A SURVEY ON LARGE SCALE MEDICAL IMAGE PROCESSING

Janardhana D R
Assistant Professor, Department of IS&E, Sahyadri College of Engineering and Management, Mangalore, India

Abstract

Nowadays, a rapid growth in amount of medical image data in hi-tech hospitals on daily basis using state-of-the-art imaging tools like magnetic resonance imaging (MRI); computed tomography (CT); and nuclear medicine, including positron emission tomography (PET) is go on increasing. These image dataset enforces to adapt novel methods to process and to get scalable solution for allowing doctors to find diseases earlier and improve patient outcomes. Many State-of-the-art methods are presently working on a few set of images. This survey provides an in-sight of various parallel processing paradigms on large scale image dataset which addresses some of the important issues like fault tolerant, load balancing, portability. The survey also provides the comparison of the various parallel processing methods with respect to the issues they addressed.

Keywords: Large-scale, Medical Imaging, Parallel Processing, MPI, OpenMP, GPGPU

I. INTRODUCTION

Advanced imaging technologies have improved significantly the quality of medical care available to patients. Non-invasive imaging modalities allow a physician to make increasingly accurate diagnoses and render precise and measured modes of treatment. Current uses of imaging technologies include laboratory medicine, surgery, radiation therapy, nuclear medicine, and diagnostic radiology. Common medical imaging methodologies may be divided grossly into two general groupings: (a) techniques that seek to image internal anatomical structures and (b) methods that present mappings of physiological function. Rapid technological advances have made the acquisition of three and four-dimensional (4D) representations of humans. Internal structures common place. A multitude of imaging modalities is available currently [1].

We are living in a revolutionary age, witnessing the next-generation of biological and medical image and information emerged in astounding volume and rich formats. Nowadays images and videos are widely used in biological and medical research and clinical applications. Manual image analysis and management is extremely time consuming, labor intensive, prone to errors, and not reproducible. Biomedical image processing has experienced dramatic expansion, and has been an interdisciplinary research field attracting expertise from applied mathematics, computer sciences, engineering, statistics, physics, biology and medicine. Computer-aided diagnostic processing has already become an important part of clinical routine.Accompanied by a rush of new development of high technology and use of various imaging modalities, more challenges arise; for example, how to process and analyse a significant volume of images so that high quality information can be produced for disease diagnoses and treatment [2].

II. RELATED WORK

In many fields of medicine, engineering and science, the volume of data generated and processed is in the order of terabytes. Providing support for storing, accessing and analysing this vast amount of datasets in a distributed environment is a challenge. Umit Catalyurek et al [7] present a compendium of run-time and compiler techniques and tools for supporting such applications. Digital imaging of pathology slides have gained an increasing interest over the last decade as hardware for digitizing tissue samples and microscopy slides has rapidly advanced.

A main challenge is storing and accessing the very large volumes of data required representing a large collection of slides. With high resolution scanners, a single focal plane of a digitized slide can be 100K × 100K pixels (30 Gigabytes). Another challenging problem is the querying and processing of large volumes of data for analysis. Analysis operations on digital slides range from simple 2D visualization and browsing of images to queries to calculate density distributions to extraction of features representing different layers of tissue or cell types to 3D reconstruction from multiple 2D scans. High performance machines are required to handle the complexity of operations and the large volume of data.

Dimitris Gerogiannis et. Al [8] discussed on image features extraction tasks using parallel implementation. Load balancing requirements in parallel image analysis are considered and results on the performance of parallel implementations of two image feature extraction tasks on the Connection Machine and the iPSC/2 hypercube are reported. A load redistribution algorithm, which makes use of parallel prefix operations and one-to-one permutations among the processors, is described and has been used in this work. The expected improvement in performance resulting from load balancing has been determined analytically and is compared to actual performance results obtained from the above implementations. The analytical results demonstrate the specific dependence of the expected improvement in performance on the computational and communication requirements of each task, characteristic machine parameters, a characterization of prior load distribution in terms of parameters which can be computed dynamically at the start of task execution, and the overhead incurred by load redistribution. Based on these results one may also define a set of rules for determining in advance the potential gains in performance to be obtained from application of this particular load balancing algorithm.

MapReduce is a programming model and an associated implementation for processing and generating large data sets. Users specify a map function that processes a key/value pair to generate a set of intermediate key/value pairs, and a reduce function that merges all intermediate values associated with the same intermediate key. Many real world tasks are expressible in this model and proposed by Jeffrey Dean and Sanjay Ghemawat [9]. Programs written in this functional style are automatically parallelized and executed on a large cluster of commodity machines. The run-time system takes care of the details of partitioning the input data, scheduling the program's execution across a set of machines, handling machine failures, and managing the required inter-machine communication. This allows programmers without any experience with parallel and distributed systems to easily utilize the resources of a large distributed system.

A prominent parallel data processing tool MapReduce is gaining significant momentum from both industry and academia as the volume of data to analyse grows rapidly. While MapReduce is used in many areas where massive data analysis is required, there are still debates on its performance, efficiency per node, and simple abstraction. This survey intends to assist the database and open source communities in understanding various technical aspects of the MapReduce framework. Kyong-Ha Lee [10] explained about parallel data processing with the MapReduce framework and discuss its inherent pros and cons.

Nasullah Khalid Alham and others[11] explained about Machine learning techniques are helped for image retrieval by automatically classifying and annotating images with keywords. Among them Support Vector Machines (SVMs) have been used extensively due to their generalization properties. However, SVM training is notably a computationally intensive process especially when the training dataset is large. Here they presents MRSMO, a MapReduce based distributed SVM algorithm for automatic image annotation. The performance of the MRSMO algorithm is evaluated in an experimental environment. By partitioning the training dataset into smaller subsets and optimizing the partitioned subsets across a cluster of computers, the MRSMO algorithm reduces the training time significantly while maintaining a high level of accuracy in both binary and multiclass classifications.

Image classification on large scale data set discussed by Yunpeng Li [12] here With the rise of photo-sharing websites such as Facebook and Flickr has come dramatic growth in the number of photographs online. Recent research in object recognition has used such sites as a source of image data, but the test images have been selected and labelled by hand, yielding relatively small validation sets. Study of image

classification on a much larger dataset of 30 million images, including nearly 2 million of which have been labelled into one of 500 categories. The dataset and categories are formed automatically from geo-tagged photos from Flickr, by looking for peaks in the spatial geo-tag distribution corresponding to frequently-photographed landmarks and learn models for these landmarks with a multiclass support vector machine, using vector-quantized interest point descriptors as features. Additionally explore the non-visual information available on modern photo sharing sites, showing that using textual tags and temporal constraints leads to significant improvements in classification rate. In some cases image features alone yield comparable classification accuracy to using text tags as well as to the performance of human observers.

MapReduce task performed on multi-node clustering using k-means clustering and discussed overview of MapReduce [14].An overview of MapReduce and common design patterns are provided for those with limited MapReduce background. Here they discussed both the high level theory and the low level implementation for several computer vision algorithms: classifier training, sliding windows, clustering, bag of-features, background subtraction, and image registration. Experimental results for the k-means clustering and single Gaussian background subtraction algorithms are performed on a 410 node Hadoop cluster.

Vladimir N. Vapnik [15] expressed his views on statistical learning theory framework for estimating dependencies on finte samples.

This theory combines fundamental concepts and principles related to learning, well-defined problem formulation, and self-consistent mathematical theory. Thus, it is (arguably) the best available theory for predictive learning, and it compares favorably to other more empirical methodologies that are often based on intuitive, asymptotic, and/or biological arguments. Unfortunately, perhaps because of its mathematical rigor and complexity, this theory is not well understood in the mainstream statistics and in the field of neural networks.

The area of content based image retrieval (CBIR) was motivated since the early 1990s by the idea to find and retrieve images independent from metadata other than extracted from the image itself. However a satisfactory solution has not been found yet, but a problem has been isolated: Researchers defined the semantic gap , which refers to the inability of a machine to fully understand and interpret images based on automatically extracted data. In current research efforts in visual information retrieval especially global features, which denote features capturing characteristics of the whole image instead of focusing for instance on segments, regions or patches, have lost part of their significance. In applied research however content based image retrieval – for instance as part of a complex system – is often relying on fast, global features at least as a foundation for further research. LIRe (Lucene Image Retrieval) [15] are an efficient library allowing researchers to integrate CBIR based on global features in an easy way.

Information technology developing rapidly, variety and quantity of image data is increasing fast. How to retrieve desired images among massive images storage is getting to be an urgent problem. Jing Zhang et al [16] developed a Distributed Image Retrieval System (DIRS), in which images are retrieved in a content based way, and the retrieval among massive image data storage is speeded up by utilizing MapReduce distributed computing model. Moreover, fault tolerance, ability to run in a heterogeneous environment and scalability are supported in our system. Experiments are carried out to verify the improvement of performance when MapReduce model is utilized. Results have shown that image storage and image retrieval based on MapReduce outperform that in centralized way greatly when total number of images is large.

Hive, an open-source data ware housing solution built on top of Hadoop [17]. Hive supports queries expressed in a SQL-like declarative language – HiveQL, which are compiled into map-reduce jobs executed on Hadoop. In addition, HiveQL supports custom map-reduce scripts to be plugged into queries. The language includes a type system with support for tables containing primitive types, collections like arrays and maps, and nested compositions of the same. The underlying IO libraries can be extended to query data in custom formats. Hive also includes a system catalog, Hive-Metastore, containing schemas and statistics, which is useful in data exploration and query optimization. In Facebook, the Hive warehouse contains several thousand tables with over 700 terabytes of data and is being used extensively for both reporting and ad-hoc analyses by more than 100 users.

Distributed storage system for structured data discussed by Fay Chang, Jeffrey Dean et al [18].Bigtable is a distributed storage system for managing structured data that is designed to scale to a very large size: petabytes of data across thousands of commodity servers. Many projects at Google store data in Bigtable, including web indexing, Google Earth, and Google Finance. These applications place very different demands on Bigtable, both in terms of data size (from URLs to web pages to satellite imagery) and latency requirements (from backend bulk processing to real time data serving). Despite these varied demands, Bigtable has successfully provided a flexible, high-performance solution for all of these Google products. Here by taking simple data model provided by Bigtable, which gives clients dynamic control over data layout and format, and describe the design and implementation of Bigtable.

In Hadoop, HDFS (Hadoop Distributed File System) [19] is designed to store very large data sets reliably, and to stream those data sets at high bandwidth to user applications. In a large cluster, thousands of servers both host directly attached storage and execute user application tasks. By distributing storage and computation across many servers, the resource can grow with demand while remaining economical at every size. The architecture of HDFS and report on experience using HDFS to manage 25 petabytes of enterprise data at Yahoo!.

QingZang Huang et al,[20] describes about Current medical image processing applications require management of huge amounts of data and executive high-performance computing. How to manage and analysis this large amount of data effectively is a major challenge. Here it shows the initial formulation of medical information integration model, it is designed to help medical workers and medical departments to share data, integrate resources and analyses data. The model using the technology of cloud computing to provide huge storage resources and powerful computing capacity. The application level of this system is on the basis of service. Hadoop cluster is the main calculation module and storage unit.

III. COMPARISION

This section provides the comparison of some of works which were carried out in the field of medical image analysis with respect to large scale dataset which are supported by them.The emergence of massive datasets in a clinical setting presents both challenges and opportunities in data storage, analysis and visualization. The MapReduce programming framework uses two tasks common in functional programming: Map and Reduce. MapReduce is a new parallel processing framework and Hadoop is its open-source implementation on a single computing node or on clusters. Compared with existing parallel processing paradigms (e.g. Message Passing interface(MPI), OpenMP, grid computing and General purpose computing on graphics processing unit (GPGPU)) like CUDA, MapReduce and Hadoop have more advantages: 1) fault-tolerant storage resulting in reliable data processing by replicating the computing tasks, and cloning the data chunks on different computing nodes across the computing cluster; 2) high-throughput data processing via a batch processing framework and the Hadoop distributed file system (HDFS); 3) Portability, Originally support distributed systems, now ported to GPU, CELL, multi-core(Phoenix, Mars, Merge etc.) and 4) Load Balancing: Skewed data appears not only on the map phase but also on the reduce phase, and can weaken the overall performance of the system. The process of partition is that transfer the result from mapper to corresponding reducer. The common partition method is applying a hash function to each (key, value)pair and assigning a partition number to each pair. The default hash function in Hadoop is Hash (HashCode (intermediate key) mod numReducer). Data are stored in the HDFS and made available to the slave nodes for computation.

Conclusion

Large scale medical image processing is a critical issue in image processing. We find many parallel processing paradigms in the literature which addresses different methods to handle large scale dataset. In this paper we compare few related works which are done based on large scale medical image data which are supported by them. The survey also provides an insight on some parallel processing methods. There is still less work done in the field of large amount of medical image analysis. To conclude, there is still lot of scope in desinging new methods which could support parallel processing on large scale and which draws more attention of the users towards medical image processing.

REFERENCES

[1]. Raj Acharya,Richard Wasserman,Jeffrey Stevens and Carlos Hinojosa " Biomedical Imaging Modalities: A Tutorial" Computerized Medical Imagmg and Graphics. Vol. 19, No. I. pp. 3-25, 1995

[2]. Hongmei Zhu "Medical Image Processing Overview" University of Calgary, 2003 – academia.edu

[3]. Dimitris Gerogiannis and Stelios C. Orphanoudakis "Load Balancing Requirements in Parallel
Implementations of Image Feature Extraction Tasks"IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS. VOL 4, NO. 9, SEPTEMBER 1999

[4]. Bhaskar Prasad Rimal, Eunmi Choi and Ian Lumb "A Taxonomy and Survey of Cloud Computing Systems" 5th International Joint Conference on INC, IMS and IDC, pp. 44-51,2009

[5]. Jeffrey Dean and Sanjay Ghemawat "MapReduce: Simplified Data Processing on Large Clusters",Google, Inc.,2004

[6]. Minakshi M. sonawane , Santosh D. Pandure and Seema S. Kawthekar "A Review on Hadoop MapReduce using image processing and cloud computing",IOSR Journal of Computer Engineering (IOSR-JCE)e-ISSN: 2278-0661,p-ISSN: 2278-8727,PP 65-68, 2016

[7]. U. Catalyurek, S. Hastings, K. Huang, V.S. Kumar, T. Kurc,S. Langella, S. Narayanan, S. Oster, T. Pan, B. Rutt and X. Zhang, J. Saltz "Supporting Large Scale Medical and Scientific Datasets" Parallel Computing: Current & Future Issues of High-End Computing,Proceedings of the International Conference ParCo,Vol. 33, ISBN 3-00-017352-8, pp. 3-14, 2006

[8]. Dimitris Gerogiannis and Stelios C. Orphanoudakis "Load Balancing Requirements in Parallel Implementations of Image Feature Extraction Tasks"IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS. VOL 4, NO. 9, SEPTEMBER 1999

[9]. Jeffrey Dean and Sanjay Ghemawat "MapReduce: Simplified Data Processing on Large Clusters",Google, Inc.,2004

[10]. Kyong-Ha Lee, Hyunsik Choi and Bongki Moon "Parallel Data Processing with MapReduce: A Survey", SIGMOD Record, (Vol. 40, No. 4), pp.11-20,December 2011

[11] Nasullah Khalid Alham, Maozhen Li and Yang Liu, Suhel Hammoud "A MapReduce-based distributed SVM algorithm for automatic image annotation",Computers and Mathematics with Applications Elsevier,pp.2801–2811, 2011

[12]. Yunpeng Li, David J. Crandall and Daniel P. Huttenlocher "Landmark Classification in Large-scale Image Collections", IEEE 12th International Conference on Computer Vision (ICCV), pp.1957-1964, 2009

[13]. Brandyn White, Tom Yeh, Jimmy Lin, and Larry Davis "Web-Scale Computer Vision using MapReduce for Multimedia Data Mining",MDMKDD'10, July 25, 2010, Washington, DC, USA

[14]. Vladimir N. Vapnik. "The Nature of Statistical Learning Theory",. Springer, New York, November 1995

[15]. Mathias Lux and Savvas A. Chatzichristofis " Lire: lucene image retrieval: an extensible java CBIR library", In Proceedings of the 16th ACM international conference on Multimedia, pages 1085–1088, October 2008.

[16]. Jing Zhang, Xianglong Liu, Junwu Luo, and Bo Lang,"DIRS: Distributed image retrieval system based on MapReduce", In Proceedings of 5th International Conference on Pervasive Computing and Applications, ICPCA'10, pages 93–98. IEEE, December 2010.

[17]. Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Suresh Anthony, Hao Liu, Pete Wyckoff, and Raghotham Murthy," Hive: a warehousing solution over a map-reduce framework."Proceedings of the VLDB Endowment, 2(2):1626 – 1629, August 2009

[18]. Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach, Mike Burrows, and Tushar Chandra,"Bigtable, a distributed storage system for structured data",ACM Transactions on Computer Systems, 26(2), June 2008.

[19]. Konstantin Shvachko, Hairong Kuang, Sanjay Radia, and Robert Chansler,"The Hadoop Distributed File System",In Proceedings of the 2010 IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST), MSST '10, pages 1–10, Washington, DC, USA, 2010. IEEE Computer Society.

[20]. QingZang Huang, Lei Ye, and MingYuan Yu,"Medical information integration based cloud computing", In International Conference on Network Computing and Information Security, pages 79–83, May 2011.