Organization of High-Performance Parallel-Hierarchical Computing Processes for Classification of Laser Beam Images

Andriy A. Yarovyy
Department for Computer Science
Vinnytsia National Technical University (VNTU)
Vinnytsia, Ukraine
axa@vinnitsa.com

Leonid I. Timchenko
Department for telecommunication technologies and automatics
State University for Transport Economy and Technologies (SUTET)
Kyiv, Ukraine	timchen@list.ru

Nataliya I. Kokriatskaia
Department for telecommunication technologies and automatics
State University for Transport Economy and Technologies
Kyiv, Ukraine
kokriatskaia@rambler.ru

Svitlana V. Nakonechna
Department for telecommunication technologies and automatics
State University for Transport Economy and Technologies
Kyiv, Ukraine

Maksym S. Mateichuk
Department for Computer Science
Vinnytsia National Technical University
Vinnytsia, Ukraine

Abstract—The paper deals with a problem of insufficient productivity of existing computer means for large image processing, which do not meet modern requirements posed by resource-intensive computing tasks of laser beam profiling. Development of a theory of parallel-hierarchic transformation allowed to produce models for high-performance parallel-hierarchical processes, as well as algorithms and software for their implementation based on the GPU-oriented architecture. The analyzed performance of suggested computerized tools for processing and classification of laser beam profile images allows to perform real-time processing of dynamic images of various sizes.

Keywords—laser beams; image processing; parallel processing; GPGPU; video card programming; parallel-hierarchical transformation.

I. INTRODUCTION

A swift transit of modern computing and control systems to digital standards resulted in a need for fast processing of huge amounts of information. Researches in this area and their results are especially important for systems performing complex signal processing, such as filtration, classification and recognition of dynamic video data, which require high-performance computing systems.

Propagation of the laser radiation in the air is distorted by numerous linear and non-linear effects, and none of those effects can be actually singled out. They fall into the following main groups: laser beam refraction; absorption of laser beam energy with atmospheric gases; dissipation of laser beam energy by aerosol particles on air density fluctuations and so on; and fluctuations of laser beam parameters caused by air turbulence. Each of those groups of effects can be observed in the areas of both linear and non-linear optics. At the same time, each of those groups has some special features that should be taken into account in the respective theoretical and experimental studies [1]-[3].

Currently, a sharp need in broader implementation of high-performance computing systems with automatic error correction of the formed light radiation exists in such technological areas as laser material treatment, laser location, optical communication, printing, and many others. Those errors may be caused by destabilizing mechanic or climatic factors, unstable characteristics of the source of radiation, obstacles in the optical path, and so on. To ensure acceptable quality of correction, characteristics of light radiation (for instance, its profile, spatial intensity distribution, as well as deviation of this distribution from the outgoing or reference distribution) should be permanently controlled in a dynamic regime [2]-[4].

II. DESCRIPTION OF THE RESEARCH PROBLEM

Our research was concentrated on one of the profiling problems, namely, real-time processing of spot images of the laser beam profile (obtained from a video route). It is based on the known methods of moment characteristics, approximation, extrapolation and so on [5],[6]. At the same time, this task is a complex computational process organization that includes
displaying results of image identification and classification, as well as forming and processing vast arrays of digital information, in particular, containing energy center coordinates. Those results are an important component of the process of laser beam profiling. Finally, this complex scientific research is being conducted for a long time in cooperation with the Scientific Development and Production Center “Astrophysica” and CISC KIA Systems (Moscow, Russia), which provided real video routes of laser beam images for the research. Those institutions are involved in the high-level research and development in the area of laser systems, laser beam profiling, optical celestial polarimetry, and some others [3],[6]-[8].

Therefore, research and development focused on the precise real-time measurement of laser beam profile characteristics and coordinates of laser route images are important for various applied problems. An interest to this problem is also high in the USA, where similar scientific research and hardware and software development is conducted by such leading corporations as Ophir-Spiricon Inc., Coherent Auburn Group, and in Europe, with its MS MacroSystems (Netherlands) and PhaseView (France) [9]-[11].

To analyze a sequence of quickly changing spot images (obtained during laser video route processing), a center of the laser beam passage should be identified. Due to various atmospheric effects, images of the laser beam video route section (in the plain perpendicular to the direction of the laser beam propagation) become blurred, with a constantly changing ‘uneven’ shape [4]-[5],[7],[12]. That is why, without having analyzed previous images in the sequence \( \Omega = (\omega_0, \ldots, \omega_{N-1}) \) of images of a given laser video route, it is almost impossible to locate an energy center (EC) \( \alpha'_i = (x'_i, y'_i) \) of its current \( (i^{th}) \) spot image, which would remain constant during the sequence analysis. The EC should not shift during an analysis of a laser video route that does not explicitly change its direction. Therefore, a center of such laser video route remains constant, i.e., \( \alpha'_i = \alpha'_{i-1} \), \( i = 1, N - 1 \), where \( N \) is a number of images in the laser video route.

If a propagation direction changes, the computed EC of the laser beam profile spot image should automatically shift and adapt to the new location and characteristics of the beam.

The following approaches and methods are used to analyze such images: detecting the image center using absolute moments of the 0th and 1st orders, finding EC coordinates of the laser beam profile using the approximation of boundaries, neural network-based classification, and some others. However, all those methods cannot process high-resolution laser beam profile images fast enough [5]-[6],[13]. That is why it is so important to develop new approaches to organization of high-performance parallel-hierarchical computing processes for classification of laser beam profile images.

III. ALGORITHMS AND SOFTWARE FOR PARALLEL-HIERARCHICAL TRANSFORMATION OF SPOT IMAGES BASED ON THE GPU-ORIENTED ARCHITECTURE

The network method of the direct parallel-hierarchical transformation (PHT) with formation of the parallel-hierarchical computing system (PHCS) described in details in [14]-[15] consists in one-time consequential application of operators of G-transformation and transposition \( (T) \) to initial sets \( \bigcup_{S \in \mu_{S}} \) with further \( (k-1) \) applications of functional \( \Phi \):

\[
\Phi \left[ \begin{array}{c}
\begin{array}{c}
T \left( G \left( \bigcup_{S \in \mu_{S}} \mu_{S} \right) \right)
\end{array}
\end{array} \right] = \bigcup_{i=1}^{\delta} a_{i1}^{\delta}
\]

where \( \mu_{S} \) is an initial set \( (S=1,2,3, \ldots) \), \( a_{i1}^{\delta} \) is an element of the initial set decomposition, with one element being obtained at each level, starting from the second one.

Let us discuss an implementation of the PHT of spot images of laser beam profile based on the GPU-oriented architecture with their further classification [16].

According to the mathematic model of the PHT for image processing described in [17], and based on the recognition method in the PHCS that uses formation of a normalizing equation, an algorithm was developed that implements the PHT and image classification based on parallel GPU systems. The following main stages are implemented in the algorithm:

1. Download an image of \( n \times m \) pixels, where \( n \) is its height, and \( m \) is its width, to the memory of the computing system.

2. In order to accelerate computations and to optimize intermediate PHT results, only data on each non-zero pixel are stored. Those data are stored in the program structure that contains three fields: value (val), row (r), and column (c):

Finally, complete content and organizational editing before formatting. Please take note of the following items when proofreading spelling and grammar:

struct elem {
    int val, r, c;
};

3. Introduce another program structure with dimensions \( [n \times m] \) that contains a maximal possible number of non-zero elements of the image matrix, and then fill it with data on each non-zero element.

4. Introduce two variables \( L \) and \( R \), left and right boundaries respectively. We will then work with our data array within those variables. Initialize those variables \( (L=0, \text{and } R \text{ is equal to a number of non-zero elements}) \).

According to expression (1), the PHT is based on three operations: transposition \( (T) \), G-transformation \( (G) \), and shift \( (S) \). Let us examine their GPU-based implementation.

4.1. Transposition \( (T) \).
In the data array, transpose row \( r \) and column \( c \) values within \( [L, R) \) limits in parallel for each element. Values \( n \) and \( m \) are transposed respectively. The obtained data are transposed; however, the array should be arranged within \( [L, R) \) limits, because a situation is possible when an order of data in the array is wrong. The array is then arranged within \( [L, R) \) limits using the following logic: element \( A \) precedes element \( B \), if a row value of element \( A \) is smaller than a row value of element \( B \). If those values are equal, \( A \) precedes \( B \), if a column value for \( A \) is smaller than a column value for \( B \). In parallel, an operation of sorting is performed using a program library method \( \text{thrust::sort}() \) and a comparator that we created.

4.2. G-transformation (G).

Arrange the array within \( [L, R) \) limits applying the logic described above to the matrix rows and performing a sorting procedure. This way all matrix rows will be arranged in an ascending order. In parallel, for each matrix row, calculate a number of elements it contains. Introduce a variable \( \text{sum} = 0 \).

In parallel, perform calculations for each matrix row that produce new elements in the transformed matrix. They are determined as a product of a matrix element by a number of non-zero elements, i.e.:

\[
\text{inttmp}_\text{val} = (\text{data}[\text{cur}].\text{val} - \text{sum}) \times (\text{row}_\text{cnt}[i] - j).
\]

If the transformation yields a non-zero element, this element is written into the matrix row and, respectively, a sum value is changed:

\[
\text{sum} += (\text{data}[\text{cur}].\text{val} - \text{sum}).
\]

Here, variable \( R \) is equal to a sum of variable \( L \) and a number of non-zero elements produced, and \( m \) is equal, respectively, to the maximal number of non-zero elements in the newly produced matrix rows. A procedure of finding a maximal value of the element in the array is performed in parallel, using the program library method \( \text{thrust::reduce}() \).

4.3. Shift (S).

Assign a value to variable \( L = L + 1 \). In the data array, for each element within \( [L, R) \) limits, add a column value to a row value and distract 1 (indexation with 0). This way, a shift of the \( i \)-th order is performed. An addition procedure is performed simultaneously for each cell of the data array. In parallel, find a maximal value of the column in the resulting data array within \( [L, R) \) limits, and assign \( m \) to this value using a program library method \( \text{thrust::reduce}() \).

5. Finally, copy obtained and downloaded data to the GPU, and perform a direct PHT. The PHT is performed until the processing yields a single element, i.e. until \( L < R - 1 \). Elements within \( [0, L) \) limits are tail elements (a result of the direct PHT).

6. To classify spot images of the laser beam profile, we use a recognition method in the PHCS based on formation of a normalizing equation. Using a fact that a sum of the PHCS head elements \( \sum_i a_i \) is invariant to the sum of tail elements \( \sum_i a_i \), we compose a system of equations to obtain tuning coefficients \( w_i + w_{k-1} \) in the form (2), which will allow to form normalizing equation (3):

\[
\begin{align*}
\sum_{i=2}^{k} a_i^{11} & = \frac{\sum_{i=2}^{k} a_i^{11} + \sum_{i=2}^{k} a_i^{j}}{(a_1^{11} + \sum_{i=2}^{k} a_i^{j})} \\
\sum_{i=2}^{k} a_i^{2} & = \frac{\sum_{i=2}^{k} a_i^{2} + \sum_{i=2}^{k} a_i^{j}}{(a_1^{2} + \sum_{i=2}^{k} a_i^{j})} \\
\sum_{i=2}^{k} a_i^{k-2} & = \frac{\sum_{i=2}^{k} a_i^{k-2} + \sum_{i=2}^{k} a_i^{j}}{(a_1^{k-2} + \sum_{i=2}^{k} a_i^{j})} \\
\sum_{i=2}^{k} a_i^{k-1} & = \frac{\sum_{i=2}^{k} a_i^{k-1} + \sum_{i=2}^{k} a_i^{j}}{(a_1^{k-1} + \sum_{i=2}^{k} a_i^{j})}
\end{align*}
\]

In (2) and (3), \( \sum_{i=2}^{k} a_i^{11} \) are reference elements, and \( \sum_{i=2}^{k} a_i^{j} \), \( a_i^{2} + a_i^{k-1} \) are current elements (characteristics) of the recognized image [14].

To normalize PHCS results with the tuning coefficients obtained from the coefficients in form (2), we will use the fact that in the PHCS, \( \sum_{i=2}^{k} a_i^{11} = \sum_{i=2}^{k} a_i^{j} \) [14]. Then the left part of normalizing equation (3) presents a ratio of the sum of products of average tuning coefficients and tail elements to the sum of tail elements of the network, and its right part in case of the correct recognition is approaching 1, or \( d \to 1 \). A degree of this closeness to 1 reflects a measure of similarity of recognized images, and in an ideal case of correct recognition \( d = 1 \).

\[
d = \frac{\sum_{i=2}^{k} a_i^{2} + \sum_{i=2}^{k} a_i^{k-2} + \sum_{i=2}^{k} a_i^{k-1} + \sum_{i=2}^{k} a_i^{11}}{\sum_{i=2}^{k} a_i^{11} + \sum_{i=2}^{k} a_i^{j} + \sum_{i=2}^{k} a_i^{11} + \sum_{i=2}^{k} a_i^{j}}
\]
Using normalizing equation (3), where \( w_1 = w_2 = \cdots = w_{k-2} = w_{k-1} = 1 \), we can easily perform a previous operation of image classification using an expert-defined \( d \), and then form tuning coefficients \( \bar{w}_1 + \bar{w}_{k-1} \) according to system (2), thus completing a procedure of finding weight coefficients for each image class.

In particular, in case of real-time classification and analysis of, say, consecutive frames of the laser video route, normalizing equation (3) takes the following form [14].

\[
\sum_{i=2}^{k} a_{1i}^{j+1} = \sum_{i=2}^{k} a_{1i}^{j+1} + \sum_{i=2}^{k} a_{1i}^{j+1} + \sum_{i=2}^{k} a_{1i}^{j+1} + \sum_{i=2}^{k} a_{1i}^{j+1}
\]

where \( \sum_{i=2}^{k} a_{1i}^{j} \) and \( \sum_{i=2}^{k} a_{1i}^{j+1} \) are, respectively, a sum of tail elements and values of image tail elements of the previous (\( j^{th} \)) and the following (\( j+1^{th} \)) frames, and \( j \) is a frame number, \( j = 1, m-1 \).

Based on the mentioned property of the PHCS, normalizing equation (4) takes form:

\[
d = \left( a_1^{j+1} \right) + \left( a_2^{j+1} \right) + \cdots + \left( a_{N-1}^{j+1} \right) + \left( a_N^{j+1} \right)
\]

In this case, for normalization of processing results of incoming data \( \left( a_1 + a_N \right) \) on the \( j^{th} \) frame, the recognition time in the PHCS may be significantly reduced for the time needed to process this frame.

In processing a sequence of laser video route images, a normalizing equation may be used for two consequential images, with the first image taken as reference. Then the normalizing equation takes the following form:

\[
d = \left( a_1^{j+1} \right) + \left( a_2^{j+1} \right) + \cdots + \left( a_{N-1}^{j+1} \right) + \left( a_N^{j+1} \right)
\]

where \( \sum_{i=2}^{k} a_{1i}^{j} \) and \( \sum_{i=2}^{k} a_{1i}^{j+1} \) are tail elements of the PHT of the current and previous images respectively.

Here we use parallel algorithms to compute a data array sum with a program library method \( \text{thrust::reduce()} \).

Processing of the image consequence yielded the following experimental data (Fig. 1).

Further processing requires selecting only certain values \( d \) that exceed some expert-defined threshold. For instance, for the processed laser video route (Fig. 1), some \( d \)-s accept small values, and their respective starting images may be excluded.

![Figure 1 – Results of classification of the laser beam profile spot images](image)

A comparative analysis shows that suggested methods allow to perform real-time processing and analysis of spot images with more accurate measurement of EC coordinates as a characteristic of the laser beam profile (up to 1.5 pixel). Their accuracy is on average 1.5 times higher than that of other known methods, for instance, based on finding a weight center using moment characteristics. In our experiments, we studied 14 laser video routes, containing 2044 images each. The conducted research shows that due to various destabilizing factors, EC coordinates of laser route images cannot be accurately measured. That is why in this research we suggested a method of coordinate measurement that analyzes a mutual position of two adjacent images of the laser beam profile. Such analysis produces a corrected value of EC coordinates of the current image. Such principle of EC coordinate measurement allows to avoid using inaccurate procedures based on various approximating operators.

Data obtained with repeated processing of spot images of the laser beam profile of various sizes (i.e. resolution) demonstrate promptitude of the PHT and image classification based on the CPU- and GPU-oriented architecture.

A computer system with multicore CPU Intel Xeon E5606 (Clockspeed: 2.1 GHz; No of Cores: 4; Instruction Set: 64-bit) was used as a CPU-oriented architecture. A computer complex with 2 NVIDIA GeForce GTX590 video adapters (each working in 1024 streams) was used as a GPU-oriented architecture.

![Figure 1 – Results of classification of the laser beam profile spot images](image)
architecture. Each of those videoadapters (with theoretic performance of 2488.3 GFLOPS) contains two 512-core GPU GTX 500, which in total comprise a 2048-core hardware. Therefore, a developed high-performance PHCS based on Quad SLI technology that contains two such video adapters allows to process information in 2048 streams with a theoretical performance limit of 4.9766 TFLOPS.

Table 1 and Fig. 2 present experimental results for the total processing time of 25 laser video route frames which meet requirements to the real-time processing, as well as dependence of the PHT performance on the image size in processing a laser video route.

Table 2 and Fig. 3 present results of experimental study of the time used to compute and process 25 laser video route frames (excluding time used for the operation of download to the computing system memory), as well as dependence of the PHT performance on the image size.

IV. CONCLUSIONS

The novelty of the given research is the development of theory of parallel-hierarchical transformation in computer systems, as well methods and models of high-performance PHCS on the base of GPGPU technology.

Therefore, the following conclusions can be made based on the discussed research results:

1) by a criterion of the total processing time, when processing small images (up to 512×512 pixels), performance of the PHT based on the GPU-oriented architecture does not exceed that of the CPU-oriented architecture (Fig. 2). This is explained by a small size of processed image files (.bmp files of no more than 300 Kb each), which can be processed efficiently enough by CPU; in case of the GPU, however, expenses of work with the computing system memory and of the computer process parallelising exceed processing effect.

2) by a criterion of total processing time, when processing large images (up to 4096×4096 pixels), performance of the PHT based on the GPU is more than twice higher than efficiency of the CPU (Fig. 2). This is caused by a bigger size of the processed files with images and bigger size of images themselves (up to 20Mb for a .bmp file), which decreases CPU processing efficiency; however, in case of the GPU, an effect of parallelising computing process exceeds expenses of work with the computing system memory and control.

3) by the criterion of the total processing time, when processing extra-large images (10000×10000 pixels and more, up to 16384×16384 in experiments), performance of the PHT based on the GPU is almost three times higher than performance of the CPU (Fig. 2). This is caused by a significant increase in the size of processed files and image sizes (a size of the .bmp file is about hundreds megabytes, and up to a gigabyte for images of 32768×32768 pixels), which significantly decreases processing efficiency for the CPU; however, for the GPU in this case an effect of parallelisation of the computational process even more exceeds expenses related to work with the computing system memory and control.

4) by a criterion of the computation-only time (without time used for operations of download to the computing system memory), when processing small images (up to 512×512 pixels) the same situation is observed as in the first conclusion: both the CPU, and the GPU provide efficient real-time processing of images. However, with a growing image size (and, respectively, bigger file sizes), the GPU performance substantially increases, with up to 3 times for images of 1024×1024 pixels, up to 6 times for 2048×2048 pixels, more than 13 times for 4096×4096 pixels, over 48 times for 8192×8192 pixels, and more than 60 times higher than CPU performance for processing images of 16384×16384 (Fig. 3).

Therefore, a conclusion can be made that the highest performance is reached when the PHT based on the GPU-oriented architecture is used to process extra-large images, which cannot be attained by classic methods (fast Fourier transform, Walsh-Hadamard, Haar), as well as based on the CPU-oriented architecture.

REFERENCES

Table 1. Experimental results of processing laser video route

<table>
<thead>
<tr>
<th>Size of spot images of the laser beam profile, pixels</th>
<th>Processing time of 25 laser video route images based on the CPU-oriented architecture, ms</th>
<th>Processing time of 25 laser video route images based on the GPU-oriented architecture, ms</th>
</tr>
</thead>
<tbody>
<tr>
<td>128×128</td>
<td>1</td>
<td>47</td>
</tr>
<tr>
<td>256×256</td>
<td>16</td>
<td>47</td>
</tr>
<tr>
<td>512×512</td>
<td>47</td>
<td>62</td>
</tr>
<tr>
<td>1024×1024</td>
<td>156</td>
<td>94</td>
</tr>
<tr>
<td>2048×2048</td>
<td>575</td>
<td>234</td>
</tr>
<tr>
<td>4096×4096</td>
<td>2418</td>
<td>1027</td>
</tr>
<tr>
<td>8192×8192</td>
<td>9641</td>
<td>3479</td>
</tr>
<tr>
<td>16384×16384</td>
<td>39228</td>
<td>13609</td>
</tr>
</tbody>
</table>

Table 2. Experimental results of study of time used for computation of laser video route processing (excluding time used for download into the computing system memory)

<table>
<thead>
<tr>
<th>Size of spot images of the laser beam profile, pixels</th>
<th>Processing time of 25 laser video route images based on the CPU-oriented architecture, ms</th>
<th>Processing time of 25 laser video route images based on the GPU-oriented architecture, ms</th>
</tr>
</thead>
<tbody>
<tr>
<td>128×128</td>
<td>1</td>
<td>37</td>
</tr>
<tr>
<td>256×256</td>
<td>16</td>
<td>41</td>
</tr>
<tr>
<td>512×512</td>
<td>47</td>
<td>46</td>
</tr>
<tr>
<td>1024×1024</td>
<td>156</td>
<td>51</td>
</tr>
<tr>
<td>2048×2048</td>
<td>575</td>
<td>94</td>
</tr>
<tr>
<td>4096×4096</td>
<td>2418</td>
<td>184</td>
</tr>
<tr>
<td>8192×8192</td>
<td>9641</td>
<td>201</td>
</tr>
<tr>
<td>16384×16384</td>
<td>39228</td>
<td>648</td>
</tr>
</tbody>
</table>

Figure 2 – Dependence of the PHT performance on the image size when processing a laser video route

Figure 3 – Dependence of computing speed with the PHT on the image size when processing a laser video route (excluding time used for operations of download to the computing system memory)