Journal of Biomedical Optics

BiomedicalOptics.SPIEDigitalLibrary.org

Toward real-time diffuse optical tomography: accelerating light propagation modeling employing parallel computing on GPU and CPU

Matthaios Doulgerakis
Adam T. Eggebrecht
Stanislaw Wojtkiewicz
Joseph P. Culver
Hamid Dehghani

SPIE.

Matthaios Doulgerakis, Adam T. Eggebrecht, Stanislaw Wojtkiewicz, Joseph P. Culver, Hamid Dehghani, "Toward real-time diffuse optical tomography: accelerating light propagation modeling employing parallel computing on GPU and CPU," *J. Biomed. Opt.* **22**(12), 125001 (2017), doi: 10.1117/1.JBO.22.12.125001.

Toward real-time diffuse optical tomography: accelerating light propagation modeling employing parallel computing on GPU and CPU

Matthaios Doulgerakis,[a,*] Adam T. Eggebrecht,[b] Stanislaw Wojtkiewicz,[a] Joseph P. Culver,[b,c,d] and Hamid Dehghani[a]
[a]University of Birmingham, School of Computer Science, Birmingham, United Kingdom
[b]Washington University School of Medicine, Department of Radiology, St. Louis, Missouri, United States
[c]Washington University in St. Louis, Department of Biomedical Engineering, St. Louis, Missouri, United States
[d]Washington University School of Medicine, Division of Biology and Biomedical Sciences, St. Louis, Missouri, United States

Abstract. Parameter recovery in diffuse optical tomography is a computationally expensive algorithm, especially when used for large and complex volumes, as in the case of human brain functional imaging. The modeling of light propagation, also known as the forward problem, is the computational bottleneck of the recovery algorithm, whereby the lack of a real-time solution is impeding practical and clinical applications. The objective of this work is the acceleration of the forward model, within a diffusion approximation-based finite-element modeling framework, employing parallelization to expedite the calculation of light propagation in realistic adult head models. The proposed methodology is applicable for modeling both continuous wave and frequency-domain systems with the results demonstrating a 10-fold speed increase when GPU architectures are available, while maintaining high accuracy. It is shown that, for a very high-resolution finite-element model of the adult human head with ~600,000 nodes, consisting of heterogeneous layers, light propagation can be calculated at ~0.25 s/excitation source.
© *2017 Society of Photo-Optical Instrumentation Engineers (SPIE)* [DOI: 10.1117/1.JBO.22.12.125001]

Keywords: NIRFAST; parallel computing; GPU; diffuse optical tomography; finite-element method.

Paper 170398R received Jun. 28, 2017; accepted for publication Nov. 6, 2017; published online Dec. 1, 2017.

1 Introduction

Functional neuroimaging provides an essential tool in the study of the brain. It has been used to detect, localize, and classify brain activations during physical and psychological events, propelling applications in a myriad of areas, such as guiding treatment and monitoring the rehabilitation progress in cases of stroke, depression, or schizophrenia.[1–3] Diffuse optical tomography (DOT) is a soft tissue imaging technique based on injecting near-infrared (NIR) light in a volume and measuring the re-emerging light. In brain-related studies, the measured alterations in the attenuation of the resurfacing NIR light reflect changes in blood oxygenation and concentration induced by tissue metabolism within the brain due to local neuron activations. Therefore, DOT has been used for functional brain imaging[4,5] and neonatal brain monitoring,[6,7] as well as for measuring absolute oxygenation values in the brain.[8] NIR light is nonionizing and requires relatively low-cost equipment, which is wearable and, therefore, allows some movement of the subject. Additionally, DOT is relatively portable and can be used in clinical applications where use of fMRI or PET is not possible, for example as a bedside monitoring tool.

DOT specifically is concerned with tomographic reconstruction of volumetric and spatially distributed optical parameters from finite boundary measurements. This is commonly solved as an optimization problem using model-based approaches, whereby accurate modeling of light propagation within the volume, known as the forward model, is required. Therefore, to allow a real-time parameter recovery from measured data, both fast and accurate forward modeling is essential.

The objective of this work is the acceleration of the forward light propagation model, while maintaining numerical accuracy. Specifically, the focus is on the application of DOT for functional imaging on an adult human head, employing the finite-element method (FEM) to solve the diffusion approximation (DA) for modeling of light propagation as implemented within the NIRFAST[9] modeling and image reconstruction software package.

The proposed acceleration approach relies on employing parallel computing to expedite the solution of the forward problem, an option that has recently become popular due to the relatively low cost of GPUs and that has attracted the attention of researchers for solving similar problems in medical imaging.[10–12] In DOT, the acceleration of the forward model with GPUs has been employed for Monte-Carlo algorithms,[13–15] where simulating the behavior of each photon can be efficiently parallelized, with reported accelerations in the scale of 10^2 to 10^3. However, millions of photons must be simulated to achieve an accurate solution, so modeling the total fluence for a geometrically large volume and multiple excitation sources still requires time in the order of tens of minutes.

Accelerating the forward model of DOT using GPU parallelization has also been reported with FEM formulation. Specifically, the acceleration of the forward model solution has been proposed for frequency-domain (FD) simulations.[16] Unfortunately, due to a lack of sparse arithmetic architecture, computationally tractable mesh sizes have been limited to

*Address all correspondence to: Matthaios Doulgerakis, E-mail: mxd552@cs.bham.ac.uk

1083-3668/2017/$25.00 © 2017 SPIE

~9000 nodes due to the excessive memory requirements. Solving the forward model employing GPU parallelization, in continuous wave (CW), with parallelization over the nodes and over the excitation sources simultaneously, using a block-based formulation of the forward linear system has also been proposed.[17] However, due to the size of the augmented block matrix, this approach is only applicable in small size meshes, with up to ~2500 nodes. Using multiple GPUs to solve the forward problem in CW, for infant brain studies has been suggested[18] but has been only evaluated qualitatively in a homogenous phantom head model, again in meshes with ~9000 nodes. An approach combining CPU and GPU parallelization was proposed[19] where the imaging domain is decomposed into overlapping subdomains, therefore, allowing a high level of parallelization for the forward problem. However, this approach was only evaluated in CW, on a simplified cylindrical geometry with homogenous optical properties where the decomposition in regular overlapping subdomains is a straightforward procedure, while in the case of a complex volume, as the adult head, such decompositions are not a trivial task. Finally, a framework for the solution of the forward problem in CW and FD, accelerated in the GPU, was proposed and evaluated in homogenous cylindrical models with up to 330,000 nodes.[20] It was shown that the relation between error of the iterative solution and optical properties of the volume was identified while the single precision numerical accuracy was found to be insufficient for solving for a wide range of optical properties. However, none of the previous work has evaluated the accuracy and computational speed of iterative solvers on anatomically realistic head models, with high-resolution meshes and the high-density (HD) DOT system.

This work provides tractable solutions to overcome the current computational time and memory limitations arising when dealing with high-resolution FEM meshes and HD source–detector (SD) pairs, both important features for high-quality functional DOT (fDOT) brain imaging. Additionally, an extended evaluation is performed on high-resolution meshes with up to ~600,000 nodes, based on a realistic anatomical head model, with five tissue layers, focusing on achieving the desired numerical accuracy for functional brain imaging. Furthermore, support for complex numbers for the cases of FD simulations is incorporated. Specifically, Sec. 2 outlines a DOT implementation using the NIRFAST package along with details highlighting the computational complexity of parameter recovery and the proposed parallelization approaches. In Sec. 3, the results are presented and discussed in the context of employing iterative solvers for the forward problem in DOT. Section 4 concludes with the remaining challenges and opportunities for further optimizations and applications.

2 Methodology

The procedure followed in DOT image reconstruction can be summarized by the following consecutive steps: modeling the light propagation through the medium, also known as the forward problem, and a parameter recovery process based on the forward model and NIR measurements, also known as the inverse problem. This section provides an overview of the underlying mathematics that directly affects the computational aspects of DOT and emphasizes the necessity of parallelization, specifically considering the existing implementation within NIRFAST. Currently, the most computationally expensive procedure is the solution of large FEM sparse linear systems, involved in estimating light propagation in the forward problem.

2.1 *Forward Problem*

The first step of the DOT algorithm, the forward problem, is the basis for the application of model-based image reconstruction; therefore, it must be as accurate as possible, numerically and geometrically, as any errors will affect the formulation of the inverse problem.

The accuracy of the numerical solutions of FEM is greatly affected by the prior knowledge of the underlying tissue geometry; therefore, the maximum potential is reached when FEM is combined with input from other standard imaging techniques[21,22] or generic atlas models.[23] Mesh generation based on structural images from other modalities, usually MRI, is a well-studied problem. There are existing algorithms that automatically segment tissue layers and create surface-based meshes.[24] When the model volume remains constant but meshes with more elements are created, what effectively changes is the resolution of the mesh. Higher resolution meshes have elements with smaller volume and, therefore, minimize partial volume effects due to mesh elements integrating over multiple segmented tissue regions. The fine complex structures, such as the brain cortex in a head model, can be modeled more accurately with a high-resolution mesh, providing optical properties assigned to each node (or element) that are more likely to represent the underlying baseline optical properties of the tissue at each position.

When the volume is meshed, FEM is employed to formulate a discretized weak form representation of the DA for each node of the mesh. It follows that, as the volume of each element tends toward zero (increasing the mesh resolution), the calculated approximation becomes more accurate; therefore, in fDOT, very high-resolution meshes with up to 600,000 nodes are used (Table 1). This is a domain size-dependent problem; a smaller volume, such as an infant's head, will require fewer nodes to achieve elements of sufficiently small volume. Additionally, dividing the volume into a higher resolution mesh minimizes the discretization error introduced by FEM. However, increasing the mesh resolution results in the requirement of solving a bigger linear system to estimate the light fluence, which, until now, has dramatically increased computational time. The focus of this work is optimizing numerical approaches employing FEM-DA to estimate light propagation—a well-studied problem that assumes the DA is valid for all tissue properties used.[9,25,26] However, the advancements described herein can be applied to any models (e.g., radiative transport equation) based on discretized approximations.

FEM is a numerical technique where a heterogeneous problem is divided into many smaller parts, creating a nonuniform mesh, consisting of elements defined by connected nodes. The diffusion equation can then be discretized and represented as a set of linear equations, describing simultaneously all the nodes and hence the entire medium. The problem thus reduces to a sparse, well-posed linear problem of the form

$$M\Phi = q, \tag{1}$$

where M is a sparse matrix with dimensions $N \times N$, with N denoting the number of nodes, q represents the sources and has dimension N by the number of sources Q of the DOT system, and Φ is the photon fluence rate for all nodes for each source, which has dimensions $N \times Q$.

Table 1 Different resolution meshes based on linear tetrahedral elements for an adult head model.

Number of nodes	Number of elements	Element volume average and standard deviation (mm^3)
50,721	287,547	9.26 ± 3.43
68,481	393,863	6.76 ± 2.29
101,046	589,658	4.51 ± 1.64
139,845	821,926	3.24 ± 1.18
205,568	1,215,434	2.19 ± 0.78
235,564	1,395,242	1.90 ± 0.68
271,135	1,609,152	1.65 ± 0.59
305,058	1,813,036	1.46 ± 0.52
324,756	1,931,374	1.37 ± 0.49
360,777	2,149,250	1.23 ± 0.43
411,567	2,454,350	1.08 ± 0.37
515,837	3,084,689	0.86 ± 0.29
610,461	3,656,890	0.72 ± 0.24

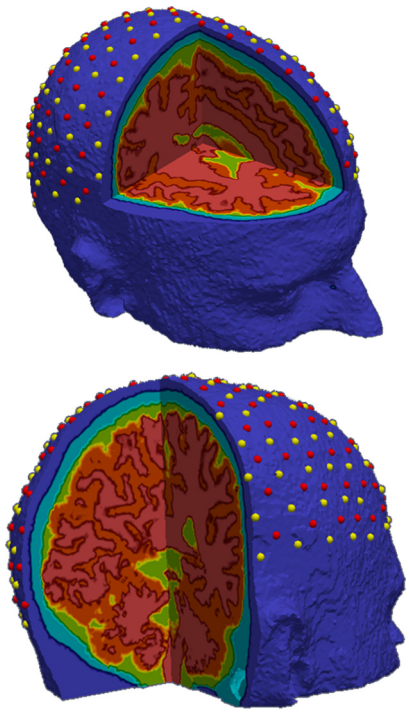

Fig. 1 The modeled HD-DOT system with 158 sources (red) and 166 detectors (yellow) on an adult head model with 5 tissue layers.

2.2 *Inverse Problem*

The second step of DOT image reconstruction, the inverse problem, estimates tissue optical properties based on the forward model and NIR boundary measurements. To acquire the required measurements, NIR light is injected into the imaging volume by optical fibers or LEDs positioned on its surface (light source) while the transmitted and reflected diffused light is measured using optical fibers or detector arrays (light detectors), as it remerges from the surface of the volume in nearby positions.

There are three categories of NIR measurement systems: CW, FD, and time resolved (TR).[27] TR and FD systems are advantageous because the measured transient time or phase-shift information allows the determination of scattering and absorption simultaneously, while CW systems are unable to make quantitative absorption measurements without *a-priori* assumptions of the scattering.[28] However, modeling of light propagation in FD has increased computational cost due to complex arithmetic, while TR requires multiple light propagation models to be estimated for consecutive time instances.

DOT acquires boundary data from multiple and overlapping SD pairs and, therefore, provides valuable spatial depth information and improves lateral image reconstruction resolution. Studies have shown that the density of the SD pairs can directly affect the spatial resolution and localization accuracy of reconstructed images.[29,30] HD-DOT, an arrangement with dense SD pairs, is considered to produce superior results and is particularly effective in brain functional imaging.[4,7,29,31–33]

In fDOT, models of estimated light propagation based on assumptions of the underlying tissue scattering and attenuation are used to create a sensitivity matrix, also known as the Jacobian. The Jacobian is the basis for solving the inverse problem, allowing the recovery of spatiotemporal changes of internal optical properties for the whole volume, using temporal derivatives of measurements obtained on the surface of the volume, known as boundary data, then performing single-step (linear) reconstruction.

The approach employed to form the Jacobian is the adjoint method,[34] where the direct fluence for each source and the adjoint fluence for each detector must be calculated; then the sensitivity is calculated as the product of the direct and the adjoint field, with regard to the basis function for each element, the result of which is a dense matrix. The construction of the Jacobian, therefore, requires the forward problem to be solved twice: once for all sources and once for all detectors.

2.3 *Computational Problem*

The solution of large sparse linear systems involved in the forward model is currently the computational bottleneck of the DOT algorithm. In this presented example, the fluence throughout the volume must be calculated for all 158 sources and for all 166 detectors to create the Jacobian for the modeled DOT system as shown in Fig. 1. To solve the linear systems arising in the forward modeling, in the form $A\mathbf{x} = \mathbf{b}$, for $\mathbf{x}$, where A has dimensions $N \times N$, where N is the number of unknowns, the inverse of A must be calculated. However, calculating a matrix inverse is computationally inefficient; therefore, a variety of algorithms have been proposed that can solve linear systems without explicitly calculating a matrix's inverse. These algorithms can be either direct, providing an exact solution, or an approximate, usually employing an iterative algorithm. The storage convention used in this work to represent sparse matrices within memory is the compressed row storage, which requires $2N_{NZ} + N + 1$ space in memory for a $N \times N$ matrix with N_{NZ} nonzero entries.

2.3.1 *Direct solvers*

The most popular direct solver is the Gaussian elimination, also known as row reduction, where the echelon form of A is calculated through row operations on the augmented matrix $(A|b)$. The echelon form of A is an upper triangular matrix, making the solution of the linear system easy through backward substitution. However, lower–upper (LU) factorization is considered the standard efficient computational approach for direct solution of a linear system. The LU factorization decomposes A into lower (L) and upper (U) triangular matrices. Substituting L and U in $A\mathbf{x} = \mathbf{b}$ gives $LU\mathbf{x} = \mathbf{b}$, and letting $U\mathbf{x} = Y$, then $LY = \mathbf{b}$. Now, it is trivial to solve $LY = \mathbf{b}$ for Y through forward substitutions and then solving $U\mathbf{x} = Y$ through backward substitutions. The advantage of LU factorization over the traditional Gaussian elimination is that decomposing A into L and U is independent of $\mathbf{b}$, also known as the right-hand side vector. This allows L and U to be used for solving for multiple right-hand side vectors. However, this approach has very large memory requirements of $N^2 + N$ and high computational cost, of $\frac{2}{3}N^3$ floating point operations (FLOPS), to solve a full linear system. A more efficient alternative, that can be used only if A is Hermitian and, therefore, symmetric when real, is the Cholesky factorization, where A is decomposed to LL^*, where * denotes the transpose conjugate operator, requiring $\frac{1}{2}N^2 + N$ memory for a full system. The linear system can be solved as with the LU method, substituting $U = L^*$, with computational cost, for the solution of full systems, of $\frac{N^3}{3}$ FLOPS. However, in the case of sparse linear systems, such as resulting from the FEM formulation, memory and computational costs are related to the number of nonzero elements of A rather than the size N; additionally, there are reordering strategies that when applied on sparse matrices allow more sparse factorizations. Specifically, in this work, the approximate minimum degree permutation algorithm was found to produce the most sparse factorizations and, therefore, was used for all the direct solvers. Nevertheless, factorization approaches rely on forward and/or backward substitutions to provide a solution; therefore, they cannot be efficiently parallelized. In MATLAB®, when solving linear systems invoking the backslash operator, the Cholesky approach is used when the matrix is Hermitian; otherwise, the LU approach is employed. The "spparams"[35] command was used to confirm that all the real-linear systems were solved with the Cholesky solver and all the complex with the LU. The MATLAB® backslash operator is considered the numerical ground truth for the solution of linear systems throughout this work.

2.3.2 *Iterative solvers*

To overcome the computational limitations of direct solvers, a variety of approximate solvers have been proposed; they can be classified into three general categories: iterative, multigrid, or domain decomposition methods.[36] Multigrid and domain decomposition methods can be very efficiently parallelized, with solving speed not greatly affected by the size of the linear system. However, these methods require additional input parameters (e.g., the range of eigenvalues of the system, restriction and prolongation parameters, and smoothing operators) that might be difficult to define and may vary for different systems to efficiently converge to adequate approximations. As such, these methods work best when they are tailored to solve a very specific problem. In contrast, iterative solvers are generic and require little or no additional input from the user; thus, they are traditionally chosen for the solution of linear systems describing light propagation.

Iterative approaches approximate a solution vector $\mathbf{x}_n$ and then attempt to minimize the residual $r_n = \|\mathbf{b} - A\mathbf{x}_n\|$ through n iterations, until r_n is lower than a user-defined residual threshold r_{th}. However, in practice, the termination criteria are defined relatively as $t_c = \frac{r_n}{r_0}$, where r_0 is the residual after the initial guess, with the initial guess $\mathbf{x}_0$ set usually as a vector of zeros. Using a relative threshold ensures that the iterations will converge with a final r_n usually within the same order of magnitude as the t_c, even when the number of unknowns is very large. Iterative approaches usually work on a projection space for increased computational efficiency. The most established projection scheme is the Krylov subspace, which is based on the Cayley–Hamilton theorem that implies that the inverse of a matrix can be found as a linear combination of its powers. The Krylov subspace generated by a $N \times N$ matrix A and a vector $\mathbf{b}$ of dimension N is the linear subspace spanned by images of $\mathbf{b}$ under the first α powers of A:

$$K_\alpha(A, \mathbf{b}) = \text{span}\{\mathbf{b}, A\mathbf{b}, A^2\mathbf{b}, \ldots, A^{\alpha-1}\mathbf{b}\}. \tag{2}$$

This formulation avoids matrix-to-matrix operations and instead utilizes matrix-to-vector operations, which can be very efficiently implemented in parallel architectures. The Krylov subspace is generated while the solver seeks to find the minimum of the projection space. Usually least square or gradient-based optimization techniques are employed to solve such problems. There are many proposed algorithms to implement a Krylov space solver, but there is no clear conclusion on which one is fastest when the same termination criteria are required.[37] The most popular approach for the Krylov space gradient optimization is the conjugate gradient (CG), but it is not guaranteed to work in non-Hermitian linear systems.[38] However, there are appropriate Krylov subspace solvers that can handle non-Hermitian systems with relatively low additional computational cost, such as the biconjugate gradient stabilized (BiCGStab).

2.3.3 *Preconditioners*

Iterative solvers do not have robust performance and can be very slow, when the condition number of the system is very large. To overcome this, preconditioned versions of the solvers have been developed. Efficient preconditioning can largely reduce the condition number of a linear system, leading to a dramatically reduced number of iterations to convergence. The preconditioner P, in effect is changing the geometry of the Krylov subspace to a simpler one, making the solution of the system much easier by providing an approximation of the matrix inverse that is easy to compute and solve. Instead of trying to minimize $\|\mathbf{b} - A\mathbf{x}_n\|$, the expression to minimize becomes $\|P^{-1}A\mathbf{x}_n - P^{-1}\mathbf{b}\|$; to be effective, the preconditioner P must be of much lower condition than A. In general, the $P^{-1}A$ product should be as close as possible to identity matrix, in other words, $P^{-1} \approx A^{-1}$.It is hard to theorize what consists a good preconditioner, and the main diagonal of A, also known as Jacobi preconditioner, can be very effective in diagonally dominant systems; however, usually an incomplete factorization of A is used as incomplete LU or incomplete Cholesky (IC) factorization. Recent research on solving linear systems focuses mainly on the choice of efficient preconditioners, emphasizing preconditioners that can be implemented in parallel architectures,[39]

rather than improving the solvers themselves. This happens because the time within each iteration is greatly reduced due to the high level of parallelism offered by GPUs,[40] while preconditioning reduces the number of iterations needed to converge. Nevertheless, in practice, choosing the best preconditioner is usually a trial-and-error procedure. There are block-based preconditioners that are favorable for GPU parallelization.[41,42] In practice, the IC with no prior permutation or pivoting scheme was found to be the best preconditioning option for fast convergence of the MATLAB®-based iterative solvers, while the factorized sparse approximate inverse (FSAI)[43] was found to be the option that produced the fastest overall result with the CUDA and OpenMP implementations. FSAI is constructed by solving local linear systems for each column of A to approximate an A^{-1} with sparsity pattern defined by powers of A. Additionally, a preconditioner inspired by FSAI, where the local linear systems were solved in parallel and only for the three larger values for each column, was implemented; it achieved similar preconditioning effectiveness while reducing the computational time for the construction of the preconditioner. This preconditioner is referred to as "FSAIP" for the rest of this work.

2.3.4 *Numerical accuracy*

The iteration residual r_n, and to an extent the realization of the termination criteria t_c, is bound to the numerical binary representation precision of numbers that the machine, the programming language, and the employed libraries allow. In modern systems, this is double precision, represented by 64 bits of memory, which in practice can represent numbers with relative differences no smaller than 2^{-52}; this is $\sim 2.22 \times 10^{-16}$, which is the minimum value defined in MATLAB®. Any difference smaller than this is lost due to the quantization involved in converting a number that belongs to the real set $\mathbb{R}$ into the binary set $\mathbb{B}_2^{64}$, where $\mathbb{B}_2 = \{0,1\}$. Therefore, requesting a termination residual lower than a scale of 10^{-16} will not result in a more accurate solution, since any additional variation would be under the double precision quantization bin size of MATLAB® and will be rounded to the nearest bin. Apart from the binary rounding errors, when solving a linear system with an iterative solver, the maximum solving precision that can be achieved is analogic to the condition number of the system. The condition number of a linear system A can be estimated as $\kappa(\mathrm{A}) = \|A\|\|A^{-1}\|$, where $\|\|$ is a matrix norm (see Sec. 2.6). The condition number of a linear system is large if there are big differences in its eigenvalues; therefore, when solving large condition linear systems with iterative solvers, good preconditioning is essential to achieve convergence to low errors.

2.3.5 *Complex numbers support*

The existing open-source libraries that provide low-level functions and primitive data structures for parallel programming support on CPUs and GPUs are the Open Multiprocessing (OpenMP)[44] language and Compute Unified Device Architecture (CUDA).[45] However, up to the current version of OpenMP4.5 does not provide native complex numbers support, and CUDA, while providing support for complex numbers, does not come with high-level mathematical functions, such as sparse iterative linear solvers and preconditioners. Therefore, implementations that allow complex support are not a trivial task. Additionally, open-source mathematical libraries that provide iterative solving of sparse linear systems on parallel architectures, such as PARALUTION[46] and ViennaCL,[47] do not provide complex numbers support. However, when formulating the forward problem for systems operating in the FD, the resulting linear system consists of complex numbers. Nevertheless, there are algebraic schemes that allow a linear system of complex numbers to be represented as an equivalent system of real numbers, solved in the real-number domain, and then the solution can be converted back to a complex representation. There are four possible formulations of equivalent real systems as described in Ref. 48; the approach chosen for this work is the K1 approach, which is formulated as

$$A_c = (x + yi) \leftrightarrow \begin{pmatrix} x & -y \\ y & x \end{pmatrix} = A_r, \tag{3}$$

where A_c is the complex form and A_r is the equivalent real representation. Generalizing, the n'th dimensional complex linear system $A_c\mathbf{x}_c = \mathbf{b}_c$ with entries

$$\begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix} \tag{4}$$

is equivalent to the real-linear system $A_r\mathbf{x}_r = \mathbf{b}_r$ with entries

$$\begin{pmatrix} \Re a_{1,1} & -\Im a_{1,1} & \Re a_{1,2} & -\Im a_{1,2} & \cdots & \Re a_{1,n} & -\Im a_{1,n} \\ \Im a_{1,1} & \Re a_{1,1} & \Im a_{1,2} & \Re a_{1,2} & \cdots & \Im a_{1,n} & \Re a_{1,n} \\ \Re a_{2,1} & -\Im a_{2,1} & \Re a_{2,2} & -\Im a_{2,2} & \cdots & \Re a_{2,n} & -\Im a_{2,n} \\ \Im a_{2,1} & \Re a_{2,1} & \Im a_{2,2} & \Re a_{2,2} & \cdots & \Im a_{2,n} & \Re a_{2,n} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \Re a_{n,1} & -\Im a_{n,1} & \Re a_{n,2} & -\Im a_{n,2} & \cdots & \Re a_{n,n} & -\Im a_{n,n} \\ \Im a_{n,1} & \Re a_{n,1} & \Im a_{n,2} & \Re a_{n,2} & \cdots & \Im a_{n,n} & \Re a_{n,n} \end{pmatrix} \begin{pmatrix} \Re x_1 \\ \Im x_1 \\ \Re x_2 \\ \Im x_2 \\ \vdots \\ \Re x_n \\ \Im x_n \end{pmatrix} = \begin{pmatrix} \Re b_1 \\ \Im b_1 \\ \Re b_2 \\ \Im b_2 \\ \vdots \\ \Re b_n \\ \Im b_n \end{pmatrix}. \tag{5}$$

The equivalent real system has the same sparsity pattern and sparsity factor as the original complex system; however, the new system has twice as many unknowns and, therefore, requires double the computations. Additionally, FEM-DA linear systems in FD are no longer Hermitian; therefore, the BiCGStab solver is employed for FD simulations.

In addition, a BiCGStab solver, based in CUDA, operating directly on the complex domain was implemented and used with

the developed parallel constrained FSAI version (FSAIP) to solve FD simulations.

2.4 *GPU/CPU Parallelization*

The proposed approach for accelerating fluence estimation relies on employing efficient libraries for linear algebra operations and performs remarkably faster when GPU-based parallel architectures are available. Over the last decade, the technical advancements in GPUs and their relatively low cost have made GPU computing a very attractive option. Specifically, many linear algebra operations can be parallelized very efficiently in GPU architectures[49] while using sparse representations, resulting in massive reductions of computational time. This can be applied on the solution of the forward model, dramatically decreasing the computational time required to estimate the Krylov subspace. Solvers based on libraries that can be used both in CPU and GPU were implemented to guarantee accessibility by all users. Additionally, the solvers were compiled as MATLAB® executable files (mex files) for both Windows and Linux 64 bit systems, to allow easy invocation from within the MATLAB® environment where the NIRFAST package is based. Specifically, to take advantage of the GPU computing power, implementations based on CUDA[45] were produced; they require the CUDA runtime that is provided with the NVIDIA drivers, and the CUDA Software Developer Kit that is free to download. For CPU environments, the OpenMP backend was used, which is also publicly available with all standard C/C++ compilers. The mathematical library employed to provide efficient implementations of high-level linear algebra operations is PARALUTION;[46] it offers a wide variety of linear solvers and preconditioners, supports sparse matrix and vector formats, and allows a high level of abstraction between code and hardware, making the code highly portable and efficiently scalable to the available hardware. The produced CUDA-based implementations will revert to OpenMP if there is no GPU available in the system.

An algorithm to distribute workload between the CPU and GPU was implemented, and the workload was distributed by balancing the right-hand side input (sources) between the CPU and GPU. Benchmarking tests were performed on all mesh resolutions to define the best workload distribution in each case. However, it was found that in all meshes above 70,000 nodes the solely GPU-based solution was faster, while with meshes with a smaller number of nodes ($\sim 50,000$), the computational time reduction was less than a second. On the other end, the CPU implementation is faster than the equivalent GPU implementation in meshes with <15.000. This is primarily due to time-consuming data transfer and device initialization procedures. Nevertheless, this is dependent on hardware, the number of right-hand side vectors, and the complexity of the imaging domain.

MATLAB® provides sparse linear solvers on the CPU, which can be easily parallelized over the right-hand side vectors using the parallel computing toolbox. However, there are overhead data transfers between memories (RAM, CPU cache memory, and GPU memory) and between computational threads and memory that do not allow computational accelerations to scale linearly with the number of available computational cores.

2.5 *Experimental Setup*

Data from MRI of an adult head were segmented and meshed into 13 different resolution meshes using the algorithm proposed by Jermyn et al.[24] The modeled DOT instrument is a high-density system with 158 NIR light sources and 166 detectors. Each detector is related to sources in separation distance configurations from 1.3 to 4.6 cm, resulting in 3500 associated SD pairs. More details about the resolution of the meshes can be found in Table 1, and the optical properties for each layer of the anatomical model are described in Table 2.[32]

Table 2 Optical properties of tissue layers at 750-nm wavelength.[32]

Tissue layer	μ_a (mm^{-1})	μ_s' (mm^{-1})	Refractive index
Scalp	0.0170	0.74	1.33
Skull	0.0116	0.94	1.33
Cerebrospinal fluid	0.004	0.3	1.33
Gray matter	0.0180	0.84	1.33
White matter	0.0167	1.19	1.33

The light propagation model was calculated for all 158 sources in all experiments, in CW mode and in FD mode at a modulation frequency of 100 MHz, for one NIR wavelength of 750 nm. All the experiments were performed on a desktop computer with 16 GB of RAM, an Intel Core I7-4790 CPU with 4 physical cores, allowing 2 threads per core, resulting in 8 logical cores at 3.6 GHz, and a NVIDIA GTX970 graphics card with 1664 logical cores at 1050 MHz with 4-GB dedicated memory.

2.6 *Metrics*

It is important to ensure that employing an iterative linear solver will not increase the error of the solution. To this end, the accuracy of the proposed solvers was compared against the direct solution, calculated with the backslash operator in MATLAB®. There is no standard way of comparing two matrices, Φ_{ref}, for the fluence calculated with a direct solver, and Φ_{ite}, for the fluence calculated with an iterative solver; however, the first step for all approaches is taking the difference $\Phi_{\mathrm{dif}} = |\Phi_{\mathrm{ite}} - \Phi_{\mathrm{ref}}|$. The most common metrics to quantify the difference Φ_{dif} are the ones induced from vector norms: the 1-norm$\|\Phi_{\mathrm{dif}}\|_1$, which is the maximum of the column sums of Φ_{dif}, the ∞-norm$\|\Phi_{\mathrm{dif}}\|_\infty$, which is the maximum of the row sums of Φ_{dif}, and the 12-norm$\|\Phi_{\mathrm{dif}}\|_2$, which is the maximum singular value of Φ_{dif}, also known as the spectral norm.

However, those metrics do not provide easily comprehensible quantities; therefore, the relative error per node r was calculated as

$$\varepsilon_{\mathrm{rel}}(r) = \frac{|\Phi(r)_{\mathrm{ref}} - \Phi(r)_{\mathrm{ite}}|}{|\Phi(r)_{\mathrm{ref}}|} \times 100\%. \tag{6}$$

This relative error representation is useful for visualization of the error on the mesh nodes and boundary data and provides more comprehensible numbers than the matrix norms.

3 Results and Discussion

The evaluation is performed on one adult head model using an HD-DOT system with 158 sources and 166 detectors (Fig. 1).

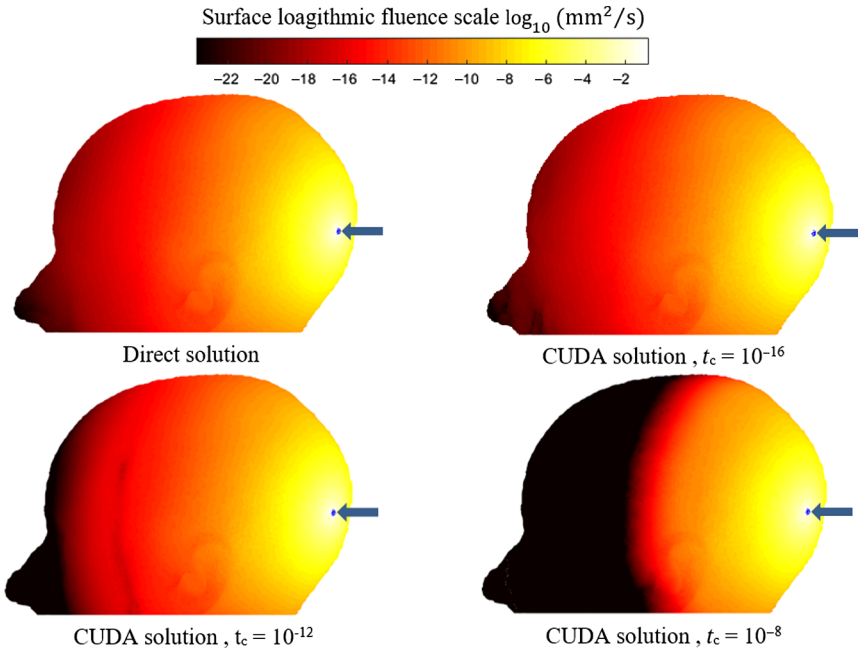

Fig. 2 Visual comparison of surface fluence while using different termination criteria. Simulation in CW, for one source indicated in blue, in a 400,000 nodes mesh, solving with CUDA CG.

The behavior of the solvers is examined under varying error demands and in different mesh resolutions, considering the accuracy and the computational time. The focus is on the relation between mesh resolution (and hence problem size), termination criteria, computational time, and solution error. The direct solutions are only able to calculate up to the 400,000 nodes mesh for CW and up to 200,000 nodes in FD systems due to high-memory requirements, so all quantitative comparisons are performed in the subset of the meshes where a direct solution is available.

3.1 *Qualitative and Quantitative Comparisons*

Considering that there is already some error introduced by the discretization of the DA within the FEM formulation, the error from solving the linear systems should be kept at a minimum. However, the amount of error that can be afforded in the modeling procedure is dependent on the error tolerance for the application. Figure 2 shows the surface fluence when utilizing the CUDA-based solver at different termination criteria; the simulated light source is near the back of the head, indicated by the blue dot and arrow.

When high-termination criteria are set, the fluence is not estimated for the distant nodes as the solution iterates to a stable solution quickly. The fluence approximately follows an exponential decay through tissue; therefore, its value dramatically decreases with distance from the source. Therefore, the required termination criteria are reached while only partially calculating the solution for the highest fluence values.

As FD simulations provide amplitude and phase information, the errors for each were examined separately. Figures 3 and 4 provide a quantification of the relationship between distance

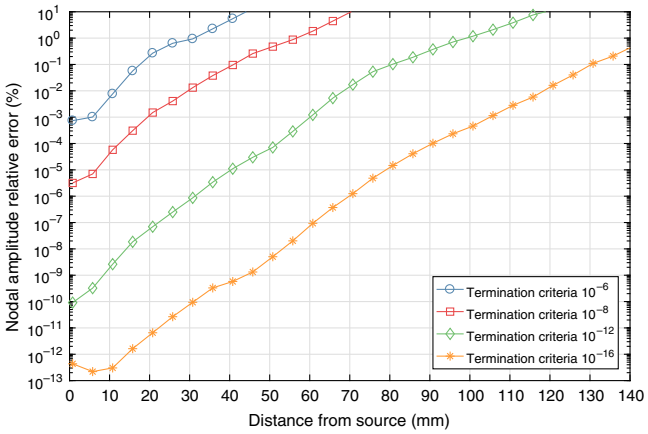

Fig. 3 Maximum relative amplitude errors per node [Eq. (6)] as a function of distance from source. Comparison between different termination criteria. Simulation in 100-MHz frequency, on a 200,000 node mesh, solving with CUDA BiCGStab.

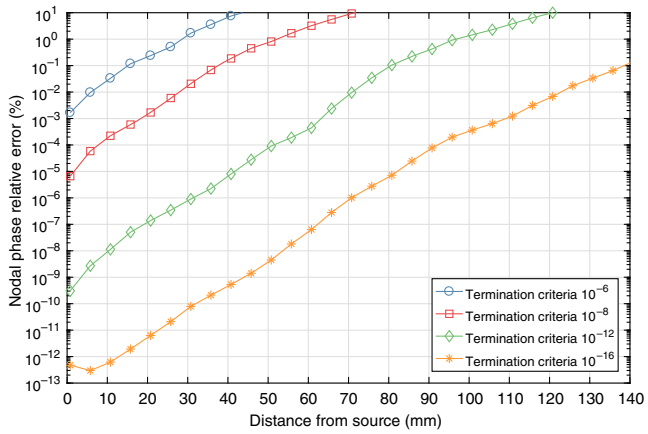

Fig. 4 Maximum relative phase errors per node [Eq. (6)] as a function of distance from sources. Comparison between different termination criteria. Simulation in 100 MHz frequency, on a 200,000 node mesh, solving with CUDA BiCGStab.

from the source and the relative nodal amplitude and phase errors arising from solving with an iterative solver. The demonstrated simulation is in the FD, at 100 MHz, for a mesh with 200,000 nodes, solving with a CUDA BiCGStab and FSAI. In Figs. 3 and 4, the maximum relative error for all nodes as a function of distance from the source is extracted for different termination criteria.

As evident, the relative errors are small and located away from the source with low-termination criteria of 10^{-16}, but they become larger and manifest nearer the source as the termination criteria rises. Similar results (not shown) were acquired for amplitude errors from CW simulations.

Lower termination criteria will provide smaller numerical errors but also slower solver convergence, as a larger number of iterations is required. In the modeled DOT system, the maximum SD separation typically considered to acquire boundary data is at 46 mm. The performed evaluation in Figs. 3 and 4 reveals that, for anatomically accurate adult head models, the termination criteria can be selected in the range of 10^{-8} or lower, for CW and FD systems, to ensure that minimal error is introduced in the parameter recovery, when acquiring measurements from SD separation distances <46 mm. The sensitivity matrix will have approximately square of the error of the forward solution. Therefore, termination criteria chosen to be large, a practise often employed to accelerate reconstructions, while the boundary data are measured in large SD separations, can lead to large errors in the sensitivity matrix and, consequently, large errors in the parameter recovery and image reconstruction.

3.2 *Computational Time Comparisons*

Three parameters mainly affect the computational speed of iterative linear solvers: the size of the problem, which in our case is the number of nodes; the number of right-hand side vectors, which in our case is the number of excitation sources; and the termination criteria. All the experiments were performed 10 times; the mean time is shown in all figures, while the standard deviation in all cases was small, at around 1 s for CPU implementations and 0.1 s for GPU and, therefore, is not shown in the figures.

Figure 5 shows the computational time for fluence estimation for 158 sources in a 400,000 node mesh as a function of termination threshold. The direct solver provides an exact solution to the linear system and, therefore, does not introduce any error. However, it is displayed as a horizontal line through all the termination criteria in Fig. 5 to serve as point of reference. The GPU-based solver yields the best termination criteria to computational time ratio. Employing implementations that do not require much additional time to converge to smaller errors can increase the accuracy of the estimated light propagation model while keeping the computational time low.

Each source is represented by one right-hand side vector in the linear system resulting from the FEM Eq. (1), and the fluence must be calculated for all sources. To achieve this, the iterative solvers must create the Krylov space under the projections of each right-hand side vector, which, as expected, increases the computational cost and therefore the computational time required. Figure 6 demonstrates how the number of sources (right-hand side vectors) affects the computational time of the solution, showing that the computational time increases linearly with the number of sources. It is interesting to note that the direct solver, which yields the exact solution relying on Cholesky decomposition followed by forward and backward substitutions, is almost as efficient for each additional source as the GPU-based solver. However, the time spent initially for the factorization is very large, which, in combination with the very high-memory requirements as discussed in Sec. 2.3.1, renders the direct solver impractical.

Nevertheless, the factor that affects computational time the most is the resolution of the mesh. The more nodes a mesh contains, the bigger the linear systems that needs to be solved is; therefore, more mathematical operations have to be applied to create the Krylov space. Figure 7 presents the computational time needed as the mesh resolution increases. The fastest of the solvers is the CUDA-based solver, which achieves computational time of ~42 s for calculating the fluence for all 158 excitation sources in a 600,000 node mesh this is ~0.25 s to calculate the fluence for one source. The CUDA-based solver performs almost 11 times faster than the MATLAB®-based iterative solver without any parallelization, which takes ~460 s for the same calculation. The direct solver can only solve up to systems with 500,000 nodes before the 16-GB hardware memory availability becomes an underlying issue.

Figure 8 shows the computational time for different mesh resolutions for FD simulations at a modulation frequency of

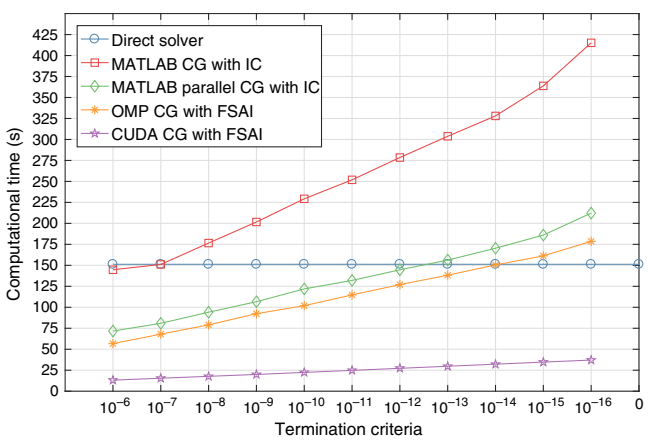

Fig. 5 Computational time as a function of termination criteria, compared between different linear solvers. Simulation in CW, for 158 sources in a 400,000 node mesh.

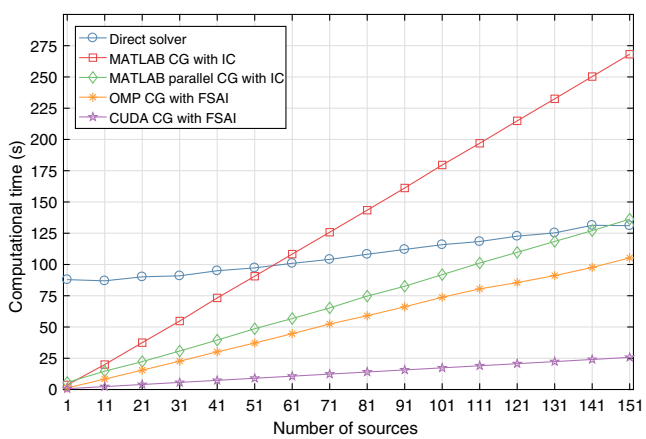

Fig. 6 Computational time as a function of excitation sources number, compared between different linear solvers. Simulation in CW, in a 400,000 node mesh, with 10^{-12} termination criteria.

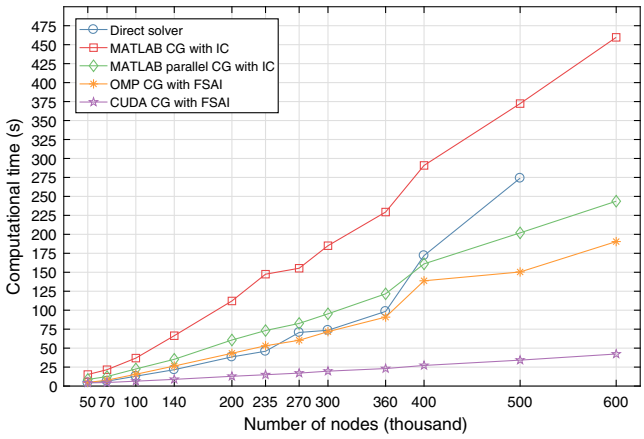

Fig. 7 Computational time with respect to mesh resolution, compared between different linear solvers. Simulation in CW, for 158 sources with 10^{-12} termination criteria.

100 MHz. The direct solver can only handle up to 200,000 nodes due to the increased memory requirements for storing complex numbers. The displayed computational time includes the computations to create the equivalent real system and transform the fluence back to complex after solving the system where is necessary.

The direct solver takes 4612 s to calculate the fluence for the 200,000 node mesh; however, Fig. 8 was limited to 1100 s to provide a better scale. The direct solver becomes intractable due to the increased memory requirements for complex arithmetic storage and because of the non-Hermitian nature of the FEM matrix, which is also reflected as increased memory and computational requirements for the required LU decomposition (in comparison with the Cholesky for the real cases). A linear system resulting from an FD FEM mesh does not have the same condition number as the same mesh in CW due to different attenuation coefficients for frequency modulated light, which makes the FD problem harder to solve. Therefore, there is not a direct analogy between their computational costs. However, one can roughly assume that for a given mesh if a CW solution requires O operations the FD will require $2O$. This is confirmed by our demonstrated results in Figs. 6 and 8.

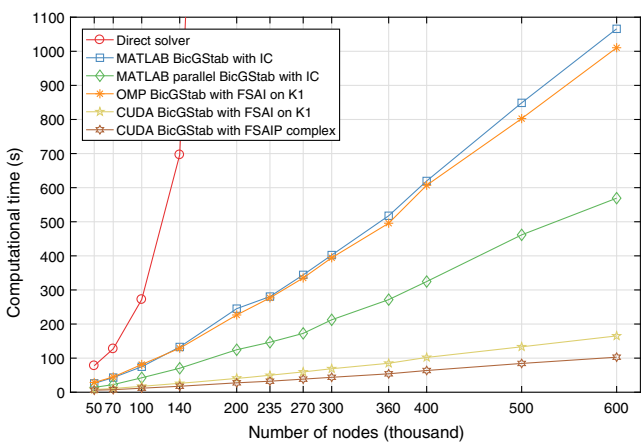

Fig. 8 Computational time with respect to mesh resolution, compared between different linear solvers. Simulation in 100 MHz frequency, for 158 sources with 10^{-12} termination criteria.

Furthermore, the "OMP BiCGStab with FSAI on K1" operates on the complex-to-real transformed (K1) matrix, resulting in double computations in comparison to the "MatlabBiCGStab with IC," which operates directly on the complex domain. As an approximation, one could assume that if a mesh in CW requires O number of operations, it requires $2O$ in the complex domain but $4O$ when the complex-to-real transformation is used. Also, the MATLAB® parallel version requires almost half the computational time of the nonparallel MATLAB® version, and the OpenMP version is slightly faster than the MATLAB® parallel version when operating in the same space (O). Then it is possible to observe the following: a solution on CW would take T seconds for MATLAB® nonparallel, $T/2$ for MATLAB® parallel, and slightly faster than $T/2$ for OpenMP (note that all these cases do O operations). In contrast, a solution on FD would take $2T$ for MATLAB® nonparallel (operates in $2O$), T for MATLAB® parallel (operates in $2O$) and slightly faster than $2T$ for OpenMP (operates in $4O$). Furthermore, the implemented complex CUDA version, which operates in $2O$, requires approximately half the computational time in comparison with the CUDA in the K1 ($4O$) space.

4 Conclusion

DOT is a promising imaging modality, steadily gaining ground among the established imaging techniques. The harmless and patient-friendly procedure enables use in applications where other techniques are inadequate. However, the DOT reconstruction algorithm, especially when employed for functional brain imaging, suffers from large computational time, mainly due to solving large sparse linear systems. This work provides fast GPU and CPU implementations of efficient and stable linear solvers, based on CUDA and OpenMP, respectively, compiled as mex files to be directly accessible from MATLAB®. It will become publicly available in the next release of the NIRFAST package. It is shown that numerical errors introduced by iterative solvers are spatially located away from the excitation source. However, the distance of the numerical errors from the excitation source is related to the termination criteria, indicating that choosing large termination criteria to accelerate the modeling procedure could negatively affect the quality of the reconstruction, dependent on the application. Nevertheless, if the application allows, the computational time of any iterative solver can be greatly reduced by increasing the termination criteria. For example, for the models examined in this work, increasing the termination to 10^{-16} from 10^{-12} will reduce the computational time by half but will increase the modeling error above 1% for the farthest SD separation. Therefore, the underlying physics and the modeling and reconstruction procedure must be considered before attempting to solve with higher termination criteria. However, it is now computationally feasible to select lower termination criteria for the iterative solvers, practically eliminating any error induced by the approximate solving or the complexity of the volume, as the GPU parallelized approach has overly significantly lower computational time. Furthermore, the proposed approaches can be very efficient for systems with a large number of sources and detectors since the computational time is not greatly affected by solving for multiple sources and, in addition, can be employed in FD simulations. Based on the performed experiments, the fastest approach is to parallelize the matrix-to-vector operations involved in iterative solvers in GPU architectures.

The produced solvers allow researchers to explore approaches in DOT that, until now, were out of reach due to the slowness of the algorithm. Simulations of light propagation that would take a long time, now, can be done in a few minutes, forging a path toward real-time DOT. The work presented here is based on systems with one GPU node; however, the same philosophy can be applied in systems with multiple GPUs and extended to cloud computing to achieve real-time solutions. Parallelization approaches can also be applied for the optimization of the inverse problem of DOT, where the creation and the inversion of the Jacobian are currently the most computationally expensive parts of the algorithm, especially when recovering absolute optical parameters. In functional brain imaging, creating sparse Jacobians enables solving the linear inverse problem directly in the GPU in real-time speed for each temporal set of measurements.

Disclosures

The authors have no relevant financial interests in this article and no potential conflicts of interest to disclose.

Acknowledgments

This work has been funded by the National Institutes of Health under Grant Nos. R01NS090874-08, RO1-CA132750, K01MH103594, and R21MH109775.

References

1. A. D. Craig, "How do you feel? Interoception: the sense of the physiological condition of the body," *Nat. Rev. Neurosci.* **3**(8), 655–666 (2002).
2. R. Sitaram et al., "Acquired control of ventral premotor cortex activity by feedback training: an exploratory real-time fMRI and TMS study," *Neurorehabilitation Neural Repair* **26**(3), 256–265 (2012).
3. S. Ruiz et al., "Acquired self-control of insula cortex modulates emotion recognition and brain network connectivity in schizophrenia," *Hum. Brain Mapp.* **34**(1), 200–212 (2013).
4. B. W. Zeff et al., "Retinotopic mapping of adult human visual cortex with high-density diffuse optical tomography," *Proc. Natl. Acad. Sci. U. S. A.* **104**(29), 12169–12174 (2007).
5. D. A. Boas et al., "Improving the diffuse optical imaging spatial resolution of the cerebral hemodynamic response to brain activation in humans," *Opt. Lett.* **29**(13), 1506–1508 (2004).
6. T. Austin et al., "Three dimensional optical imaging of blood volume and oxygenation in the neonatal brain," *NeuroImage* **31**(4), 1426–1433 (2006).
7. S. M. Liao et al., "High-density diffuse optical tomography of term infant visual cortex in the nursery," *J. Biomed. Opt.* **17**(8), 081414 (2012).
8. T. S. Leung et al., "Measurement of the absolute optical properties and cerebral blood volume of the adult human head with hybrid differential and spatially resolved spectroscopy," *Phys. Med. Biol.* **51**(3), 703–717 (2006).
9. H. Dehghani et al., "Near infrared optical tomography using NIRFAST: algorithm for numerical model and image reconstruction," *Int. J. Numer. Methods Biomed. Eng.* **25**(6), 711–732 (2008).
10. H. Scherl et al., "Fast GPU-based CT reconstruction using the common unified device architecture (CUDA)," in *Nuclear Science Symp. Conf. Record (NSS'07)*, Vol. **6**, pp. 4464–4466, IEEE (2007).
11. S.-H. Yoo, J.-H. Park, and C.-S. Jeong, "Accelerating multi-scale image fusion algorithms using CUDA," in *Int. Conf. of Soft Computing and Pattern Recognition (SOCPAR'09)*, pp. 278–282 (2009).
12. Z. Zhang, Q. Miao, and Y. Wang, "CUDA-based Jacobi's iterative method," in *Int. Forum on Computer Science-Technology and Applications (IFCSTA'09)*, Vol. **1**, pp. 259–262 (2009).
13. N. Ren et al., "GPU-based Monte Carlo simulation for light propagation in complex heterogeneous tissues," *Opt. Express* **18**(7), 6811–6823 (2010).
14. Q. Fang and D. A. Boas, "Monte Carlo simulation of photon migration in 3D turbid media accelerated by graphics processing units," *Opt. Express* **17**(22), 20178–20190 (2009).
15. E. Alerstam, T. Svensson, and S. Andersson-Engels, "Parallel computing with graphics processing units for high-speed Monte Carlo simulation of photon migration," *J. Biomed. Opt.* **13**(6), 060504 (2008).
16. J. Prakash et al., "Accelerating frequency-domain diffuse optical tomographic image reconstruction using graphics processing units," *J. Biomed. Opt.* **15**(6), 066009 (2010).
17. T. Zhang et al., "Towards real-time detection of seizures in awake rats with GPU-accelerated diffuse optical tomography," *J. Neurosci. Methods* **240**, 28–36 (2015).
18. M. J. Saikia and R. Kanhirodan, "High performance single and multi-GPU acceleration for diffuse optical tomography," in *Int. Conf. on Contemporary Computing and Informatics (IC3I'14)*, pp. 1320–1323 (2014).
19. X. Yi et al., "Full domain-decomposition scheme for diffuse optical tomography of large-sized tissues with a combined CPU and GPU parallelization," *Appl. Opt.* **53**(13), 2754–2765 (2014).
20. M. Schweiger, "GPU-accelerated finite element method for modelling light transport in diffuse optical tomography," *Int. J. Biomed. Imaging* **2011**, 10 (2011).
21. C. M. Carpenter et al., "Image-guided optical spectroscopy provides molecular-specific information in vivo: MRI-guided spectroscopy of breast cancer hemoglobin, water, and scatterer size," *Opt. Lett.* **32**(8), 933–935 (2007).
22. B. Brooksby et al., "Magnetic resonance-guided near-infrared tomography of the breast," *Rev. Sci. Instrum.* **75**(12), 5262–5270 (2004).
23. J. Mazziotta et al., "A probabilistic atlas and reference system for the human brain: international consortium for brain mapping (ICBM)," *Philos. Trans. R. Soc. London B Biol. Sci.* **356**(1412), 1293–1322 (2001).
24. M. Jermyn et al., "Fast segmentation and high-quality three-dimensional volume mesh creation from medical images for diffuse optical tomography," *J. Biomed. Opt.* **18**(8), 086007 (2013).
25. K. D. Paulsen and H. Jiang, "Spatially varying optical property reconstruction using a finite element diffusion equation approximation," *Med. Phys.* **22**(6), 691–701 (1995).
26. S. R. Arridge et al., "A finite element approach for modeling photon transport in tissue," *Med. Phys.* **20**(2), 299–309 (1993).
27. D. M. Hueber et al., "Non-invasive and quantitative near-infrared haemoglobin spectrometry in the piglet brain during hypoxic stress, using a frequency-domain multidistance instrument," *Phys. Med. Biol.* **46**(1), 41–62 (2001).
28. S. R. Arridge and W. R. B. Lionheart, "Nonuniqueness in diffusion-based optical tomography," *Opt. Lett.* **23**(11), 882–884 (1998).
29. Y Zhan et al., "Image quality analysis of high-density diffuse optical tomography incorporating a subject-specific head model," *Front. Neuroenerg.* **4**, 6 (2012).
30. B. R. White and J. P. Culver, "Quantitative evaluation of high-density diffuse optical tomography: in vivo resolution and mapping performance," *J. Biomed. Opt.* **15**(2), 026006 (2010).
31. X. Wu et al., "Quantitative evaluation of atlas-based high-density diffuse optical tomography for imaging of the human visual cortex," *Biomed. Opt. Express* **5**(11), 3882–3900 (2014).
32. A. T. Eggebrecht et al., "A quantitative spatial comparison of high-density diffuse optical tomography and fMRI cortical mapping," *NeuroImage* **61**(4), 1120–1128 (2012).
33. A. T. Eggebrecht et al., "Mapping distributed brain function and networks with diffuse optical tomography," *Nat. Photonics* **8**(6), 448–454 (2014).
34. S. R. Arridge and M. Schweiger, "Photon-measurement density functions. Part 2: finite-element-method calculations," *Appl. Opt.* **34**(34), 8026–8037 (1995).
35. Mathworks, "spparams," 2017 https://uk.mathworks.com/help/matlab/ref/spparms.html (5 October 2017).
36. Y. Saad, "Parallel iterative methods for sparse linear systems," *Stud. Comput. Math.* **8**, 423–440 (2001).
37. M. Nachtigal, S. C. Reddy, and L. N. Trefethen, "How fast are nonsymmetric matrix iterations?," *SIAM J. Matrix Anal. Appl.* **13**(3), 778–795 (1992).

38. V. Faberf and T. Manteuffel, "Necessary and sufficient conditions for the existance of a conjugate gradient method," *SIAM J. Numer. Anal.* **21**(2), 352–362 (1984).
39. O. Axelsson, S. Farouq, and M. Neytcheva, "Comparison of preconditioned Krylov subspace iteration methods for PDE-constrained optimization," *Numer. Algorithms* **73**(3), 631–663 (2016).
40. M. Benzi, "Preconditioning techniques for large linear systems: a survey," *J. Comput. Phys.* **182**(2), 418–477 (2002).
41. H. Elman et al., "Block preconditioners based on approximate commutators," *SIAM J. Sci. Comput.* **27**(5), 1651–1668 (2006).
42. Y. Saad and J. Zhang, "Enhanced multi-level block ILU preconditioning strategies for general sparse linear systems," *J. Comput. Appl. Math.* **130**(1), 99–118 (2001).
43. L. Y. Kolotilinat and A. Y. Yeremin, "Factorized sparse approximate inverse preconditionings I. Theory," *SIAM J. Matrix Anal. Appl.* **14**(1), 45–58 (1993).
44. L. Dagum and R. Menon, "OpenMP: an industry standard API for shared-memory programming," *IEEE Comput. Sci. Eng.* **5**(1), 46–55 (1998).
45. J. Nickolls et al., "Scalable parallel programming with CUDA," *AMC Queue* **6**(2), 40–53 (2008).
46. PARALUTION Labs, "PARALUTION - v1.0.0," 2015, https://www.paralution.com/ (06 October 2017).
47. P. Tillet and K. Rupp, "Towards performance-portable, scalable, and convenient linear algebra," HotPar'13, 2013, http://0b4af6cdc2f0c5998459-c0245c5c937c5dedcca3f1764ecc9b2f.r43.cf2.rackcdn.com/11423-hotpar13-tillet.pdf (06 June 2017).
48. D. Day and M. A. Heroux, "Solving complex-valued linear systems via equivalent real formulation," *SIAM J. Sci. Comput.* **23**(2), 480–498 (2001).
49. J. R. Humphrey et al., "CULA: hybrid GPU accelerated linear algebra routines," *Defense* **7705**(1), 770502 (2010).

Matthaios Doulgerakis received his BEng degree in electronic engineering from Alexander Technological Educational Institute, Thessaloniki, Greece, in 2013, and his MRes in medical imaging from Kingston University, London, UK. He is a PhD candidate working in developing real-time algorithms for functional diffuse optical tomography (DOT) at the University of Birmingham, UK. His research interests are inverse problems and optimization algorithms in medical imaging, particularly in diffuse optics.

Adam T. Eggebrecht, PhD, is an instructor of radiology at Washington University School of Medicine in the Neurophotonics Lab of the Optical Radiology Labs. His research focuses on developing analytical and optical tools to provide portable and wearable systems for minimally constrained methods to map brain function that extend functional neuroimaging beyond current limitations. These methods allow brain imaging in populations unsuited to traditional methods due to discomfort with current technology or implanted devices that are not able to be studied with MRI due to safety concerns.

Stanislaw Wojtkiewicz, PhD, is a research associate at the School of Computer Science, University of Birmingham. His research focuses on development of algorithms and instruments for recovery of hemodynamic parameters of tissues. His research interests include Monte Carlo and finite-element methods for light transport modeling, DOT, time-resolved and continuous wave near-infrared spectroscopy, diffuse correlation spectroscopy, speckle contrast tomography, and laser Doppler flowmetry.

Joseph P. Culver, PhD, is a professor of radiology, physics, and biomedical engineering at Washington University in Saint Louis. His lab explores ways of exploiting noninvasive optical measurements for both functional- and molecular-biological imaging. Specifically, his group develops subsurface optical tomography for imaging intrinsic, hemoglobin-sensitive contrasts, and exogenous, molecularly targeted-fluorescent contrasts. His lab also develops fluorescence tomography systems and methods to image the biodistribution of molecularly targeted probes in small animal models of disease.

Hamid Dehghani received his BSc degree in biomedical and bioelectronic engineering from the University of Salford, Salford, United Kingdom, in 1994, his MSc degree in medical physics and clinical engineering, and his PhD in medical imaging from Sheffield Hallam University, Sheffield, United Kingdom, in 1999. He is currently a professor of medical imaging at the School of Computer Science, University of Birmingham, United Kingdom. His research interests include development of biophotonics-based methods for clinical and preclinical applications.