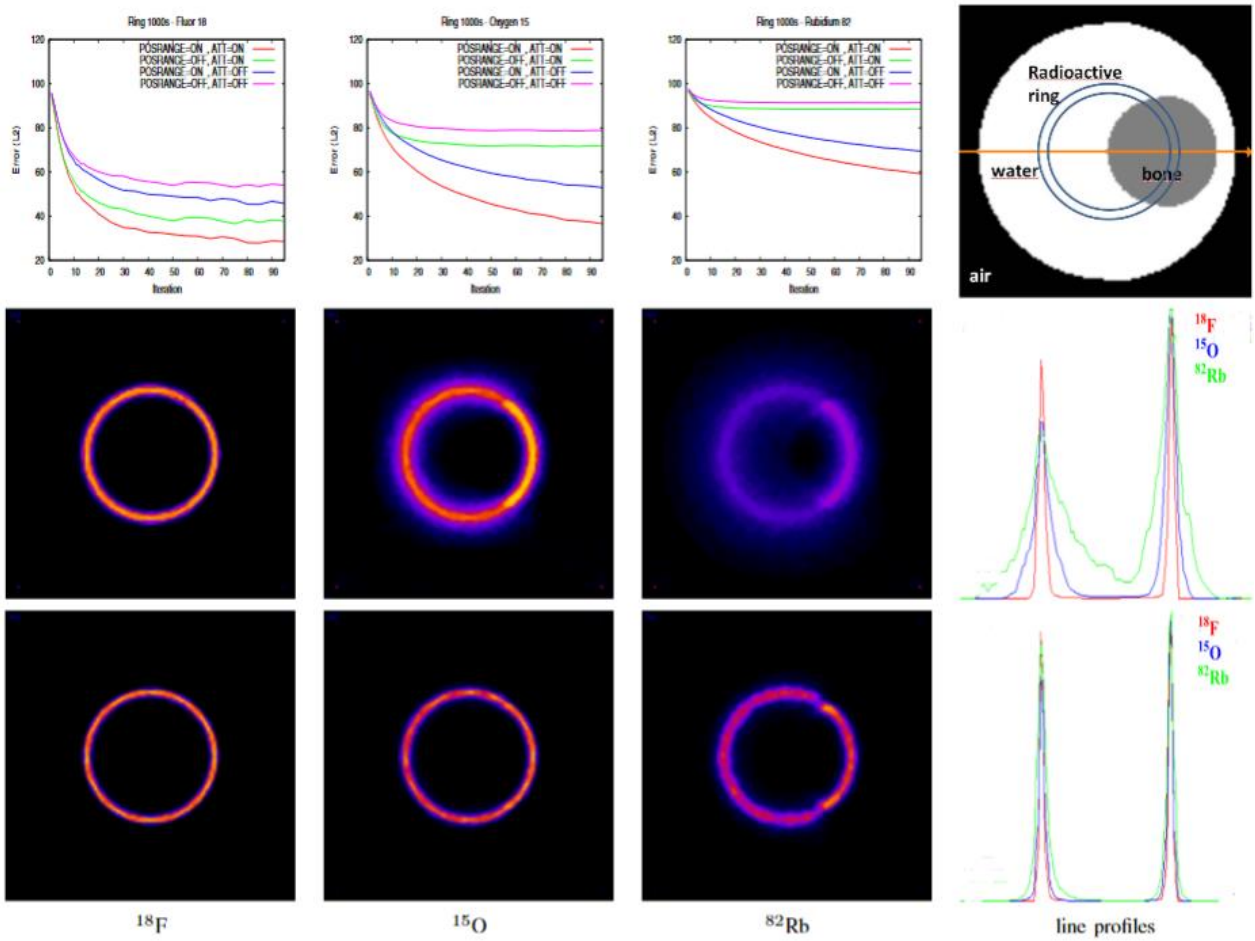


Figure 2. Comparison of the reconstruction without (middle row) and with (bottom row) our positron range compensation algorithm for a ring phantom and three different radiotracer isotopes. Note the significantly reduced blurring. The error curves shown by the upper row and the profile lines in the right column also validate the increased accuracy.

[3] Milán Magdics: GPU-BASED PARTICLE TRANSPORT FOR PET RECONSTRUCTION. PhD dissertation, Budapest University of Technology and Economics, 2014.

2.4. Combined direct/adjoint simulation in PET reconstruction[4]

We also considered direct physical simulation of the photon transport, which is less effective in parallel systems if it is used alone since it leads to scattering type algorithms. However, combining direct and adjoint methods using the concept of *multiple importance sampling*, we can obtain solution strategies that preserve the advantages of both approaches. Simulation algorithms are also examined from the point of view of factoring, which allows the re-use of results obtained on other parallel processors without excessive communication overhead. 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 good 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.

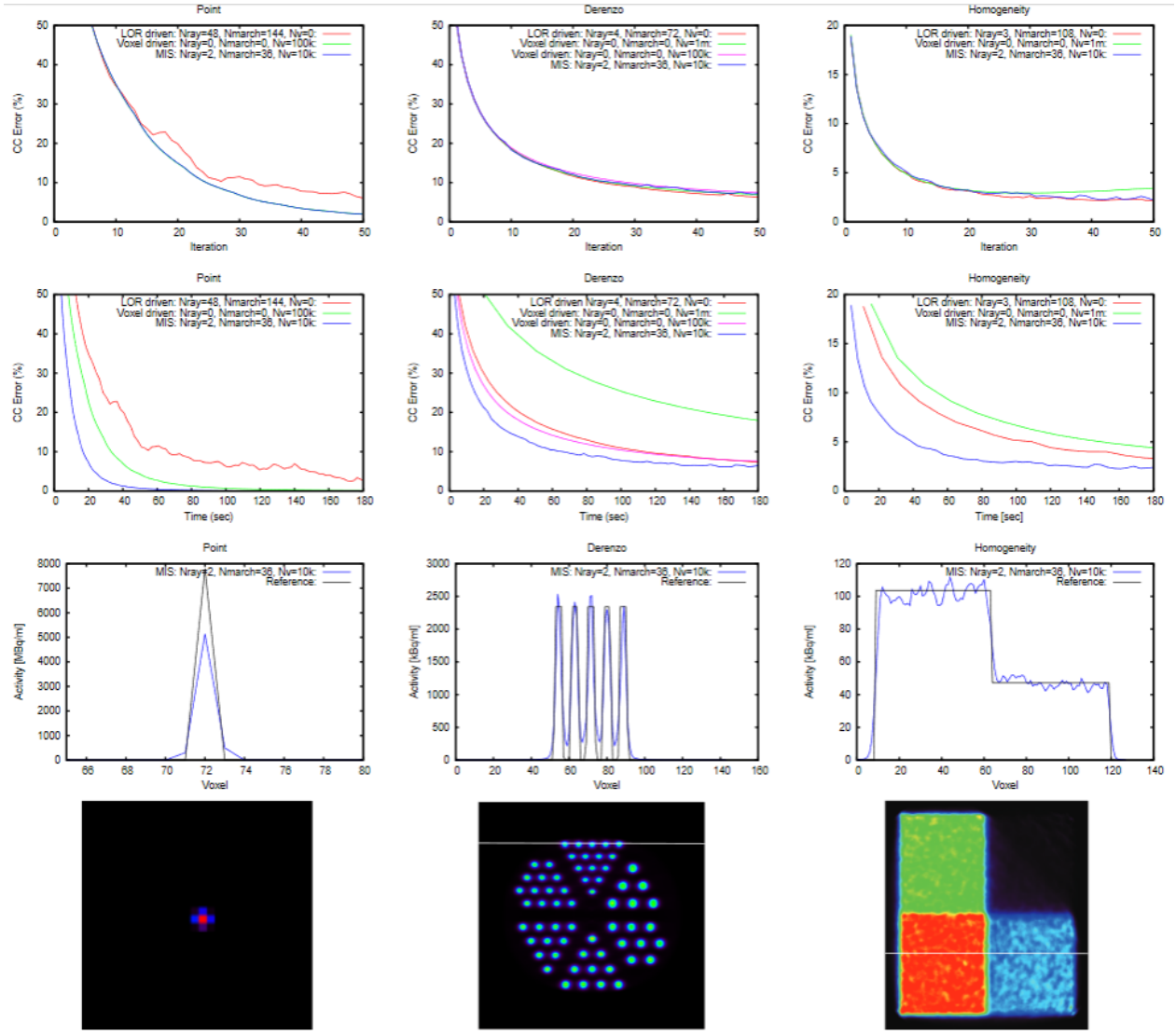


Figure 3. Reconstruction with combined direct/adjoint simulation showing the error curves as functions of the number of iterations (first row), calculation time (second row), profile curves (third row) and slices of the reconstructed 3D object obtained with the new method. Note that if the simulation is accurate enough, then the error curves are similar (upper row). However, there is a significant difference in the computation time (second row) since the proposed algorithm can deliver the same accuracy taking much less samples.

[4] László Szirmay-Kalos, Milán Magdics, Balázs Tóth. Multiple Importance Sampling for PET. IEEE TRANSACTIONS ON MEDICAL IMAGING. 33:(4) pp. 970-978. Paper TMI2300932. (2014)

2.5. Inverse problem solution with on-line parallel Monte Carlo forward projection[5]

Positron Emission Tomography (PET) reconstruction computes a series of projections between the voxel space and the LOR space, which are mathematically equivalent to the evaluation of multi-dimensional integrals. Monte Carlo (MC) quadrature is a straightforward method to approximate these multi-dimensional integrals, and remains the only feasible alternative when the scattering should be accurately compensated. As the numbers of voxels and LORs can be in the order of hundred millions and the projection also depends on the actually measured object, the quadratures cannot be pre-computed, but Monte Carlo simulation should take place on-the-fly during the iterative reconstruction process.

We have presented modifications of the Maximum Likelihood, Expectation Maximization (ML-EM) iteration scheme to reduce the reconstruction error due to the on-the-fly MC approximations of forward and back projections. If the MC sample locations are the same in every iteration step, then the approximation error will lead to a modified reconstruction result[6]. However, when random estimates are statistically independent in different iteration steps, then the iteration may either diverge or fluctuate around the solution. Our goal was to increase the accuracy and the stability of the iterative solution while keeping the number of random samples and therefore the reconstruction time low. We first analyzed the error behavior of ML-EM iteration with on-the-fly MC projections, then proposed two modifications of the classical ML-EM scheme: 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. We demonstrated that these techniques can significantly reduce the fluctuations in the reconstructed data, which would otherwise be present due to the noise of MC integration.

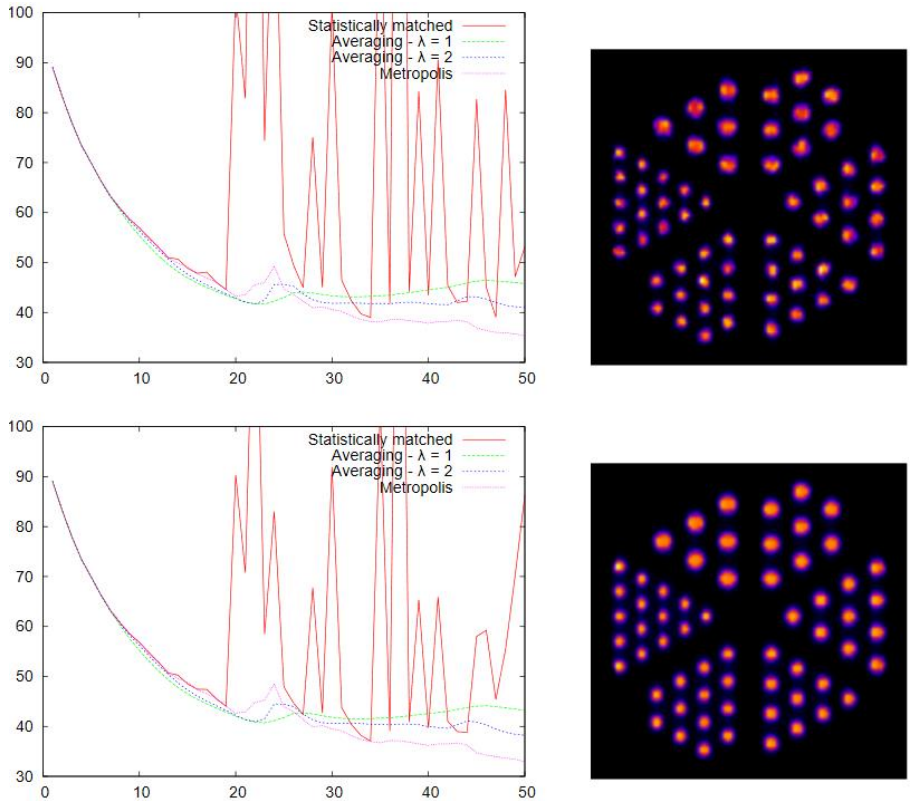


Figure 4. Comparison of the error curves of the classical ML-EM reconstruction using randomized projections (red curves) and our averaging and Metropolis algorithms. The classical scheme is unstable unless the number of samples and thus the computation time are very high. The new methods deliver numerically stable results with low sample numbers as well.

[5] László Szirmay-Kalos, Milán Magdics, Balázs Tóth, Tamás Bükki: Averaging and Metropolis Iterations for Positron Emission Tomography. IEEE TRANSACTIONS ON MEDICAL IMAGING (ISSN: 0278-0062) 32:(3) pp. 589-600, 2013.
[6] Milán Magdics, László Szirmay-Kalos, Balázs Tóth, Anton Penzov: ***Analysis and Control of the Accuracy and Convergence of the ML-EM Iteration***, LECTURE NOTES IN COMPUTER SCIENCE 8353: pp. 147-154., 2014

2.6. L1 regularization schemes and their efficient parallel implementation[7]

Positron Emission Tomography reconstruction is ill posed. The result obtained with the iterative ML-EM algorithm is often noisy, which can be controlled by regularization. Common regularization methods penalize high frequency features or the total variation, thus they compromise even valid solutions that have such properties. Bregman iteration offers a better choice enforcing regularization only where needed by the noisy data. Bregman iteration requires a nested optimization, which poses problems when the algorithm is implemented on the GPU where storage space is limited and data transfer is slow. Another problem is that the strength of the regularization is set by a single global parameter, which results in overregularization for voxels measured by fewer LORs. To handle these problems, we proposed a modified scheme that merges the two optimization steps into one, eliminating the overhead of Bregman iteration. The benefits over TV regularization are particularly high if the data has higher variation and point like features.

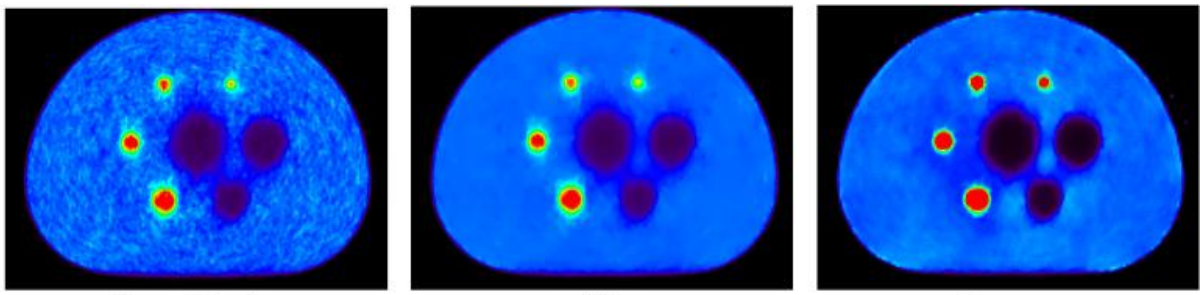


Figure 5. Reconstruction of the human chest phantom with no regularization, with Total Variation (TV) regularization, and with our method based on Bregman distance.

2.7. Volume processing with controlled anisotropic diffusion[8]

We have also developed a method to enhance volumetric data using anisotropic diffusion controlled by another voxel array representing the same object with different physical quantities. The main application of this approach is to enhance volumetric functional data (obtained e.g. with PET or SPECT) based on anatomic (e.g. CT or MRI) information. Enhancement includes noise removal, sharpening and resolution upsampling. As different modalities measure different physical quantities that may or may not be correlated, enhancement must be carefully designed not to introduce spurious features that are present only in one modality. Forward diffusion working with non-negative diffusivity guarantees this kind of causality but also limits the potential of enhancement. To allow the preservation or even the increase of the dynamic range, diffusion should also go backwards. Therefore, we proposed a forward–backward diffusion scheme for the enhancement where stability and the avoidance of spurious features are provided by the automatic determination of parameters controlling the diffusion process. Sophisticated filtering schemes can also be used in processing noisy depth images[9].

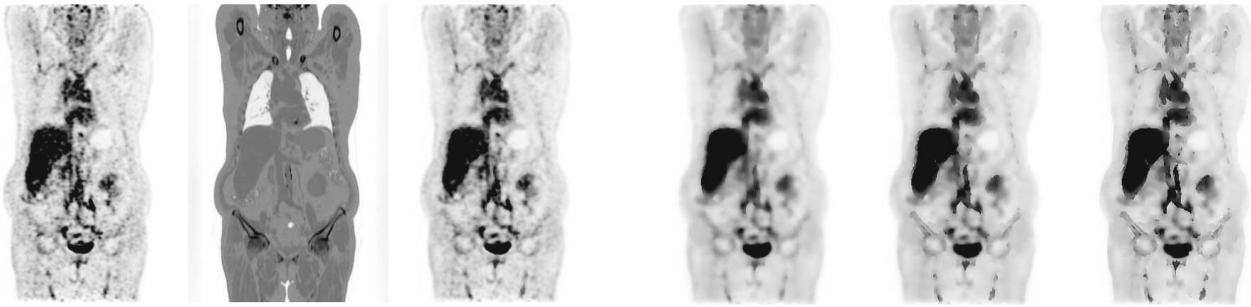


Figure 6. Corresponding slices of a PET reconstruction, CT reconstruction, filtering the PET data with a multi-lateral filter controlled by the CT data, and the results of our algorithm with three different settings.

[7] László Szirmay-Kalos, Balázs Tóth, Gábor Jakab: Efficient Bregman Iteration in Fully 3D PET. IEEE Nuclear Science Symposium and Medical Imaging Conference, Seattle, 2014.
[8] László Szirmay-Kalos, Milán Magdics, Balázs Tóth: Volume Enhancement with Externally Controlled Anisotropic Diffusion, The Visual Computer (DOI: 10.1007/s00371-015-1203-y), 2017.
[9] László Szirmay-Kalos: ***Filtering and Gradient Estimation for Distance Fields by Quadratic Regression***, PERIODICA POLYTECHNICA-ELECTRICAL ENGINEERING AND COMPUTER SCIENCE 59:(4) pp. 175-180. (2015), 2015

2.8. Multiple scattering simulation[10]

We have proposed a new stochastic particle model for efficient and unbiased Monte Carlo rendering of heterogeneous participating media. We randomly add and remove material particles to obtain a density with which free flight sampling and transmittance estimation are simple, while material particle properties are simultaneously modified to maintain the true expectation of the radiance. We show that meeting this requirement may need the introduction of light particles with negative energy and materials with negative extinction, and provide an intuitive interpretation for such phenomena. Unlike previous unbiased methods, the proposed approach does not require a-priori knowledge of the maximum medium density that is typically difficult to obtain for procedural models. However, the method can benefit from an approximate knowledge of the density, which can usually be acquired on-the-fly at little extra cost and can greatly reduce the variance of the proposed estimators. The introduced mechanism can be integrated in participating media renderers where transmittance estimation and free flight sampling are building blocks. We demonstrated its application in a multiple scattering particle tracer, in transmittance computation, and in the estimation of the inhomogeneous air-light integral.

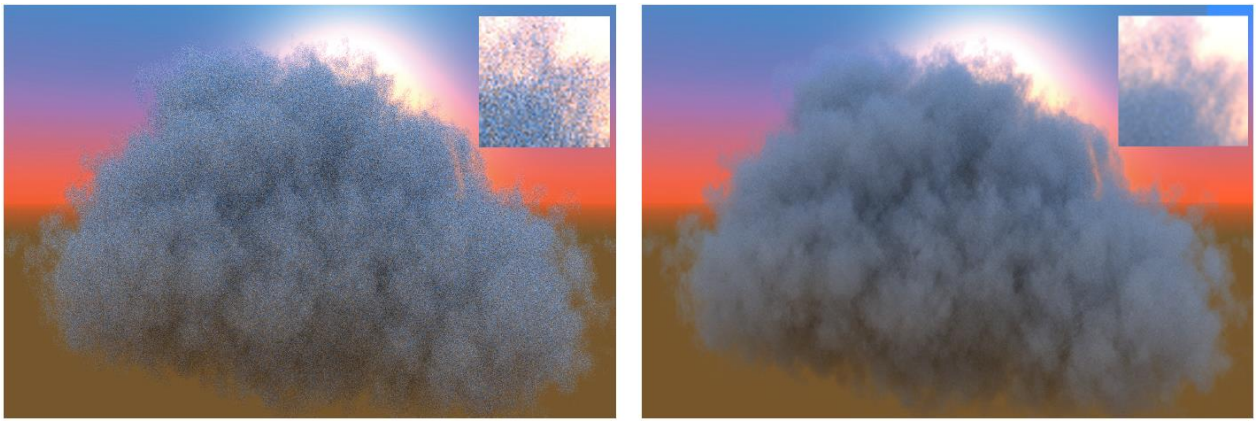


Figure 7. Comparison of classical Woodcock tracking (left) and our method (right) taking the same number of samples.

3. Utilization

The theoretical results have been utilized in practical applications. Our PET related results have been built into the Tera-Tomo software system developed for PET reconstruction, and European and American Patents have been accepted. Our achievement got Innovation Prize in 2013.

The multiple scattering simulation algorithm has become a part of the popular Arnold system. Tóth Balázs, Magdics Milán, Áfra Attila have submitted and successfully defended their Ph.D. dissertations. Magdics Milán received the Best PhD dissertation of NJSZT/KEPAF in 2014.

[10] László Szirmay-Kalos, Iliyan Georgiev, Milán Magdics, Balázs Molnár and Dávid Légrády: Unbiased Light Transport Estimators for Inhomogeneous Participating Media. COMPUTER GRAPHICS FORUM, Vol 36, No 2, 2017