PERFORMANCE ANALYSIS OF MORPHOLOGICAL OPERATIONS IN CPU AND GPU FOR ACCELERATING DIGITAL IMAGE APPLICATIONS

T. Kalaiselvi, P. Sriramakrishnan and K. Somasundaram

Department of Computer Science and Applications, Gandhigram Rural Institute –Deemed University, Gandhigram

ABSTRACT

In this paper, we evaluate the performance of morphological operations in central processing unit (CPU) and graphics processing unit (GPU) on various sizes of image and structuring element. The languages selected for algorithm implementation are C++, Matlab for CPU and CUDA for GPU. The parallel programming approach using threads for image analysis is done on basic entities of images. The morphological operations namely dilation and erosion are purely depends upon local neighborhood information of each pixel and thus independent. GPU capable to create more number of threads. Here thread per pixel of the image is created to execute the algorithms with the neighbors of relative pixels. Finally the speed performance of all algorithms on conventional processor CPU and parallel processor GPU are computed and compared. Dilation operation in GPU is up to 5 times faster than CPU C++ code and up to 4 - 11 times faster than CPU MATLAB code, likewise Erosion operation in GPU is up to 2 times faster than CPU C++ code and up to 6-12 times faster than CPU MATLAB code when image size varying from 256 × 256 to 1024 × 1024. Further it shows that the performance of GPU implementation is gearing up when the image size is increased. While changing the structuring element size with 1024 × 1024 image the Dilation operation in GPU is up to 3-5 times faster than CPU C++ code and up to 10 - 35 times faster than CPU MATLAB code, likewise Erosion operation in GPU is up to 2 – 6 times faster than CPU C++ code and up to 12-46 times faster than CPU MATLAB code.

KEYWORDS

Graphics processing unit, CUDA, Dilation and Erosion, Parallel processing, Morphological operations

1. INTRODUCTION

Graphics processing unit (GPU) is a multi-core computer chip that performs rapid mathematical calculations, allowing very efficient manipulation of large blocks of data, primarily for rendering images, animations and video for the computer’s screen [1]. A GPU is able to render images more quickly than a central processing unit (CPU) as it performs multiple calculations at the same time using its parallel processing architecture. It is also used for smooth decoding and rendering of 3D animations and video. GPU’s have lot of computational hardware resources. Most of the time the large resource of GPU are unused when graphics applications like games are not run. Computation using GPU has a huge edge over CPU in speed. Hence it is one of the most interesting areas of research in the field of modern industrial research and development for accelerating computing speed. Image processing algorithms in general are good candidates for exploiting GPU capabilities. The parallelization is naturally provided by basic entity of an image.
To process these images, parallel processing capabilities in GPU can be used. The most common GPU programming model is compute unified device architecture (CUDA).

Ghorpade et al., [2] gave an analysis on the architecture of CUDA, CPU and GPU performance. In 2014, Thurley and Danell [3] worked in open source CUDA with morphological operations. They tested 8 bit and 32 bit images with morphological operations in CUDA. They realized CUDA-GPU does not have sufficient shared memory for process the 32 bit images. Jagannathan et al., [4] implemented vHGW algorithm for dilation and erosion independent of structuring element size has been implemented for different types of structuring elements of an arbitrary length and along arbitrary angle on CUDA programming environment with GPU hardware as GeForce GTX 480. They showed the results with maximum performance gain of 20 times than the conventional serial implementation of algorithm in terms of execution time. Harish and Narayanan [5] has implemented breadth first shortest path and single source shortest path algorithm in GPU. Their result showed that GPU is 20-50 times faster than CPU.

In MRI processing, Somasundaram and Kalaiselvi [6] showed that it requires more time to implement an algorithm for automatic segmentation of MRI scans of human brain in CPU. They realized CPU computation power not sufficient to process the large volume of medical data with complex algorithms. Yadav et al., [7] implemented texture based similarity function using GPU. They found that GPU is 30 times faster than CPU. ELEKS [8] has developed post processing of MRI images using GPU. Additionally they used parallel imaging methods to reduce the scan time in singular value decomposition (SVD). They achieved 155X than CPU. In 2015, Jing et al., [9] developed a fast parallel implementation of group independent component analysis (PGICA) for functional magnetic resonance imaging (FMRI). This proposed work demonstrated the speed accuracy of their experiments. But they realized the device memory constraints for large amounts of subject’s data processing.

In this paper, we evaluate the performance of morphological operations in CPU and GPU. The languages selected for algorithm implementation are C++, Matlab for CPU and CUDA for GPU. The parallel programming approach using threads for image analysis is done on basic entities of images. This analysis is based on the nature of information required to process the image. Local processing includes a single pixel and its neighbors. Dilation and erosion comes under this category, because these algorithms purely depend upon the local neighborhood information of each pixel and thus independent. A thread for every pixel of the image is created to execute the algorithms with the neighbors of relative pixels. Here the number of threads creation depends on the nature of processor architecture. GPU has more cores and threads then conventional CPU. When this GPU model invoked with the possible threads, it’s execute the threads in parallel and produce the results for the images quickly. Finally the speed performance of all algorithms on conventional processor CPU and parallel processor GPU on various size of images and structuring elements are computed and compared.

The remaining part of this paper is organized as follows. Section 2 describes the feature of GPU-CUDA programming then section 3 explains the implementation details of the morphological operations in CPU and GPU. Section 4 gives experimental results and discussion and section 5 concludes the paper.

2. GPU-CUDA PROGRAMMING

There are varieties of GPU programming models available in the market for accelerating computational capability of a complex system. CUDA is one such model. It is a parallel computing platform and programming model introduced by NVIDIA in late 2006. NVIDIA
released the documentation about CUDA programming. CUDA is an open source and minimal extension of the C and C++ programming languages [10]. CUDA is a new hardware and software architecture for issuing and managing computations on the GPU as a data-parallel computing device without the need of mapping them to a graphics API. Other GPU programming languages are openCL (open Computing Library) and DirectCompute.

2.1. CUDA Programming Model

For the programmer the CUDA model is a collection of threads running in parallel and its architecture as shown in Figure 1. A kernel is a function or routine that executes in parallel across a set of parallel threads. The programmer organizes these threads into a hierarchy of grids of thread blocks. A thread block is a set of concurrent threads that can cooperate among themselves through barrier synchronization and shared access to a memory space private to the block. A grid is a set of thread blocks that may each be executed independently and thus may execute in parallel. When invoking a kernel, the programmer specifies the number of threads per block and the number of blocks making up the grid. Each thread is given a unique thread ID number threadIdx within its thread block, numbered 0, 1, 2, ..., blockDim–1, and each thread block is given a unique block ID number blockIdx within its grid. CUDA supports thread blocks containing up to 1024 threads. For convenience, thread blocks and grids may have one, two, or three dimensions, accessed via .x, .y, and .z index fields.

![Figure 1: Programming model of GPU- CUDA](image)

2.2. CUDA Memory Model

The GPU-CUDA memory model is given in Figure 2. Threads may access data from multiple memory spaces during their execution. Each thread has a private local memory. CUDA uses this memory for thread-private variables that do not fit in the thread’s registers, as well as for stack frames and register spilling. Each thread block has a shared memory visible to all threads of the
block that has the same lifetime as the block. Finally, all threads have access to the same global memory. GPU have some special set of memories are cache, texture, constant memory [11]. GPU have small cache memory compared than CPU. Texture memory specililzed in fetching and cacheing data from 2D and 3D textures. Constant memory useful for storing data that remains unchanged. GPU have faster memory hierarchical as follows registers, shared memory, cache memory, constant memory, global memory and finally host memory (CPU). The execution flow of GPU-CUDA is a five step process given in Figure 3.

Figure 2: Memory model of GPU- CUDA

Figure 3. Execution Flow of GPU- CUDA
3. MORPHOLOGICAL OPERATIONS

Morphological operations are methods for processing images based on their shapes [12] [13]. It is very often used in applications where the shape of the object is an issue. In mathematical morphology, shape is controlled by “structuring element”. The structuring element determines the precise details of the effect of the operators on the image. The structuring element consists of a pattern, specified as the coordinates of a number of discrete points relative to some origin. The structuring element shape may vary based on the applications [14]. Some different shapes of structuring elements are given in Figure 4. They are diamond, disk, octagon, square, arbitrary and line. The rounded cell in the structuring element denoted as origin. The changes in the size or shape of the structuring element gives reflection in the output image.

![Figure 4. Types of structuring elements (a) Diamond (b) Disk (c) Octagon (d) Square (e) Arbitrary (f) Line](image)

The primary morphological operations are dilation and erosion. Dilation and erosion are dual operations in that they have opposite effects in the resulting image. Dilation adds pixels to the boundaries of the objects while erosion removes a larger of pixels on the object boundaries. Neither erosion nor dilation is an invertible transformation. The implementation details of morphological operations in CPU and GPU is given below.

**Dilation**

Dilation is the operation used to extent the thickness is controlled by a shape referred by various structuring elements. In mathematically, dilation \(\ominus\) combines two sets using vector addition [15]. The two sets referred to be image and corresponding structuring elements. The dilation \(X \ominus B\) is
the point set of all possible vector additions of pairs of elements, one from each of the sets X and B,

\[ X \oplus B = \{ p \in \mathbb{R}^2 : p = x + b, \ x \in X \text{ and } b \in B \} \] (1)

where, X is the image with M × N size and B is the structuring element. It is typically applied to binary images, but work also on grayscale images. The basic effect of the operator on a binary image is to gradually enlarge the boundaries of regions of foreground pixels. Thus areas of foreground pixels grow in size while holes within those regions become smaller.

**Erosion**

Erosion is an operation used to shrink the object of the binary image. In mathematically, Erosion \( \ominus \) combines two sets using vector subtraction of set elements [15]. The erosion \( X \ominus B \) is the point set of all possible vector subtractions of pairs of elements, one from each of the sets X and B

\[ X \ominus B = \{ p \in \mathbb{R}^2 : p = x + b \in X \text{ for every } b \in B \} \] (2)

where, X is the image with M × N size and B is the structuring element or kernel. The basic effect of the operator on a binary image is to erode away the boundaries of regions of foreground pixels. Thus areas of foreground pixels shrink in size, and holes within those areas become larger.

**CPU Implementation**

Morphological CPU implementation done by using C++ and Matlab. Matlab is an interpreted language so it is usually slower than compiled languages like C and C++. That is the reason behind we take this two languages for CPU implementation. This operations implemented in CPU using conventional single threaded approach with the configurations of CPU as given below:

- **Processors Name**: Intel - i5 2500
- **Speed**: 2.9 GHz
- **RAM**: 4 GB
- **Languages**: Matlab and Visual C++

**GPU Implementation**

The GPU implementation of morphological operations done by CUDA. CUDA is capable to create more number of threads per image. Here we can create thread per pixel and each thread have unique thread id [16]. Each thread executes the kernel code simultaneously and produce the output. CUDA 7.0 is used for this implementation. GPU hardware Nvidia Quadro K5000 is used and its complete specification is given below:

- **Processors name**: NVIDIA Quadro K5000
- **Speed**: 1.4GHz
- **Multiprocessor count**: 8
- **Number of cores**: 1536 (8 x 192)
- **Memory**: 4GB
- **Threads**: 1024 per Block
CUDA_Dilation

Read a binary image size of $\text{ROW} \times \text{COL}$ and store the two dimensional values to one dimensional $\text{IN}$ input array row wise. Create same size array $\text{OUT}$ to store the resultant image. Copy binary values of $\text{IN}$ to $\text{OUT}$ array. Define $M \times N$ structuring elements $\text{SE}$ for further process. Allocate $\text{IN}$, $\text{OUT}$ arrays in GPU memory. Transfer all data from CPU to GPU. Create $\text{ROW} \times \text{COL}$ threads and call kernel function. Execute all threads parallel using thread id. If value of $\text{SE}$ array equal to 1 then corresponding pixel positions of $\text{OUT}$ array is set to one. Get back resultant $\text{OUT}$ array from GPU to CPU. CUDA implementation of dilation is given in Algorithm 1.

Algorithm 1: CUDA_Dilation with $M \times N$ Structuring Element

```c
CUDA_DILATION (IN, OUT, SE)

Create a one dimensional intensity arrays $\text{IN}$ and $\text{OUT}$ with the size of $\text{ROW} \times \text{COL}$
Read a binary image with $\text{ROW} \times \text{COL}$ pixels and copy to $\text{IN}$ in row wise
Define $M \times N$ structuring element $\text{SE}$
Transfer the $\text{IN}$, $\text{OUT}$, $\text{SE}$ array from host to device memory
Create $\text{ROW} \times \text{COL}$ threads and call kernel for parallel execution
for all $\text{ROW} \times \text{COL}$ do in parallel
  tid = get thread id
  if $\text{tid} < \text{ROW} \times \text{COL}$ then do
    if $\text{IN}[\text{tid}]$ equal to 1 then
      if $\text{SE}[0]$ equal to 1 then
        $\text{OUT}[\text{tid} - \text{COL}-1] = 1$;
      end
      if $\text{SE}[1]$ equal to 1 then
        $\text{OUT}[\text{tid} - \text{COL}] = 1$;
      end
      if $\text{SE}[2]$ equal to 1 then
        $\text{OUT}[\text{tid} - \text{COL} + 1] = 1$;
      end
      if $\text{SE}[3]$ equal to 1 then
        $\text{OUT}[\text{tid}-1] = 1$;
      end
      if $\text{SE}[4]$ equal to 1 then
        $\text{OUT}[\text{tid}] = 1$;
      end
      if $\text{SE}[5]$ equal to 1 then
        $\text{OUT}[\text{tid}+1] = 1$;
      end
      if $\text{SE}[6]$ equal to 1 then
        $\text{OUT}[\text{tid} + \text{COL}-1] = 1$;
    end
  end
```
end
if SE [7] equal to 1 then
    OUT [tid + COL] = 1;
end
if SE [8] equal to 1 then
    OUT [tid + COL + 1] = 1;
end
end
end
Transfer OUT array from device to host memory.

CUDA Erosion

Create ROW x COL size of one dimensional IN array. Read a binary image with the size of ROW x COL to IN row wise. Create same size OUT array for resultant image. Copy binary values of IN to OUT array. Define M x N structuring elements SE with all 1’s and Mat array for further processing. Allocate same size memory to IN, OUT, SE and Mat arrays in GPU. Transfer the data from CPU memory to GPU memory. Create ROW x COL threads and call kernel function. Execute all threads parallel using thread id. If the value of IN array to thread id equal to 1 then assign corresponding M x N neighbouring pixels to Mat array. Next compare Mat with SE if each value equals then set OUT [tid] to zero. Finally copy the resultant OUT array from GPU to CPU. CUDA implementation of erosion is given in Algorithm 2.

Algorithm 2: CUDA Erosion with M x N Structuring Element

CUDA_EROSION (IN, OUT, SE, Mat)

Create a one dimensional intensity arrays IN and OUT with the size of ROW x COL
Read a binary image with ROW x COL pixels and copy to IN in row wise
Define M x N structuring element SE and initialize Mat array with zero
Transfer the IN, OUT, SE, Mat arrays from host to device
Create ROW x COL threads and call kernel for parallel execution
for all ROW x COL do in parallel
    tid = get thread id();
    if tid < ROW x COL then do
        if IN [tid] equal to 1 then
            assign values (IN [tid - COL-1], IN [tid - COL], IN [tid - COL + 1], IN [tid - 1], IN [tid], IN [tid+1],
            IN [tid + COL-1], IN [tid + COL], IN [tid+COL+1]) to Mat
            if Mat equal to SE then do
                OUT [tid] =0;
            end
        end
    end
end
Transfer OUT array from GPU device memory to CPU host memory

4. RESULTS AND DISCUSSION

The qualitative results of dilation and erosion using 3 x 3 structuring elements on a sample image as shown in Figure 5. We carried out GPU implementation in Algorithm 1 and 2. The square shaped structuring element (Figure 4 d) used for this CPU and GPU implementation. The execution time taken by morphological operations are recorded and are given in Table 1. Table 1
shows the time taken by each processor for performing pixel based local approaches namely dilation and erosion. Here the performance is given by three sizes of image 256 × 256, 512 × 512 and 1024 × 1024 on 3 × 3 square shaped structuring element. Computing times of CPU implementation algorithms in C++ (A) and Matlab (B) also compared with GPU (C). For comparison, results obtained in this proposed work by GPU- CUDA are shown in the Figure 6.

Table 1. Execution time for CPU and GPU implementation using 3 × 3 square shaped structuring element on various size of images

<table>
<thead>
<tr>
<th>Operations</th>
<th>Image size = 256 × 256</th>
<th>= 512 × 512</th>
<th>= 1024 × 1024</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Processor</td>
<td>Speed up</td>
<td>Processor</td>
</tr>
<tr>
<td>CPU + Matlab (B)</td>
<td>CPU (C)</td>
<td>B/C</td>
<td>CPU + Matlab (B)</td>
</tr>
<tr>
<td>Dilation</td>
<td>0.001</td>
<td>0.00349</td>
<td>0.00073</td>
</tr>
<tr>
<td>Erosion</td>
<td>0.001</td>
<td>0.00650</td>
<td>7</td>
</tr>
</tbody>
</table>

Figure 5. Morphological operations using 3 × 3 square shaped structuring element (a) Sample image (b) Dilation (c) Erosion
Figure 6. Results for CPU and GPU implementation a) Dilation b) Erosion on various sizes of images using 3 × 3 structuring element

The experiment of morphological operations done on various sizes of structuring element on fixed image size. Table 2 shows the time taken by various sizes of square shaped structuring element on image size 1024 × 1024. Computing times of CPU implementation algorithms in C++ (A) and Matlab (B) also compared with GPU (C). For comparison, results obtained in this proposed work by GPU- CUDA are shown in the Figure 7.

From Table 1 and Figure 6, we observed that GPU implementation can access every pixel in parallel and handling the ROW × COL number of threads as independent process. GPU is 4 - 10 times faster than CPU Matlab code and up to 3 times faster than CPU C++ code in dilation operation when image sizes are different on static size structuring element. Erosion gives 6 – 12 times faster than CPU Matlab and up to 2 times better than CPU C++. Further it shows that the performance of GPU implementation is gearing up when the image size is increased.

From Table 2 and Figure 7, CPU C++ and GPU are almost static time taken when handling various sizes of structuring element on same size image. But Matlab taking more time for various sizes of structuring elements. GPU in dilation gives 3 – 5 times faster than C++ and 10 – 35 times faster than Matlab. GPU in erosion gives 2 – 6 times faster than C++ and 12 – 46 times faster than Matlab. GPU is much faster than Matlab compared with C++ because Matlab is an interpreted language so its usually slower than C++. More number of data parallelization gives attractive results. GPU not suitable for very less number of threads. Here we used light weight mathematical calculations and thus GPU gives very less speedup (up to 46 times). If we use heavy complex mathematical calculations like brain tumour extraction and visualization then GPU will give better results.
Table 2. Execution time for CPU and GPU implementation on 1024 × 1024 size image using various size of structuring element.

<table>
<thead>
<tr>
<th>Operation</th>
<th>Image 1024 × 1024</th>
</tr>
</thead>
<tbody>
<tr>
<td>Structuring element</td>
<td>3 × 3</td>
</tr>
<tr>
<td><strong>Execution Time in sec</strong></td>
<td></td>
</tr>
<tr>
<td><strong>CPU</strong></td>
<td><strong>GPU</strong></td>
</tr>
<tr>
<td>C++ + (A)</td>
<td>Matlab (B)</td>
</tr>
<tr>
<td><strong>Dilation</strong></td>
<td>0.016</td>
</tr>
<tr>
<td><strong>Erosion</strong></td>
<td>0.015</td>
</tr>
</tbody>
</table>

Figure 7. Results for CPU and GPU implementation a) Dilation b) Erosion on various sizes of structuring element on image size 1024 × 1024.

5. CONCLUSION

In this work, we have implemented morphological operations using GPU - CUDA model. Independent pixels can be accessed parallel for morphological operations. The GPU based coding yielded speedup values in the order of 46 times compared to conventional processor CPU with its conventional single threaded approach. Further it showed that the performance of GPU implementation is gearing up when the image size as well as structuring element size is increased.
ACKNOWLEDGEMENTS

We gratefully acknowledge the support of NVIDIA Corporation Private Ltd, USA with the donation of the Quadro K5000 GPU used for this research.

REFERENCES

Authors

T. Kalaiselvi is currently working as an Assistant Professor in Department of Computer Science and applications, Gandhigram Rural Institute, Dindigul, Tamilnadu, India. She received her Bachelor of Science (B.Sc.) degree in Mathematics and Physics in 1994 & Master of Computer Applications (M.C.A) degree in 1997 from Avinashilingam University, Coimbatore, Tamilnadu, India. She received her Ph.D. degree from Gandhigram Rural University in February 2010. She has completed a DST sponsored project under Young Scientist Scheme. She was a PDF in the same department during 2010-2011. An Android based application developed based on her research work has won First Position in National Student Research Convention, ANVESAN-2013, organized by Association of Indian Universities (AUI), New Delhi, under Health Sciences Category. Her research focuses on MRI of Human Brain Image Analysis to enrich the Computer Aided Diagnostic process, Telemedicine and Teleradiology Technologies.

P. Sriramakrishnan received his Bachelor of Science (B.Sc.) degree in 2011 from Bharathidasan University, Trichy, Tamilnadu, India. He received Master of Computer Application (M.C.A) degree in 2014 from Gandhigram Rural Institute-Deemed University, Dindigul, Tamilnadu, India. He worked as Software Developer in Dhvani Research and Development Pvt. Ltd, Indian Institute of Technology Madras Research Park, Chennai during May 2014 – March 2015. He is currently pursuing Ph.D. degree in Gandhigram Rural Institute-Deemed University. His research focuses on Medical Image Processing and Parallel Computing. He has qualified UGC-NET for Lectureship in June 2015.

K. Somasundaram received his Master of Science (M. Sc) degree in Physics from the University of Madras, Chennai, India in 1976, the Post Graduate Diploma in Computer Methods from Madurai Kamaraj University, Madurai, India in 1989 and the Ph.D degree in theoretical Physics from Indian Institute of Science, Bangalore, India in 1984. He is presently working as Professor at the Department of Computer Science and Applications, Gandhigram Rural Institute, Dindigul, India. From 1976 to 1989, he was a Professor with the Department of Physics at the same Institute. He was senior Research Fellow of Council Scientific and Industrial Research (CSIR) Govt. of India, in 1983. He was previously a Researcher at the International Centre for Theoretical Physics, Trieste, Italy and Development Fellow of Commonwealth Universities, at Edith Cowan University, Perth, Australia. His research interests are image processing, image compression and Medical imaging. He is Life member of Indian Society for Technical Education, India and Life member in Telemedicine society of India. He is also a member of IEEE USA.