GPCALMA: An Italian Mammographic Database of Digitized Images for Research

Adele Lauria[1], Raffaella Massafra[2], Sabina Sonia Tangaro[2], Roberto Bellotti[2,3], MariaEvelina Fantacci[4], Pasquale Delogu[4], Ernesto Lopez Torres[5], Piergiorgio Cerello[6], Francesco Fauci[7], Rosario Magro[7], and Ubaldo Bottigli[8]

[1] Università di Napoli "Federico II", Dipartimento di Scienze Fisiche, and Istituto Nazionale di Fisica Nucleare, Sezione di Napoli, via Cinthia, I-80126 Napoli, Italy
adele.lauria@na.infn.it
[2] Università di Bari and Istituto Nazionale di Fisica Nucleare, Sezione di Bari, Italy
[3] TIRES, Center of Innovative Technologies for Signal Detection and Processing, Bari, Italy
[4] Università di Pisa, Pisa, and INFN, Sezione di Pisa, Italy
[5] CEADEN, Habana, Cuba
[6] INFN, Sezione di Torino, Torino, Italy
[7] Università di Palermo, Palermo, and INFN, Sezione di Catania, Italy
[8] Università di Siena, Siena, and Istituto Nazionale di Fisica Nucleare, Sezione di Cagliari

Abstract. In this work the implementation of a database of digitized mammograms is described. The digitized images were collected since 1999 by a community of physicists in collaboration with radiologists in several Italian hospitals, as a first step in order to develop and implement a Computer Aided Detection (CAD) system. 3369 mammograms were collected from 967 patients; they were classified according to the type and the morphology of the lesions, the type of the breast tissue and the type of pathologies. A dedicated Graphical User Interface was developed for mammography visualization and processing, in order to support the medical diagnosis directly on a high-resolution screen. The database has been the starting point for the development of other medical imaging applications such as a breast CAD, currently being upgraded and optimized for the use in conjunction of the GRID technology in the framework of the INFN-funded MAGIC-5 project.

1 Introduction

A medical images dataset is considered the starting point for important epidemiological and statistical studies and also to develop and test algorithms for CAD systems, but also for teaching and training of medical students and as an archive of rare cases. In 1995 Osuch et al. proposed a mammography database for a national mammography inspection and to monitor patients through a centralized system [1]. Technological improvements in digitizing scanners make now possible to digitize radiographic films with no significant loss of information. At the moment many large datasets of digitized mammograms are available on the web [2,3]. Other databases, also "GRID compliant", are described in the literature [4-6]. The development of a CAD system is strictly tied to the collection of a large dataset of selected images.

Susan M. Astley et al. (Eds.): IWDM 2006, LNCS 4046, pp. 384 – 391, 2006.

In this work a full description of the GPCALMA (*Grid Platform for Computer Aided Library in Mammography*) database is given.

2 Method

Images were acquired in various mammographic centers using different mammographic screen/film systems and settings (all with molybdenum anode) and in the framework of different applications, including both clinical routine carried out on symptomatic women, and screening programs addressed to asymptomatic women. Moreover, many images come from an archive of particularly meaningful clinical cases collected in the previous years at the Bari hospital. Unfortunately at the moment of the digitization the information about acquisition settings were no more available, thus making impossible normalization procedures. A workstation, composed of a PC running the Linux operating system and a film scanner, was installed at each site involved in the program. Digitized images are stored in a dedicated hard disk, which presently stores the whole GPCALMA database of mammographies. All the mammograms of the database were digitized using the same digitizer model and under the same conditions in order to avoid fake features caused by variations in the digitization step. A CCD scanner was used, choosing [7] a pixel size of 85 μm and a 12 bit depth. The typical scan time is 20s. The acquisition software provided with the scanner was modified to scan and save images in a special format (called CALMA format) consisting of a long vector of numbers corresponding to the pixel intensities and two other numbers representing the image dimension. These numbers are used to transform the vector in a matrix: each pixel of the image can be represented by a triplet (*x, y, I*), where *x* is the row number, *y* is the column number and *I* is the intensity of the pixel, ranging from 0 (black) to 4095 (white). Such workstations have been continuously operative in various collaboration sites for several years without problems. In sites where clinical studies were performed, the PC was connected to a high resolution and high luminosity B/W LCD monitor.

3 Description of the Database

The database is composed of 3369 mammographic images, each including data and clinical information. Images were collected from 967 patients. The age distribution is reported in figure 1. Each patient has from one to six views, according to the distribution shown in figure 2. The repartition of the database in left/right breast images is 1835 (51%) and 1734 (49%) respectively, while for the craniocaudal/oblique/lateral views is 1601 (48%), 1456 (43%) and 312 (9%) respectively. The image size is 2067 x 2657 pixels, 85μm of pitch, 12 bit/pixel (4096 grey levels); each image file is about 8 Mbytes. All the mammographic images with other information related to the patient (follow up, age of patients and interesting cases) were collected in the Italian hospitals involved in the collaboration from 1997 to 2002. The geographic provenience of the images is shown in figure 3.

Prior to being processed all images were anonymized. All the images of the database containing one (or more) lesions were characterized according to the kind of