Developing Real-Time Implementations of Non-Linear Beamformers for Enhanced Optical Ultrasound Imaging

Fraser T. Watt*,†, Paul C. Beard*,†, Erwin J. Alles*,†
*Wellcome / EPSRC Centre for Interventional and Surgical Sciences, University College London, London, UK
†Department of Medical Physics & Biomedical Engineering, University College London, London, UK

***Abstract*—Free-hand optical ultrasound (OpUS) imaging is an emerging ultrasound imaging paradigm that utilises an array of fiber-optic sources and a single fiber-optic detector to achieve video-rate, real-time imaging with a flexible probe that is immune to electromagnetic interference. Due to the use of only a single detector, such probes have limited channel counts, resulting in significant imaging artefacts and limited contrast when imaging is performed with a conventional Delay-and-Sum (DAS) beamformer. Non-linear beamforming can help improve the imaging quality by exploiting cross-channel coherence across the aperture, at the expense of significantly increased computational complexity. In this work, GPU implementations of different non-linear beamformers were implemented and tailored specifically to OpUS array devices and tested on both simulated and experimental data.**

***Index Terms*—Optical ultrasound, Non-linear beamforming, GPU programming, Low Channel Count Reconstruction**

I. Introduction

Optical Ultrasound (OpUS) is an imaging paradigm in which light is used for both generation and detection of ultrasound, as opposed to the use of conventional ultrasound devices. In an OpUS system ultrasound is generated by the selective application of the photoacoustic effect in an optically absorbing material [1], which is typically either deposited at the tip of an optical fiber [2] or formed into a membrane [3]. Back-scattered ultrasound is then detected by an optically resonant structure such as Fabry-Pérot cavities [4], [5] or ring resonators [6], which are mounted on an optical fiber tip. Typically these components are used to form two-fiber devices, with a single fiber for ultrasound generation and another for detection mounted together. These probes function on a sub-millimetre scale, can emit and detect broadband ultrasound (commonly 20-30 MHz around a 10-15 MHz center frequency) and are immune to electromagnetic interference [4]. To form an imaging aperture, two-fiber probes require some form of mechanical translation [7] and as a result typically require long acquisition times.

Recently a hand-held OpUS imaging probe was presented that performed real-time, video rate imaging, using an array of 64 fiber-optic sources and a single fiber-optic Fabry-Pérot detector [8]. This probe achieved frame rates of up to 11 Hz, however the low channel count combined with the applied Delay-and-Sum (DaS) reconstruction algorithm resulted in significant artefacts and thus limited image contrast. The DaS algorithm is the standard beamformer for biomedical ultrasound imaging and OpUS devices, and operates under the assumption that actual pulse-echo signal originating from the location corresponding to an image pixel will sum coherently across the imaging aperture, while other signal components (such as noise, interference and out-of-plane artefacts) are zero-mean and hence will average out. This method is effective for systems with a large number of channels, but the assumption does not hold for systems with low channel counts such as the hand-held OpUS probe [8].

Non-linear beamformers such as the Delay-Multiply-and-Sum (DMaS) algorithm [9] and the Short-lag-spacial-coherence (SLSC) algorithm [10] have been shown to be effective at improving image quality in situations with low channel counts, at the expense of increased computational complexity. Such methods exploit cross-channel coherence to distinguish true signals from artefacts and noise, thus resulting in improved image contrast. However, for freehand OpUS imaging, reconstruction algorithms need to run in real-time to be clinically relevant.

Here, we present graphics processing unit (GPU)-enabled implementations of several non-linear beamforming algorithms, developed with the NVIDIA CUDA toolkit [11]. These implementations were specifically taylored for use with hand-held OpUS imaging probes, compirisng a single detector and an array of OpUS sources. Implementations designed for use with an OpUS system of the DaS, DMaS and SLSC algorithms are presented and tested with simulated and experimental data, and both the reconstruction speed and image quality were assessed. In addition a novel variation to the DMaS algorithm is presented that exploits the cross-channel coherence only of near-neighbour pairs; presented as a more efficient alternative to DMaS with comparable performance.

II. Methods

A. Delay and Sum

A GPU-enabled implementation of the DaS beamformer was developed to act as a baseline to judge both relative image quality and as a reference point for targeted reconstruction

This work was supported by the Wellcome Trust (203145Z/16/Z), the Engineering and Physical Sciences Research Council (NS/A000050/1, EP/N021177/1, EP/S001506/1,EP/N509577/1,EP/T517793/1), and the Rosetrees Trust (PGS19-2/10006).

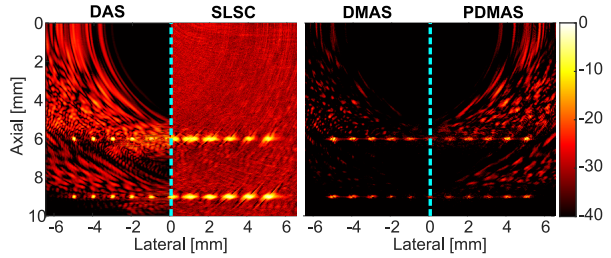

Fig. 1. **Beamforming with simulated data** Comparison images of simulated data reconstructed by each of the four beamforming implementations developed. The simulated phantom used was comprised of two rows of nine point scatterers at 0.5 mm spacing, positioned in the imaging plane at 6 mm and 9 mm from the probe face.

speed. In this work a version of the DaS algorithm specifically tailored to the OpUS probe configuration was used [3]. The image signal $I(\vec{r})$ at image location $\vec{r}$ at a time t_i is given by

$$I(\vec{r}) = \sum_{i=1}^{N} S_i \left(t_i = \frac{|\vec{r} - \vec{r}_{s,i}| + |\vec{r}_d - \vec{r}|}{c} \right), \quad (1)$$

where S_i is the time delayed signal from source i, $\vec{r}_{s,i}$ is the position of source i, $\vec{r}_d$ is the location of the single, stationary receiver, c is the speed of sound and N is the number of acoustic sources. A CUDA kernel was written to perform the computations required for a single image pixel, and the algorithm was parallelised by distributing the computation of each pixel to separate threads. A linear grid of 1024 thread blocks was used (empirically determined to maximise performance) to distribute a total of $(N_{img} + 1024 - 1)/1024$ blocks, which ensured all pixels were computed correctly regadless of the number of pixels N_{img} in the image.

B. Short Lag Spatial Coherence Beamformer

The SLSC beamformer was proposed by Lediju *et al.* as a method for directly exploiting cross-channel correlation in ultrasound beamforming [10]. The normalised spatial correlation calculated by SLSC for a given lag M in a single-detector, multiple source OpUS system is given by

$$R_{SLSC} = \sum_{m=1}^{M} \frac{1}{N-m} \hat{R}(m), \quad (2)$$

where spatial correlation $\hat{R}$ is given by

$$\hat{R} = \sum_{i=1}^{N-m} \frac{\sum_{n=0}^{w} S_i(t_{i+n}) S_{i+m}(t_{(i+m)+n})}{\sqrt{\sum_{n=0}^{w} S_i(t_{i+n})^2 \sum_{n=0}^{w} S_{i+m}^2(t_{(i+m)+n})}}. \quad (3)$$

Where N is the total number of sources, $S_i(t_i)$ is the signal for source i at time t_i and t_{i+n} includes an additional time delay of n samples used to compute cross-channel cross-correlations across a temporal window of length w. Optimal values for M and w were determined empirically and were set to 12 and 4 respectively to balance image quality and reconstruction speed effectively. The GPU implementation of the SLSC algorithm used two separate CUDA kernels to perform the SLSC calculation for each pixel. The first kernel calculated the element-wise square of all RF data, and the second calculated the normalised spatial correlation for each pixel. The same grid scheme used for DaS was used for SLSC, with each thread calculating the value for a single pixel in the image field. Additional white noise was added to all simulated RF data to avoid high amplitude coherence artefacts arising from the coherent background as previously discussed by Lediju *et al.* [10].

TABLE I
IMAGE RECONSTRUCTION PARAMETERS FOR GPU IMPLEMENTATIONS OF NON-LINEAR BEAMFORMERS

Method	Computation Time (ms)	Axial Res. (μm)	Lateral Res. (μm)	Contrast (dB)
DaS	67	150	209	28
SLSC	200	800	966	11
DMaS	88	120	228	52
PDMaS	70	79	195	45

C. Delay Multiply and Sum Beamformer

DMaS was originally proposed as an improved beamformer for confocal microwave imaging to improve clutter rejection and noise levels, especially targeting the effects of side lobes in devices with low channel counts [12]. First applied to biomedical ultrasound imaging by Matrone *et al.* [9], DMaS has become a widely recognised non-linear beamformer in the research community.The DMAS beamformed signal at at pixel $\vec{r}$ is given by

$$I_{DMAS}(\vec{r}) = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \hat{S}_{ij}(t_i, t_j), \quad (4)$$

with $\hat{S}_{ij}(t_i, t_j)$ the amplitude corrected DMAS equivalent RF signal at the point $\vec{r}$ for sources i and j is given by

$$\hat{S}_{ij}(t_i, t_j) = \text{sign}(S_i(t_i) S_j(t_j)) \cdot \sqrt{|S_i(t_i) S_j(t_j)|}. \quad (5)$$

Here time delays t_i,t_j are as defined in equation (1). The resulting DMaS signal is then bandpass filtered (the so-called Filtered-DMaS) to remove sub- and super-harmonic components arising from the multiplication operations [9].

Two forms of CUDA kernel were used for the DMaS implementations tested here. The first was based on the DaS implementation described above, with each thread calculating a single pixel value. The second method took advantage of the small channel counts to unroll one of the summations in the DMaS process. The latter kernel format developed unrolled the summation over i in equation (4), with one thread running the summation over j for each value of i. Each thread then saved the calculated value to the shared memory on the GPU, and the final thread performed a warp unrolled summation over all data points to complete the summation. Whilst this method forced some threads to stall due to varying lengths of the summation over j, an overall decrease in reconstruction time was observed. This resulted in the use of $(N_{src} - 1)^* N_{pixel}$ threads arranged such that each block in the computation grid could calculate a certain number of pixels. For an OpUS array comprising 64 sources, a block size of four pixels was emperically determined to yield the best performance.

D. Pseudo-DMaS beamformer

To improve the speed of the DMAS algorithm, and bring the reconstruction speeds closer to those seen with the DAS algorithm, a novel extension of the DMAS algorithm was proposed and implemented. Referred to as pseudo- or windowed-DMAS (PDMaS), this scheme applies a windowing scheme to the summations in the DMAS algorithm, selecting only the cross-correlation between close-neighbour pairs. This is motivated by the limited omni-directionality of the sources in the OpUS array probe, where each source insonifies only part of the imaging volume. As a consequence the number of iterations needed in each calculation loop is reduced, reducing the overall computational complexity. For a 16-element window that is symmetrical around the current source the PDMAS signal is given by

$$I_{PDMaS}(\vec{r}) = \sum_{i=1}^{N} \left(\sum_{j=i+1}^{\min(N,i+8)} \hat{S}_{ij} + \sum_{j=\max(1,i-8)}^{i-1} \hat{S}_{ij} \right), \tag{6}$$

where $\hat{S}_{ij}$ is the time delayed, scaled, DMAS signal for sources *i* and *j* as given by Eq. (5). The limits of the innermost summations are adjusted to correctly handle the edges of the aperture. The PDMaS implementation was developed in both the single and dual kernel configurations discussed for DMaS above. However, as for PDMaS signals originating from fewer elements are summed, a one-thread-per-pixel scheme was found to achieve highest efficiency.

E. Testing

All GPU implementations were written in C++ using the NVIDIA CUDA platform (CUDA toolbox version 11.7.64) [11], and used to generate a CUDA-enabled dynamic-link-library (DLL) that could be used with other applications such as MATLAB or LABVIEW. Each GPU implementation was used to reconstruct both simulated and experimental datasets.

Simulated data was generated using a two stage simulation: the forward propagation and interaction with isotropic point scatterers was modelled using the fast-nearfield method using the FOCUS MATLAB toolbox [13], the detection of reflected ultrasound was then modelled using the free-space Green's function for a point detector [14] (figure 1). Simulated data from this scheme has been previously validated against experimental results [3], [15], [16]. Each GPU implementation was then used to reconstruct experimental data sets acquired with a previously-reported free-hand OpUS probe [8]. A static image of a tungsten wire phantom (figure 2) in a waterbath was used to determine point scatterer performance. In addition a video data set of a needle inserted into a tissue-mimicking gelwax phantom, originally recorded at 11 Hz frame rate, was reconstructed with each GPU implementation to demonstrate performance of the algorithms in video-rate imaging. All reconstructed data sets had 64 channels, 5001 time samples and a 250 MHz sample rate. All simulations and image reconstructions presented were performed on a Dell Inspiron 7501 PC, with an Intel Core i7-10750H CPU, 16GB RAM and an NVIDIA GeFORCE GTX 1650 Ti GPU.

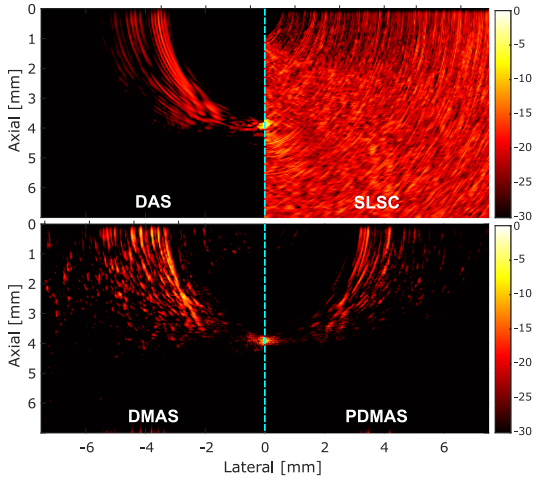

Fig. 2. **Beamforming with experimental data of a wire phantom.** Comparison images of single-frame data collected by a free-hand OpUS probe [8] of a tungsten wire, diameter 27 µm, placed at an axial depth of 3.9 mm.

III. Results

All GPU-implementations were capable of reconstructing image data with a reconstruction time significantly less than the equivalent CPU-based implementation. This SLSC implementation demonstrated a contrast reduction of 17 dB when compared to DaS beamforming, however the distinction between point targets and background is more apparent, with greatly reduced artefacts. However the longer reconstruction times for SLSC severely limit the possible frame rates achievable, and limit the possibility for real-time OpUS imaging.

In contrast, the DMaS and PDMaS algorithms were capable of image contrasts of 52 dB and 42 dB, respectively, whilst maintaining reconstructions speeds that would enable real-time video imaging. This presents a significant improvement over DaS reconstruction, which achieved 28 dB contrast, whilst maintaining a reconstruction rate that would enable real-time video-rate imaging.

IV. Discussion and Conclusion

This work demonstrates the power of non-linear beamforming when applied to situations with low channel counts. This is particularly relevant to the OpUS probes discussed here as expanding the number of channels in a free-hand OpUS device is not feasible due to the prohibitively high costs of the interrogation optics required for multiple fibre-optic detectors. The DMaS-based algorithms considered here achieve improved image quality through reduction in grating- and side-lobe artefacts, and out-of-plane-clutter that dominate linear DaS beamformed images, and yield images with increased dynamic range despite similar reconstruction times (particularly for PDMaS).

For both simulated and experimental point target data, SLSC significantly reduced the level of "wing-shaped" side-lobe artefacts. However, when reconstructing experimental data of a phantom study, SLSC imaging was found to be less effective in improving the image quality: vessel walls that are distinguishable in DaS are unclear, and during the needle insertion

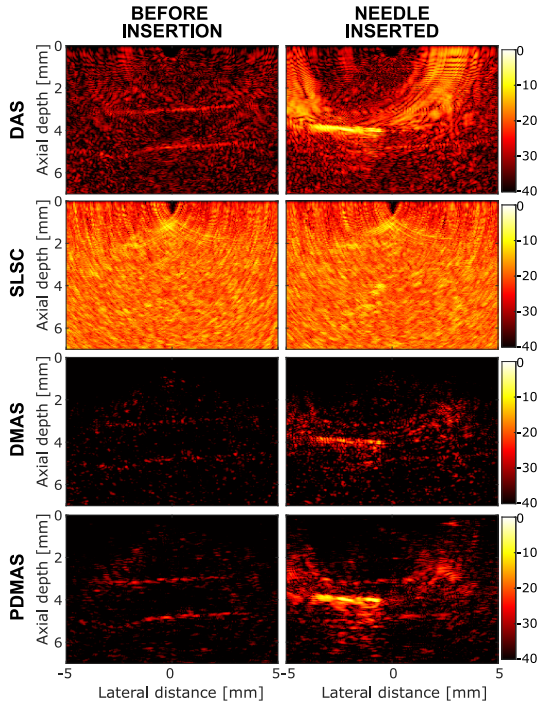

Fig. 3. **Video reconstruction** Comparison images taken from real-time imaging of a tissue mimicking phantom, containing a wall-less structure mimicking a blood vessel. The phantom was imaged whilst a needle was inserted into the phantom and then withdrawn [8]

only the tip of the needle is visible, and can only really be seen in videos. In addition, with the method presented here, SLSC was not achievable at a reconstruction rate that would enable it to be used for video-rate imaging in real-time. SLSC is routinely used as a mask or weighting factor alongside other beamformers, indicating a possible use when post-processing OpUS data, however due to the comparably low numerical efficiency this aspect was not further investigated.

The DMaS implementation presented here achieved significant reductions in side lobe artefacts when compared to DaS reconstruction of the same data. This implementation exhibited contrast gains of up to 24 dB (when compared to DaS reconstruction), whilst maintaining a reconstruction time that would enable imaging at 11 Hz with a free-hand OpUS imaging probe. One possible route to improve the performance of the DMaS algorithm presented here could be to use a re-factored format of the DMaS algorithm to reduce computational complexity [17].

The novel PDMaS algorithm implemented here demonstrated an effective middle ground between DaS and full DMaS reconstruction. Contrast gains of up to 17 dB when compared to DaS are less effective than DMaS, however the faster computation time would enable imaging at 14 Hz with a free-hand OpUS probe, matching the performance of the equivalent GPU enabled DaS implementation. In addition, the PDMaS algorithm also demonstrated improved imaging of wall structures running parallel to the probe face, as demonstrated in figure 3, which is explained by PDMaS rejecting widely spaced channels that exhibit low cross-channel coherence.

The methods presented in this paper demonstrate the capabilities of non-linear beamforming for improving the image quality of freehand OpUS array probes. The implementations developed here may also improve imaging of other low-channel count systems such as sparse arrays. Notably, DMaS based beamformers approached the image reconstruction speeds of conventional DaS methods, whilst achieving significant contrast gains. The results presented here were achieved using GPU acceleration on consumer-grade hardware readily available in personal computers. High-end workstation GPU cards, with significantly higher clock speeds, memory capacity, and throughput will further accelerate the reconstructions, and could readily achieve frame rates exceeding 50 Hz.

REFERENCES

[1] Paul Beard. Biomedical photoacoustic imaging. *Interface focus*, 1:602–631, 2011.

[2] Sacha Noimark, Richard J. Colchester, Radhika K. Poduval, et al. Polydimethylsiloxane composites for optical ultrasound generation and multimodality imaging. *Advanced Functional Materials*, 28:1–16, 2018.

[3] Erwin J Alles, Sacha Noimark, Efthymios Maneas, et al. Video-rate all-optical ultrasound imaging. *Biomedical optics express*, 9:3481–3494, 2018.

[4] B T Cox, E Z Zhang, J G Laufer, and P C Beard. Fabry perot polymer film fibre-optic hydrophones and arrays for ultrasound field characterisation. *Journal of Physics: Conference Series*, 1:32–37, 2004.

[5] James A. Guggenheim, Jing Li, Thomas J. Allen, et al. Ultrasensitive plano-concave optical microresonators for ultrasound sensing. *Nature Photonics*, 11:714–719, 2017.

[6] Wouter J. Westerveld, Md Mahmud-Ul-Hasan, Rami Shnaiderman, et al. Sensitive, small, broadband and scalable optomechanical ultrasound sensor in silicon photonics. *Nature Photonics*, 15:341–345, 2021.

[7] Richard J. Colchester, Callum Little, George Dwyer, et al. All-optical rotational ultrasound imaging. *Scientific Reports*, 9:1–8, 2019.

[8] Erwin J Alles, Eleanor C Mackle, Sacha Noimark, Edward Z Zhang, Paul C Beard, and Adrien E Desjardins. Freehand and video-rate all-optical ultrasound imaging. *Ultrasonics*, 116:106514, 2021.

[9] Giulia Matrone, Alessandro Stuart Savoia, Giosue Caliano, and Giovanni Magenes. The delay multiply and sum beamforming algorithm in ultrasound b-mode medical imaging. *IEEE Transactions on Medical Imaging*, 34:940–949, 2015.

[10] Muyinatu A. Lediju, Gregg E. Trahey, Brett C. Byram, and Jeremy J. Dahl. Short-lag spatial coherence of backscattered echoes: imaging characteristics. *IEEE transactions on ultrasonics, ferroelectrics, and frequency control*, 58:1377–88, 7 2011.

[11] Shane Cook. *CUDA Programming: A Developer's Guide to Parallel Computing with GPUs*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1st edition, 2012.

[12] Hooi Been Lim, Nguyen Thi Tuyet Nhung, Er Ping Li, and Nguyen Duc Thang. Confocal microwave imaging for breast cancer detection: Delay-multiply-and-sum image reconstruction algorithm. *IEEE Transactions on Biomedical Engineering*, 55:1697–1704, 2008.

[13] D. Chen and R. J. McGough. A 2d fast near-field method for calculating near-field pressures generated by apodized rectangular pistons. *Journal of the Acoustical Society of America*, 124(5):1526–1537, 2008.

[14] M.D. Verweij, B.E. Treeby, K.W.A. van Dongen, and L. Demi. *Simulation of Ultrasound Fields*, volume 2. 2014.

[15] Erwin J. Alles and Adrien E. Desjardins. Source density apodization: Image artifact suppression through source pitch nonuniformity. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 67:497–504, 2020.

[16] Arttu Arjas, Erwin J Alles, Efthymios Maneas, et al. Neural network kalman filtering for 3-d object tracking from linear array ultrasound data. *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, 69(5):1691–1702, 2022.

[17] Alessandro Ramalli, Alessandro Dallai, Luca Bassi, et al. High dynamic range ultrasound imaging with real-time filtered-delay multiply and sum beamforming. *IEEE International Ultrasonics Symposium, IUS*, 2017.