Journal of Biomedical Optics

BiomedicalOptics.SPIEDigitalLibrary.org

On the possibility of producing true real-time retinal cross-sectional images using a graphics processing unit enhanced master-slave optical coherence tomography system

Adrian Bradu
Konstantin Kapinchev
Frederick Barnes
Adrian Podoleanu

SPIE.

On the possibility of producing true real-time retinal cross-sectional images using a graphics processing unit enhanced master-slave optical coherence tomography system

Adrian Bradu,[a,*] Konstantin Kapinchev,[b] Frederick Barnes,[b] and Adrian Podoleanu[a]
[a]University of Kent, Applied Optics Group, School of Physical Sciences, Canterbury, CT2 7NH, United Kingdom
[b]University of Kent, Programming Languages and Systems, School of Computing, Canterbury, CT2 7NF, United Kingdom

Abstract. In a previous report, we demonstrated master-slave optical coherence tomography (MS-OCT), an OCT method that does not need resampling of data and can be used to deliver *en face* images from several depths simultaneously. In a separate report, we have also demonstrated MS-OCT's capability of producing cross-sectional images of a quality similar to those provided by the traditional Fourier domain (FD) OCT technique, but at a much slower rate. Here, we demonstrate that by taking advantage of the parallel processing capabilities offered by the MS-OCT method, cross-sectional OCT images of the human retina can be produced in real time. We analyze the conditions that ensure a true real-time B-scan imaging operation and demonstrate *in vivo* real-time images from human fovea and the optic nerve, with resolution and sensitivity comparable to those produced using the traditional FD-based method, however, without the need of data resampling. [DOI: 10.1117/1.JBO.20.7.076008]

Keywords: master-slave; cross-section; optical coherence tomography; graphics processing units; CUDA.

Paper 150193R received Mar. 23, 2015; accepted for publication Jun. 24, 2015; published online Jul. 21, 2015.

1 Introduction

Both implementations of the spectral or Fourier domain optical coherence tomography (OCT), spectrometer based OCT and swept source OCT, can be used to produce cross-sectional (B-scan) images with high speed and high sensitivity. We will refer to spectral/Fourier or spectrometer based or swept source based implementations as spectral domain (SD) from now on for simplicity. Traditionally, in order to produce a B-scan, in both implementations, each channeled spectrum acquired while scanning the probing beam over the sample is subject to a fast Fourier transform (FFT). However, before the FFT, several preparatory signal processing steps are necessary, such as zero padding, spectral shaping, apodization, dispersion compensation, or data resampling. Without these preparatory steps, axial resolution and sensitivity suffer.[1] As all these steps can only be sequentially executed, the production of the B-scan images in real-time is limited. So far, several techniques involving both hardware and/or software solutions have been demonstrated to successfully eliminate or diminish the execution time of the preparatory steps. Correct resampling and compensation for dispersion mismatch are extremely important as an incorrect k-mapping or dispersion left unbalanced in the system broadens the coherence peak.[2] To eliminate the resampling step in spectrometer based OCT, a solution using a prism after the diffraction grating was proposed.[3] However, this complicates the optics hardware, requires careful adjustment, and introduces losses. In swept source OCT, the swept sources are often equipped with a supplementary clock signal (k-clock)[4] that not only adds to the cost of the source, but also requires a specialized, more sophisticated digitizer that in the end does not yield a perfect resampling and also limits the axial range of the image.[5] Other techniques such as using an additional light source that produces several spectral lines in the region of interest of the spectrometer,[6] parametric iteration methods,[6] phase linearization techniques,[7,8] and automatic calibrations[9] have been proposed. All these methods are normally computationally expensive and limit the real-time operation of OCT systems.

Dispersion compensation is another important limiting factor in obtaining high-resolution SD-OCT images. Both hardware and software (numerical) methods have been implemented to overcome this limitation.[10] Compared to hardware methods, which involve physically matching the dispersion of the reference and the sample arms, the numerical dispersion compensation is more cost-effective and flexible. Various numerical techniques to compensate for dispersion, such as autofocusing,[10,11] have been developed. However, all the numerical algorithms used to compensate for dispersion involve Hilbert transformations, phase corrections, filtering, etc.; hence, a heavy computational loading and, therefore, numerical dispersion compensation has to be performed as a postprocessing step.

As the computational requirements for high-speed SD-OCT image processing usually exceed the capabilities of most computers, the display rates of B-scan images rarely match the acquisition rates for most devices. After the preparatory steps, most image analysis and diagnosis become a postprocessing operation. A true real-time (TRT) display of processed SD-OCT images when a B-scan is produced in the next frame

*Address all correspondence to: Adrian Bradu, E-mail: a.bradu@kent.ac.uk

time could vastly benefit applications that require instant feedback of image information, such as image guided surgery or endoscopy,[12] or ophthalmology, where it could aid with patient alignment and reduce session time.[13] TRT display is also a prerequisite for three-dimensional (3-D) real-time imaging.[14]

Solutions involving parallel computing hardware enable dramatic increases in performance, taking advantage of devices such as graphics processing units (GPUs) or field-programmable gate arrays (FPGAs). By significantly reducing the computation time of preparatory steps, true real-time production of B-scan images becomes possible. However, as some of the preliminary steps are often based on iterative calculations, errors in correctly preparing the data are unavoidable.

In MS-OCT, where the FFT is replaced by cross-correlation, some of the preparatory steps (resampling and compensation for dispersion mismatch between the two arms of the interferometer) are no longer necessary.[15–17] Thus, the MS-OCT, being based only on mathematical operations, is able, in principle, to produce better images in terms of their axial resolution and sensitivity than their FFT-based counterparts. Moreover, as MS-OCT involves mathematical operations that can be performed in parallel, it makes sense to take advantage of tools already harnessed by the OCT community to produce images fast, such as GPU (Refs. 18–26) or FPGA (Refs. 27–30) hardware. In terms of computation speed, GPU is an efficient computational engine for very large data sets. By contrast, the FPGA can be as efficient as the GPU for large data sets, but FPGAs are better suited for applications where the size of the data to be processed is small.[31] However, on FPGAs, memory is a scarce resource as the block RAM capacity of the FPGA is generally limited. In addition, when performing cross-correlations, which are required by the MS method, data from both the current channeled spectrum and the mask must be available simultaneously as both have to be loaded from the memory in the same clock cycle, which makes the FPGA solution highly dependent on fast memory operations. This favors GPUs as the technology of choice, instead of the FPGAs, to produce TRT MS based B-scan images. Apart from the superiority of the GPU in terms of memory availability, the developers should also consider in their choice that in the foreseeable future, the multicore architecture of the CPU and the many-core architecture of the GPU are likely to merge. This trend can be seen in the introduction of computing parallel solutions used as coprocessors, such as Intel Xeon Phi and NVIDIA Tesla. Visual Studio (Windows) and GCC (Linux) both offer programming environment and language support for these architectures. A GPU combines both fine-grained (threads) and coarse-grained (blocks) parallel architectures. This feature offers an optimal mapping for the specific structure of the signal generated by the OCT systems, where data points are processed by parallel threads and spectra are processed by thread blocks. In addition, the GPU is financially and technologically a more accessible solution. Finally, there are some other immediate benefits of using GPUs over FPGAs in OCT, such as no need for extra hardware (standard PC components are sufficient), availability of a free programming environment, language support, and additional libraries (NVIDIA CUDA C, CUDA FFT).

2 Methodology of Producing MS Based B-Scans

In SD-OCT, a cross sectional image is produced by assembling A-scans (longitudinal reflectivity profiles) into a B-scan, with each A-scan being the result of a single FFT operation. For simplicity, let us consider that no preparatory steps are required before FFT. In this case, mathematically, we can describe a B-scan image as

$$B_{\text{FD-OCT}} = \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_r \end{bmatrix} = \begin{bmatrix} |\text{FFT}(\text{CS}_1)| \\ |\text{FFT}(\text{CS}_2)| \\ \vdots \\ |\text{FFT}(\text{CS}_r)| \end{bmatrix}. \tag{1}$$

In Eq. (1), $\text{CS}_1, \text{CS}_2, \dots \text{CS}_r$ are channeled spectra collected at r lateral positions (pixels along a line perpendicular to the axis connecting the system with the object) of the scanning beam on the sample, while $A_1, A_2, \dots A_r$ are their corresponding A-scans. All the q components of a particular i A-scan $(A_{i,1}, A_{i,2}, \dots A_{i,q})$ from lateral pixel i are obtained simultaneously by a single FFT operation [q is, in principle, equal to half the number of sampling points of each channeled spectrum $\text{CS}_i (i = 1 \dots r)$, but very often, in practice, q is increased by zero padding]. To obtain a B-scan image of size $r \times q$, a number of r independent FFT operations need to be performed. As these operations can be executed in parallel, theoretically, the time to produce a B-scan image can be reduced to the time to perform a single FFT operation. In this case, the maximum time to deliver a B-scan, $T_\text{B}(\text{FFT})$, is made of the time to acquire all r channeled spectra CS_i plus the processing time, which, using parallel processing, can be reduced to the time for a single FFT.

$$t_\text{B}(\text{FFT}) = t_\text{acq} + t_\text{FFT}. \tag{2}$$

To keep the comparison of different regimes of operation simple, let us consider that the acquisition and processing are not interleaved.

As explained in Ref. 17, in MS-OCT, it is possible to generate B-scan images in two ways: by assembling A-scans as in conventional SD-OCT and also by assembling T-scans (transversal reflectivity profiles). Each component of the A-scan (or T-scan) is obtained by integrating the result of the cross-correlation between a channeled spectrum CS_i and a reference channeled spectrum (denoted from now on as mask) $M_j (j = 1 \dots p)$, recorded for an optical path difference (OPD_j) between the sample and reference arms of the interferometer when a highly reflective mirror is used as an object.

$$A_{ij} = \sum_{k=-(w-1)}^{k=+(w-1)} |\text{CS}_i \otimes M_j|_k. \tag{3}$$

The integration is done over a window of size $W = 2w - 1$ points around the maximum value of the correlation result. The masks should be recorded at OPD values in the set $\text{OPD}_1, \text{OPD}_2, \dots \text{OPD}_p$, separated by half the coherence length of the source or denser. To obtain a B-scan image of the same size as that obtainable in SD-OCT, a number of $p = q$ masks need to be recorded. In a matrix representation, an MS-based image can be presented as

$$B_{\text{MS-OCT}} = \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_r \end{bmatrix} = \begin{bmatrix} M_1 \otimes \mathrm{CS}_1 & M_2 \otimes \mathrm{CS}_1 & \cdots & M_p \otimes \mathrm{CS}_1 \\ M_1 \otimes \mathrm{CS}_2 & M_2 \otimes \mathrm{CS}_2 & \cdots & M_p \otimes \mathrm{CS}_2 \\ \vdots & \vdots & \ddots & \vdots \\ M_1 \otimes \mathrm{CS}_r & M_2 \otimes \mathrm{CS}_r & \cdots & M_p \otimes \mathrm{CS}_r \end{bmatrix}. \tag{4}$$

Each column of the bidimensional array in Eq. (4) represents a T-scan, while each row contains the information to assemble an A-scan. As all the elements of this matrix can be computed independently, the time to produce, for example, an A-scan based B-scan can, in principle, be reduced to the time to perform a single cross-correlation.

$$A_{ij} = i\mathrm{FFT}[\mathrm{FFT}(\mathrm{CS}_i) \cdot \mathrm{FFT}(M_j)^*]. \tag{5}$$

In the equation above, i is the lateral pixel index, $j = 1\ldots p$ signifies the depth, iFFT denotes the inverse fast Fourier transform, while the * operator is the complex conjugate. As suggested in a previous report,[17] to speed up the process, instead of recording the values of the masks M_j, the $\mathrm{FFT}(M_j)^*$ values are precomputed and recorded instead. This reduces the time to work out the reflectivity A_{ij} of a point in the A-scan to the time to sequentially execute two FFTs. To obtain a B-scan image of size $r \times p$, a number of $r \times p$ independent cross-correlations have to be performed. As these operations can be executed in parallel, the time to produce an MS-based B-scan image, $T_{\mathrm{B}}(\mathrm{MS})$, can, in principle, be reduced to the time to perform a single cross-correlation.

$$T_{\mathrm{B}}(\mathrm{MS}) = t_{\mathrm{acq}} + 2t_{\mathrm{FFT}}. \tag{6}$$

For a spectrum acquisition in 10 μs, considering $r = 200$ lateral pixels, $t_{\mathrm{acq}} = 2$ ms. The time for a single FFT in the SD-OCT and for two FFT processes in the MS-OCT case can be absorbed into the acquisition time, as they are below a couple of microseconds. This means that $T_{\mathrm{B}}(\mathrm{MS}) \approx T_{\mathrm{B}}(\mathrm{FFT})$.

As a consequence of the above analysis, it appears that it is possible to produce MS-based B-scan images at a rate similar to that of producing FFT-based B-scans.

In case data in the SD-OCT need to be prepared before FFT, then the processing time exceeds the time for a single FFT operation. Depending on the processing involved, which also depends on the nonlinearities to be corrected, the evaluation of a single A-scan via linearization/calibration followed by an FFT may exceed the time for p correlations (considering implementation in parallel in the time for a single correlation in our discussion).

As these extra processes are not needed in MS-OCT, MS based B-scan images can then be produced at similar rates with, or even faster than, the FFT based method.

In all this discussion and in Eq. (6), we did not consider the time to transfer data between hardware components or between memories, therefore, in practice, $T_{\mathrm{B}}(\mathrm{MS})$ includes more terms than shown in Eq. (6). This means that MS-OCT cannot be faster than SD-OCT in producing a B-scan OCT image when no linearization/calibration is performed.

However, B-scans can still be produced in a TRT regime using the MS-OCT method.

3 GPU Implementation and Experimental Setup

The MS principle can be applied to any SD-OCT technology. This is illustrated here on a conventional swept source OCT setup. The experimental setup employed for this paper is identical to that used in Ref. 15. Figure 1 illustrates the flow of data and the structure of trigger/clock signals required to produce a B-scan image.

A LabVIEW project is created to perform most of the tasks in the system. The choice for a LabVIEW implementation is dictated by its versatility, covering acquisition, processing, and visualization of data. LabVIEW is also highly popular among OCT research groups.[19,28,30] It facilitates easy interaction with the hardware via data acquisition (DAQ) boards and acquisition of data via digitizers or image acquisition (IMAQ) boards. This platform offers an excellent user interface and, in terms of development time, is superior to any C/C++ implementation. Our LabVIEW application generates a triangular waveform via the DAQ (NI PCI 6110) to drive the galvo-scanner (SX), acquires and stores the data into a buffer via a fast digitizer (Alazartech, Quebec, Canada, model ATS9350) according to timings set by the DAQ and the trigger signal provided by the swept source (SS, Axsun Technologies, Billerica, Massachusetts) and, finally, communicates with a GPU application via a dynamic link library (DLL). The GPU application performs the signal processing and generates the resulting image using OpenGL drawing primitives. The DLL and the GPU application exchange the buffered data as well as a number of control values required to manipulate the contrast and brightness of the image or the value of the window W over which the correlation signal is integrated. The use of a standalone application that facilitates the LabVIEW project to communicate with CUDA code is not unique. The DLL could

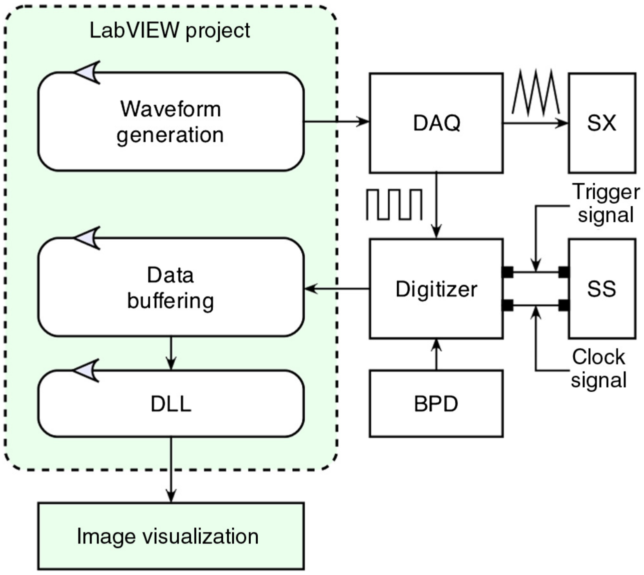

Fig. 1 Overall structure of the graphics processing unit (GPU) enhanced optical coherence tomography (OCT) system illustrating the operation required for the production of a [master-slave (MS) or fast Fourier transform (FFT) based] B-scan image. DAQ, data acquistion board; SX, galvo-scanner; SS, swept source; BPD, balanced photodetector; DLL, dynamic link library.

encapsulate the CUDA code (CUDA kernels). However, in this case, there are two main issues: (1) the GPU context needs to be created every time the DLL is called, which significantly increases the processing time and prevents the system from working in real time and (2) the DLL cannot load OpenGL to display the processed images.[32]

For our experiments, a commonly available NVIDIA GeForce GTX 780 Ti (3 GiB on-board memory, 2880 CUDA cores) is used, installed in a Dell Precision 7500 equipped with dual CPUs (E5503, 4 cores @ 2.0 GHz).

To make a fair comparison between the MS and FFT methods, the internal clock signal equipping the swept source was used to sample the data, though it is not necessary for the MS-OCT method. According to the laser manufacturer, the maximum frequency of the clock signal is $\sim$330 MHz, and the duty cycle of the laser emission is 45%, so for a sweeping rate of 100 kHz, a maximum of 1531 sampling points can be used to digitize each channeled spectrum. In practice, as the digitizer requires a number of data points evenly divisible by 32, 1472 sampling points are employed to digitize the entire spectral range of each channeled spectrum, which can be used to sample up to 736 cycles.

A number of $r = 500$ A-scans are used to build each cross-sectional image; hence, a maximum image size of 500×736 pixels^2 could be produced at a frequency of $F_x = 100$ Hz, while the galvo-scanner is driven with a triangular waveform at 500 Hz. In this way, the data acquisition time of a single frame (B-scan) is 5 ms. The time to buffer the necessary data to produce a B-scan image is depicted as process A in Fig. 2.

To ensure an acceptable real-time operation of the system, the data processing time required to produce the final B-scan image has to be <10 ms. In a conventional non-GPU enhanced application, after the acquisition of each buffer, data are directly accessed and processed by the CPU. The processing of data on the GPU is obviously faster than on the CPU; however, as the acquired data are not directly accessible by the GPU, the time to transfer data to the CPU memory (shared memory) accessible by the GPU has to be taken into account (process B in Fig. 2). Once in the shared memory, data can be processed either traditionally via FFTs (process C in Fig. 2) or using the MS method (process D). After processing, data are available to OpenGL for image generation and visualization (process E or F in Fig. 2 for displaying FFT or MS based B-scans, respectively).

To ensure a TRT operation of the system, the time to transfer data to the shared memory has to be less than the acquisition time (5 ms in our case), while the time allocated to processing or visualization has to be less than the period of the galvo-scanner (10 ms).

In Fig. 3, the entire chain of processes and operations required to display an image is presented. This chain starts from the moment data are available in the shared memory to be accessed by the GPU application for both FFT (A) and MS-OCT (B) methods. The simplest FFT based case is considered, where

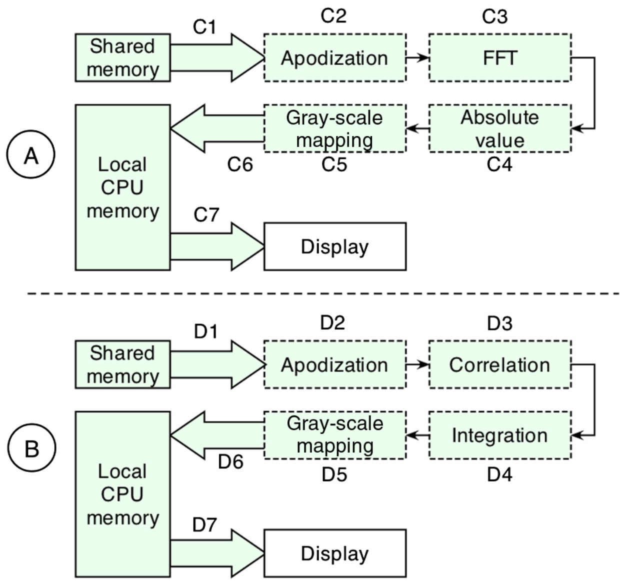

Fig. 3 GPU threads execution in FFT-based OCT (A) and MS-OCT (B). C1 and D1 signify the process of data transfer from the shared memory to the GPU memory, C6 and D6 represent the transfer of data from the GPU memory back to the local CPU memory, and C7 and D7 represent data transfer to OpenGL for image generation and visualization.

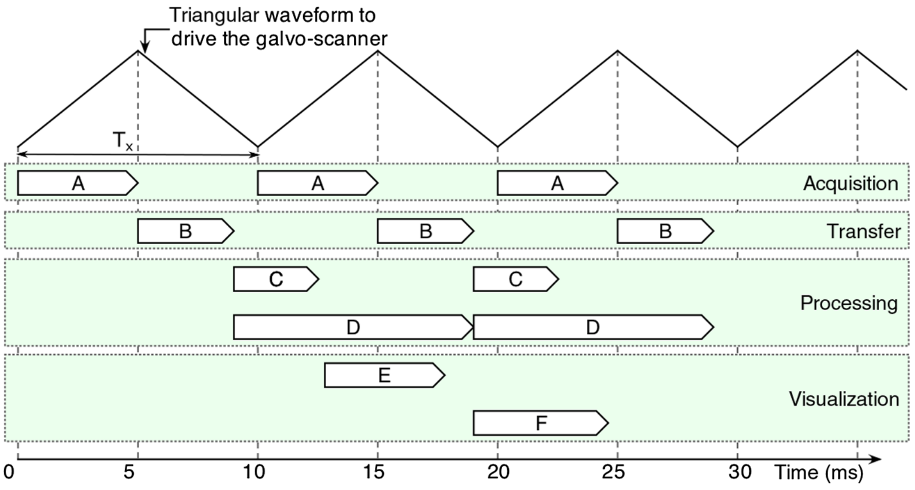

Fig. 2 Flow chart showing the timing of data acquisition (A), transfer (B), processing (C, D), and visualization (E, F) in a GPU enhanced OCT system. C and E refer to the spectral domain OCT method, and D and F to the MS-OCT method. The time intervals shown on the figure are compatible with true real-time operation of the system, practically obtained for $q = 1024$ and $p = 144$.

there is no need for dispersion compensation and no need for resampling of data. Processes C1 and D1 designate the transfer of data from the shared memory to the GPU memory. As the same amount of data is transferred in both C1 and D1 cases, it is expected that they take the same amount of time: $T_{\mathrm{C1}} = T_{\mathrm{D1}}$. Processes C6 and D6 designate the transfer of data from the GPU memory back to the local CPU memory (of the GPU application). In this case, as typically $p \neq r$, it is expected that $T_{\mathrm{C6}} = T_{\mathrm{D6}}$. Finally, C7 and D7 refer to the transfer of data to the OpenGL primitives for image generation and display.

4 Results

In order to establish the conditions to produce TRT based B-scan images, the time required for process B shown in Fig. 2, as well as the time required for the components of Fig. 3 are measured. Once the best settings that ensure a (true) real-time operation are established, *in vivo* images of the human eye are produced.

4.1 *Benchmarking*

To ensure TRT operation of the system, the following conditions need to be satisfied:

(1) transfer data to the shared memory within half the period of the triangular waveform applied to the galvo-scanners (T_x), i.e., 5 ms;

(2) perform data processing (including the time required to transfer data to and from the GPU) within T_x;

(3) generate and display an image within T_x.

If any of these three conditions are not fulfilled, the system may still work in real time, but some acquired frames may be skipped (data may not have been transferred to the shared memory, not processed, or not displayed). Let us look at these three conditions in detail.

(1) When using GPUs for parallel computing, attention needs to be paid not only to the size of data to be transferred, but to the way the memory is allocated as well. As the aim is to build a B-scan from $r = 500$ sweeps, the amount of data buffered by the LabVIEW project and transferred to the shared memory when 1472 sampling points are used is 2.944 MiB (we allocated 4 bytes for each sampling point). Our hardware and software managed this in ∼4.5 ms, satisfying the first condition.

(2) In terms of allocating memory on the GPU, the two cases (FFT and MS) have to be considered separately. When producing an FFT based image, only 8.42 MiB are needed to be allocated, which is much less than the 3 GiB available on the GPU. In MS-OCT, if the same image size needs to be produced, with $p = 736$ depth pixels, the memory requirement increases to a little over 3 GiB, which is more than the capacity of the GPU (GeForce GTX 780 Ti).

(3) Lowering the number of pixels in depth to $p = 384$, the GPU memory requirement is reduced to 2.2 GiB, which is still insufficient for an optimum operation, considered when the memory left free exceeds half of the total GPU memory. Not all of the entire 3 GiB is available to the CUDA application (some being used by the operating system and/or windowing system for frame buffers). When reserving such a huge amount of memory, the CUDA software produces images, however, with occasional interruptions. Further decrease of p is not recommended as this would translate into deterioration of the axial resolution (the p points sampling the A-scan in the MS method have to be spread over an axial range of ∼2 mm when imaging the retina, in case at least a separation of 10 microns is targeted; this limits the number of axial pixels to a minimum of $p = 200$). One solution to maintain a sufficiently high number of points in the A-scan with the limitation imposed by the memory is to decrease the number q of sample points used to digitize each channeled spectrum.

Benchmark results showing the time intervals required for different numbers of sampling points q (relevant for the FFT case) and different number of p masks (relevant for the MS case) are presented in Table 1. All numbers q chosen are multiple of 32, as required by the digitizer, within a range from the maximum $q = 1472$ to half of the maximum, 736. For $q = 1472$ and $q = 1280$, with $p = 384$, the remainder of the GPU memory left when using the MS method is not sufficient to produce an image; therefore, the cells corresponding to these values in Table 1 are empty. Consequently, the production of an image size of 500 × 384 is not possible with our hardware configuration for these cases.

By lowering the number of samples to $q = 1024$, the GPU memory required becomes half of that available and an MS-based image of size 500 × 384 pixels can be generated. For this image size, in SD-OCT, the axial range of the B-scan is sampled every 7.2 μm (obtained by dividing the axial range of 3.7 mm, limited by the clock in the swept source to half of the q number). If $p = 384$ masks separated by 7.2 μm are used in MS-OCT, a more than sufficient axial range of 2.77 mm is covered. However, the time to produce an image (process D in Fig. 2) is $T_{\mathrm{D}} = 21.9$ ms, longer than T_x, so TRT operation is not possible. Only by reducing the number of axial pixels to $p = 144$, $\mathrm{T_{DZ}}$ is reduced below 10.0 ms, in which case TRT MS B-scan images can be produced, with size 500 × 144 pixels^2 and covering an axial depth of 1.04 mm.

As the CuFFT library used by the GPU application is highly optimized to work with power of 2 signal lengths, a condition satisfied by $q = 1024 = 2^{10}$, decreasing q from this value to 736 does not make the process of producing B-scan images faster, as shown by the numerical values in Table 1.

Not sampling the entire spectral range has the consequence of a reduction in depth resolution. In Table 1, we also show experimentally measured values for depth resolution. The depth resolution obtained using the MS method is slightly larger than that achieved with the FFT technique as a quite large integration window ($W = 10$) was used to average each cross-correlated signal to obtain better sensitivity.[6,8] In both cases, when $q = 1024$ sampling points are used instead of 1472, a deterioration of the depth resolution of ∼1.4 to 1.5 μm is observed.

Irrespective of the number of sampling points, q, for a maximum image size of 500 × 768 pixels, we found out that the time required by the OpenGL to display an image is ∼5.1 ms

Table 1 Benchmarks showing the time required to produce a B-scan image of $r = 500$ lateral pixels using the two methods, by each of the multiple processes involved, such as the time to transfer data from the CPU to the shared memory (TB), from the shared memory to graphics processing unit (GPU) (T_{C1} and T_{D1}), and from GPU to the local CPU memory (T_{C6} and T_{D6}). Timings for generating images and display (T_{C7} and T_{D7}) and the throughput of data for each case are presented as well. Results are shown for different numbers of sampling points. The depth resolution experimentally measured and the amount of memory allocated for each case using the fast Fourier transform (FFT) and master-slave (MS) techniques are also shown.

Sampling points (q):	1472	1280	1024	736
T_B (ms)	4.5	4.3	4.0	3.6
$T_{C1} = T_{D1}$ (ms)	2.6	2.2	1.7	1.3
T_{D6} ($p = 384$) (ms)	—	—	1.7	1.6
T_{D6} ($p = 144$) (ms)	1.7	1.6	1.5	1.2
T_{C6} (ms)	1.2	1.2	1.2	0.8
$T_{C7} = T_{D7}$ (ms) ($p = 768$)	5.1	5.1	5.1	5.1
Total time to produce an FFT-based B-scan image (ms)	3.8	3.5	3.4	2.3
Throughput of FFT-based OCT data (Mpixels/s)	96.84	91.43	75.30	83.47
Total time to produce an MS-based B-scan image ($p = 384$) (ms)	—	—	21.9	26.8
Throughput of MS-based OCT data ($p = 384$) (Mpixels/s)	—	—	8.76	7.16
Total time to produce an MS-based B-scan image ($p = 144$) (ms)	21.1	14.6	10.0	11.3
Throughput of MS-based OCT data ($p = 384$) (Mpixels/s)	3.41	4.93	7.20	6.37
Axial resolution δz (FFT) (μm)	8.7	9.1	10.1	14.3
Axial resolution δz (MS) (μm)	8.9	9.3	10.4	14.8
Memory allocation spectral domain optical coherence tomography (MiB)	8.42	7.32	5.86	4.21
Memory allocation MS (GiB)	2.17	1.89	1.50	1.09

remaining approximately constant when q is decreased to 384, so the third condition for a TRT regime is fulfilled in any of the situations mentioned above.

For the hardware employed, there are no differences in terms of calling the DLL from either within the LabVIEW project or from a C/C++ based project. By using a C/C++ software implementation, the process B in Fig. 2 is eliminated; hence, $T_B = 0$ ms. As a result, processes B and C shown in Fig. 2 start sooner by ~4 ms according to the values in the second row in Table 1. The memory size when using a large number of sampling points is still a limitation when using C/C++; the only important advantage of such an implementation is that the delay between the end of the acquisition and the display of the B-scan image can be reduced by T_B.

To summarize, for our hardware configuration, TRT operation is secured when the size of B-scan images to be produced is $<500 \times 144$ pixels^2 and the number of sampling points q is 1024. The obvious drawbacks are a decreased axial range, which may be tolerated depending on applications, and a small reduction in the axial resolution.

4.2 *Fast B-Scan Imaging*

Ethical approval to image researchers' retinas was obtained from the Faculty of Sciences at University of Kent ethics committee, application "Tests of optical instruments for imaging the eye, assembled by the Applied Optics Group (AOG), on AOG researchers." The images were collected from the eyes of two of the authors, Adrian Bradu and Adrian Podoleanu. Both co-authors have signed consent forms. In this manuscript, only images from AB's eye are presented.

The capability of the system to produce MS based B-scan images sufficiently fast, of comparable resolution and sensitivity as their FFT based counterpart, is demonstrated by two movies (Videos 1 and 2) showing B-scans side by side using both methods. Figures 4 and 5 show *in vivo* B-scan images of the human fovea and optic nerve area, respectively, obtained using the two methods (frames from the generated videos). To produce both images, $q = 1024$ sampling points for each channeled spectrum are used, resulting in 7.2 μm depth between consecutive points in the FFT based A-scan. As a consequence, $p = 384$ masks separated by 7.2 μm are recorded in order to obtain a nearly perfect axial correspondence between the horizontal lines of the two images. As illustrated in Table 1, in this case, TRT is not achieved (the time to produce an MS based B-scan image is 21.9 ms, longer than T_x). Therefore, although both MS and FFT based B-scans are produced simultaneously, and the images in the movies are displayed at a rate close to that demanded to display the MS based images only.

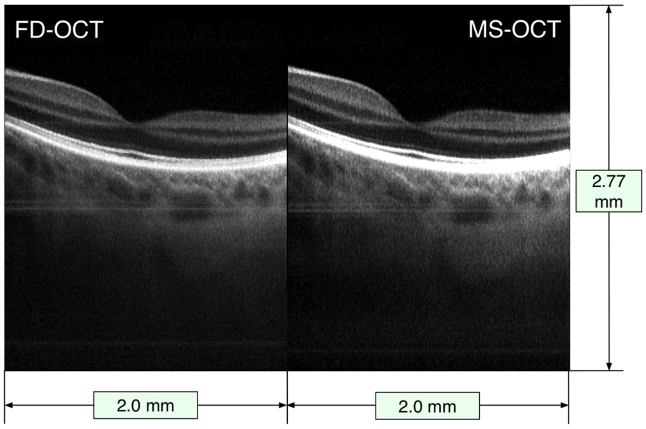

Fig. 4 B-scan images of size 500(lateral) × 384(axial) pixels^2 of the optic nerve area of AB showing lamina cribrosa obtained using the conventional Fourier domain technique (left) and the MS method (right). The axial size of images is evaluated in air. The images are extracted from a movie (Video 1) produced in real time (~50 fps) (MOV, 14.6 MB) [URI: http://dx.doi.org/10.1117/1.JBO.20.7.076008.1].

The fact that the two images are not perfectly axially aligned can be seen better in the two movies (Videos 1 and 2) from which the images are extracted. There, it is easily observable that the horizontal lines (artifacts the movies are not corrected for) originate from slightly different depths in the two images. The FFT based image is cropped to cover a 384 points subset from each A-scan. To record the two movies, no fixation lamp was used. Also, no (post)processing step was applied to the images to correct for movements, and no image registration procedure was used. The only processing operation applied to the images was a manual adjustment of the contrast and brightness, which explains the small difference between the gray levels in the movies shown side by side. The movies are displayed in real time at ~50 fps. To produce the images in Figs. 4 and 5, the frames of the movies are corrected for movement before averaging 25 consecutive B-scans. It is obvious that the images produced by MS are similar in terms of contrast and resolution with those produced by conventional SD-OCT. To produce the movies, the optical power at the cornea was limited to 2.2 mW, below the ANSI (Ref. 33) maximum allowed level for the 1060-nm region.

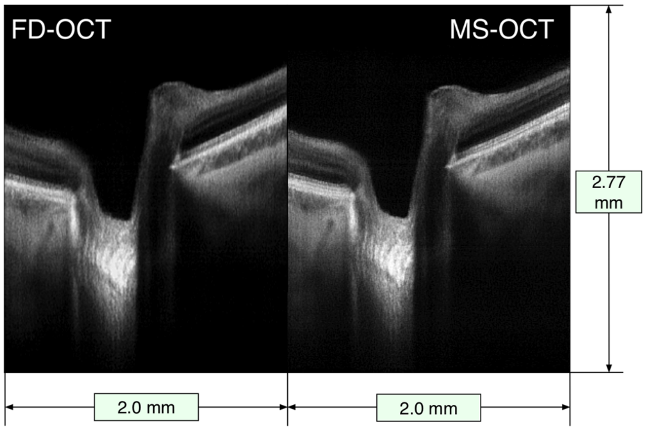

Fig. 5 B-scan images of size 500(lateral) × 384(axial) pixels^2 of the foveal area of AB obtained using the conventional FFT-based technique (left) and the MS method (right). The lateral size of images is evaluated in air. The images are extracted from a movie (Video 2) produced in real time (~50 fps) (MOV, 14.4 MB) [URI: http://dx.doi.org/10.1117/1.JBO.20.7.076008.2].

5 Conclusion

Fast display of MS based B-scan OCT imaging is demonstrated by harnessing the GPU capabilities to perform the entire signal processing load that the MS method requires. While the MS method exhibits sensitivity and axial resolution similar to that provided by the FFT based technique, for the particular GPU used in this work, the production of the B-scan retinal images in a TRT regime is not possible due to the limited memory of the GPU. TRT imaging regime is achievable, however, for a reduced size of 500 × 144 pixels^2 for the B-scan images. The axial size of 144 pixels is not sufficient when imaging the retina. A GPU with a larger memory is necessary to increase the number of depths p processed in parallel. However, an ~50 Hz frame rate was achieved for the display of B-scans with 384 depths, showing that by harnessing the GPU cards, the bottleneck of the MS method in producing B-scans[17] can be successfully addressed.

When the GPU application is neither computationally nor memory bounded, the TRT operation is secured. The GPU method proposed here is computationally and memory bounded. As specified in Table 1, for the MS case, a large number of sampling points such as $q = 1472$ or $q = 1280$ with $p = 384$ masks require a memory size larger than that available. As soon as q exceeds 1024, the GPU method becomes computationally bounded and a TRT operation cannot be ensured.

As demonstrated in previous studies,[15–17] the MS method is highly suitable for 3-D images. Each point A_{ij} in the 3-D image is the result of a cross-correlation operation. Preferably, 3-D via the MS method would assemble p *en face* images into a volume. As this report is about B-scans, let us consider a 3-D assembly of B-scans into a 3-D volume. As the swept source employed in the study sweeps at 100 kHz, considering a number of lateral pixels as in previous reports,[15,16] of $r = 200$, a 3-D dataset of size 200 × 200 × 384 pixels^3 can be acquired in 0.4 s, while $p = 384$ masks determine the number of axial points in the volume. Due to the limitations on the shape and frequency of the signal that can be sent to the galvo-scanners, the time to acquire data is $t_{\mathrm{acq}} = 0.8$ s. As a consequence, to ensure a 3-D TRT operation of the system, once data are acquired, the 200 × 200 × 384 pixels^3 operations have to be produced within 0.8 s. For our hardware implementation, this is feasible, as the time to transfer data to the shared memory is evaluated as <0.45 s. However, especially due to the limited amount of memory available on the GPU, the TRT production of the full size 3-D images is not possible yet.

Operating the setup in the MS-OCT regime cannot deliver a B-scan as fast as its SD-OCT counterpart when no linearization is required, so the advantage of using the MS-OCT may not be apparent. However, if resampling of data is needed prior to FFT processing, or numerical calculations to eliminate the dispersion mismatch between the arms of the interferometer are required, then the balance swings in favor of the MS-OCT method. FFT-based B-scan images obtained with the clock disabled are not shown here. In such a case, they would display axial resolutions $>50\ \mu$m with a barely distinguishable definition at larger depths. The B-scan OCT images obtained using the MS-OCT look exactly the same as those generated with the conventional FFT-based method applied to linearized data. If, let us say, a swept source with no clock is employed, then a software linearization calibration procedure needs to be implemented. Such a procedure would extend the time for the FFT-based method by at least 50% while the correction may not be perfect at all depths. The images are not shown, as they have been already reported by other groups. The imperfections in linearization/calibration may swing the balance again in favor of the MS-OCT via GPU. Better GPUs with more memory that would allow production of larger image sizes are already available. Also, computers with faster interconnects can reduce the time to transfer data between different memories, improving both latency (data available to the GPU application sooner) and visualization throughput (at higher frame rates).

The time to produce an MS-based B-scan demonstrated here when considered in combination with other advantages in terms of hardware cost makes the MS-OCT method worth considering for imaging the eye. As there is no need for data resampling, MS-OCT can operate with a simpler swept source, not equipped with a k-clock, or even with potentially highly nonlinear tuneable lasers. An MS-OCT system can operate in terms of its axial resolution and sensitivity decay with depth at the level of a perfectly corrected SD-OCT setup. Therefore, MS-OCT has the potential to provide better sensitivity, resolution, and better axial range than practical systems implementing conventional FFT technology, which depart from perfectly corrected systems in terms of linearization and dispersion compensation.

With a speed in displaying B-scans comparable to those typically reported by SD-OCT systems, as detailed here, MS-OCT can become the technique of choice for TRT imaging in ophthalmology and other applications that require instant feedback, such as in surgery.

Acknowledgments

K. Kapinchev and F. Barnes acknowledge the support of the School of Computing at the University of Kent. K. Kapinchev also acknowledges the support of EPSRC DTA funding. A. Bradu and A. Gh. Podoleanu acknowledge the support of the European Research Council (ERC) (http://erc.europa.eu) COGATIMABIO 249889. A. Podoleanu is also supported by the NIHR Biomedical Research Centre at Moorfields Eye Hospital NHS Foundation Trust and the UCL Institute of Ophthalmology.

References

1. R. Leitgeb et al., "Ultrahigh resolution Fourier domain optical coherence tomography," *Opt. Express* **12**(10), 2156–2165 (2004).
2. M. Wojtkowski et al., "In vivo human retinal imaging by Fourier domain optical coherence tomography," *J. Biomed. Opt.* **7**(3), 457–463 (2002).
3. Z. Hu and A. M. Rollins, "Fourier domain optical coherence tomography with a linear-in-wavenumber spectrometer," *Opt. Lett.* **32**(24), 3525–3527 (2007).
4. B. Potsaid et al., "Ultrahigh speed 1050 nm swept source / Fourier domain OCT retinal and anterior segment imaging at 100, 000 to 400, 000 axial scans per second," *Opt. Express* **18**(19), 20029–20048 (2010).
5. B. Liu, E. Azimi, and M. E. Brezinski, "True logarithmic amplification of frequency clock in SS-OCT for calibration," *Biomed. Opt. Express* **2**(6), 1769–1777 (2011).
6. B. Park et al., "Real-time fiber-based multi-functional spectral-domain optical coherence tomography at 1.3 μm," *Opt. Express* **13**(11), 3931–3944 (2005).
7. X. Liu et al., "Towards automatic calibration of Fourier-Domain OCT for robot-assisted vitreoretinal surgery," *Opt. Express* **18**(23), 24331–24343 (2010).
8. Y. Yasuno et al., "Three-dimensional and high-speed swept-source optical coherence tomography for in vivo investigation of human anterior eye segments," *Opt. Express* **13**(26), 10652–10664 (2005).
9. M. Mujat et al., "Autocalibration of spectral-domain optical coherence tomography spectrometers for in vivo quantitative retinal nerve fiber layer birefringence determination," *J. Biomed. Opt.* **12**(4), 041205 (2007).
10. M. Wojtkowski et al., "Ultrahigh-resolution, high-speed, Fourier domain optical coherence tomography and methods for dispersion compensation," *Opt. Express* **12**(11), 2404–2422 (2004).
11. D. L. Marks et al., "Autofocus algorithm for dispersion correction in optical coherence tomography," *Appl. Opt.* **42**(16), 3038–3046 (2003).
12. W. Kuo et al., "Real-time three-dimensional optical coherence tomography image-guided core-needle biopsy system," *Biomed. Opt. Express* **3**(6), 1149–1161 (2012).
13. T. Klein et al., "Multi-MHz retinal OCT," *Biomed. Opt. Express* **4**(10), 1890–1908 (2013).
14. Y. Huang, X. Liu, and J. U. Kang, "Real-time 3D and 4D Fourier domain Doppler optical coherence tomography based on dual graphics processing units," *Biomed. Opt. Express* **3**(9), 2162–2174 (2012).
15. Gh. Podoleanu and A. Bradu, "Master-slave interferometry for parallel spectral domain interferometry sensing and versatile 3D optical coherence tomography," *Opt. Express* **21**(16), 19324–19338 (2013).
16. A. Bradu and Gh. Podoleanu, "Imaging the eye fundus with real-time en-face spectral domain optical coherence tomography," *Biomed. Opt. Express* **5**(4), 1233–1249 (2014).
17. A. Bradu and Gh. Podoleanu, "Calibration-free B-scan images produced by master/slave optical coherence tomography," *Opt. Lett.* **39**(3), 450–453 (2014).
18. K. Zhang and J. U. Kang, "Real-time 4D signal processing and visualization using graphics processing unit on a regular nonlinear-k Fourier-domain OCT system," *Opt. Express* **18**(11), 11772–11784 (2010).
19. S. Van der Jeught, A. Bradu, and Gh. Podoleanu, "Real-time resampling in Fourier domain optical coherence tomography using a graphics processing unit," *J. Biomed. Opt.* **15**(3), 030511 (2010).
20. Y. Jian, K. Wong, and M. V. Sarunic, "Graphics processing unit accelerated optical coherence tomography processing at megahertz axial scan rate and high resolution video rate volumetric rendering," *J. Biomed. Opt.* **18**(2), 026002 (2013).
21. J. Probst et al., "Optical coherence tomography with online visualization of more than seven rendered volumes per second," *J. Biomed. Opt.* **15**(2), 026014 (2010).
22. K. Zhang and J. U. Kang, "Graphics processing unit accelerated nonuniform fast Fourier transform for ultrahigh-speed, real-time Fourier-domain OCT," *Opt. Express* **18**(22), 23472–23487 (2010).
23. K. Zhang and J. U. Kang, "Real-time intraoperative 4D full-range FD-OCT based on the dual graphics processing units architecture for microsurgery guidance," *Biomed. Opt. Express* **2**(4), 764–770 (2011).
24. J. U. Kang et al., "Realtime three-dimensional Fourier-domain optical coherence tomography video image guided microsurgeries," *J. Biomed. Opt.* **17**(8), 081403 (2012).
25. M. Sylwestrzak et al., "Four-dimensional structural and Doppler optical coherence tomography imaging on graphics processing units," *J. Biomed. Opt.* **17**(10), 100502 (2012).
26. Y. Huang, X. Liu, and J. U. Kang, "Real-time 3D and 4D Fourier domain Doppler optical coherence tomography based on dual graphics processing units," *Biomed. Opt. Express* **3**(9), 2162–2174 (2012).
27. T. E. Ustun et al., "Real-time processing for Fourier domain optical coherence tomography using a field programmable gate array," *Rev. Sci. Instrum.* **79**(11), 114301 (2008).
28. E. Desjardins et al., "Real-time FPGA processing for high-speed optical frequency domain imaging," *IEEE Trans. Med. Imaging* **28**(9), 1468–1472 (2009).
29. V. Bandi et al., "FPGA-based real-time swept-source OCT systems for B-scan live-streaming or volumetric imaging," *Proc. SPIE* **8571**, 85712Z (2013).
30. D. Choi et al., "Spectral domain optical coherence tomography of multi-MHz A-scan rates at 1310 nm range and real-time 4D-display up to 41 volumes/second," *Biomed. Opt. Express* **3**(12), 3067–3086 (2012).
31. Altera Corporation, "Radar processing: FPGAs or GPUs?," 2013, https://www.altera.com/content/dam/altera-www/global/en_US/pdfs/literature/wp/wp-01197-radar-fpga-or-gpu.pdf (23 May 2015)
32. K. Kapinchev et al., "Approaches to general purpose GPU acceleration of digital signal processing in optical coherence tomography systems," in *IEEE Int. Conf. on Systems, Man, and Cybernetics*, pp. 2576–2580 (2013).

33. *American National Standard for Safe Use of Lasers (ANSI Z136.1)*, Laser Institute of America, Orlando, FL 2014, https://www.lia.org/PDF/Z136_1_s.pdf (26 June 2015).

Adrian Bradu received his PhD in physics from the "Joseph Fourier" University, Grenoble, France. He joined the applied optics group at the University of Kent, Canterbury, United Kingdom, in 2003. Currently, he is a research associate on the ERC grant "combined time domain and spectral domain coherence gating for imaging and biosensing." His research is mainly focused on optical coherence tomography, confocal microscopy, adaptive optics, and combining principles of spectral interferometry with principles of time domain interferometry to implement novel configurations up to proof of concept, applicable to biosensing and cell, tissue imaging, or imaging of different organs.

Konstantin Kapinchev received his MSc degree in computer science from the Technical University of Varna, Bulgaria, in 2002. He worked there as an assistant lecturer until 2012, when he started a PhD study at the University of Kent, United Kingdom. His research is focused on parallel computing, process and thread management, and scalability. He works on general purpose GPU optimization in digital signal processing and optical coherence tomography.

Frederick Barnes is a senior lecturer in School of Computing, University of Kent, Canterbury, United Kingdom. His main research interests are centered around the CSP model of parallel processing, encapsulated by the occam-pi multiprocessing language. This is a message-passing based model of concurrency, but he is also interested in the others (e.g., shared variable, including lock-free and wait-free algorithms, join calculus-based abstractions, and so on). Also he is interested in parallel computing across clusters and on GPUs, programming language design and implementation, operating systems, and embedded systems.

Adrian Podoleanu is a professor of biomedical optics in the School of Physical Sciences, University of Kent, Canterbury, United Kingdom. He was awarded a Royal Society Wolfson Research Merit Award in 2015, an ERC Advanced Fellowship 2010–2015, Ambassador's Diploma, Embassy of Romania/United Kingdom 2009, a Leverhulme Research Fellowship 2004–2006, and the Romanian Academy "Constantin Miculescu" prize in 1984. He is a fellow of SPIE, OSA, and the Institute of Physics, United Kingdom.