A Bag-of-Tasks Approach to Speed Up the Lung Nodules Retrieval in the BigData age

Marcelo Costa Oliveira, José Raniery Ferreira

Computing Institute, Lab of Telemedicine and Medical Informatics of the University Hospital(HUPAA)
Federal University of Alagoas (UFAL)
Maceió, Brazil
oliveirame@ic.ufal.br

Abstract— The Content-Based Image Retrieval (CBIR) has received great attention in the medical community because it is capable of retrieving similar images that have known pathologies. However, the sheer volume of data produced in radiology centers has precluded the use of CBIR in the daily routine of hospitals. The volume of medical images produced in medical centers has increased fast. The annual data produced from exams in the big radiology centers is greater than 10 Terabytes. Therefore, we have reached to an unprecedented age of “BigData”. We here present a bag-of task approach to speed up the images retrieval of lung nodules stored in a large medical images database. This solution combines texture attributes and registration algorithms that together are capable of retrieving images of benign lung nodules with greater-than-72% precision and greater-than-67% in malignant cases, yet running in a few minutes over the Grid, making it usable in the clinical routine.

Keywords—Content-Based Image Retrieval, Grid Computing, BigData, Computer-Aided Diagnosis

I. INTRODUCTION

The volume of medical images produced in hospitals and medical centers has increased fast. The annual data produced from exams in the big radiology centers is greater than 10 Terabytes. This is due to the growing importance and the use of exams based on images [1]. Besides that, never before have population scientists had the capacity to collect, share, and analyze data as they have today. Proposals to link disease registers to other surveillance systems, to clinical care, a larger public health informatics infrastructure, and to data on the human genome are ushering in a new age of analytic exploration. Therefore, we have reached to an unprecedented age, which we have named of “BigData”.

In the clinical routine, the big volume of medical images needs to be stored safely and analyzed in appropriate time with precision. Besides that, medical images have complex nature and require heavily computational techniques to aid medical diagnosis. The purpose of computer-aided diagnosis (CAD) is to improve the accuracy and consistency of medical image diagnosis through the computational support used as reference [2]. However, the diagnosis support systems are still limited, because they are developed to help the diagnosis and detect specific diseases. Therefore, physicians still prefer to analyze the images in conventional way, based on their knowledge and individual experience [3]. Moreover, some CAD tools have shown great results but are not being used in the clinical routine because they incur in high computational costs. These difficulties have motivated businesses and research institutions find new solutions able to automate the diagnostic aid process and to be capable to store, process and retrieval images in large database [4].

In the last 10 years, the Content-Based Image Retrieval (CBIR) has been one of the most studied computational vision technique. The CBIR has a future potential as a CAD technique and is expected to help the physician in the clinical decision process as a second opinion [5]. Through CBIR, it is possible to use a reference image to find similar images using only the medical image intrinsic attributes. The reason for presenting similar images is based on the fact that radiologists learn diagnostic skills by observing many clinical cases during their training and clinical practice, and their knowledge obtained from visual impression of images with various diseases constitutes the foundation for their diagnosis [6].

Therefore, based on this expectations, the CBIR roused fast interest in the medical image community, due to its capability of retrieve diagnosed images in relation to an image with disease suspicion, which gives directions to aid the specialists [7]. Although part of the information about the diagnosis can be in the DICOM (Digital Imaging and Communication in Medicine) image header, this textual tag frequently shows high error rate. Literature has reported cases with 16% of error [8]. This way, the medical community has emphasized the adoption of alternative methods in relation to the traditional methods based on the data manual insertion into the DICOM files.

In the future, the CBIR and other CAD techniques will be implemented as a part of the PACS (Picture Archiving and Communication System) and it will be possible to search and retrieve similar cases to an unknown case. Nowadays, the majority of the clinical images in PACS have not been used for clinical purposes, except for comparing images from the same patients, such as a current image with previous image [1].

Image Registration is another important technique in medical image processing. The main applications have been to overlap different image modalities and to verify changes or
disease evolution coming from images acquired at different times [9]. The Image Registration is made of methods capable of finding spatial transformations necessary to map the homologous points between two images. These images in the registration process are called reference and target images. The group of methods capable to find the dependence of points between the images is obtained from four main operations: (1) the transformation of the target image coordinates to the image reference space; (2) a new image is made using interpolation of the reference space; (3) the new image is compared with the reference image using the similarity measurement; and (4) the transformation parameters are adjusted in agreement with the best match obtained. The transformations are done in an iterative way until the best match is obtained between the reference and target images. The similarity measurement is maximized or minimized by optimization process to obtain the best alignment [10].

Although CBIR and Image Registration techniques have the same goal of finding optimum similarity between two images, they are lead by the literature in an isolated way. This is because the methodologies used in Image Registration algorithms have high computational costs and the success of the CBIR applications in clinical routine are very dependent of their performance [4]. Therefore, to apply Image Registration as a CBIR technique is impracticable by using only one computer to process large medical image databases.

The Grid Computing (GC) technology represents one of the most recent and promising tool in distributed computing. GC is the integration of many computers distributed geographically, making it possible to create a virtual computing platform, giving to users and institutions a virtually unlimited capacity to solve problems related to the storage and access of data, and also to process applications with high computational costs. Moreover, GC creates an infrastructure to support remote and secure access to the distributed resources [11].

The shared processing capacity of GC allows the study of large data quantities using image processing algorithms that require high computational power [12]. The use of GC also allows that regional health networks combine resources into one large distributed image database, making it possible for medical communities to share data as well as applications and knowledge, which make a greater interaction among the medical centers possible [13].

Among the various CAD approaches, techniques focused on the Content-Based Medical Images Retrieval are a major beneficiary of the Grid Computing technology due to their characteristics and necessities: high processing and large storage. Besides, GC is a low cost solution for public hospitals and small clinics, because, it is able to use the idle recourses of computers [4].

Since the first project capable to identify the potential in applying the GC to the biomedicine proposed by Breton, et al. [14], a growing number of papers has revealed the importance in applying the GC to the medical science area. Try to finding the benefits offered by the recent technology, many federations were composed by academic institutions and governments to develop projects for explore the benefits of the GC in medicine. The MammoGrid is a GC project from the European Union to make a structure to share mammography and to use the CAD to detect microcalcifications and masses among their institutions [15]. The approach taken by Gurcan et al. [16] created a Grid-aware (GridIMAGE) to share medical images, compare interpretations between specialists and to use the CAD tools in a distributed way. However, the GC technology is still in the beginning of its applications in medical informatics [1]. The main goal of this work was to develop a CBIR algorithm capable of using the GC to process a big lung nodules database using the Image Registration technique. This work also shows that Image Registration techniques are accurate and can be applied to help physicians in classify lung nodules making use of a BigData of images as target. This is possible by using a methodology capable to filter an image database using Texture Attributes and also by using the GC techniques to amortize the high computation cost of the Image Registration algorithms.

II. MATERIALS AND METHODS

The algorithm was developed in the operation system GNU/Linux Debian using the Java 1.7 programming language. The prototype PACS server was developed with free software tools (PostgreSQL 8.4-13 and Hibernate 4.1.6) and it followed the DICOM compliances. The PACS server was ran on a PC using GNU/Linux Debian 3.1 with an Intel Pentium Xeon 3.2 Ghz processor and 4 Gbyte of RAM memory.

The Image Registration algorithms are an implementation of the InsightToolkit [17]. For its assessment, we have used 20,000 images from the Lung Image Database Consortium (LIDC), which is a big database of lung cancer [18]. In the LIDC each nodule is manually segmented and then classified in benign or malignant by physicians. A XML file describes the nodule’s texture characteristics and if a nodule is benign or malignant. To use as reference, we selected 100 benign nodules and 100 malignant nodules.

The classification of nodule in benign or malignant by specialists is a very challenging task, because the nodules have low contrast, are small (< 3.00 mm) and are attached to other complex lung structures. Furthermore, medical images are extremely complex and its diagnosis consists of a meticulous task that must be done by qualified specialists. However, errors in detect or classify a lung cancer are not uncommon.

A. Application description

The overall application is described in Fig. 1 and a detailed description is given in the sections below. All the images inserted in the PACS have removed the patients’ information in respect to the HIPAA (“Health Insurance Portability and Accountability Act Security Standard”) specifications [19].

The images used in this work have granted ethics board permission. For the better performance of the system, images’ binaries are not stored in the server, but only the images’ complete path. The Core of the application has easy access to the DICOM images using the Network File System (NFS).

The application has two CBIR modules. The first module uses the second-order Texture Analysis (co-occurrence matrix) to filter the 1000 most similar images into the second
module. The second module uses the Image Registration algorithms to find the similarity between an image defined by the user as a reference and the images filtered by the first module.

The application has a graphic interface that requires specialists to authenticate themselves. After authentication, the specialists obtain permission to seek and visualize the exam of interest. They can easily select a reference nodule image and start the first CBIR module. When the first module is finished, the nodule images are classified according to the smallest value of Manhattan Distance between the characteristic vector of reference image and the database images.

The second module starts when the specialists select the registration module. The Image Registration algorithm is executed in parallel on the grid machines using the reference image and the images classified by the first module. The application sorts in a list the most similar nodules according to the Cross Correlation values. Based on the DICOM protocol information, the application can also retrieve the information stored on the LIDC to aid the physician.

B. Grid middleware

Because of the high computational cost related to the Image Registration algorithms, the second module is processed on the OurGrid computational grid (www.ourgrid.org). This grid is free-to-join and cooperative, where sites share their idle computing resources and, when necessary, receive idle resources from other sites (Figure 2). OurGrid assumes that the parallel applications that run on it are Bag-of-Tasks (BoT), i.e., those tasks are independent from each other. Besides this, OurGrid ensures priority over the local resources for the local users [20]. Currently, the OurGrid is composed of nearly 500 computers.

The MyGrid is the OurGrid’s broker and its main function is to manage the applications that run on the Grid. MyGrid offers a set of abstractions that hide heterogeneity and complexity of the Grid Computing from the users. MyGrid uses task replication in order to obtain a better performance without relying on the information of the Grid or application [21]. In order to protect the data and the local resources OurGrid uses the SWAN (Sandboxing Without A Name). SWAN is an application based on the Xen virtual machine that isolates the users Grid applications in a sandbox, this way it does not allow Grid tasks to access the network or local data. Only the necessary basic resources to run the remote tasks are available to the Grid tasks.

MyGrid allows the data to be stored in the Grid machines or removed after the application has finished running. In our application, the images are not left in the Grid machines in order to assure the safety of the images. However, in order to optimize the application process time, the Image Registration algorithms libraries are stored in the Grid machines. In order to give further easiness and dynamism to the specialist in the application, MyGrid API was used to hide the complexities of the computational grid. The user only sees the application user interface. After all the tasks finished, MyGrid tells the application that the second module has finished.

III. RESULTS

The results of the CBIR tool developed in this work were assessed by the leave-one-out technic. The tests were repeated five times for each module and we have used different kinds of nodules for each test (ten benign nodules and ten malignant nodules). The nodules were selected in a random way. The algorithm precision was obtained by the equation:

\[ Precision = \frac{\text{type number of nodules retrieved}}{\text{number of nodules retrieved}}. \]

The first module’s time processing average was 1.9 minutes. This time was related to the calculation of the Manhattan Distance between the characteristic vector of each DICOM server’s image and the characteristic vector of the reference image. The average precision of retrieving obtained by the first module was 0.73 (benign nodules) and 0.76 (malignant nodules), considering the ten most relevant nodules retrieved. To analyze the image retrieval capacity of the second module, a group of 1000 most similar nodules filtered by the first module has used in the second module. The average precision of retrieving obtained by the second module was 0.72 (benign nodules) and 0.67 (malignant nodules). In all the experiments, the results produced by the Image Registration algorithms were very close to the traditional Texture Analysis technique.

The Bag-of-Task approach was able to greatly reduce the high processing time of the CC algorithm. The GC was able to amortize the total time of the algorithm in 116.97 minutes, compared to the processing time obtained in one machine. The experiments used 50 Grid processors and 100 MB/s LAN. The application total time, considering the time to calculate the
Manhattan Distance and the time to execute the CC algorithm, was 5.02 min. We divided the application into 20 tasks consisting of 50 images each. The images were compacted before being sent to the Grid and the average size of the files with 50 images was 4 MB.

The SWAN ensured the necessary security to process the medical images in the Grid’s machines and after the images had been processed they were removed from the Grid’s machines. Moreover, the OurGrid is capable to ensure the transmission of the medical images to the Grid’s machines using cryptography.

The MyGrid allowed that the Image Registration algorithm libraries (7.8 MBytes) to be stored in the Grid’s machines, which reduced the time process of the application and avoided the unnecessary data traffic in the network, which is one of the aggravating time consumptions in the application.

IV. DISCUSSION

The Grid Computing is still being confirmed as a new opportunity to establish even more the use of the CBIR technology as a CAD tool and also allows the algorithms with a high computational cost, which were only being used in researches until today, to be applied in the clinical routine [4].

The actual clinical scenario demands for new CBIR methods capable to process huge images databases. The applications used in the clinical routine should be precise, have low computational cost and be compatible for the medical information system and the PACS system [1].

This work showed a mixed CBIR techniques solution using Image Registration and Texture Analysis. These techniques are based on the nodule’s texture and its border, which are the main descriptors used by the specialists to classify the lung nodules [22]. Oliveira MC and colleagues [4] showed a preliminary work using Image Registration as CBIR technique. However, the authors did not focus his attention in a specific disease. To our knowledge the results using Image Registration and Texture Analysis to retrieve lung nodules have never been published. Furthermore, our results evidenced that Image Registration is precise and effective in retrieve similar lung nodules, besides that, the Image Registration has showed very close results to the traditional Texture Analysis technique.

A CBIR system needs to be precise in the first images retrieved to really aid the specialist in diagnosis [7]. The IR mean precision in the ten images retrieved was 0.72 (benign nodules) and 0.67 (malignant nodules). This little gap precision between the results is due the complex structure of the nodule used as reference.

The retrieval processing time was more than 100 minutes in only one computer using Image Registration techniques. This time is impracticable for a CBIR to be applied in the clinical routine. In the grid, however, this time was reduced to less than 3 minutes in mean, making it affordable for clinical use.

V. CONCLUSION

The CAD acting as second opinion and helping physicians in the detection of subtle diseases has become part of the clinical work. In the future, methods such CBIR will be incorporated into the CAD to help physicians in the patient’s diagnosis. Besides this, it is also expected that the CAD schemes will be incorporate into the PACS being a useful tool for the diagnosis in the clinical routine. Nowadays, the CBIR methods are still limited to detect subtle diseases. Therefore, the CBIR methods needs that new techniques of image processing being integrated into the traditional methods, besides that, the CBIR must be able to process large medical image database.

A computational algorithm was developed in this work, capable of using the high processing power of the Grid Computing technology to make feasible the CBIR using the Image Registration algorithms against a BigData of lung cancer images. The CBIR and Image Registration techniques are well established in the medical image area; however, they are covered by the literature in an isolated way. This is because the Image Registration algorithms have high computational costs making them impracticable for the CBIR that must ensure fast results for specialists in the processing of medical image stored in BigData. Therefore, a Bag-of-Task approach was fundamental to amortize the total processing time of the Image Registration algorithms against a BigData of lung cancer images. Besides this, we have shown a new methodology to evolve CBIR’s state of art techniques through the Image Registration techniques.

ACKNOWLEDGMENT

Special thanks to CNPQ for the financial support granted to the project.

REFERENCES


