applied optics

Graphics processing unit accelerating compressed sensing photoacoustic computed tomography with total variation

Mingjie Gao,[1,†] Guangtao Si,[1,†] Yuanyuan Bai,[2] Lihong V. Wang,[3] Chengbo Liu,[2] and Jing Meng[1,*]

[1]School of Information Science and Engineering, Qufu Normal University, Rizhao 276826, China
[2]Institute of Biomedical and Health Engineering, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China
[3]Department of Electronic Engineering, Andrew and Peggy Cherng Department of Medical Engineering, California Institute of Technology, California 91125, USA
*Corresponding author: qrnumj@gmail.com

Received 20 September 2019; revised 20 November 2019; accepted 10 December 2019; posted 10 December 2019 (Doc. ID 378466); published 15 January 2020

Photoacoustic computed tomography with compressed sensing (CS-PACT) is a commonly used imaging strategy for sparse-sampling PACT. However, it is very time-consuming because of the iterative process involved in the image reconstruction. In this paper, we present a graphics processing unit (GPU)-based parallel computation framework for total-variation-based CS-PACT and adapted into a custom-made PACT system. Specifically, five compute-intensive operators are extracted from the iteration algorithm and are redesigned for parallel performance on a GPU. We achieved an image reconstruction speed 24–31 times faster than the CPU performance. We performed *in vivo* experiments on human hands to verify the feasibility of our developed method. © 2020 Optical Society of America

https://doi.org/10.1364/AO.378466

1. INTRODUCTION

Currently, photoacoustic imaging (PAI) has emerged as a novel biomedical imaging modality because of its ability to simultaneously provide high contrast of pure optical imaging and high resolution of ultrasound imaging. The added benefits of PAI are that it is noninvasive, nonionizing, and low cost compared to other conventional imaging modalities, such as magnetic resonance imaging (MRI) and computer tomography (CT). PAI has been used for many applications in the biomedicine field, such as the early diagnosis of cancers in the breast and prostate, the early detection of vulnerable plaques in atherosclerosis, and experiments in small animals [1–3].

Photoacoustic computed tomography (PACT) is one of the major forms of PAI. PACT has great potential for many preclinical and clinical applications because of its large imaging depth and large field of view. The ultrasound (US) transducer used in PACT typically consists of hundreds of densely packed phased array elements to guarantee high-quality PAI [4,5]. For example, in the Twente photoacoustic mammography system, the number of elements used was 588 [4]. The photoacoustic data are acquired from each of these elements through data acquisition (DAQ) boards and are used to reconstruct the photoacoustic image. In many PACT systems, due to the high cost of the DAQ boards, the number of DAQs used are usually only a fraction (e.g., 1/2 or 1/4) of the total number of phased array elements used in the US transducer. For example, Xia *et al.* used 64 DAQs versus 512 transducer elements [5]. DAQs are multiplexed to acquire data from all elements present in the US transducer. Therefore, multiple laser pulses are required to form one B-scan image, increasing the image acquisition time significantly. Besides, the DAQ speed is also affected by the slow laser repetition rate used in PACT (typically 10–20 Hz). Thus, slow repetition rate, as well as the requirement of multiple laser pulses, increases the acquisition time required to form an image significantly, directly influencing the diagnostic ability of the system. For example, in the system discussed by Xia *et al.*, to form one image, they needed eight laser transmissions with their 64 DAQs for a 512-channel system, resulting in nonreal-time 1.25 frames/s. Such low frame rate DAQ would not capture fast-moving objects, limiting its application where high frame rate DAQ is needed. For example, in the oxy-metabolic imaging of the brain [6], the imaging at a high frame rate is required so that the oxy-metabolic parameters in the tissues do not show an apparent change when the light is switched between different wavelengths. To image the hemodynamic activities of the mouse

brain, Tao *et al.* achieved 400 Hz 2D frame rate over a 3 mm scanning range. Another example where high frame rate DAQ is necessary is in the study of the neural activities (utilizing voltage-sensitive dye). For monitoring the neural dynamics, a higher frame rate is required to capture the rapid cellular-resolved neuronal activities [7]. In addition, the slow frame rate also results in undesired motion artifacts, degrading the image quality significantly. To decrease the acquisition time, new techniques are being explored, e.g., forming high-quality images from a fewer number of laser transmissions while not increasing the cost of the system.

Compressed sensing (CS) is a novel information theory technique that recovers signals even when sampling is far less than Nyquist sampling theory. Many researchers have investigated PACT with the CS technique. Guo *et al.* implemented a CS-based PAI of a rat brain and subcutaneous blood vessels *in vivo* [8], and Meng *et al.* proposed an advanced CS reconstruction model with partially known support for acoustic and optical-resolution PACT [9,10]. Haltmeier *et al.* introduced a different concept for CS-based photoacoustic tomography. In this approach, they used the fact that the typical photoacoustic sources consist of both smooth parts and singularities along interfaces, with the Laplacian of the source being sparse (or at least compressible) [11]. Sandbichler *et al.* also proposed a new scheme based on CS to simultaneously reduce both acquisition time and the system costs of PACT [12].

All the works listed above accelerated DAQ and improved the image quality of PACT with compressed sensing (CS-PACT) with sparse sampling. However, computing requirements for CS-PACT (thus, time for processing) increased significantly because of the inherent iterative characteristic of the algorithm and large computations required within each iteration. Thus, to achieve high-speed image reconstruction and real-time display, new advanced computation methods to accelerate the image reconstruction process are needed to expand the application field. This is the focus of our paper.

Graphics processing unit (GPU)-based parallel computation techniques have become increasingly popular in recent years. The GPU is being used widely in gaming, big-data mining, and artificial intelligence [13–15]. Nowadays, GPU technology is also being used in the medical imaging fields. Yu *et al.* presented a compute unified device architecture (CUDA)-based CT image reconstruction using the algebraic reconstruction technique (ART), and the results show that their approach can achieve up to 6.8×, 7.2×, and 5.4× speedups over counterparts CUDA sparse matrix library (cuSPARSE), Berkeley research computing (BRC), and compressed sparse row-5 (CSR5), respectively [16]. Ha *et al.* developed a GPU-accelerated multivoxel update (MVU) scheme in statistical iterative CT reconstruction and achieved 2× speedup for reconstruction [17]. Inam *et al.* presented a novel GPU-accelerated self-calibrating GRAPPA operator gridding (SC-GROG) for radial acquisitions in MRI and achieved speedup of 6× to 30× [18]. Xu *et al.* accelerated dynamic respiratory correction for MRI-guided cardiac interventions using a GPU and achieved image registration in 176.9 ± 14.0 ms, which was 139× faster than a CPU implementation [19]. Wen *et al.* proposed a GPU-based adaptive kernel regression method for freehand three-dimensional (3D) ultrasound reconstruction, achieving a 288× speedup on GPU [20].

The use of GPU in PAI has also been reported in the literature. Kang *et al.* performed parallel processing using GPU to enable real-time display in optical-resolution photoacoustic microscopy (OR-PAM), and the speedup achieved on GPU was 60× and 30× faster than on the CPU [21]. Peng *et al.* implemented 3D photoacoustic tomography based on GPU-accelerated finite-element method, and the computational cost reduced significantly, by a factor of 38.9 [22]. Wang *et al.* proposed a parallelization strategy to accelerate the filtered back-projection algorithm, and two pairs of projection/back-projection operators and the computation efficiency were improved by factors of 1000, 125, and 250, respectively [23]. Shan *et al.* proposed a finite-element 3D quantitative PAI reconstruction algorithm using GPU, and the imaging speed was 38.9 times faster than the CPU [24]. Recently, Reza *et al.* utilized the GPU parallel computation technique to accelerate the double-stage delay-multiply-and-sum (DS-DMAS) reconstruction method for fast photoacoustic tomography, and the imaging frame rates were improved dramatically [25]. These previous works have demonstrated the feasibility of accelerating PAI using a GPU-based parallel computation method.

However, so far, GPU parallel computing in PAI has been used either for accelerating the OR-PAM display or for accelerating the back-projection, delay-and-sum, and finite-element methods in the PACT reconstruction. No research, to the best of our knowledge, has been published on the use of GPU for CS-PACT reconstruction. As stated before, the CS-PACT has computations different from traditional approaches, including computationally expensive iterative techniques. In this study, we present a GPU-based CS-PACT implementation based upon total variation for high-speed clinical use. We incorporated this algorithm into a custom-made PACT with a high-frequency ultrasonic array to accelerate the CS-PACT reconstruction. We verified the feasibility of our proposed method in *in vivo* experiments using vascular imaging of two human hands. In the next few sections, we elaborate on our CS-PACT algorithm developed for GPU parallel architecture. We utilize the heterogeneous architecture of the computer containing both GPU and CPU to divide the CS-PACT tasks for high performance.

2. METHOD

A. CS-PACT Reconstruction Model

In PAI, the generated acoustic pressure propagates through the tissue and is detected by ultrasonic sensors placed on the tissue surface. The optical absorption accumulation images are then reconstructed using a reconstruction strategy, such as back-projection and model-based methods [26–28].

In this work, our focus is the model-based reconstruction with CS. If X is the image to be reconstructed, Y is the measurement data, and Ψ is the sparse transform, then the objective function of CS-PACT can be expressed as

$$\min F = \|KX\text{-}Y\|_2^2 + \lambda\|\Psi X\|_1. \tag{1}$$

In Eq. (1), the first term represents the square error between the estimated measurements from the reconstructed signal and

the experimentally acquired measurements, and the second part represents the l_1 norm of the sparse signals in the sparse domain. For the L1 norm computation of the image matrix, it is first transformed into a column vector. Then we define L1 norm as the sum of the absolute value of all elements in this vector, i.e., $\|X_{n\times n}\|_1 = \|(\mathrm{vec}(X)_{N\times 1}\|_1 = \sum_{j=1}^{N} |\mathrm{vec}(X)_j|$. The λ is the regularization parameter to determine the trade-offs between data fidelity and sparsity. K is the system matrix and is computed by the principle of the back-projection method. Each row (m, t) of this matrix represents which pixels' signals will be detected by the mth transducer at time t; then the corresponding locations of these pixels will be set at a constant value, and other locations are set to zero. The detailed definition of K can be found in Ref. [8].

Total variation (TV) is a typical sparse transform used in the reconstruction model of the CS-PACT [29,30]. The CS-PACT with TV is written as

$$\min F = \|KX\text{-}Y\|_2^2 + \lambda \|TV(X)\|_1, \tag{2}$$

where TV consists of two parts: row differential transform (represented by T_r) and column differential transform (represented by T_c). So the $TV(X)$ in Eq. (2) is expressed as $\|TV(X)\|_1 = \|T_r * X\|_1 + \|X * T_c\|_1$.

Many methods have been discussed in the literature to solve Eq. (2) [31]. The gradient descent (GD) is a popular method because of its simplicity and effectiveness and is also the preferred method in this paper. The gradient of the objective function with respect to X is calculated using the following equation:

$$\nabla F(X) = 2K^{\mathbf{T}}(KX - Y) + \lambda \nabla \|TV(X)\|_1. \tag{3}$$

Generally, the l_1 norm here is not smooth and not differentiable [30]. To implement the gradient computation on an l_1 norm, a method proposed by Lustig is adopted in our paper. In this method, the absolute value of the l_1 norm of vector X is approximated with a smooth function by using the relation $|X| \approx \sqrt{X * X + \mu}$, where μ is a positive smoothing parameter. With this approximation, $d|X|/dX \approx \frac{X}{\sqrt{X*X+\mu}}$. Let W be a diagonal matrix with the diagonal elements $\mathbf{W}(i, i) = \sqrt{(\Psi X)_i^*(\Psi X)_i + \mu}$; Ψ is a sparse transform. Then the gradient of the l_1 norm of X can be computed with the formulation $\nabla\|\Psi X\|_1 = \Psi^* \mathrm{W}^{-1}\Psi X$. A more detailed description can be found in Ref. [30]. To compute the gradient $\nabla\|TV(X)\|_1$ in Eq. (3), two diagonal matrices, $W_r(i, i) = \sqrt{(T_r * X)_i((T_r * X)_i)^*}$ and $W_c(i, i) = \sqrt{(X * T_c)_i((X * T_c)_i)^*}$, were constructed. Using these two matrices, $\nabla\|TV(X)\|_1$ is calculated as follows:

$$\nabla\|TV(X)\|_1 = \big(T_r' * (W_r)^{-1} * T_r * X + X * (T_c * (W_c)^{-1} * T_c')\big). \tag{4}$$

The iterative image reconstruction process for CS-PACT with TV is illustrated in Fig. 1. The major steps of this flow chart are as follows:

Step 1. Inputs and initialization. Inputs include the system matrix K and the measurement data Y. The gradient, iteration termination condition ξ, and the image are all initialized in this step.

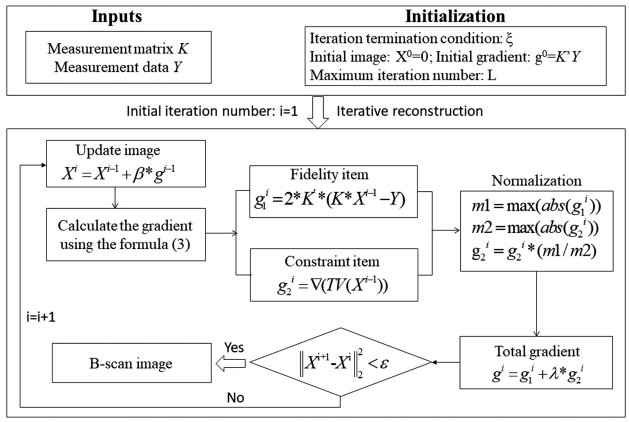

Fig. 1. Flow chart of the iterative image reconstruction for CS-PACT.

Step 2. Image update. With the gradient calculated in the $(i-1)$th iteration, the image in the ith iteration is updated using $x^i = x^{i-1} + \beta * g^{i-1}$, where β is the updating step. In our experiments, the parameter β is chosen optimally after several trials. If β is set too large, the reconstructed image becomes unstable, and if it is set too small, it leads to slow convergence speed. We set $\beta = 0.1$ for optimal convergence.

Step 3. Gradient computation. Calculating the gradient of the objective function with respect to image X includes two parts: fidelity-item gradient and the constraint-item gradient. These two gradients are computed using Eqs. (3) and (4). Gradients are normalized in our algorithm to make the constraint gradient term occupy a reasonable proportion of the total gradient.

Step 4. Judgment of the iteration termination. If the reconstruction error between two images from the successive iterations is smaller than ξ, then terminate the iteration and output the reconstructed image; else, return to Step 2.

As shown in the flow chart, the model-based PACT reconstruction requires many iterations to recover high-quality photoacoustic images. In addition, in each iteration, many matrix–matrix and matrix–vector calculations are needed. Therefore, to accelerate the model-based PACT reconstruction, GPU parallel techniques are needed for both iterative computations as well as matrix operations. The analysis, design, and parallel implementation of computations included in the reconstruction process are presented in the following few sections.

B. Parallel Computation Architecture of CS

In this paper, we implemented the GPU-based parallelized code using NVIDIA's CUDA program model, which provides a unified hardware and software platform for parallel computing [32]. The parallel architecture for GD-based CS-PACT is illustrated in Fig. 2. The CPU tasks include: the DAQ from the PACT system, the construction of the system matrix, and the image display. The system matrix K and the image data Y are copied from the CPU to the global memory of the GPU for

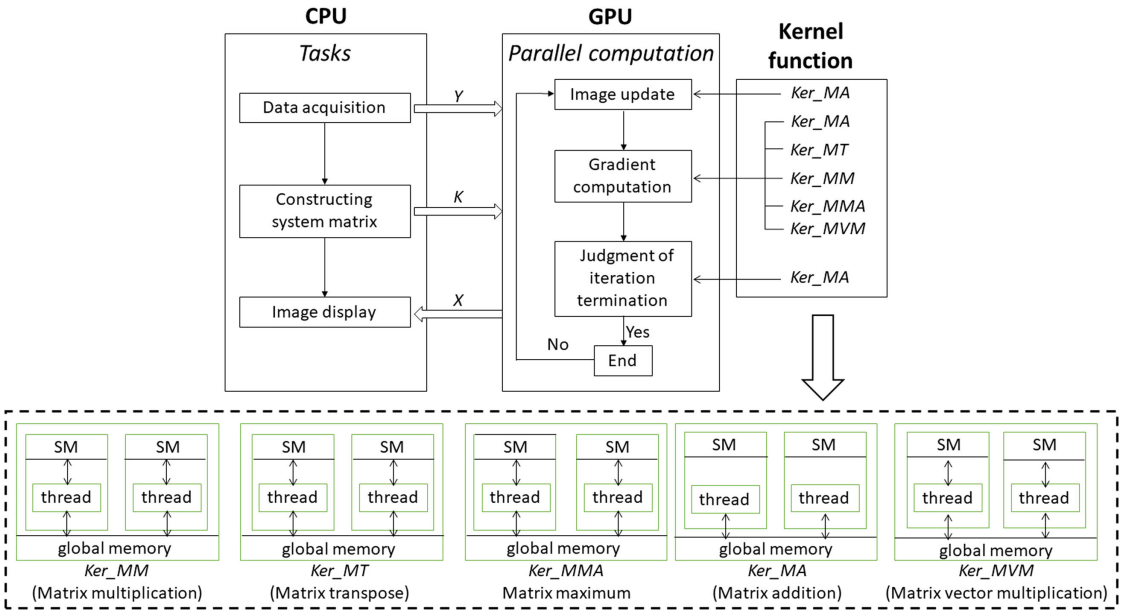

Fig. 2. Parallel computation architecture of CS-PACT.

the computation of the iterative image reconstruction. The five main computations in each iteration of the loop are: (1) matrix multiplication (used to calculate the constraint-term gradient in the objective function); (2) matrix transpose (used in the gradient computation of the fidelity item); (3) matrix maximum (used in the normalization on the gradient of the TV item); (4) matrix addition (used to update the image X and judge the iteration termination condition in each iteration); (5) matrix–vector multiplication (This operator, incorporated with the matrix addition, is used to calculate the fidelity-term gradient in the objective function). All data required for the operator (4) are directly read from the global memory of the GPU by individual threads without using the shared memory. Data required for the other four operators are read from the global memory and then stored in shared memory for sharing data across all threads in a block. When the iterations are completed, the reconstructed photoacoustic images are transferred from the global memory in the GPU to the CPU for image display.

C. Imaging System and In Vivo Experiments

Noninvasive PAI of human hands was performed using a linear-array PACT platform, illustrated in Fig. 3. The major components of the PACT system include: (1) a tunable dye laser (Cobra, Sirah Laser-und Plasmatechnik GmbH, Germany) pumped by a Q-switched Nd:YLF laser (INNOSLAB, Edgewave GmbH, Germany); (2) a custom-built linear ultrasonic array with a center frequency of 30 MHZ; and (3) an 8-core PC (Dell Precision 490) equipped with an eight-channel high-speed peripheral component interconnection (PCI) DAQ card (Octopus CompuScope 8389, GaGe, USA). The system is equipped with a container filled with deionized water, and a low-density polyethylene film-sealed window is placed underneath the container for laser irradiation and signal acquisition. In this system, the DAQ for each two-dimensional (2D) image requires

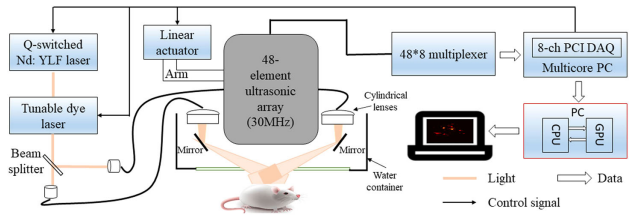

Fig. 3. Schematic of the custom-made PACT system.

six laser pulses, and the 3D data are obtained by mechanical scanning of the 2D ultrasonic probe.

For *in vivo* imaging experiments, the dye laser output was tuned to 584 nm—an isosbestic point at which the oxy- and deoxy-hemoglobin absorb equally. For human hand imaging, the optical fluence on the skin surface was set to $\sim 0.5\ \mathrm{mJ/cm^2}$ per pulse, well below the American National Standard Institute (ANSI)-recommended maximum permissible exposure (MPE) of $20\ \mathrm{mJ/cm^2}$ for a single pulse in the visible spectral range. As the ANSI safety limit for this pulse width region is based on the thermal mechanism, our compliance with the ANSI standards guarantees no thermal damage to the tissue. The human experiments described here were carried out in compliance with Washington University-approved protocols.

3. RESULTS

To evaluate the developed method, we performed experiments on two data sets acquired from two human hands, named hand-1 and hand-2. For each reconstructed 3D volume of a data set, 166 B-scan frames were acquired. For hand-1, each reconstructed B-scan image consists of 128 pixels × 128 pixels (corresponding to a cross section of ~ 6.4 mm × 1.6 mm). We reconstructed maximum amplitudes projection (MAP) (along the depth direction) images for the acquired 3D volume by

using different reconstruction methods with different sampling rates (SRs); results are shown in Fig. 4.

The MAP results using measurements from all 48 transducers with back projection (BP), CS-PACT (CPU), and CS-PACT (GPU) reconstruction methods are shown in Figs. 4(A)–4(C), respectively. Representative B-scan images reconstructed using the above three methods, along the horizontal dashed lines in MAP images, are shown in Figs. 4(a1)–4(c2). From these figures, we can observe that (1) our proposed GPU-based PACT method reconstructs the image equivalent to traditional BP, demonstrating the accuracy of our method; (2) higher-quality photoacoustic images with fewer artifacts are achieved by the CS-PACT method [Figs. 4(B)–4(C), 4(b1)–4(c2)], demonstrating the advantages of this reconstruction method. Reconstructed results from only 24 transducer elements are also shown in this figure. Figures 4(D)–4(F) are MAP results obtained with BP, CS-PACT (CPU), and CS-PACT (GPU), respectively, when using results from only 24 channels. Figures 4(d1)–4(f2) are representative B scans reconstructed by the above three methods, along the horizontal dashed lines shown in MAPs. When the SR is decreased, the reconstructed images with BP became worse due to reconstruction artifacts from sparse sampling [Figs. 4(D), 4(d1), and 4(d2)]. However, when the CS-based PACT method was used, the quality of reconstructed images improved significantly, and most of the artifacts are suppressed [Figs. 4(E), 4(F), 4(e1)–4(f2)]. Moreover, the reconstructed results of CS-PACT on GPU are similar to those on CPU, demonstrating the accuracy of our GPU algorithms. These results establish that the proposed GPU-based CS-PACT method can reconstruct high-quality photoacoustic images accurately with fewer measurements.

Reconstructed photoacoustic images for hand-2 are shown in Fig. 5. In this data set, a B-scan image consists of 256 pixels × 128 pixels (corresponding to a cross section of ~ 6.4 mm × 3.2 mm), whose size is $2x$ the first data set. MAP images reconstructed with full measurements are listed in Figs. 5(A)–5(C), and those results obtained from sparse sampling are shown in Figs. 5(D)–5(F). Representative B-scan images indicated by the dashed lines in MAP images are also shown in this figure. The hand-2 results are similar to the hand-1 results. The CS-PACT can reconstruct high-quality PACT images, compared to the traditional BP method when using the same measurements. Our developed GPU-based CS-PACT can recover the same images as those reconstructed on the CPU. The

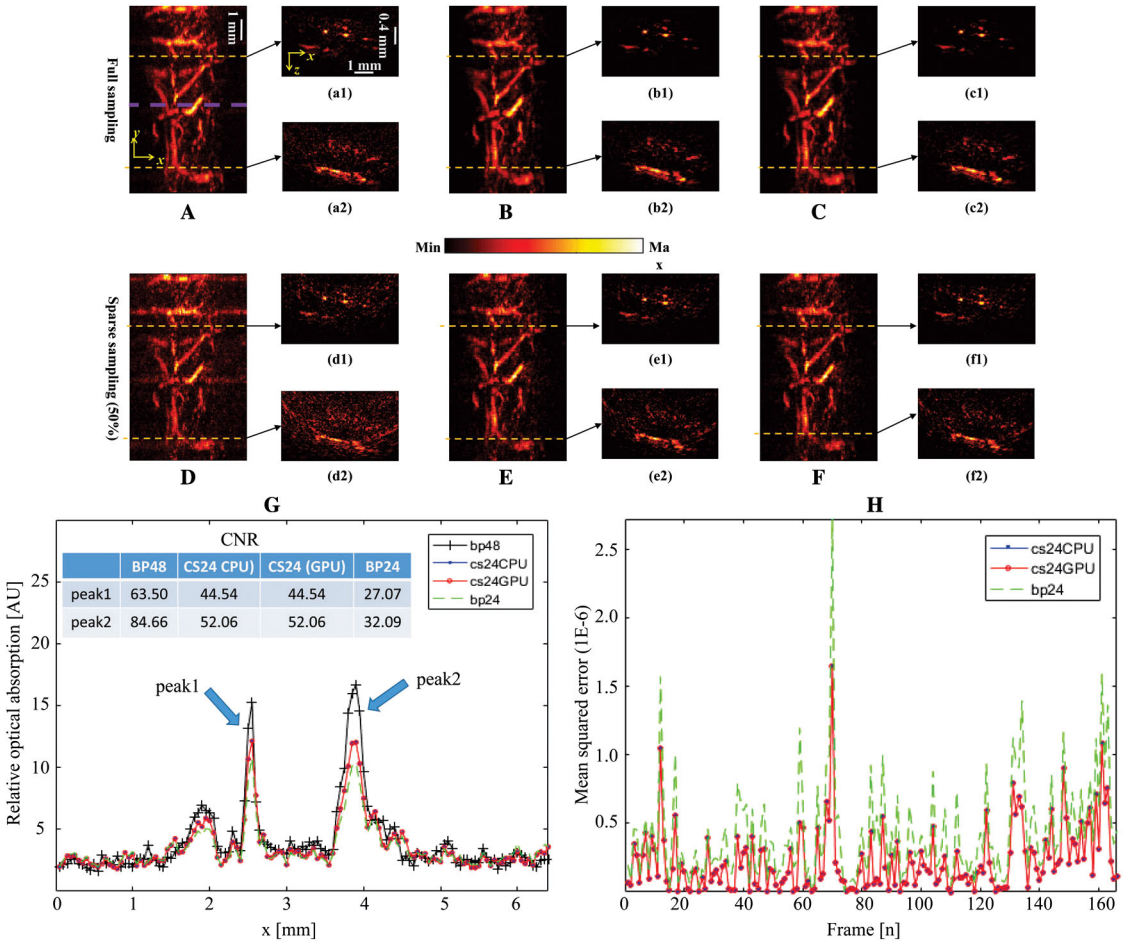

Fig. 4. Reconstructed photoacoustic images of a human hand-1. (A)–(C) MAP images reconstructed by BP, CS-PACT (CPU) and CS-PACT (GPU), respectively, with data from 48 transducer elements; (D)–(F) MAP images reconstructed using three methods with data from 24 transducer elements; (a1)–(f2) representative B-scan images reconstructed by different methods, along the dashed lines shown in MAPs; (G) photoacoustic amplitudes (relative optical absorption) along the chosen thicker dashed line in MAP image of (A); (H) MSE curves of the 166-frame photoacoustic images reconstructed by different methods with a 50% SR.

experiments further verified the feasibility and advantages of our developed GPU-accelerating PACT method.

To compare the results between different reconstruction methods, two quantitative analyses are presented when the SR is 50%. The first quantitative parameter is the contrast-to-noise ratio (CNR) for a single image, and the second is the mean squared error (MSE) between the reconstructed image and the reference image (here, the reference image is the photoacoustic image reconstructed by BP with full data).

The reconstructed hand-1 and hand-2 data set at 50% SR was selected to compute the CNRs. The relative optical absorption amplitudes of different MAP images along the thicker dashed lines in the reference images Figs. 4(A) and 5(A) are plotted in Figs. 4(G) and 5(G), respectively. For each plot, two representative peak signals (indicated by the arrows) are selected to compute the CNRs, and the results for different methods are listed in Table 1. From the table, the CNRs of the images recovered by CS-PACT (CPU and GPU) are about 1.7X more than the images reconstructed by BP. The calculated MSEs using different methods are shown in Figs. 4(H) and 5(H) for reconstructed 166-frame photoacoustic images of data set hand-1 and hand-2 at 50% SR, respectively. From these curves, we can see that the MSEs of CS-PACT (CPU and GPU) are much lower than the traditional BP method. In addition, the differences between CS-PACT images on CPU and CS-PACT images on GPU are very small, and the curves match, demonstrating the accuracy of our proposed GPU-based CS-PACT.

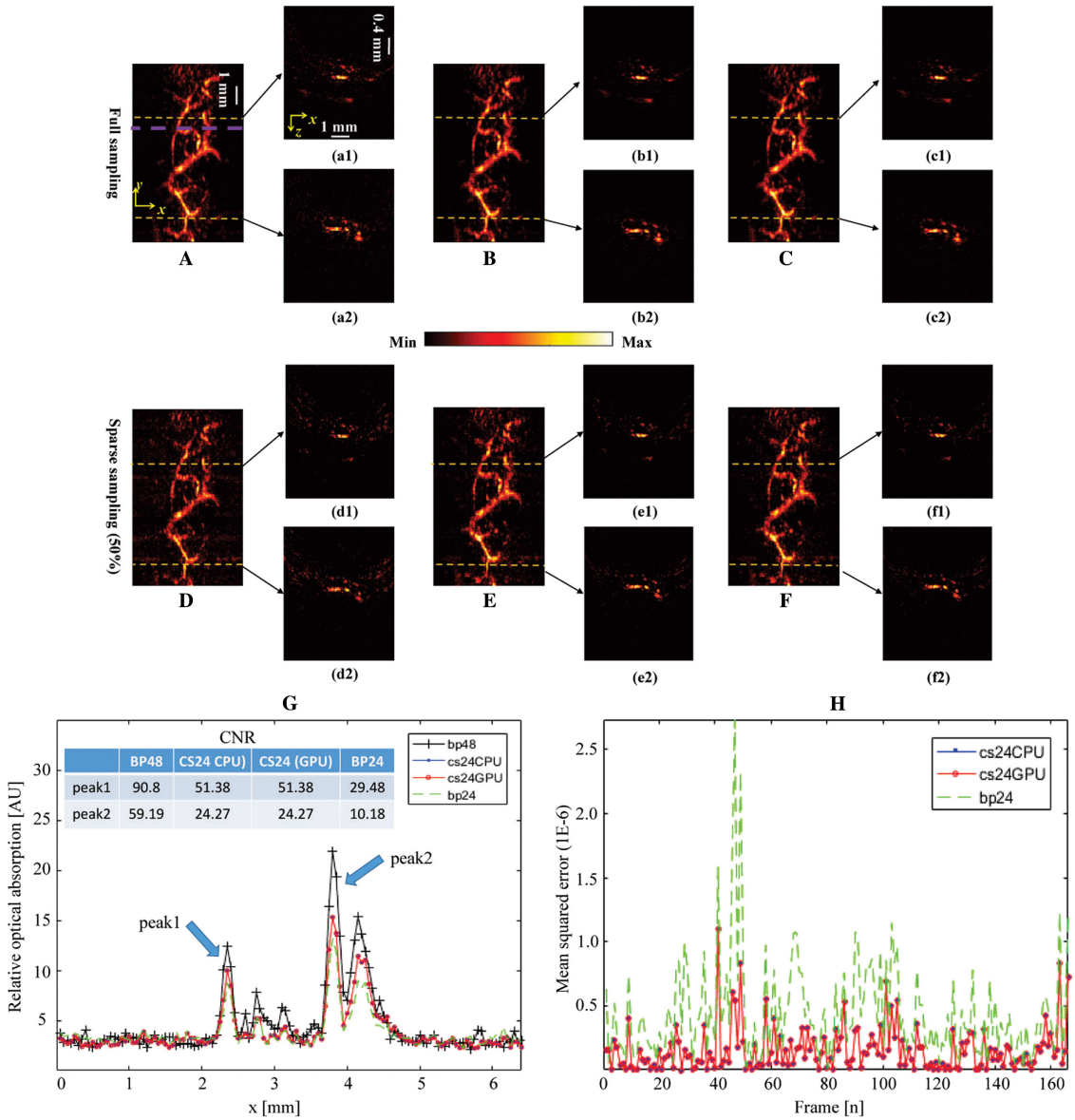

Fig. 5. Reconstructed photoacoustic images of a human hand-2. (A)–(C) MAP images reconstructed by BP, CS-PACT (CPU) and CS-PACT (GPU), respectively, with data from 48 transducer elements; (D)–(F) MAP images reconstructed using three methods with data from 24 transducer elements; (a1)–(f2) representative B-scan images reconstructed by different methods, along the dashed lines shown in MAPs; (G) photoacoustic amplitudes (relative optical absorption) along the chosen thicker dashed lines in MAP images of (A); (H) MSE curves of the 166-frame photoacoustic images reconstructed by different methods with a 50% SR.

Table 1. Comparisons on the Performance of CS-PACT between GPU and CPU[a]

	Hand-1				Hand-2			
Operators	**GPU (48/24)**	**CPU(48/24)**	**A.F.**	**Err**	**GPU(48/24)**	**CPU(48/24)**	**A.F.**	**Err**
B scan	2560/1900	80,000/45,000	31/24	$<1e-7$	8000/5000	224,000/130,000	28/26	$<1e-7$
Matrix multiplication	0.2/0.2	10/10	50	$<1e-7$	1/1	20/20	20	$<1e-7$
Matrix transpose	22.0/16.3	540/260	25/16	$<1e-7$	85/57	9469/3720	111/65	$<1e-7$
Matrix addition	0.003/0.003	0.05/0.05	17	$<1e-7$	0.07/0.07	1/1	14	$<1e-7$
Matrix vector multiplication	10.0/7.2	351/231	35/32	$<1e-7$	50/21	1650/690	33	$<1e-7$
Matrix maximum	0.05/0.05	0.973/0.973	19	$<1e-7$	0.08/0.08	2/2	25	$<1e-7$
Data transfer	80/47	—	—		376/217	—	—	

[a]Unit, millisecond; A.F., acceleration factor; Err, error.

The performance comparisons (for two data sets) between the CPU and GPU are shown in Table 1. In this table, both the B-scan reconstruction time of CS-PACT for the CPU and GPU with data from 48 and 24 transducer elements and the reconstruction errors between CPU and GPU are shown. The computation times of major operators in one iteration are listed separately for comparison. When compared to the CPU speed, the imaging speed of B scan on the GPU improved by 24–31 times (depending on the data sets). The other operators in the iteration improved by about 14–110 times. Since the matrix addition, matrix multiplication, and a maximum of a matrix are performed on the TV and image matrices; their performance is independent of several transducer elements used to reconstruct the image. The reconstruction error between CPU and GPU is less than 1e-7, thus demonstrating the equivalence of parallel algorithms performed on GPU compared to CPU. In addition, the data transfer time from CPU to GPU is also listed in this table. The data transfer time occupied about $\sim 2\%-4\%$ of the total B-scan image reconstruction time on GPU. Thus, the data transfer time can be ignored when evaluating the performance of the computation. Here, all algorithms are executed on a PC with a 3.4 GHz Intel Core i3-3240 CPU, 10 GB memory and a GPU with NVIDIA GeForce GTX 1060. The performance depends upon the GPU used, and the comparisons in this table are just to showcase the performance improvement that can be achieved on a GPU platform using parallel programming.

4. DISCUSSION AND CONCLUSIONS

Our experimental results successfully demonstrate the feasibility of a GPU-based CS-PACT for high performance. Even though the CS-PACT algorithms were incorporated into a custom-made PACT system, it can be easily integrated into a commercial system containing the GPU resources.

The major contributions of our work presented here are summarized as follows: (1) we are the first, we believe, to present a GPU-based CS-PACT reconstruction based upon TV for high-speed clinical use; (2) we incorporated this algorithm into a custom PACT and demonstrated the *in vivo* feasibility by conducting experiments on two human hands; (3) MSE and CNR using CS-PACT are better than conventional methods; (4) the improved processing speed frees up the computational resources on the system for developing more advanced reconstruction algorithms or for incorporating new image processing routines to improve the image quality. The developed GPU-accelerating CS-PACT will be an effective imaging tool for PAI fields requiring high imaging frame rates, such as the metabolic imaging of the body or the study on brain activity, as discussed in the Introduction.

One of the challenges for integrating algorithms into GPU is limited memory size. For CS-PACT, storing K matrix is one of the major considerations. When the image to be reconstructed is large, the size of K could become too large to be stored in the GPU memory. As an example, assuming we have a PACT equipped with a 512-element ultrasonic array, it is used to image a region of 1000 pixels $\times$ 1000 pixels. If the number of sampling time points for each transducer is 500, then the system matrix K will occupy about 1 T memory when using the float data type, which is very large for GPU storage. Two strategies can be adopted to resolve this issue: one is to use the sparse-storage method to store the sparse matrix K as proposed by Refs. [33,34]; the other is to divide a large image to be reconstructed into several small subimages and reconstruct each subimage independently on the GPU or on the host processor. In our experiments, the image size is relatively small (128 pixels $\times$ 128 pixels for hand-1 and 256 pixels $\times$ 128 pixels for hand-2) and the maximum K size is about 1.5 G. The GPU on our custom-PACT is equipped with sufficient memory (8 G) and thus could accommodate the K matrix in our experiments easily.

The performance we presented here is for the GPU used in this experiment. However, when a different GPU is used, performance will vary based upon the capability of the respective GPU. The algorithms implemented using CUDA are portable across any NVIDIA GPU. Thus, if a more powerful GPU is used, e.g., GeForce RTX 2060, performance will also improve significantly. In addition, the performance on the GPU may be affected by the use of memories available on the GPU, the grid division on the data, the number of threads supported per block, etc. The methodology of implementation we have discussed on NVIDIA GPU are generic and thus can be implemented on other architecture as well, such as AMD's Radeon and or Intel's IvyBridge architecture [35]. The CUDA software used here is for convenience on NVIDIA GPU and thus can be extended to other GPU software, e.g., openCL software.

The sparse data used in the image reconstruction are extracted from the full data, which were not acquired on a real sparse-array PACT. Thus, our parallel image reconstruction implementation is executed offline and is not integrated into the whole PACT system. In the near future, we will integrate our parallel

algorithm into the PACT system and apply the GPU-based CS-PACT reconstruction to *in vivo* data for real-time visualization. We will also conduct *in vivo* blood-flow imaging to exhibit the utility of improved performance on a PACT platform integrated with GPU and a sparse ultrasonic array.

Funding. National Natural Science Foundation of China (61308116, 81427804, 81522024, 91739117); Shenzhen Science and Technology Innovation grant (JCYJ20160531175040976, JCYJ20170413153129570).

Disclosures. The authors declare no conflicts of interest.

[†]These authors contributed equally to this work.

REFERENCES

1. J. Yao and L. V. Wang, "Breakthrough in photonics 2013: photoacoustic tomography in biomedicine," IEEE Photon. J. **6**, 0701006 (2014).
2. X. Li, C. D. Heldermon, L. Yao, L. Xi, and H. Jiang, "High resolution functional photoacoustic tomography of breast cancer," Med. Phys. **42**, 5321–5328 (2015).
3. J. Krista, V. S. Gijs, and A. F. W. van der Steen, "Intravascular photoacoustic imaging: a new tool for vulnerable plaque identification," Ultrasound Med. Biol. **40**, 1037–1048 (2014).
4. M. Heijblom, W. Steenbergen, and S. Manohar, "Clinical photoacoustic breast imaging," IEEE Pulse **6**, 42–45 (2015).
5. J. Xia, M. R. Chatni, K. Maslov, Z. Guo, K. Wang, M. Anastasio, and L. V. Wang, "Whole-body ring-shaped confocal photoacoustic computed tomography of small animals in vivo," J. Biomed. Opt. **17**, 050506 (2012).
6. J. Yao, L. Wang, J. M. Yang, K. I. Maslov, T. T. Wong, L. Li, C. H. Huang, J. Zou, and L. V. Wang, "High-speed label-free functional photoacoustic microscopy of mouse brain in action," Nat. Methods **12**, 407–410 (2015).
7. S. Gottschalk, O. Degtyaruk, B. Mc Larney, J. Rebling, M. A. Hutter, X. L. Dean-Ben, S. Shoham, and D. Razansky, "Rapid volumetric optoacoustic imaging of neural dynamics across the mouse brain," Nat. Biomed. Eng. **3**, 392–401 (2019).
8. Z. Guo, C. Li, L. Song, and L. V. Wang, "Compressed sensing in photoacoustic tomography in vivo," J. Biomed. Opt. **15**, 021311 (2010).
9. J. Meng, C. Liu, J. Zheng, R. Lin, and L. Song, "Compressed sensing based virtual-detector photoacoustic microscopy in vivo," J. Biomed. Opt. **19**, 036003 (2014).
10. J. Meng, L. V. Wang, D. Liang, and L. Song, "Compressed-sensing photoacoustic computed tomography *in vivo* with partially known support," Opt. Express **20**, 16510–16523 (2012).
11. M. Haltmeier, M. Sandbichler, T. Berer, J. Bauer-Marschallinger, and L. Nguyen, "A new sparsification and reconstruction strategy for compressed sensing photoacoustic tomography," J. Acoust. Soc. Am. **143**, 3838–3848 (2018).
12. M. Sandbichler, F. Krahmer, T. Berer, P. Burgholzer, and M. Haltmeier, "A novel compressed sensing scheme for photoacoustic tomography," Siam. J. Appl. Math. **75**, 2475–2494 (2015).
13. R. L. Davidson and C. P. Bridges, "Error resilient GPU accelerated image processing for space applications," IEEE Trans. Parallel Distrib. Syst. **29**, 1990–2003 (2018).
14. M. Kim, L. Ling, and W. Choi, "A GPU-aware parallel index for processing high-dimensional big data," IEEE Trans. Comput. **67**, 1388–1402 (2018).
15. F. Garcia, L. Ubeda-Medina, and J. Grajal, "Real-time GPU-based image processing for a 3-D THz radar," IEEE Trans. Parallel Distrib. Syst. **28**, 2953–2964 (2017).
16. X. Yu, H. Wang, W.-C. Feng, H. Gong, and G. Cao, "GPU-based iterative medical CT image reconstructions," J. Signal Process. Syst. **91**, 321–338 (2019).
17. S. Ha and K. Mueller, "A GPU-accelerated multivoxel update scheme for iterative coordinate descent (ICD) optimization in statistical iterative CT reconstruction (SIR)," IEEE Trans. Comput. Imag. **4**, 355–365 (2018).
18. O. Inam, M. Qureshi, S. A. Malik, and H. Omer, "GPU-accelerated self-calibrating GRAPPA operator gridding for rapid reconstruction of non-Cartesian MRI data," Appl. Magn. Reson. **48**, 1055–1074 (2017).
19. R. Xu and G. A. Wright, "GPU accelerated dynamic respiratory motion model correction for MRI-guided cardiac interventions," Comput. Methods Programs Biomed. **136**, 31–43 (2016).
20. T. Wen, Y. Feng, G. Jia, S. Chen, W. Lei, and Y. Xie, "An adaptive kernel regression method for 3D ultrasound reconstruction using speckle prior and parallel GPU implementation," Neurocomputing **275**, 208–223 (2018).
21. H. Kang, S. W. Lee, E. S. Lee, S. H. Kim, and T. G. Lee, "Real-time GPU-accelerated processing and volumetric display for wide-field laser-scanning optical-resolution photoacoustic microscopy," Biomed. Opt. Express **6**, 4650–4660 (2015).
22. K. Peng, L. He, Z. Zhu, J. Tang, and J. Xiao, "Three-dimensional photoacoustic tomography based on graphics-processing-unit-accelerated finite element method," Appl. Opt. **52**, 8270–8279 (2013).
23. K. Wang, C. Huang, Y. J. Kao, C. Y. Chou, A. A. Oraevsky, and M. A. Anastasio, "Accelerating image reconstruction in three-dimensional optoacoustic tomography on graphics processing units," Med. Phys. **40**, 023301 (2013).
24. T. Shan, J. Qi, M. Jiang, and H. Jiang, "GPU-based acceleration and mesh optimization of finite-element-method-based quantitative photoacoustic tomography: a step towards clinical applications," Appl. Opt. **56**, 4426–4432 (2017).
25. S. R. M. Rostam, M. Mozaffarzadeh, M. Ghaffari-Miab, A. Hariri, and J. Jokerst, "GPU-accelerated double-stage delay-multiply-and-sum algorithm for fast photoacoustic tomography using LED excitation and linear arrays," Ultrason. Imaging **41**, 301–316 (2019).
26. C. Gong and L. Zeng, "Adaptive iterative reconstruction based on relative total variation for low-intensity computed tomography," Signal Process. **165**, 149–162 (2019).
27. X. L. Dean-Ben, E. Mercep, and D. Razansky, "Hybrid-array-based optoacoustic and ultrasound (OPUS) imaging of biological tissues," Appl. Phys. Lett. **110**, 203703 (2017).
28. M. Xu and L. V. Wang, "Universal back-projection algorithm for photoacoustic computed tomography," Phys. Rev. E **71**, 016706 (2005).
29. Y. Tsaig and D. L. Donoho, "Extensions of compressed sensing," Signal Process. **86**, 549–571 (2006).
30. M. Lustig, D. Donoho, and J. M. Pauly, "Sparse MRI: the application of compressed sensing for rapid MR imaging," Magn. Reson. Med. **58**, 1182–1195 (2007).
31. D. L. Donoho and Y. Tsaig, "Fast solution of l_1-norm minimization problems when the solution may be sparse," IEEE Trans. Inf. Theory **54**, 4789–4812 (2008).
32. NVIDIA CUDA Programming Guide 2.0 (NVIDIA, 2008).
33. F. S. Smailbegovic, G. N. Gaydadjiev, and S. Vassiliadis, "Sparse matrix storage format," in *Proceedings of the 16th Annual Workshop on Circuits, Systems and Signal Processing* (2005), pp. 445–448.
34. I. Simecek and D. Langr, "Tree-based space efficient formats for storing the structure of sparse matrices," IJDPS **15**, 1–20 (2014).
35. M. Mantor, "AMD Radeon™ HD 7970 with graphics core next (GCN) architecture," in *IEEE Hot Chips 24 Symposium (HCS)* (2012), pp. 1–35.