Medical Image Processing on the GPU
–
Past, Present and Future

Anders Eklund, PhD
Virginia Tech Carilion Research Institute
andek@vtc.vt.edu
Outline

• Motivation – why do we need GPUs?
• Past - how was GPU programming done 10 years ago?
• Present - how are GPUs used in medical imaging today?
• Future - what challenges do we face?
What is medical imaging?

• Creating images of the interior of the human body, for research and clinical purposes

• The three most common modalities are

• Computed tomography (CT)
• Magnetic resonance imaging (MRI)
• Ultrasound (US)
Why do we need GPUs in medical imaging?

- The medical data explosion
- Demanding algorithms for image reconstruction and data analysis
- Visualization & interactivity
The medical data explosion

• Medical image data have evolved from 2D to 4D

• Temporal and spatial resolutions continue to improve

• The number of subjects being scanned are increasing
The medical data explosion
From 2D to 4D data

2D
One image of size
512 x 512

3D
One volume of size
512 x 512 x 512

4D
Many volumes
128 x 128 x 64 x 100
The medical data explosion
From 2D to 4D data

• 512 x 512 image 1 MB
• 512 x 512 x 512 volume 512 MB
• 128 x 128 x 64 x 100 dataset 420 MB
• 512 x 512 x 512 x 20 dataset 10.7 GB
What about the computational complexity for 2D, 3D and 4D data?
The medical data explosion
From 2D to 4D data

• Filtering is one of the most important operations in (medical) image processing

• Can be performed as convolution in the spatial domain or as multiplication in the frequency domain

• \( s = \text{signal (image)}, \quad f = \text{filter} \)

\[
(s * f)[x, y] = \sum_{f_x = -N/2}^{f_x = N/2} \sum_{f_y = -N/2}^{f_y = N/2} s[x - f_x, y - f_y] \cdot f[f_x, f_y]
\]
Filtering – Edge detection

- Apply two filters to detect edges along x and y
The medical data explosion
From 2D to 4D data

• Convolving a 512 x 512 image with a 11 x 11 filter requires \(~32\) million multiply add operations
The medical data explosion
From 2D to 4D data

• Convolving a 512 x 512 x 512 volume with a 11 x 11 x 11 filter requires \( \sim 179 \) billion multiply add operations
The medical data explosion
From 2D to 4D data

- Convolving a $512 \times 512 \times 512 \times 20$ dataset with a $11 \times 11 \times 11 \times 11$ filter requires $\sim 39$ trillion multiply add operations
The medical data explosion
From 2D to 4D data

• Data size from 1 MB to 10.7 GB, increase of a factor \(\approx 10,000\)

• Computational complexity from 32 million operations to 39 trillion operations, increase of a factor \(\approx 1\) million
The medical data explosion
Higher temporal and spatial resolution

- The temporal and spatial resolution of all medical imaging modalities continue to improve
- Better hardware, compare with digital cameras
- More complex sampling patterns
Magnetic resonance imaging (MRI)

- No ionizing radiation
- Can measure different properties (fMRI, DTI, SWI)
- Good for soft tissue
- Can generate 2D, 3D, 4D data
- Expensive
- Significantly slower compared to CT
Computed tomography (CT)

- Extremely quick
- High spatial resolution
- Good for hard tissue
- Can generate 2D, 3D, 4D data
- Expensive
- Ionizing radiation
Ultrasound

• Cheap
• Mobile
• Very high temporal resolution (20-30 Hz)
• Can generate 2D, 3D, 4D data
• Lower spatial resolution
• Noisy images
How to get a higher spatial resolution

• MRI: Stronger magnetic fields or longer scan times (expensive and difficult)

• CT: More radiation (not so good for the subject)
The medical data explosion
Higher temporal and spatial resolution

• More complex sampling techniques to further improve spatial and temporal resolution

• Compressed sensing, sample data in a smarter way

• Parallel imaging, sample more data at the same time

• More complex image reconstruction algorithms
The medical data explosion
More subjects

• 1980, 5 million CT scans in the US
• 2007, 65 million CT scans in the US

Brenner DJ. Should we be concerned about the rapid increase in CT usage? Reviews on Environmental Health, 25, 63–68, 2010
The medical data explosion
More subjects

- Functional magnetic resonance imaging (fMRI) can be used to study brain activity

- A small fMRI study involves some 20 subjects

- fMRI data collection is expensive

- The human connectome project will share fMRI and DTI data from 1200 subjects

http://www.humanconnectome.org/
Demanding algorithms

- Image reconstruction, to convert the collected data to an image or volume
- Image registration, to align two images or volumes
- Image segmentation, to extract a specific part of an image or volume
- Image denoising, to suppress noise and improve the image quality
Demanding algorithms

• The human connectome project will collect and share fMRI data from 1200 subjects

• 12 GB of data per subject

• Apply a permutation test with 10,000 permutations to each dataset (statistical analysis)

• Equivalent to analyze 144,000,000 GB of data
Visualization & Interactivity

• Hard to look at 3D/4D data as 2D images

• 512 x 512 x 512 x 20 dataset = 10 240 images

• ~3 hours if you look at every image for 1 second

• Use volume rendering techniques instead

• Interactive algorithms, combined with visualization
Past
Why GPUs?

• GPUs are very popular for image processing

• Computer graphics; render all the pixels in the same way

• Image processing; apply the same operation to all pixels

• GPUs have hardware support for (linear) interpolation
Eklund et al., Medical image processing on the GPU – Past, present and future, Medical Image Analysis, 2013
Eklund et al., Medical image processing on the GPU – Past, present and future, Medical Image Analysis, 2013
How was GPU programming done 10 years ago?

- Do image processing through computer graphics languages
- OpenGL, Open Graphics Language
- Direct X
- HLSL, High Level Shading Language
- Cg, C for graphics
- GLSL, OpenGL Shading Language
How was GPU programming done 10 years ago?

• Only a few experts knew how to use these programming languages for image processing

• Hard to optimize the performance

• Very hard to debug the code
Present
How is GPU programming done today?

- C programming of GPUs
- CUDA, Compute Unified Device Architecture
- OpenCL, Open Computing Language
- Possible to debug GPU code as regular C code
- Possible to improve performance by using tools like the Nvidia visual profiler
How are GPUs used in medical imaging today?

• Image reconstruction

• Image registration

• Image segmentation

• Image denoising
Image reconstruction - MRI

• MRI data is sampled in the frequency domain

• Most common reconstruction; apply an inverse fast Fourier transform (FFT)

• CUFFT (CUDA), clFFT (OpenCL)

• More advanced sampling patterns result in more complex image reconstruction algorithms
Non-cartesian sampling

• Non-cartesian sampling is sometimes better
• The FFT requires cartesian sampling
**fMRI**

- *Functional* magnetic resonance imaging (fMRI)

- Collect volumes of the brain while the subject is performing a task

- Used to study brain *activity*

- Standard fMRI dataset: 64 x 64 x 30 x 400 (voxels are 4.0 x 4.0 x 4.0 mm, sampling rate 0.5 Hz)

- High resolution dataset: 128 x 128 x 60 x 1200 (voxels are 2.0 x 2.0 x 2.0 mm, sampling rate 1.5 Hz)
fMRI = pattern matching in time

High correlation with paradigm (brain activity)

Low correlation with paradigm (no brain activity)

200 time points
High-resolution fMRI for mapping fingers

1 mm isotropic functional MRI at 3 T. Bilateral finger tapping blocked study, Red is index finger and Blue is pinky or fifth finger. Contrast map formed by subtracting (Red: index – pinky) and (Blue: pinky-index)

**Challenge:** fMRI has MANY TIME POINTS AND SLICES. In this data we had 16 slices and 200 time points = 3200 images to be reconstructed. On CPU this is not feasible with total reconstruction time reaching 1 month.

**Used IMPATIENT reconstruction on GPU.**

Total reconstruction time of 40 hours instead of 1 month.

http://impact.crhc.illinois.edu/mri.aspx

University of Illinois at Urbana-Champaign  Brad Sutton  mrfil.bioen.illinois.edu
DTI

• Diffusion tensor imaging (DTI)

• Measure diffusion of water in different directions

• Combine the measurements to a diffusion tensor (a 3 x 3 matrix in each voxel)

• Often used to study brain connectivity
High-resolution Diffusion Tensor Imaging of neural pathways

1 mm isotropic DTI MRI at 3T. DTI allows for a non-invasive characterization of neural integrity. Our technique corrects for field inhomogeneity and performs SENSE reconstructions on high resolution data.

Challenge: Multiple slabs and multiple directions in a single data set. Diffusion Tensor Imaging (DTI) requires many reconstructions for one data set.

Used IMPATIENT reconstruction on GPU.
Reduced reconstruction time from 18 hours to 5 minutes.
http://impact.crhc.illinois.edu/mri.aspx

Brad Sutton  mrfil.bioen.illinois.edu

University of Illinois at Urbana-Champaign
Fiber tracking for DTI

- DTI can be used for tracking of fibers in the brain
- Place a seed somewhere in the brain
- Follow the main orientation of the diffusion tensor in each voxel, gives the path of each fiber
- The main orientation is given by the eigenvector corresponding to the largest eigenvalue
Performing Real-Time Interactive Fiber Tracking

Adiel Mittmann*, Tiago H. C. Nobrega, Eros Comunello, Juliano P. O. Pinto, Paulo R. Dellani, Peter Stoeter, Aldo von Wangenheim

adiel@inf.ufsc.br

Image registration

• Image registration is needed whenever you want to align two images or volumes

• Compare a subject before and after surgery

• Combine different medical imaging modalities

• Make a group analysis of fMRI data (transform all subjects to a brain template)
Image registration - Example

Subject's brain volume  Brain template
Image registration - Algorithm

1. Calculate similarity measure between images
2. Calculate a new set of transformation parameters (using some optimization algorithm)
3. Apply transformation using interpolation
4. Go to 1
Image registration – Non-linear

• Linear image registration, optimize a few parameters like rotations and translations

• Non-linear image registration, use 100,000 – 1,000,000 parameters (three parameters per voxel)

• Non-linear registration often gives a better result, at the cost of a longer processing time
Image registration - Algorithm

• 1. Calculate similarity measure between images GPU

• 2. Calculate a new set of transformation parameters

  Linear registration: CPU
  Non-linear registration: GPU

• 3. Apply transformation using interpolation GPU

• 4. Go to 1
Image registration

- Non-linear registration of 200 MRI volumes to a brain template (182 x 218 x 182)

<table>
<thead>
<tr>
<th>Software</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>FSL</td>
<td>116 hours</td>
</tr>
<tr>
<td>AFNI</td>
<td>110 hours</td>
</tr>
<tr>
<td>AFNI OpenMP</td>
<td>31 hours</td>
</tr>
<tr>
<td>BROCCOLI Intel CPU</td>
<td>3.5 hours</td>
</tr>
<tr>
<td>BROCCOLI Nvidia GPU</td>
<td>15 minutes</td>
</tr>
<tr>
<td>BROCCOLI AMD GPU</td>
<td>20 minutes</td>
</tr>
</tbody>
</table>

Eklund et al., BROCCOLI: Software for fast fMRI analysis on many-core CPUs and GPUs, Frontiers in Neuroinformatics, 8:24, 2014
Image segmentation

• Image segmentation is needed whenever you want to study a specific part of a dataset

• What is the size of hippocampus?

• How big is the brain tumour?

Has it grown since a previous timepoint?
Image segmentation - Example
Image segmentation

- Most image segmentation algorithms can run in parallel
- Easy to visualize data in GPU memory
- Easy to create interactive interfaces for image segmentation
Roberts et al., A work-efficient GPU algorithm for level-set segmentation, High performance graphics, 123-132, 2010
Image denoising

• Image denoising is used to suppress noise and to improve the image quality

• Makes it easier for a medical doctor to do a diagnosis

• Often used before image registration or image segmentation, to improve the result
Adaptive filtering of a 4D CT dataset

- A CT dataset of the size 512 x 512 x 512 x 20

- Adaptive filtering; apply 11 non-separable filters of the size 11 x 11 x 11 x 11

- CPU: 2.5 days
- GPU: 26 minutes

Eklund et al., True 4D image denoising on the GPU, International journal of biomedical imaging, 2011
A beating heart
Future
What challenges do we face?

• Large datasets, how can we analyze them on a GPU?

• Code optimization, do not want to re-optimize code for each GPU generation

• Easier programming & usage, such that GPUs can solve more problems
Large datasets

• CPUs normally have access to more memory
• Easier to increase the amount of memory

• A GPU has to process a small part of the dataset at a time, gives smaller speedup

• The amount of GPU memory needs to increase faster than the temporal and spatial resolution of medical imaging data
Code optimization

• Frustrating to have to re-optimize code for each GPU generation

• Use libraries as often as possible, wait for libraries to be re-optimized

• Future compilers will hopefully be better at optimizing code
Easier programming & usage

- Many persons do not have time to learn GPU programming

- Accelerate existing C/C++ code using the PGI accelerator model, the HMPP workbench compiler or C++ AMP

- Important that more programming interfaces are created
Easier programming & usage

• Many Adobe products have GPU support (e.g. for image processing and video editing)

• The parallel computing toolbox for Matlab supports GPU computing

• Important that more softwares provide GPU support
Thank you for your attention

Questions?