ACCELERATED ITERATIVE RECONSTRUCTION OF THREE-DIMENSIONAL IMAGES USING THE OS-SART AND OS-HBIR METHODS

Sergei A. Zolotarev¹, e-mail: zolotarev@iapb.bas-net.by, ¹NAS of Belarus Institute of Applied Physics, Minsk, Belarus
Valery L. Vengrinovich¹, veng@iapb.bas-net.by, ¹NAS of Belarus Institute of Applied Physics, Minsk, Belarus
Mietee M. Mukhtar ² email: mietee@gmail.com, ²Belarussian National Technical University, Minsk, Belarus

Abstract: Statistical method of maximum likelihood (EM) and algebraic reconstruction technique with simultaneous iterations (SART) are the two iterative tomographic reconstruction techniques. These algorithms are often used when the projection data contains a large amount of random noise has been obtained from a limited angle range, or have a limited number of angles. One popular approach used to improve the speed of convergence of the algorithm is to perform a correction of the reconstructed approximation of the current object on the projection data subsets. This led to the creation of the method of ordered subsets of the maximum likelihood method for EM (OS-EM) and algebraic reconstruction technique with simultaneous iterations SART (OS-SART). Both of these methods have been accelerated by the use of OpenGL graphics library, through its incorporation in the GPU graphics card architecture.

Keywords: cone beam computed tomography, iterative image reconstruction, os-sart, os-hbir

1. Introduction

Methods of computer tomography reconstruction can be conditionally divided into two classes: analytical algorithms and iterative algorithms. All analytical algorithms are based on two fundamental mathematical principles - the Radon transform and the central section theorem. It is easy to show that the Fourier image of the projection is the central section of the two-dimensional Fourier transform of the two-dimensional function. In the literature this property is called the central layer theorem or the central section. Thus, by collecting a sufficient number of one-dimensional X-ray projections from the two-dimensional object under test and performing a one-dimensional direct Fourier transform for each of them, we can obtain a two-dimensional transformation from the scanned object by interpolation. After completing the inverse two-dimensional Fourier transform from the two-dimensional Fourier image, we finally get a two-dimensional image with the distribution of the absorption coefficients of the scanned object in the reconstruction area, that is, we reconstruct the required two-dimensional function [1]. The most popular methods are filtered in the case of reverse projection (FBP) [2] for the two-dimensional case, the method of Feldkamp (FDC) also provides for filtering in the implementation of back projection [3] for the three-dimensional case.

When the number of scanned X-ray projections sufficiently large, the above methods can produce an accurate or approximate reconstruction of the image of the scanned object. The simplicity of these methods guarantees the efficiency of computational operations, for example, makes them popular for many clinical applications. For a different category of algorithms, the so-called iterative reconstruction methods, the image reconstruction problem is reduced to the problem of solving a very large and simultaneously sparse system of linear equations, that is, to the problem of linear algebra. In this case, therefore, the reconstruction procedure reduces to solving the system of equations [1]. Direct methods for solving systems of linear equations can not be applied because of the immense dimension of the system under consideration.

The most reasonable approach is to use various numerical optimization methods to solve this system. Thus, to solve the above system, iterative methods for solving large systems of equations will be used. First, an initial assumption is set that will be iteratively altered in order to achieve the minimum difference between the experimental projections and model projections that are calculated at each step of the iterative process for the current three-dimensional image, that is, when the current approximation is directly projected onto the virtual matrices of two-dimensional detectors. Iterative methods can be further divided, for example, into projection methods on convex sets (POCS) algorithms such as SART [4], SIRT [5], and POCS [6] and statistical algorithms such as EM [7], OS-EM [8] and MAP [9,10]). Because of the need to continuously implement large cycles, the computational load of these algorithms is quite high.

Nevertheless, in a number of cases they work much better than analytical methods of reconstruction. For example, when x-ray scanning is performed at small anode currents of the X-ray tube to provide small doses of radiation that patients receive.

Iterative methods can be further divided, for example, into projection methods on convex sets (POCS) algorithms such as SART [4], SIRT [5], and POCS [6] and statistical algorithms such as EM [7], OS-EM [8] and MAP [9,10]). Because of the
need to continuously implement large cycles, the computational load of these algorithms is quite high.

In addition, they are most suited to include a priori (preliminary) knowledge about the object of reconstruction for inclusion in the iterative process of tomographic reconstruction.

Theory

1. Technology for accelerating reconstruction using graphics processors

Rapid speed growth and the ability of programmable general-purpose graphics cards (GPUs) have moved high-performance computing programs to conventional desktop computers, increasing computing speed to the level of cluster systems. High-quality graphics cards, such as the NVIDIA GeForce GTX 680, for example, show a performance of up to 3,000 Giga Flops and, moreover, are now available for prices not exceeding $ 500, and their performance is constantly growing according to Moore’s law. Moore’s law is an empirical observation originally made by Gordon Moore, according to which (in the modern formulation) the number of transistors placed on an integrated circuit chip doubles every 24 months. The often quoted interval of 18 months is related to the forecast of David House from Intel, according to which the processor performance should be doubled every 18 months due to a combination of the increase in the number of transistors and the speed of each of them. Acceleration of 1-2 orders of magnitude can be achieved by mapping the computational algorithms to the architecture of the graphics processors that make up the graphical video card of the computer.

The variety of areas in which the mapping of computational algorithms to the graphics processor architecture is used is quite large [11]. These areas include, including tomographic reconstruction of medical images. Such impressive successes due to the highly parallel architecture of the graphics processors SIMD (one instruction - a lot of data). The architecture of the graphics processor has a high memory access bandwidth. So, for example, NVIDIA 8800 GTX has 128 such SIMD floating-point processors, while the majority of new video cards from NVIDIA have significantly more stream processors. For example, the GTX 680 graphics card has 1536 stream processors. It is important to note, however, that the high acceleration values achieved with the use of graphics processors do not come by themselves. They require that the software developer carefully compare the target algorithm of the transition from single-threaded programming models for each individual stream processor to a multi-threaded SIMD software model for graphical computing. The high performance of graphics processors is a consequence of their highly parallel architecture. The huge computational potential of graphics cards, which can be used for high-performance general computing, has recently generated a trend in the widespread use of computations on graphics processors (GPGPUs).

In the recent past, GPU programming was possible only with the help of graphical interfaces created using the programming languages CG, GLSL and HDSL, which required programmers to have a lot of experience in the field of computer graphics. In order to make hardware graphics cards more accessible to ordinary programmers, a C-like parallel programming interface was developed, called the CUDA (Compute Unified Device Architecture) technology, which was recently introduced by the graphics card manufacturer, NVIDIA. At once it is necessary to note the main disadvantage of the proposed computing technology, which consists in the fact that it works only on video cards from the company NVIDIA. You can use a similar, but more general API called OpenCL, which is now quite affordable. However, instead of using the CG, GLSL and CUDA technologies for the implementation of tomographic reconstruction programs, we will use the OpenGL graphics library, which avoids low-level programming in CG, GLSL and HDSL languages, as well as the use of CUDA technology and allows using any video card, Not just video cards from the company NVIDIA.

2. Algebraic reconstruction method with simultaneous iterations (OS-SART)

The statistical maximum likelihood (EM) method and the algebraic reconstruction method with simultaneous iterations (SART) are two methods of iterative tomographic reconstruction. These algorithms are often used when projection data contain a large amount of statistical noise, have been obtained from a limited range of angles, or have a limited number Perspectives. One of the popular approaches used to increase the rate of convergence of these algorithms is to correct the current approximation of the reconstructed object on subsets of the projection data. This led to the creation of an ordered-subset method for the maximum likelihood method EM (OS-EM) and for the algebraic reconstruction method with simultaneous iterations of SART (OS-SART).

Graphic video cards of general purpose have found great prospects for counteracting the large computational loads that are characteristic for iterative reconstruction methods. Nevertheless, we find that the special architecture and model of GPU programming adds additional performance constraints in real time using ordered subsets of projections, counteracting the increase in performance by using small subsets of projections, which was previously observed when executing such algorithms on a conventional central processor. This feature leads to the appearance of new regularities for determining the optimal number of subsets, as well as to new ways of adjusting the relaxation coefficient to obtain the shortest possible reconstruction time.

3. Heuristic method of statistical iterative reconstruction (HBIR)

Iterative methods for tomography are usually based on a solution of a system of linear algebraic equations

$$Ax = p.$$  (1)

In the past we used for the solution a heuristic iterative algorithm inherently based on theoretical basis close to the algorithm published in [12], but which primarily differs by the fact that we use here the correction of the this approach by minimizing the residuals, not intensities, for the radial integrals. The proposed reconstruction algorithm (Bayesian Inference Engine, BIR) can be written as

$$x^{(k)} = x^{(i)} + \lambda^{(i)} \sum_{j=0} A_{ij} \left( p_j - \sum_{k=0} A_{kj} x^{(i)}_k \right),$$  (2)

where $x_j^{(k)}$ – the value of the $j$-th component of the vector of unknowns at the $k$-th iteration, $A_{ij}$ – elements of the projection matrix, $p_j$ – the value of the measured radial integral in the pixel number $i$, $\lambda^{(i)}$ – sequence relaxation parameter values.
(0 < λ < 1).

Corrected algorithm represented by the formula (2) is convenient for parallelization using OpenGL graphics library, as well as the famous SART algorithm [4]. Direct projecting of a three-dimensional texture, containing an image of the current approach, is the same as for the SART algorithm, which is described in detail in [1]. The difference lies in the fact that its implementation must use two additional two-dimensional textures. A corrected image is contained in one of them. The second texture contains an image of layer of the corrected three-dimensional texture. By correcting an image, it is first projected perspective to the place where the corrected layer is, and then multiplied with the corrected image of layer and recorded in its place. For multiplication of two-dimensional texture, texture mapping is applied to the object within GL_MODULATE mode.

4. Results

We programmed both OS-HBIR and OS-SART algorithms, and also carried out a series of reconstructions of a special phantom "Rozi", an X-ray survey of which was carried out in Heidelberg, Germany, at the German Cancer Research Center-DKFZ. A total of 360 X-ray projections were taken through the angular interval of 10. We briefly describe the method of dividing a complete set of projections into a given number of ordered subsets of projections. The text of the reconstruction program specifies a two-dimensional array:

```c
int os_prj[180][180];
float os_alfa[180][180];
```

In this case, the maximum number of ordered subsets should not exceed 180. This array is filled before the reconstruction process begins:

```c
if((num_os%2)==0)nmx1=rint1((float)0.5*(float)(num_os));
else                         nmx1=rint1((float)0.5*(float)(num_os))+1;
else                         nmx2=nmx1 - 1; .
if((num_os%2)==0)nmx2=nmx1;
else
for(ii=0;ii<nprj_os;ii++)
  j2=2*jj;
for(ii=0;ii<nprj_os;ii++)
  os_prj[j1][ii]=nprj_os*j2+ii;
  os_alfa[j1][ii]=alfa[nprj_os*j2+ii];
}

for(jj=0,j1=nmx1;jj<nmx2;j1++,jj++)
  j2=2*jj+1;
for(ii=0;ii<nprj_os;ii++)
  os_prj[j1][ii]=nprj_os*j2+ii;
  os_alfa[j1][ii]=alfa[nprj_os*j2+ii];
}
```

Where the identifier `num_os` is the number of ordered subsets of the projections, and the `nprj_os` identifier specifies the number of projections that enter into each ordered subset of the projections, and the two-dimensional array `os_alfa [num_os] [nprj_os]` specifies the projection angle for each particular projection belonging to the specified ordered subset of the projections.

In this case, ordered subsets of projections are processed in the following order:
0, 2, 4, ..., 2 (nmx1-1); 1, 3, 5, ..., 2 (nmx1-1) + 1.

Of course, one can change the order of the ordered subsets of projections one after another if the index `jj` on the right side of the cycle is changed according to a predetermined rule, that is, if instead of the index `jj` there is some integer function of this index `Fun (jj)`, and the value `Fun` will be determined by the number of some of the ordered subsets of the projections. To check the effect of the number of ordered subsets of projections on the rate of convergence of the iterative process, 12 variants of the partitioning of 360 projections into ordered subsets of projections presented in Table 1 were considered.

Table 1. Variants of partitioning into ordered subsets of projections

<table>
<thead>
<tr>
<th>Variant number</th>
<th>Number of subsets</th>
<th>The number of projections in each of the subsets</th>
</tr>
</thead>
<tbody>
<tr>
<td>1)</td>
<td>3</td>
<td>120</td>
</tr>
<tr>
<td>2)</td>
<td>4</td>
<td>90</td>
</tr>
<tr>
<td>3)</td>
<td>5</td>
<td>72</td>
</tr>
<tr>
<td>4)</td>
<td>6</td>
<td>60</td>
</tr>
<tr>
<td>5)</td>
<td>8</td>
<td>45</td>
</tr>
<tr>
<td>6)</td>
<td>9</td>
<td>40</td>
</tr>
<tr>
<td>12)</td>
<td>24</td>
<td>15</td>
</tr>
<tr>
<td>13)</td>
<td>30</td>
<td>12</td>
</tr>
<tr>
<td>14)</td>
<td>36</td>
<td>10</td>
</tr>
<tr>
<td>15)</td>
<td>40</td>
<td>9</td>
</tr>
<tr>
<td>16)</td>
<td>45</td>
<td>8</td>
</tr>
<tr>
<td>17)</td>
<td>60</td>
<td>6</td>
</tr>
</tbody>
</table>

We introduce the following notation:

\[ \hat{F} = \{ f_{j,n} \} : n = 1, N; \]

Where the function \( f_{j,n} \) - measured data on the n-th projection,
\[ \{ p_{j,n} : j = 1, J \} \] - positions of pixels of the n-th projection,
\( \delta \) - projection angle for the n-th projection,
\[ J = \sum_{n} J_{n} \] - total number of pixels on all projections.

Functional

\[ \delta_{j,n} = \sum_{i} \left| f_{j,n} - f_{j,n}(p) \right| \]

will determine the error for the n-th projection after the completion of the next iteration.

The average value of the error functional is

\[ \delta_{\text{avg}} = \sum_{n} \delta_{j,n} / N. \]

We present in Table 2 the average value of the error functional for each iteration of the iterative reconstruction process for each of the 20 variants of partitioning into ordered subsets of projections. In total, 7 iterations were done for each variant of the partition into ordered subsets of projections. The value of the relaxation parameter was chosen to be the same for all 20 reconstructions. Its value was equal to \( \lambda = 0.15 \). In view of the
smallness of the values of the error functional, all values for clarity were multiplied by 1000.0.

Table 2. The mean value of the error functional for 20 variants of partitioning into ordered subsets of projections

<table>
<thead>
<tr>
<th>№ Iter.</th>
<th>Var. №1</th>
<th>Var. №2</th>
<th>Var. №3</th>
<th>Var. №4</th>
<th>Var. №5</th>
<th>Var. №6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>12.67</td>
<td>12.50</td>
<td>13.17</td>
<td>13.24</td>
<td>13.54</td>
<td>13.75</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>№ Iter.</th>
<th>Var. №12</th>
<th>Var. №13</th>
<th>Var. №14</th>
<th>Var. №15</th>
<th>Var. №16</th>
<th>Var. №17</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.</td>
<td>7.974</td>
<td>8.085</td>
<td>8.094</td>
<td>8.056</td>
<td>8.082</td>
<td>8.118</td>
</tr>
<tr>
<td>3.</td>
<td>7.558</td>
<td>7.674</td>
<td>7.696</td>
<td>7.649</td>
<td>7.678</td>
<td>7.719</td>
</tr>
<tr>
<td>4.</td>
<td>7.357</td>
<td>7.480</td>
<td>7.503</td>
<td>7.449</td>
<td>7.491</td>
<td>7.541</td>
</tr>
<tr>
<td>5.</td>
<td>7.245</td>
<td>7.348</td>
<td>7.379</td>
<td>7.320</td>
<td>7.364</td>
<td>7.414</td>
</tr>
<tr>
<td>6.</td>
<td>7.144</td>
<td>7.249</td>
<td>7.283</td>
<td>7.225</td>
<td>7.262</td>
<td>7.310</td>
</tr>
<tr>
<td>7.</td>
<td>7.057</td>
<td>7.169</td>
<td>7.205</td>
<td>7.151</td>
<td>7.175</td>
<td>7.225</td>
</tr>
</tbody>
</table>

5. Conclusions

Considering Table 2, you can pay attention to the fact that the reconstruction error increases, and hence the rate of convergence of the iterative process deteriorates with increasing number of subsets. It follows that when using the algebraic reconstruction method with simultaneous iterations of SART (OS-SART) to achieve optimal reconstruction, it is necessary that the number of ordered subsets be less than some optimal number, and the number of iterations exceeds a certain predetermined boundary. Thus, we came to the conclusion that the special architecture and model of GPU programming adds some limitations to real-time performance when using ordered subsets, countering the increase in performance by using small subsets that was previously observed when executing sequential algorithms on a conventional CPU. This feature leads to the appearance of new regularities for determining the optimal number of subsets, as well as to new ways of adjusting the relaxation coefficient to obtain the shortest possible reconstruction time.

References