The MAGIC-5 Project:
Medical Applications on a Grid Infrastructrure Connection

Ivan De Mitri
Dipartimento di Fisica - Università di Lecce, Italy
and Istituto Nazionale di Fisica Nucleare - Sezione di Lecce, Italy

On behalf of the MAGIC-5 Collaboration

Abstract

The main purpose of the MAGIC-5 collaboration is the development of Computer Aided Detection (CAD) software for **M**edical **A**pplications on distributed databases by means of a **G**RID **I**nfrastructure **C**onnection. A prototype of the system, based on the AliEn GRID Services is already available with a central Server running common services and several clients connecting to it. It has been already successfully used for applications in mammography together with a specific CAD developed within the collaboration. Applications to the case of malignant nodule detection in lung CT scans are now being implemented, while a use of the GRID services is also being applied to PET image analysis aiming at early diagnosis of Alzheimer disease. In this work the status of the project and its future prospects will be given, with particular attention to the data management and processing aspects. Medical applications carried on by the collaboration will be also described together with the analysis of the results so far obtained.

Keywords: Medical Image Processing , CAD system, Grid, Virtual Organization

1 Introduction

It has been shown that screening programs are of paramount importance for early cancer diagnosis in asymptomatic subjects and consequent mortality reduction. The development of Computer Aided Detection (CAD) systems would significantly improve the prospects for the screenings, by working either as second reader to support the physicians's diagnosis or as first reader to select images with highest cancer probability.

The amount of data generated by such periodical examinations is so large that it can not be managed by a single computing center. As an example, let us consider a mammographic screening program to be carried out in Italy: it should check a target sample of about 6.8 millions of women in the 49-69 age range, thus implying 3.4 millions of mammographic exams/year. For an average data size of 60 MB/exam, the amount of raw data would be of the order of 200 TB/year. On a European scale, the data source would be comparable to one of the next generation High Energy Physics (HEP) experiments (1-2 PB/year).

In addition, the image collection in a large scale screening program intrinsically creates a distributed database, involving both hospitals, where the data are recorded, and diagnostic centers, where the radiologists should be able to query and analyze all the images and the related data. This amount of data grows with time and a full transfer over the network would be large enough to saturate the available connections.

On the other hand, the need for making the whole database available, regardless of the data distribution, would provide several advantages. For example, a new CAD system could be trained on a much larger data sample, with an improvement of its performances in terms of both sensitivity and specificity. The CAD algorithms could be used as real time selectors of images with high cancer probability, with a remarkable reduction of the delay between acquisition and diagnosis. Moreover, the data associated to the images, also known as *metadata*, would be available to select the proper input for epidemiology studies or for the training of young radiologists.

This framework requires huge distributed computing efforts as for the case of the HEP experiments, e.g. the CERN/LHC collaborations. The best way to tackle these demands is to use the GRID technologies to manage distributed databases and to allow real time remote diagnosis. This approach would provide access to the full database from everywhere, thus making possible large-scale screening programs.

The MAGIC-5 project perfectly fits in this framework, as it aims at developing medical applications that make use of GRID Services, starting from a data model similar to that adopted by the ALICE collaboration [1]. In particular, the project is an evolution of a former activity - GP-CALMA [2] - that was mainly devoted to large scale mammographic screening. MAGIC-5 include those aspects together with new efforts in order to develop a CAD system for the detection of malignant nodules in lung CT scans and the management of the related distributed database. A new application devoted to the Alzheimer desease is now being implemented in the MAGIC-5 GRID infrastructure.

From this point of view, the collaboration can be seen as one or more Virtual Organizations (VO), with common services (Data and Metadata Catalogue, Job Scheduler, Information System) and a number of distributed nodes providing computing and storage resources. There are three main differences with respect to the model applied to the HEP experiments, that can be summarized as follows: some of the use cases require interactivity; the network conditions do not allow the transfer of large amounts of data; the local nodes (the hospitals) do not agree on the raw data transfer to other nodes.

According to these constrains, the MAGIC-5 approach to the implementation of a prototype is based on two basic tools: AliEn [3] for the management of the common services, and PROOF [4] for the interactive analysis of remote data without transfer.

Data management and processing will be discussed in Sec.2, while medical applications will be shown in the rest of the paper together with the description of the implemented CAD algorithms and their performances. Finally, the present status and the future plans of the project will be given in the Sec.4.

2 Data management and processing

A dedicated AliEn Server for the MAGIC-5 Collaboration has been configured [2], with a central Server running common services and several clients connected to it.

The images can be acquired in any hospital belonging to the Collaboration: the data are stored on local resources and registered to a common service, known as *Data Catalogue*, together with other related information, the *metadata*, required to select and access them at any future time. The result of a query can be used as input for the analysis through the various CAD systems, which are executed on nodes that are, usually, remote to the user, thanks to the ROOT/PROOF facility. A selection of the cancer candidates can be quickly performed and only images with high cancer probability would be transferred to the diagnostic sites and interactively analyzed by the radiologists. This approach avoids data transfers for images with a negative CAD response and allows an almost real time diagnosis for the images with high cancer probability.

In order to make the images available to a remote Diagnostic Center, a mechanism able to identify the data corresponding to the exam in a site-independent way is used: the images are selected by means of a set of requirements on the attached metadata and identified through a Logical Name which is independent on the physical name on the local hard drive where they are stored. AliEn [3] implements these features in its Data Catalogue Services, run by the Server: data are registered making use of a hierarchical namespace for their Logical Names and the system keeps track of their association to the actual name of the physical files. In addition, it is possible to attach metadata to each level of the hierarchical namespace. The Data Catalogue can be browsed from the AliEn command line as well as from its Web portal. The metadata associated to the images can be divided into several categories: patient and exam identification data, results of the CAD algorithm analysis, radiologist's diagnosis, histological diagnosis, and so on.

Both tele-diagnosis and tele-training require interactivity in order to be fully exploited. The PROOF (Parallel ROOt Facility) system [4] provides the functionality required to run interactive parallel processes on a distributed cluster of computers. A dedicated cluster of several PCs was configured and the remote analysis of digitized mammograms without data transfer was recently run. The basic idea is that, whenever a list of input Logical Names is selected, it generates a list of physical names, one per image, consisting of the node name corresponding to the Storage Element where it is located, and the physical path on its file-system. The information is used to dynamically generate a C++ script driving the execution of the CAD algorithm, which is sent to the remote node. Its output is a list of positions and probabilities corresponding to the image regions identified as pathological by the CAD algorithm. Based on that, it is possible to decide whether the image retrieval is required for immediate analysis or not.

3 Medical applications

The medical applications of the MAGIC-5 Project cover two main fields: breast cancer detection in mammographic images and nodule detection in lung Computed-Tomography (CT) images.

While the analysis of mammographic images started some years ago, the detection of malignant nodule in CT scans represents a new activity of the collaboration. In the following sections a brief review of the CAD systems till now developed in both applications will be given. Moreover, a GRID implementation for the diagnosis in Alzheimer disease (AD) will be mentioned as a neuro-application which is going to be implemented in a GRID environment by the collaboration.

3.1 Mammographic CAD systems

A database of mammographic images was acquired in the hospitals belonging to the collaboration. Pathological images have been diagnosed by experienced radiologists and confirmed by histological exam; they contain a full description of the pathology including radiological diagnosis, histological data, type and location. This information provides the truth, the CAD results will be compared with. Images with no sign of pathology were considered as healthy and included in the database after a follow up of three years.

The images were digitized by means of a Linomed CCD scanner with 85 μm pitch and 12 bits per pixel. Each image is thus described by 2657×2067 pixels with $G = 2^{12} = 4096$ grey level tones.

Two different kinds of structures could mark the presence of a breast neoplasia: massive lesions (ML) and microcalcification clusters (MC). Massive lesions are rather large (diameter of the order of centimeters) objects with very different shapes, showing up with faint contrast. Microcalcification clusters consist in groups of rather small (approximately from 0.1 to 1.0 mm in diameter) but very brilliant objects. The database composition is reported in the table 1.

images with ML	images with MC	healthy images
1153	287	2322

Table 1: Composition of the MAGIC-5 mammographic image database.

Different CAD systems have been developed for ML and MC detection. For both cases, the algorithms consist in three main steps:

1. **segmentation**: to perform an efficient detection in a reasonable amount of time, a reduction of the image size is required, without missing any pathology; to this purpose, some portions of the mammogram, having the highest probability to contain the pathology, are selected with a demand of efficiency as close as possible to 100%.

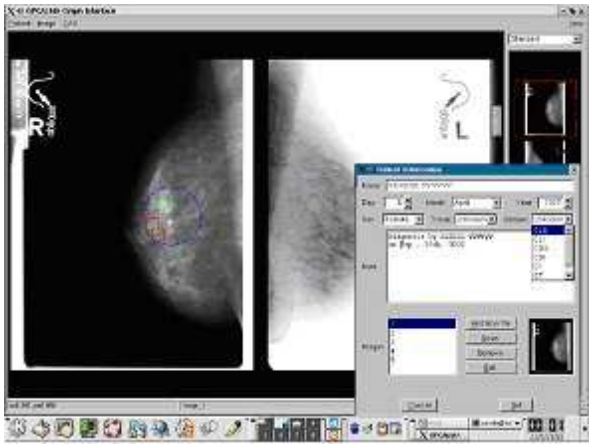

Figure 1: The Graphic User Interface. Three menus allow to browse the Patient, the Image and the CAD diagnosis levels. On the left mammogram, the CAD results for microcalcifications and masses are shown in red squares and green circles, respectively, together with the radiologist's diagnosis (blue circle).

2. **feature extraction**: the portions of the mammogram extracted by the segmentation step are characterized by proper sets of features;

3. **classification**: the selected regions are used as inputs to a supervised two-layered feed-forward neural network whose output provides a degree of suspiciousness for the corresponding region.

A detailed description of the CAD algorithms for massive lesion and microcalcification cluster detection is given in [5] and [6], respectively, where the results obtained, in terms of sensitivity and fraction of false positives, are also reported.

3.1.1 The CAD station

The hardware requirements for the CAD station consists of a PC running Linux connected to a planar scanner and to a high resolution monitor. The station allows human or automatic analysis of the digital mammogram which can be directly acquired by the scanner or from a file. The software configuration for the local mode use requires the installation of ROOT [4] and GPCALMA, which can be downloaded in the form of source code from the respective CVS servers.

A Graphic User Interface (GUI) has been developed (see figure 1) to drive the execution of three basic functionalities related to the Data Catalogue:

1. registration of a new patient, based on the generation of a unique identifier, which could be easily replaced by the *Social Security* identification code;

2. registration of a new exam associated to an existing patient;

3. query to the Data Catalogue to retrieve all physical file names of the exams related to a patient and, eventually, analyze them.

The images are displayed according to the standard format required by the radiologists: for each image, it is possible to insert or modify diagnosis and annotations, and to manually select the portion of the mammogram corresponding to the radiologist indication. An interactive procedure allows a number of operations such as zooming, windowing, gray levels and contrast selection, image inversion, luminosity tuning. The human analysis produces a diagnosis of the breast lesions in terms of kind, localization on the image, average dimensions and, if present, histological type. The automatic CAD procedure finds the Regions of Interest (ROIs) of the image with a probability of containing a pathological area larger than a pre-defined threshold value.

3.2 Malignant nodule detection in CT scans

The detection of malignant nodule in lung Computed-Tomography (CT) images represents the new-born activity of the MAGIC-5 Collaboration. Two initial steps for the development of a CAD system have been implemented:

1. the automated extraction of the pulmonary parenchyma;

2. the detection of nodule candidates based on a dot-enhancement filter.

The first step aims at removing from the CT image all pixels located outside the chest. It is based on a combination of image processing techniques to identify the pulmonary parenchyma, such as threshold-based segmentation, morphological operators, border detection, border thinning, border reconstruction, and region filling.

The nodule detection step relies on the application of a dot-enhancement filter [7] to the 3D matrix of voxel data. A simple threshold-based peak-detection algorithm is then applied to the filter output. The above mentioned processing steps have been applied to sets of lung multi-slice CT scans acquired at high (standard setting: 120 kV, 120 mA) and low (screening setting: 120 kV, 20 mA) dose and with different slice thickness (1 mm and 5 mm). Each scan consists of a sequence of about 300 slices (for 1 mm slice thickness series) stored in the **DICOM** (**D**igital **I**maging and **CO**mmunications in **M**edicine) format.

Preliminary results show that the filter identifies spherical nodules, thus being acceptable as a pre-processing stage for nodule detection. Moreover, the filter is effective for low-dose scans, a fact which is desirable in the view of a screening. Yet, the algorithm is sensible to some saddle-like configurations of blood vessels too, thus generating a high number of false positive peaks. Further processing stages are under study to eliminate this drawback.

3.3 A GRID implementation for the Alzheimer disease diagnosis

The Alzheimer disease (AD) is the leading cause of dementia in elderly people. Clinically, AD is characterized by a progressive loss of cognitive abilities and memory.

One of the most widely used tool for the analysis of medical imaging volumes for neurological applications is the SPM (Statistical Parametric Mapping) software which has been developed by the Institute of Neurology at the University College in London. SPM provides a number of functionalities related to image processing and statistical analysis, such as segmentation, co-registration, normalization, parameter estimation, statistical mapping. The quantitative comparison, through the SPM software, of PET images from suspected AD patients with the ones included in a database of normal cases, allows powerful suggestions to an early diagnosis of AD.

To this purpose, the use of an integrated GRID environment for the remote and distributed processing of the PET images at a large scale, is strongly desirable. This application is now being implemented in the MAGIC-5 GRID infrastructure.

4 Present status and future plans

The GRID approach to the analysis of distributed medical data is very promising. The AliEn Server managing the VO services has been installed and configured, and some AliEn Clients are in use. The remote analysis of mammographic images successfully works, thanks to the PROOF facility. Presently, all blocks required for the implementation of the tele-diagnosis and screening use cases are integrated into a prototype system. The ROOT functionality has improved the GUI, which is now considered satisfactory by the radiologists involved in the project, due to the possibility to manipulate the image and the associated metadata.

The MAGIC-5 GRID philosophy relies on the principle that the images are collected in the hospitals and analyzed by means of the CAD systems; only the images with a high probability to carry a pathology are moved over the network to the diagnostic centres, where the physicians can analyze them, almost in real time, by taking advantage of the CAD selection.

A future prospect of our project is the migration from the AliEn middleware to the EGEE/gLite middleware which is likely to become a European standard and will certainly provide more sophisticated tools with respect to the present AliEn functionality.

The medical applications are continuously under development. Both new algorithms (pulmonary CAD) and improvements of the existing ones (mammographic CADs) are under study. At the same time, part of the future work will be focused on the collection of a CT image database (at present, a limited number of scans is already available) and the implementation of the VO related to the PET image analysis for the early AD diagnosis.

References

[1] http://alice.cern.ch

[2] U. Bottigli et al, *GPCALMA: a tool for mammography with a GRID connected distributed database*, Proc. of the Seventh Mexican Symp. on Medical physics 2003, vol.682/1, pag.67
also e-preprint *physics/0410084*

[3] http://alien.cern.ch

[4] http://root.cern.ch

[5] F. Fauci et al., *Mammogram Segmentation by Contour Searching and Massive Lesion Classification with Neural Network*, Proc. IEEE Medical Imaging Conference, October 16-22 2004, Rome, Italy.

[6] C. S. Cheran et al., *Detection and Classification of Microcalcifications Clusters in Digital Mammograms*, Proc. IEEE Medical Imaging Conference, October 16-22 2004, Rome, Italy.

[7] M. Aoyama, Q. Li, S. Katsuragawa, F. Li, S. Sone, K. Doi, *Computerized scheme for Determination of the Likelihood measure of malignacy for pulmonary nodules on low-dose CT images*, Medical Physics 30 (3), 387-441, 2003.