ACCELERATING REGULARIZED ITERATIVE CT RECONSTRUCTION ON COMMODITY GRAPHICS HARDWARE (GPU)

Wei Xu *Klaus Mueller*

Center for Visual Computing, Computer Science Department, Stony Brook University

ABSTRACT

Iterative reconstruction algorithms augmented with regularization can produce high-quality reconstructions from few views and even in the presence of significant noise. In this paper we focus on the particularities associated with the GPU acceleration of these. First, we introduce the idea of using exhaustive benchmark tests to determine the optimal settings of various parameters in iterative algorithm, here OS-SIRT, which proofs decisive for obtaining optimal GPU performance. Then we introduce bilateral filtering as a viable and cost-effective means for regularization, and we show that GPU-acceleration reduces its overhead to very moderate levels.

Index Terms— Iterative Reconstruction, Computed Tomography, Ordered Subsets, GPU, Bilateral Filter

1. INTRODUCTION

The high radiation dose delivered to a patient in multi-slice X-ray CT has become a source of growing concern, especially in pediatrics. Therefore, optimizing the dose to obtain the lowest quality acceptable for a given diagnostic purpose (the ALARA principle) is the overall theme in many efforts to lower these exposures. Effective methods here include limiting either the dose per projection, or the number of projections overall, or both. However, while the former decreases SNR, the latter can lead to prominent streak artifacts in the reconstruction. Both can obliterate the features of interest and generally make the CT image hard to read. Exact or approximate exact CT reconstruction methods do not work well under these conditions. Iterative optimization methods, on the other hand, can produce acceptable results but they suffer from high computational effort. This has prevented a deployment in routine clinical applications so far since these computational demands cannot be met by reasonable CPU-based platforms.

Fortunately, the maturing of high-performance graphics chips (GPUs) into commodity massively parallel processors has now begun to enable a wide variety of computationally challenging tasks to be performed inexpensively on the desktop. In the context of medical imaging, we have shown that with just a single such board one can filter and back-project cone-beam projections faster (at 50 projections/s) than they can be produced by a modern flat-panel gantry, enabling *streaming CT* [5]. Further, in earlier work [3] we have also shown that reconstruction algorithms, both iterative and analytical, can typically be broken down into blocks, which can be accelerated individually on these platforms using dedicated programs (called *shaders*).

In this current work we specifically address the acceleration of iterative optimization algorithms for the purpose of low-dose CT with reduced sets of noisy projections. Our framework alternates projection-space prediction-correction with object-space regularization. The former ensures adherence of the solution to the data, while the latter seeks to drive the former to a more plausible solution. Our prominent aim is to make this procedure amenable to GPU-acceleration.

Our paper is structured as follows. Section 2 discusses related work. Section 3 describes our framework. Section 4 present results, and Section 5 ends with conclusions.

2. RELATED WORK

We chose algebraic reconstruction as the predictor-corrector method. In expectation maximization (EM), ordered subsets (OS) have long been known to speed up convergence speed, with larger numbers of subsets converging faster. In recent work, we have introduced the idea of using ordered subsets also for algebraic settings, giving rise to OS-SIRT. In this scheme SIRT and SART form two extremes, with SIRT having just one and SART having M subsets (M being the number of projections). While on the CPU there is little difference in the running time per iteration, on the GPU an iteration with SART is typically the slowest, due to the many projection-backprojection context switches, disturbing parallelism and data flow [4]. This has significant implications for the overall reconstruction wall clock time, where SART, in the noise-free case, is no longer the fastest method (which it is on the CPU) [4]. In contrast, in our current work we address the issue of noise, and thus revisit GPU OS-SIRT under these new circumstances.

For few-view, limited-angle, and noisy projection scenarios, the application of regularization operators between reconstruction iterations seeks to tune the final or intermediate results to some *a-priori* model. A simple regularization scheme is to enforce positivity. In [1] the method of total variation (TV) was proposed for additional

regularization (in conjunction with POCS reconstruction). TV minimization (TVM) has the effect of flattening the density profile in neighborhoods and thus is well suited for noise and streak artifact reduction. Based on the assumption of a relatively sparse gradient object, the method worked quite well under a variety imperfect imaging situations [1]. This assumption may not be realistic in general, but more importantly in the context of high performance computing, the iterative procedure of TVM is quite time-consuming, even when accelerated on GPUs.

3. METHODOLOGY

We aim to devise a method that is not iterative but has the same goals than TVM, that is, the reduction of local variations (noise, streaks) while preserving coherent features. The bilateral filter [2] is such a method. It combines both a range filter and a domain filter, constituting a non-linear filter designed for edge-preserving smoothing. When based on the Gaussian function, two parameters are required, σ_r and σ_d, to control the weight of each filter.

A second important aspect of OS-EM is that it balances noise suppression with convergence speed – typically in the presence of noise using a smaller number of subsets leads to faster convergence and better results, due to the inherent smoothing provided by the larger projection sets. These issues are also relevant for our GPU-accelerated OS-SIRT, but with the added constraints imposed by the GPU hardware architecture. Finally, in contrast to EM, algebraic methods also offer a relaxation factor λ which has a great effect on convergence speed. In [5], a simple linear selection scheme for λ (as a function of subset size) was used, which we found sub-optimal in the current work. We therefore propose a scheme that determines the optimal setting of λ (and subset number) based on an exhaustive set of benchmark tests under different noise conditions.

3.1. OS-SIRT

The correction update for projection-based algebraic methods is computed by the following equation:

$$v_j^{(k+1)} = v_j^{(k)} + \lambda \sum_{p_i \in OS_s} \frac{p_i - r_i}{\sum_{l=1}^{N} w_{il}} \qquad r_i = \sum_{l=1}^{N} w_{il} \cdot v_l^{(k)} \quad (2)$$

where the weight factor w_{ij} determines the contribution of a voxel v_j to a ray r_i starting from a projection pixel p_i and is given by the interpolation kernel. This equation is a generalization of the original SART and SIRT equations to support any number of subsets [4]. The p_i are the pixels in the *M/S* acquired images that form a specific (ordered) subset OS_s where $1 \le s \le S$ and S is the number of subsets.

3.2. Bilateral Filter

The bilateral filter non-linearly averages similar and nearby pixels values. It consists of two filter components: the domain filter and the range filter:

$$h(x) = \frac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(\varepsilon)c(\varepsilon,x)s(f(\varepsilon),f(x))d\varepsilon}{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} c(\varepsilon,x)s(f(\varepsilon),f(x))d\varepsilon} \quad (4)$$

Here, ε and x represent the spatial variables, f is the input image, and c and s are the measured closeness and pixel value similarity, respectively. The geometric closeness function acts as the domain filter controlling the contribution according to spatial distance. Conversely, the pixel value similarity function acts as a range filter, generating very low weights for dissimilar pixel values. Normalization forces the sum of pixel weights to 1. In our work, we model the closeness and similarity functions as Gaussians:

$$c(\varepsilon,x) = e^{-\frac{\|\varepsilon - x\|^2}{2\cdot\sigma_d^2}} \qquad s(\varepsilon,x) = e^{-\frac{(f(\varepsilon)-f(x))^2}{2\cdot\sigma_r^2}} \quad (5)$$

where σ_r and σ_d control the amount of smoothing. Each pixel collects contributions from all image pixels. This is quite expensive and also not effective since the contributions of remote pixels are very subtle. A mask window (centered about an updated pixel) improves performance by restricting operations only to pixels inside the window.

The implementation of GPU-accelerated bilateral filtering is as follows. The rendering target is a texture of the size of the reconstructed image, with image texture and other parameters (size of image, σ_r, σ_d, etc.) passed into the GPU. We avoid the expensive evaluation of the exponential function by pre-computing both closeness and similarity functions, and store them into two 1-D lookup textures. We implemented bilateral filtering both in 2D and 3D.

3.3. OS-SIRT with Bilateral Filter regularization

In our regularized OS-SIRT bilateral filtering is applied after each iteration (after the backprojection of all subsets). This removes artifacts at the very beginning when the errors are just generated and thus steers the reconstruction towards more plausible and favorable solution regions. Since the target texture (to be filtered) is already in GPU memory, this operation does not require any expensive texture uploading or downloading operations between the CPU and GPU.

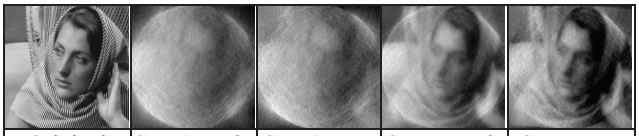
4. RESULTS

Our experiments were conducted on a NVIDIA 8800GT GPU, programmed with GLSL. We group the results into two sections: (1) the OS-SIRT results showing the relationship between noise levels and parameters settings,

and (2) the performance of our GPU-accelerated bilateral filter and the corresponding reconstruction results.

4.1. OS-SIRT with noisy data

We used the 2D *Barbara* test image (size 256^2) to evaluate the performance of the different reconstruction schemes. We obtained 180 projections at uniform angular spacing of [-90˚, +90˚] in a parallel projection viewing geometry. We then added different levels of Gaussian noise to the projection data, to obtain SNRs of 15, 10, 5, and 1. Figure 1 presents the best reconstruction results obtained (using the correlation coefficient CC between original and reconstructed image), for each SNR, at the smallest wall-clock time.



Original	SNR1 rand	SNR1 seq	SNR5 rand	SNR5 seq
CC	0.6	0.7	0.85	0.85
Time (s)	0.069	0.068	0.08	0.09

SNR 10	SNR 15	Noise-free	Noise-free
0.88	0.9	0.95	0.97
0.081	0.094	0.244	0.725

Figure 1: Reconstructions obtained with different SNR levels for the Barbara image

Since the *Barbara* test image has a high level of fine detail, which cannot be recovered at a high level of noise, the CC's upper bound is limited. For small SNR, convergence to the optimal CC is quickly followed by divergence, which explains the reduced wall-clock time at these SNR levels. In this context we also observed that a sequential ordering of the projections in the subsets improved the CC that could be obtained (over a random arrangement).

The optimal settings greatly depend on the particular imaging situation at hand, such as SNR, total number of projections and their angular range, the imaged object, scanner, etc. Figure 2 presents results on the influence of SNR. The plots give quantifying hints on how to pick the best-performing λ and number of subsets (to obtain the best possible quality within the smallest time), for each expected SNR level. For example, we observe that low SNR requires a low numbers of subsets. The figure also indicates that for the noise-free case SART is the best method – even though a single iteration takes the longest, the time performance is dominated by the fast convergence speed. On the other hand, for increasing levels of noise, where SART converges slowly, the optimal number of subsets decreases, favoring OS-SIRT with its better de-noising properties. Also, at the same time, with growing subset sizes, the relaxation factor increases in a non-linear fashion. This is a strong departure from the linear model used on [4] – a higher λ will lead to faster convergence and via our exhaustive benchmark tests we are safe that it also leads to more accurate results. It shows that damping via relaxation is replaced by damping via intra-subset averaging.

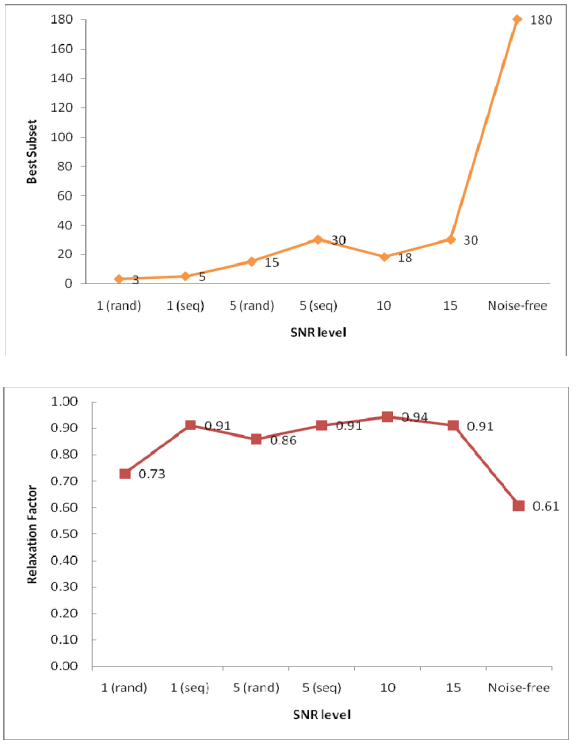


Figure 2: Best time and optimal relaxation factor as a function of imaging condition, here SNR.

4.2. Performance with Bilateral Filter

We tested the speed of both 2D and 3D bilateral filters with different sizes of images and windows on both CPU and GPU. Table 1 shows that speedups of more than two orders of magnitude can be obtained by using the GPU:

Test size	Window size	CPU time	GPU time
256×256	91×91	67.626	0.109
	61×61	32.219	0.047
	31×31	8.829	0.015
512×512	41×41	59.313	0.093
	21×21	15.875	0.032
	11×11	4.453	0.016
256×256×256	9×9×9	> 100	3.953
	7×7×7	> 100	1.984
	3×3×3	75.563	0.469

Table 1: Wall clock time (in s) of GPU vs. CPU bilateral filter

To gauge the performance of regularized reconstruction for both the few-view and the noise (SNR=10) scenario, we used the Visible Human dataset at 512^3 resolution. Here we employed SART with 8 iterations for the noise-free few-view case. The filter window size filter was fixed to 11. Figure 3 shows one slice of the reconstructed volume before and after filtering, respectively. No streaking artifacts are present, which would have dominated the reconstructions otherwise, and the features are well preserved. We tested a number of combinations of representative σ_r and σ_d and selected the best results.

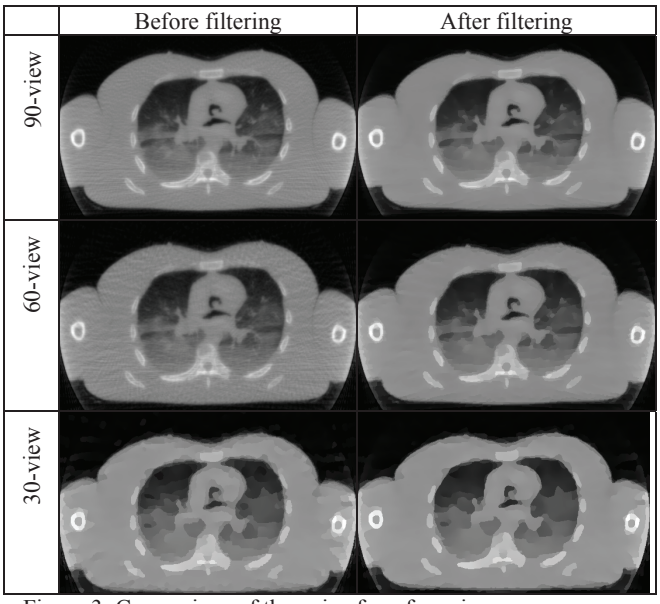


Figure 3: Comparison of the noise-free, few-view case.

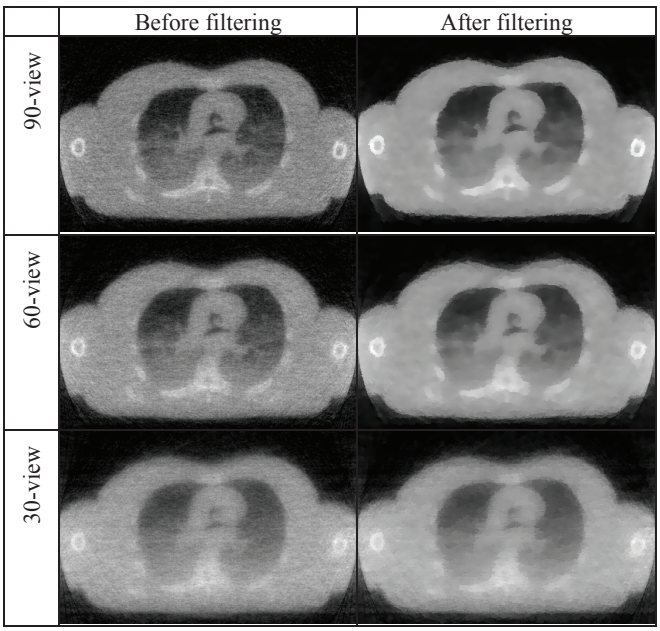


Figure 4: Comparison for the noisy (SNR=10), few-view case.

#proj	1-ch	1-ch w/ BF	4-ch	4-ch w/ BF
180	121.159	133.105	47.97	50.686
30	30.896	39.598	12.429	15.908

Table 2: Wall clock time for one GPU-accelerated SART iteration of the 512^3 volume. The timings for 1-ch and 4-ch are results obtained when accelerating the reconstruction with 1 (R) or 4 RTGBA) color channels, respectively. BF is the bilateral filter.

Figure 4 shows the results for the noisy few-view case, after 5 iterations. Like in noise-less case we observe that the salient features are well preserved in both size and shape.

Finally, Table 2 lists the GPU-accelerated reconstruction time required for one SART iteration, for the Visible Human dataset at 512^3 resolution for 180 and 30 projections. The 1-ch time uses only the R-channel of the GPU hardware, while the 4-ch time uses all 4 (RGBA) channels in parallel. We observe a 2.5-fold speedup. We further observe that the regularization via bilateral filtering adds only a moderate time overhead (about 30%).

5. CONCLUSIONS

We have shown that benchmark-based parameter selection in GPU-accelerated iterative reconstruction can make iterative reconstruction a clear option for CT for noisy and/or few-view scenarios. We also showed that bilateral filtering represents a viable option for regularization, with the added advantage that it accelerates very well on GPUs. Further study is clearly needed to thoroughly compare the bilateral filter with TVM in terms of accuracy. Future work will consider local adaptive parameter tuning to better preserve local features and to study the acceleration performance using CUDA. For OS-SIRT we are currently working on an extended study to determine if the optimal benchmark-tested parameters generalize to certain object classes and scanning scenarios.

11. REFERENCES

[1] E. Y. Sidky, C.-M. Kao, X. Pan, "Accurate image reconstruction from few-views and limited-angle data in divergent-beam CT," *J. X-ray Sci.Tech*. 14: pp. 119–139, 2006.

[2] C. Tomasi, R. Manduchi, "Bilateral filtering for gray and color images," *ICCV*, pp. 839–846, 1998.

[3] F. Xu, K. Mueller, "Accelerating popular tomographic reconstruction algorithms on commodity PC graphics hardware," *IEEE Trans. on Nuclear Science,* 52: pp. 654-663, 2005.

[4] F. Xu, K. Mueller, M. Jones, B. Keszthelyi, J. Sedat, D. Agard, "On the efficiency of iterative Ordered Subset Reconstruction algorithms for acceleration on GPUs," *MICCAI Workshop on High-Performance Medical Image Computing & Computer Aided Intervention*, New York, September, 2008.

[5] F. Xu, K. Mueller, "Real-time 3D computed tomographic reconstruction using commodity graphics hardware," *Physics in Medicine and Biology*, vol. 52, pp. 3405–3419, 2007.