A Distributed Model for Automated Diagnosis of Whole-Slide H&E Stained Prostate Tissue Images

Safa'a N. Al-Haj Saleh
Department of Software Engineering
The Hashemite University
Zarqa, Jordan
alhaj.safa2@gmail.com

Omar S. Al-Kadi
King Abdullah II School for IT
The University of Jordan
Amman, Jordan
o.alkadi@ju.edu.jo

***Abstract*— Analysis of large amounts of medical images exceeds storage capacity and computation capability of a single workstation. Distributed computing employs a set of connected machines to solve a single problem by dividing it into number of solvable sub-problems. Analyzing and processing of large-scale medical images demand employing distributed architectures to overcome the limitations of memory space and execution time. Current analysis of digitized large-scale prostate tissue images depends on ordinary sequential techniques running on a single machine. This paper presents a proposed distributed model based on Hadoop framework for automated diagnosis of digitized large-scale H&E prostate tissue images to carry out segmentation, feature extraction, and classification tasks. The proposed model is based on partitioning input images into segments and distributing them across number of slaves to perform analysis task simultaneously. Analysis task aims at segmenting and labeling Regions of Interest (ROIs) in input images to extract initial features. Initial features are combined at master side to get final features for each input image. Finally, master node classifies images into the corresponding grade based on a grading system such as Gleason Grading system. The proposed distributed model would achieve high speed performance when applied in advanced medical applications.**

Keywords— Medical distributed computing; Hadoop framework; H&E prostate tissue image; gland segmentation; Gleason grading

I. INTRODUCTION

Big data concept has recently emerged referring to the massive amounts of data that are difficult to handle and manage using traditional sequential approaches; as different aspects should be considered such as speed, time and accuracy. Nowadays, big data processing is considered among the top ten strategic technologies in the world according to [1]. Dealing with this type of data was initially carried out using high-storage multi-processor systems. Due to the rapid growth of data volumes and the need for concurrent processing, parallel and distributed computing models were introduced for the purpose of big data management and processing [2].

Distributed computing model solves one large problem by splitting it into many small tasks. This model consists of number of computational components that are network-connected computers communicating through message passing. Number of architectures are suggested to implement distributed model such as Client-Server Architecture and Peer-to-Peer Architecture. Sub-tasks can be also processed simultaneously in minimum time and maximum efficiency [3]; this is defined as distributed parallel processing. The power of this model may exceed the power of a single high performance machine [4].

Many applications require applying distributed computing model, particularly the applications that deal with processing large-size images – that is generally a time consuming process that exceeds storage capacity of a single machine– such as medical [4], geographical [5], and remote-sensing [5] applications. Many frameworks have been implemented particularly for distributed image processing. Hadoop framework is an excellent example. It facilitates storage, processing, and analysis of large amounts of distributed images in efficient, reliable, and scalable manner [5].

Diagnosing medical cases based on large-size images is an emerging trend; as it is vital to have a comprehensive overview regarding patient situation. It is the time to invest distributed computing for processing this type of medical images. In hospitals and medical centers, many aspects should be considered for distributed medical applications including getting quick results – as it is common that medical results have to be ready-made within a certain time – and accurate processing as well. Distributed model offers the potential for processing and analyzing medical images in a flexible way for the purpose of recognizing patterns and making proper medical diagnosis. It also facilitates the deployment of advanced medical applications and saves days of waiting for results [4]; as many scenarios demand high computing power and storage capacity [6].

In this paper, a distributed model based on Hadoop framework for automated diagnosis of digitized large-scale H&E prostate tissue images is presented. The proposed model performs segmentation and feature extraction tasks in a distributed manner. Thus, lumen objects and tissue glands are segmented at slave node for each image segment. Lumen objects are segmented by empirical thresholding technique, while tissue glands are segmented by a k-means clustering

978-1-5090-5969-0/17/$31.00 ©2017 IEEE

approach applied on extracted a* color channel image. Initial features are extracted at each slave and then combined to get final features at master side that classifies tissue image into the corresponding Gleason grade.

The paper is organized as follows: section II presents a brief overview of Hadoop and a medical background. Related work is discussed in section III. The proposed distributed model for automated diagnosis of prostate tissue images is illustrated in section IV. Finally, section V presents final conclusion and future work.

II. A Brief Overview of Hadoop and a Medical Background

A. Hadoop: A Distributed Image Processing Framework

Hadoop is an open source infrastructure written in Java used for storing, processing and analyzing huge datasets in a distributed way. The roots of Hadoop get back to search engines companies [1,7,8] including Yahoo and Google. Hadoop can be deployed on a single machine. Though, deploying it on number of machines achieves the required goals of distribution [2]; as it is specifically designed to scale up from a single machine to several [3].

As illustrated in Fig. 1, Hadoop is composed of number of modules [1,7] including: Hadoop Distributed File System (HDFS) and Hadoop MapReduce. HDFS is a distributed file system that stores data on machines (i.e. each file is stored as a sequence of blocks that are replicated to facilitate fault tolerance), while MapReduce is a programming model for large scale data processing.

HDFS consists of two elements [9]: NameNode and DataNode. *NameNode* is a master server that manages file system namespace and executes some operations including opening and closing. It also maps data blocks to DataNodes. *DataNode* manages the storage on the node in each slave and performs block creation, deletion, and replication. MapReduce consists of two elements [9]: JobTracker and TaskTracker. *JobTracker* manages resources and job scheduling, while *TaskTracker* is a slave node that accepts tasks from JobTracker and returns results. Based on this, MapReduce has two main functions [7]: *Map* which is the process of splitting the inputs and *reduce* which is the process of integrating the output of the map.

Hadoop framework has number of advantages [1,3] including: The ability of processing and analyzing large amounts of data; simple expansion and horizontal scalability (i.e. new nodes can be added without the need to change data formats); and low set up cost as it is free and open source.

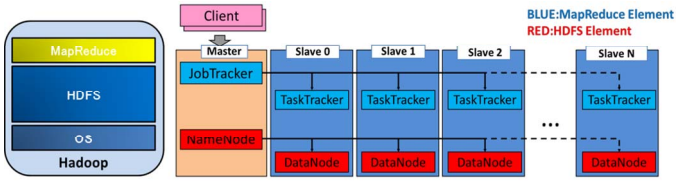

Fig. 1. Main Structure of Hadoop [9]

Nevertheless, Hadoop has some limitations [3], including: Programming model is very restrictive; and having single master requires care and may limit scaling especially when datasets are very large.

Hadoop framework supports dealing with images via image processing library [10]. Images are initially stored in HDFS, after that map reduce algorithm is applied to extract the features from images. Finally, reducer collects all results from all map functions and stores them back to HDFS.

B. Medical background

Prostate cancer is a considerable medical problem as it is classified among the top ten cancer types in the world [11]. Early detection of this serious cancer type plays a significant role in the effective treatment. Many tests are carried out to detect prostate cancer including Digital Rectal Examinations (DRE), Prostate-Specific Antigen (PSA) test, CT and MRI tests [12]. These tests provide an indicator of having cancer or not, but do not provide an accurate diagnosis. Because of that, pathologists ask their patients to take a biopsy.

Prostate biopsy is a sample of tissue taken in order to examine prostate more closely. After prostate biopsy is taken, slides of the tissue are prepared and stained with a biomarker such as Hematoxylin and Eosin (H&E) to be ready for microscope examination. Slides can be also digitized into high resolution images to be examined. If infected area is found, pathologist classifies its degree of aggressiveness using grading system like Gleason Grading system that is considered the most commonly employed grading system [13]. It classifies prostatic carcinoma into five grades based on morphological features of glandular patterns in the tissue [14], where grade 1 is a well-differentiated benign pattern and grade 5 is a poorly-differentiated advanced carcinoma pattern [11].

Classifying prostate cancer into the suitable grade is based on examining gland units in prostate tissue to check their features such as size, shape and number of glands in tissue. Prostate gland unit consists of four main components: Stroma, lumen, epithelial nuclei, and epithelial cytoplasm. Each component appears in a specific color in the tissue stained by H&E method: Stroma is pink, lumen is white, nuclei are dark blue (i.e. appearing as dots), and cytoplasm is purple. According to [11], in benign tissues (grades 1 and 2) glands are large-sized, oval or crunchy, separated, having large lumen components, and thick nuclei boundaries. In grade 3 tissues, glands are smaller and more circular with smaller lumen and thin nuclei boundaries. In grade 4 tissue, glands are not well-separated and fused together to create a mass of glands. In grade 5 tissue, glands have no clear formation and undistinguishable. Fig. 2 shows samples of benign, grade 3, and grade 4 prostate tissues.

III. Related Work

Many methodologies were proposed for distributed processing of different types of medical images including CT images, MRI images, and tissue images. Info et. al.[15] presented a distributed parallel algorithm to align large-scale

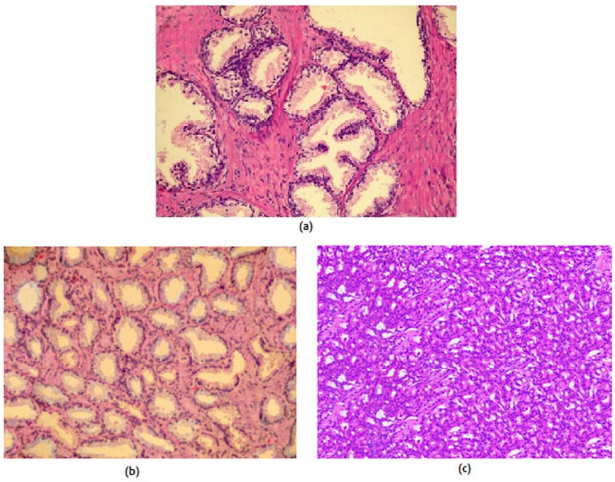

Fig. 2. Samples of Prostate Tissue: (a) Benign Tissue, (b) Grade 3 Tissue, and (c) Grade 4 Tissue

3D liver CT images of deformable objects using Schnabel's registration algorithm. Distributed processing was performed using 128-CPU cluster of interconnected PCs. Execution time required for aligning dataset images was reduced from hours to minutes. A distributed methodology for training multi-layer perceptrons (MLPs) neural networks for the detection of lesions in colonoscopy images was demonstrated in [4]. The proposed approach was based on partitioning the training set across multiple processors relying on virtual machine implementation. Also, Shen et. al. [16] presented a grid-based peer to peer distributed architecture for processing large-scale images. The distributed architecture was tested using CT medical images. Passerat-Palmbach et. al. [17] proposed a distributed OpenMOLE platform with added features to handle the analysis of medical images. The proposed platform allowed using many distributed infrastructures such as clusters and computing grids. A distributed framework for image guided neurosurgery was proposed in [18]. The proposed framework was applied on real-time image fusion for brain MRI using landmark tracking across the entire image volume.

Few works presented distributed methodology for processing digitized prostate biopsies such as the work presented by Krefting et. al. [6] where a distributed grid was implemented for applying segmentation and registration algorithms for transrectal ultrasound guided prostate biopsies. It is worthy to mention that no works were presented for distributed processing of digitized prostate tissue images, specifically H&E stained tissues.

IV. Proposed Distributed Model for Automated Diagnosis of Prostate Tissue Images

Analyzing digitized H&E prostate tissue images aims at a labeling Regions of Interest (ROIs) automatically including gland and lumen regions to extract specific features that helps in classifying the tissue into the relevant Gleason grade. Segmentation stage is the foundation for analyzing and preparing datasets (i.e. extracted features) needed for training and testing purposes (i.e. classification).

The main aim of the proposed model is to apply distributed architecture on large-scale digitized H&E prostate tissue images by splitting each input image into smaller sub-images (i.e. segments), distributing segments on number of processors (i.e. slaves) using Hadoop framework. The slaves perform segmentation process to extract gland and lumen regions –in parallel way– and extract initial features for each segment. Initial extracted features are gathered from all slaves and merged at master side (i.e. extracted initial features for segments that form the original partitioned image are merged together). Finally, master node performs training, testing and classification tasks to get final results. It is worthy to mention that distribution is necessary for segmentation stage not for training and classification stage; as segmentation task includes processing images that results in a heavy-weight work load.

The proposed distributed model illustrated in Fig. 3 performs segmentation and initial feature extraction through the following main steps:

1. Specific number of large-scale H&E prostate tissue images is stored in file system.

2. The bundle of images (i.e. training or testing images) is fed to Hadoop Distributed File System (HDFS) on master side.

3. At master node, each input image is split into smaller segments: All input images should have the same size. Images are chosen of large size (i.e. images with dimensions starting from 16000 x 12000 are considered to be large-scale images). Size of input image is based on number of slaves; so when number of slaves is 4, input image size is chosen to be divisible by 4 such as 20000 x 16000, and so on.

4. Each input image is represented by a key-value pair to facilitate distribution process and collecting results from slaves. Sequence number of each input image is considered as the key and number of segments for each image is considered as the value.

5. Master node distributes partitioned segments and assigns them to slave nodes: This is done using NameNode that maps partitioned segments to DataNodesin slaves and JobTracker that manages the job of the slaves.

6. TaskTracker in each slave node accepts tasks from JobTracker and performs segmentation job for each segment by running Map function to apply gland and

lumen regions segmentation methodology mentioned in [19] that has the following main steps:

a) Preprocessing RGB prostate tissue images by applying Gaussian Filter to smooth the images as segmentation methodology is region-based not boundary-based.

b) Segmenting lumen objects starts with converting the filtered RGB tissue image into grayscale image. After that, manual thresholding is applied to divide the tissue image in two classes: lumen objects (foreground) and other tissue components (background). Intensity threshold value is chosen based on trial and error approach. Finally, size constraint is applied to keep true lumen objects and eliminate false lumen objects (i.e. too small detected objects).

c) Segmenting glands starts with converting the filtered RGB tissue image into L*a*b* color model and separating a* channel for the reasons clarified in [19]. After that, k-means clustering is applied on the extracted a* tissue image to classify pixels into two clusters: A cluster that represents lumen and cytoplasm (i.e. inner gland region) and another cluster that represents nuclei and stroma. Then, morphological closing operation is applied to fill small holes and gaps. Finally, size constraint is applied to keep true glands.

7. When segmentation is completed for each segment, TaskTracker in each slave node extracts number of initial features for labeled glands and lumen regions. Initial features are intermediate results that are stored on slave nodes. They are considered to be local and incomplete as they depend on data available on one slave node. Initial Features are the output of each map function. *Initial lumen objects features* include: average lumen area, maximum lumen area, and average lumen eccentricity. *Initial glands features* include: average glands area, maximum gland area, average glands diameter, average glands perimeter, average glands eccentricity, and glands density.

8. JobTracker in master node sends the Reduce request to slave nodes to integrate the outputs of all map functions. The reduce function performs these two phases:

   – Sorting: The reducers sort the results (i.e. initial features) of image segments based on key-value pairs.

   – Shuffling: The reducers copy the sorted results from mappers to HDFS on master side.

   Based on this, initial features for all segments related to the same input image are gathered from all slaves and merged at master side. Initial features are merged to get final features for each input image.

9. Finally, training, testing, and classification tasks are performed at master side using Naive Bayes classifier, to get final results (i.e. classification of testing tissue images into grades).

Fig. 4 provides a simplified pseudo code that summaries the main steps of the proposed distributed model.

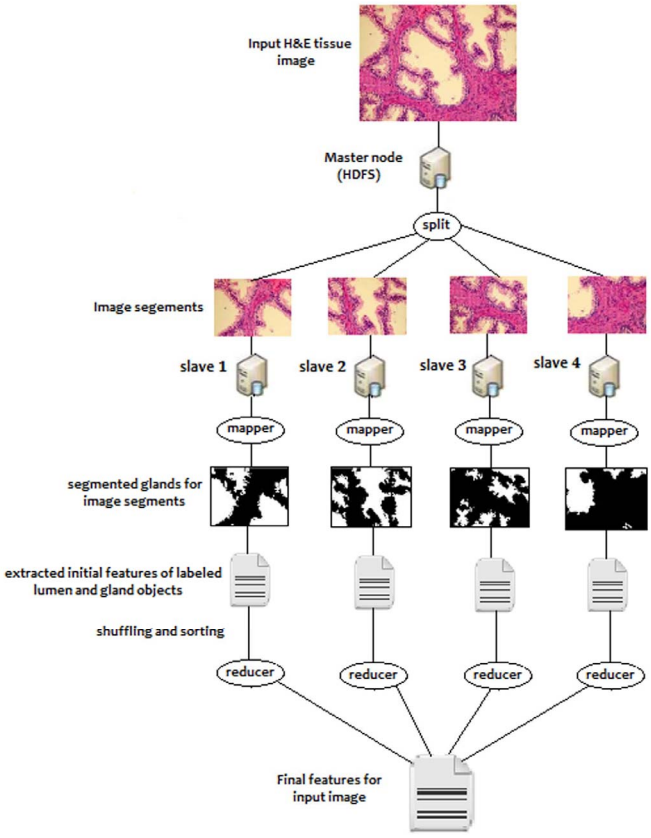

Fig. 3. The Proposed Distributed Model for Automated Diagnosis of Prostate Tissue Images

```
DISTRIBUTED PROCESSING OF PROSTATE TISSUE IMAGES PSEUDO-CODE

Input N large-scale prostate tissue images to HDFS at master node M

for all N images
    M node splits n1,n2,....N into K segments
    M node represents n1,n2,....N by a key-value pair <key,value>

for all K segments
for all S slave nodes
    NameNode in M node distributes and maps k1,k2,....K to DataNodes in s1,s2,....S
    TaskTracker in si applies segmentation algorithm on k1,k2,....kj by running map function Map

        Apply Gaussian Filter to preprocess RGB image kj

        Segment lumen objects in kj
            Convert filtered kj to grayscale image
            Apply manual thresholding using intensity threshold value
            Apply size constraint to keep true lumen objects

        Segment gland objects in kj
            Convert filtered kj to L*a*b* color model
            Seperate a* channel image
            Apply k-means clustering
            Apply morphological closing operation
            Apply size constraint to keep true glands

    TaskTracker in si extracts number of initial features I (i1,i2,....I) for k1,k2,....kj

JobTracker in M node sends the Reduce request to s1,s2,....S to integrate the outputs of Map1,Map2,....Mapx
JobTracker in M node gathers I for k1,k2,....ky that belong to the same nx to get final features F (f1,f2,....fy)
M node performs training, testing, and classification tasks to classify N into grades
```

Fig. 4. Pseudo Code of the Proposed Distributed Model

V. CONCLUSION

In this work, a proposed distribute model for automated diagnosis of large-scale digitized H&E prostate tissue images has been presented. The proposed model was based on Hadoop framework to split each input image into smaller segments to be distributed on number of slaves. The slaves extracted and labeled regions of Interest (i.e. gland and lumen objects) using region-based methodology to extract initial features via map function. Reduce function integrated the output of all map functions (i.e. initial features) to be merged at master side to get final features for each input image. Classifying tissue images into the corresponding Gleason grade was carried out on master side to get final results.

Testing the proposed distributed model on a dataset containing different sizes of large-scale H&E prostate tissue images on variable number of slaves is the next step for future work. Sample of suggested image sizes to be tested on 4 slaves is: 16000 x 13000, 21000 x 16000, 24000 x 20000.

REFERENCES

[1] S. Banaei and H. Moghaddam ,"Hadoop and Its Role in Modern Image Processing", Open Journal of Marine Science, vol. 4, issue 4, 2014.

[2] T. da Silva Morais, "Survey on Frameworks for Distributed Computing: Hadoop, Spark and Storm", In Proceedings of the 10th Doctoral Symposium in Informatics Engineering-DSIE, vol. 15, January, 2015.

[3] S. Saxena, S. Sharma, S., and N. Sharma, "Parallel Image Processing Techniques, Benefits and Limitations", Research Journal of Applied Sciences, Engineering and Technology, vol. 12, issue 2, pp. 223-238, 2016.

[4] V.P. Plagianakos, G. D. Magoulas , and M. N. Vrahatis, "Distributed Computing Methodology for Training Neural Networks in an Image-Guided Diagnostic Application", Computer Methods and Programs in Biomedicine, vol. 81, issue 3, pp. 228-235, 2006.

[5] R. Yadav and M. Padma, "Processing of Large Satellite Images using Hadoop Distributed Technology and Mapreduce: A Case of Edge Detection" , International Journal on Recent and Innovation Trends in Computing and Communication,vol. 3 ,Issue 5, May 2015.

[6] D. Krefting, M. Vossberg , and T. Tolxdorff , "Simplified Grid Implementation of Medical Image Processing Algorithms using a Workflow Management System", In Proceeding of the Workshop on Medical Imaging on Grids, pp. 23-32 ,September 2008.

[7] M. Ali, and J. Kumar, "Implementation of Image Processing System using Handover Technique with Map Reduce Based on Big Data in the Cloud Environment", The International Arab Journal of Information Technology, vol. 13, no. 2, pp. 326-331,March 2016.

[8] D. Sarade Shrikant, B. Ghule Nilkanth, P. Disale Swapnil, and R. Sasane Sandip, "Large Scale Satellite Image Processing Using Hadoop Distributed System", International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 3 , issue 3, 2014.

[9] M. Yamamoto and K. Kaneko, "Parallel Image Database Processing with MapReduce and Performance Evaluation in Pseudo Distributed Mode", International Journal of Electronic Commerce Studies, vol. 3, issue 2 ,2012.

[10] M. sonawane, S. Pandure , and S. S. Kawthekar, "A Review on Hadoop MapReduce using Image Processing and Cloud Computing", International Conference On Recent Advances In Computer Science, Engineering And Technology, 2017.

[11] K. Nguyen, A. Jain , and R. Allen, "Automated Gland Segmentation and Classification for Gleason Grading of Prostate Tissue Images", In Pattern Recognition (ICPR), 2010 20th International Conference, IEEE, pp. 1497-1500,August 2010.

[12] J. Vidal, G. Bueno, J. Galeotti, M. García-Rojo,F. Relea , and O. Déniz, 2011, "A Fully Automated Approach to Prostate Biopsy Segmentation Based on Level-Set and Mean Filtering", Journal of Pathology Informatics, vol 2, issue 2, 2011.

[13] J. Kwak, S. Hewitt, A. Kajdacsy-Balla, S. Sinha, and R. Bhargava , "Automated Prostate Tissue Referencing for Cancer Detection and Diagnosis", BMC Bioinformatics, vol 17, issue 1,2016.

[14] L. Gorelick, O. Veksle, M. Gaed, J. Gómez,M. Moussa, M., G. Bauman, A. Fenster, and A. Ward, "Prostate Histopathology: Learning Tissue Component Histograms for Cancer Detection and Classification", IEEE Transactions on Medical Imaging, vol. 32, issue 10, pp.1804-1818, 2013.

[15] F. Info, K. Ooyama, and K. Hagihara , "A Data Distributed Parallel Algorithm for Non-rigid Image Registration", Parallel Computing, vol. 31, issue 1 , pp.19-43, 2005.

[16] L. Shen, J. Ni, C. Zhu and S. Huang, "Framework of Distributed Medical Images Library for Medical Research and Education", In Internet Computing for Science and Engineering (ICICSE), 2012 Sixth International Conference, IEEE, pp. 180-187, April, 2012.

[17] J. Passerat-Palmbach, M. Leclaire , R. Reuillon, Z. Wang, and D. Rueckert, "OpenMOLE: a Workflow Engine for Distributed Medical Image Analysis", In International Workshop on High Performance Computing for Biomedical Image Analysis , September 2014.

[18] N. Chrisochoides, A. Fedorov, A. Kot, N. Archip, P. Black, O. Clatz, A. Golby, R. Kikinis and S. Warfield , "Toward Real-Time Image Guided Neurosurgery using Distributed and Grid Computing", In Proceedings of the 2006 ACM/IEEE Conference on Super Computing ,ACM, November, 2006.

[19] S. Al-Haj Saleh, O. Al-Kadi and M. Al-Zoubi, "Histopathological Prostate Tissue Glands Segmentation for Automated Diagnosis", In Applied Electrical Engineering and Computing Technologies (AEECT), 2013 IEEE Jordan Conference, IEEE, pp. 1-6, 2013.