GPU computing to data visualization in data mining

Liang Xu
Department of Electrical and Computer Engineering
Virginia Commonwealth University
Richmond, Virginia
Email: xul4@vcu.edu

***Abstract*—This paper survey is the second assignment for CS691 High performance distributed system in VCU. The topic of this survey is applications of GPU computing to data visualization in data mining. A growing number of data scientists are using GPUs for big data analytics to make better, real-time business decisions. More recently, data visualization has coughed attention by data mining field. How to transform data from high dimension to 2D or 3D which can be recognized by human visualizations becomes a very popular research topic. For example, MRI and CT scanners, data become too large to be effectively visualized on standard workstations in commonplace. One solution to this problem is to employ a visualization cluster, a small to medium scale cluster dedicated to performing visualization. The cluster is designed for transform high data to human understandable graph, which GPU has dramatic contribution[1]**

In this paper, we survey the-state-of-the-art algorithms and applications in GPU computing to data visualization in data mining and discuss the future research opportunities.

***Index Terms*—GPU, Data Visualization, Data Mining**

I. Introduction

DATA visualization is human and machine collaboration to mass data where extract the most informative features and it is become an essential component of data analysis. It involves the creation and study of the visual representation of data. A primary goal of data visualization is to communicate information clearly and efficiently via statistical graphics. Data visualization is both an art and a science and it provides a bridge from human understanding to the information from a mass data set. This information includes relations between features, structure of the data and revealing the underling relationship between features.[2], [3].

General-purpose computing on graphics processing units is the use of a graphics processing unit (GPU), which typically handles computation only for computer graphics, to perform computation in applications traditionally handled by the central processing unit (CPU)[4], [5]. GPU is initially designed for game application, however, GPUs demonstrate impressive computing power and high levels of parallelism and are now being used for may applications which not limited by the graph computation. The GPU provides the computing power especially for the algorithm can be parallelize. The GPU assisted algorithms are playing an important role in data vitalization, which include simple matrix manipulation[6], [7], [8], FFT[9], [10], and wavelet transform[11], [12] and so on[13].

It is often desirable to visualize a data set the items of which are described by more than three features. Therefore, we have multidimensional data and our goal is to make some visual insight into the data set analyzed. For human perception, the data must be represented in a low-dimensional space, usually of two or three dimensions. The goal of visualization methods is to represent the multidimensional data in a low-dimensional space so that certain properties of the structure of the data set were preserved as faithfully as possible. Such a visualization of data is highly important in data mining because recent applications produce a large amount of data that require specific means for knowledge discovery. The dimensionality reduction or visualization methods are recent techniques to discover knowledge hidden in multidimensional data sets[14]. Visualization algorithm require extensive compute power relative to data size. One possible solution is to use large scale supercomputer, which has required computational power. One alternative approach is to use a smaller scale computer equipped with GPUs, which can provide the needed computational power [1].

II. latest development

In this section, several latest ideas about using GPU for data visualization in data mining are going to discussed in the following paper.

A. hardware solution

One solution to solve large data visualization is to employ a 'visualization cluster', a small to medium scale cluster dedicated to performing visualization and analysis of massive data sets[1]. These clusters are designed to fit a different need than traditional supercomputer, and therefore their design mandates different hardware choices, such as GPUs. Their system is divided into three stages: 1. An intelligent per-partitioning which is designed to make combining results from different nodes easy. 2, A GPU volume render to perform the per-frame volume rendering work at interactive rates.3, MPI-based compositing based on a sort-last compositing framework. They implement their remote rendering system inside of VisIt(VisIt is an Open Source, interactive, scalable, visualization, animation and analysis tool), which is capable of rendering data in parallel on remote machines. The system is comprised of a 'viewer' client application, connected over TCP to a server which employs GPU cluster nodes.All rendering is performed on the cluster, composited via MPI, and images are sent back to the viewer for display.

This research demonstrate that a multiple GPU mode is a grate foundational building block to compose larger systems capable of rendering very large data. More importantly, the

performance-price ratio of a GPU is higher than CPU based solutions, this work makes the case for spending more visualization supercomputing capital on hardware acceleration, and acquiring smaller yet more performance clusters.

B. algorithm solution

High dimensional data challenge to many data mining and data visualization algorithms. The performance of those algorithms are cursed by dimensionality and high computational cost.

1) matrix multiplication: One of the important computational cost is the matrix multiplication. Fatahalian and Sugerman published one famous paper discussed about matrix and matrix multiplication. [15], and Rech Et.al also discuss about the efficiency about the efficiency of matrix multiplication. Some of the research works are talk about the sparse matrix multiplication[16]. From the result we could found out the performance of GPU. Matrix multiplication is the building block of the high dimensional algorithm.

2) dimension reduction method: It is difficult to perceive the data structure using the direct visualization methods, particularly when we deal with large data sets or data of high dimensionality. Another group of method is based on reduction of dimensionality of data. Their advantage is that each n-dimensional object is represented as a point in the space of low-dimensionality d [14]. Principal Component Analysis(PCA) is a well known data analysis technique. The property of the set is that the largest part of its information content is stored in the first few coordinates (components) of points. In Liu and Zhang's paper, they proposed a large scale PCA method based on GPU. Specifically, They construct parallel implementations of the four optimization formulations for the GPU, and compare this with a CPU implementation . Using real-world data, they experimentally validate the effectiveness of PCA and demonstrate that the parallel GPU implementation of PCA can significantly improve performance. Wavelet decomposition and reconstruction is usually implemented by applying multiple convolution and down-up-sampling steps to the volume data[11], [17], [18]. Franco et.al proposed a research about CUDA implementation for the 2D Fast Wavelet Transform on GPUs. which uses a pair of Quadrature Mirror Filters and provides insight into techniques used to wrench the full performance out of the resources for a large class of important scientific applications. Experimental results show great success for this techniques: Data coalescing saves one-third of the execution time when performing 1D-FWTs over rows and columns, and accessing data in shared memory free of bank conflicts produces a similar benefit.

C. real time data visualization

Some of the research works are focus on the real time applications. Liu Et.al proposed a method called imMeans for real time big data visualization using GPU. They contribute methods for real-time visualization[19]. To enable real-time interaction, they integrate multivariate data tiles and parallel processing using GPU.They implement those methods in imMens, a browser-based system using WebGL. WebGL is currently the only standardized way to access GPU processing in a web browser. Future browser-based GPGPU or parallel computing (multi-core) features will permit alternative approaches. Their general approach of generating multivariate data tiles and leveraging parallel query processing is also directly applicable in standard (non-web) desktop contexts.

III. Conclusion

This literature survey is focus on the data visualization in data mining by using GPU. Since the data visualization involve high dimensional data transform to 2 or 3 dimensional data, which is computational expensive. GPU provide hight computational power can improve the performance. The survey list several latest research topic both from hardware and software solution to improve the performance. It is hoped that by providing a birds eye view of the research field, this paper will offer clear directions for future developments in the area.

Liang Xu Ph.D candidate in Electrical and Computer Engineer Department.

REFERENCES

[1] T. Fogal, H. Childs, S. Shankar, J. Krüger, R. D. Bergeron, and P. Hatcher, "Large data visualization on distributed memory multi-gpu clusters," in *Proceedings of the Conference on High Performance Graphics*. Eurographics Association, 2010, pp. 57–66.

[2] I. Icke and A. Rosenberg, "Multi-objective genetic programming for visual analytics," in *European Conference on Genetic Programming*. Springer, 2011, pp. 322–334.

[3] S. Ramírez-Gallego, S. García, H. Mouriño-Talín, D. Martínez-Rego, V. Bolón-Canedo, A. Alonso-Betanzos, J. M. Benítez, and F. Herrera, "Data discretization: taxonomy and big data challenge," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 6, no. 1, pp. 5–21, 2016.

[4] J. Fung, F. Tang, and S. Mann, "Mediated reality using computer graphics hardware for computer vision," in *Wearable Computers, 2002.(ISWC 2002). Proceedings. Sixth International Symposium on*. IEEE, 2002, pp. 83–89.

[5] NVIDIA, "Cuda c programming guide," 2013.

[6] E. S. Larsen and D. McAllister, "Fast matrix multiplies using graphics hardware," in *Proceedings of the 2001 ACM/IEEE conference on Supercomputing*. ACM, 2001, pp. 55–55.

[7] J. D. Hall, N. A. Carr, and J. C. Hart, "Cache and bandwidth aware matrix multiplication on the gpu," 2003.

[8] J. Bolz, I. Farmer, E. Grinspun, and P. Schröoder, "Sparse matrix solvers on the gpu: conjugate gradients and multigrid," in *ACM Transactions on Graphics (TOG)*, vol. 22, no. 3. ACM, 2003, pp. 917–924.

[9] K. Moreland and E. Angel, "The fft on a gpu," in *Proceedings of the ACM SIGGRAPH/EUROGRAPHICS conference on Graphics hardware*. Eurographics Association, 2003, pp. 112–119.

[10] T. J. B. von Rymon-Lipinski and N. H. E. Keeve, "Fourier volume rendering on the gpu using a split-stream-fft," in *Vision, Modeling, and Visualization*, 2004, p. 395.

[11] M. Hopf and T. Ertl, "Hardware based wavelet transformations," in *Vision, Modeling, and Visualization*, vol. 99, 1999, pp. 317–328.

[12] C. Tenllado, J. Setoain, M. Prieto, L. Piñuel, and F. Tirado, "Parallel implementation of the 2d discrete wavelet transform on graphics processing units: filter bank versus lifting," *IEEE Transactions on Parallel and Distributed Systems*, vol. 19, no. 3, pp. 299–310, 2008.

[13] S. Guha, S. Krisnan, and S. Venkatasubramanian, "Data visualization and mining using the gpu," *Tutorial at ACM SIGKDD*, vol. 5, 2005.

[14] G. Dzemyda, O. Kurasova, and J. Zilinskas, "Multidimensional data visualization," *Methods and applications series: Springer optimization and its applications*, vol. 75, p. 122, 2013.

[15] K. Fatahalian, J. Sugerman, and P. Hanrahan, "Understanding the efficiency of gpu algorithms for matrix-matrix multiplication," in *Proceedings of the ACM SIGGRAPH/EUROGRAPHICS conference on Graphics hardware*. ACM, 2004, pp. 133–137.

[16] S. Dalton, L. Olson, and N. Bell, "Optimizing sparse matrixmatrix multiplication for the gpu," *ACM Transactions on Mathematical Software (TOMS)*, vol. 41, no. 4, p. 25, 2015.

[17] A. Garcia and H.-W. Shen, "Gpu-based 3d wavelet reconstruction with tileboarding," *The Visual Computer*, vol. 21, no. 8-10, pp. 755–763, 2005.

[18] J. Franco, G. Bernabé, J. Fernández, and M. Ujaldón, "The 2d wavelet transform on emerging architectures: Gpus and multicores," *Journal of Real-Time Image Processing*, vol. 7, no. 3, pp. 145–152, 2012.

[19] Z. Liu, B. Jiang, and J. Heer, "imMens: Real-time visual querying of big data," in *Computer Graphics Forum*, vol. 32, no. 3pt4. Wiley Online Library, 2013, pp. 421–430.

[20] H. Liu, F. Hussain, C. L. Tan, and M. Dash, "Discretization: An enabling technique," *Data mining and knowledge discovery*, vol. 6, no. 4, pp. 393–423, 2002.

[21] W. Song, D. Wu, R. Wong, S. Fong, and K. Cho, "A real-time interactive data mining and visualization system using parallel computing," in *Digital Information Management (ICDIM), 2015 Tenth International Conference on*. IEEE, 2015, pp. 10–13.

[22] P. Rech, C. Aguiar, C. Frost, and L. Carro, "An efficient and experimentally tuned software-based hardening strategy for matrix multiplication on gpus," *IEEE Transactions on Nuclear Science*, vol. 60, no. 4, pp. 2797–2804, 2013.