# The Database for Digital Breast Sonography

**Ying-Tao Zhang[1], H. D. Cheng[2, 1], Xiang-Long Tang[1], Yan-Hui Guo[1], Jia-Wei Tian[3]**

[1] Department of Computer Science, Harbin Institute of Technology, Harbin, China
[2]Computer Science Department, Utah State University, Logan, Utah, USA
[3]Department of ultrasound, Second Affiliated Hospital of Harbin Medical University, Harbin, China.

## Abstract

We build a digital breast ultrasound image base. The aim of this work is to provide a large set of digital breast ultrasound images to evaluate and compare the performance of the CAD (computer-aided diagnosis) algorithms and systems. This database would be useful for many purposes including research, education, quality assurance, demonstration, etc. This paper provides an overview of the database, including the background of the database, the categorization and information of the stored cases.

**Keywords**: ultrasound; Sonography; breast cancer; database; benchmark;

## 1. Introduction

Breast cancer is the most prevalent cancer among women. It is the main cause leading to death by cancer among woman of ages 15-54[1]. Many computer-aided diagnosis (CAD) algorithms for breast cancer were studied in the past five decades. For an algorithm, different testing images may produce different results; hence, it is very difficult to compare two algorithms without a common testing database. To evaluate algorithms fairly, some benchmarks in many research fields were built up in the past [2-4]. However, to authors' knowledge, there is no a benchmark for breast ultrasound images yet. We are building a database for digital breast ultrasound images (DDBUI), which could be used to develop and test computer-aided diagnosis (CAD) algorithms and systems. Currently, the database contains more than 60 cases (156 images in all). All the images were taken and selected during diagnostic breast examinations at the Second Affiliated Hospital of Harbin Medical University, Harbin, China.

## 2. The DDBUI

The image base may provide a useful means for developing new algorithms, comparing the performance of the algorithms and systems developed by different research groups, evaluating the functionality of CAD systems, and training the students and researchers. Currently, the cases in the database are adequate for such purpose, and we will increase the number of cases up to 100.

### 2.1. Background

We had studied breast cancer detection using mammograms for many years, developed several some algorithms and published several articles [5-9]. Due to the limitations of mammograms and the advantages of ultrasound imaging, more and more researchers pay attention to sonograph. However, there is no any public data available yet, which may withhold the progress in developing CAD algorithms and systems. We are trying to fulfill such task.

### 2.2. Case Gathering

The database has the following features:

- The age of patients mainly is 18 to 55 years old.
- The cases were collected retrospectively and all had been biopsied.
- The images were acquired with an ATL 3000 unit (GE, VIVID7), using an L5-14MHz linear probe, and captured directly from the video signals.
- The vertical dimension of most images was 3 to 4 cm, and the horizontal dimension was 5 to 6 cm.
- The images were quantified into 256 gray levels. The average size was 450X450.
- The number of images taken from each patient varied from one to ten, and most cases have two to five images.
- The database includes a wide range of lesions with different sizes and different characters.
- Six radiologists take part in marking all lesions. Their experience spans from 2 to 26 years.

The procedure of image collection is described below. The patient lies in a supine or oblique position, with ipsilateral arm above the head. The breast was scanned in sagittal or transverse and radial directions.

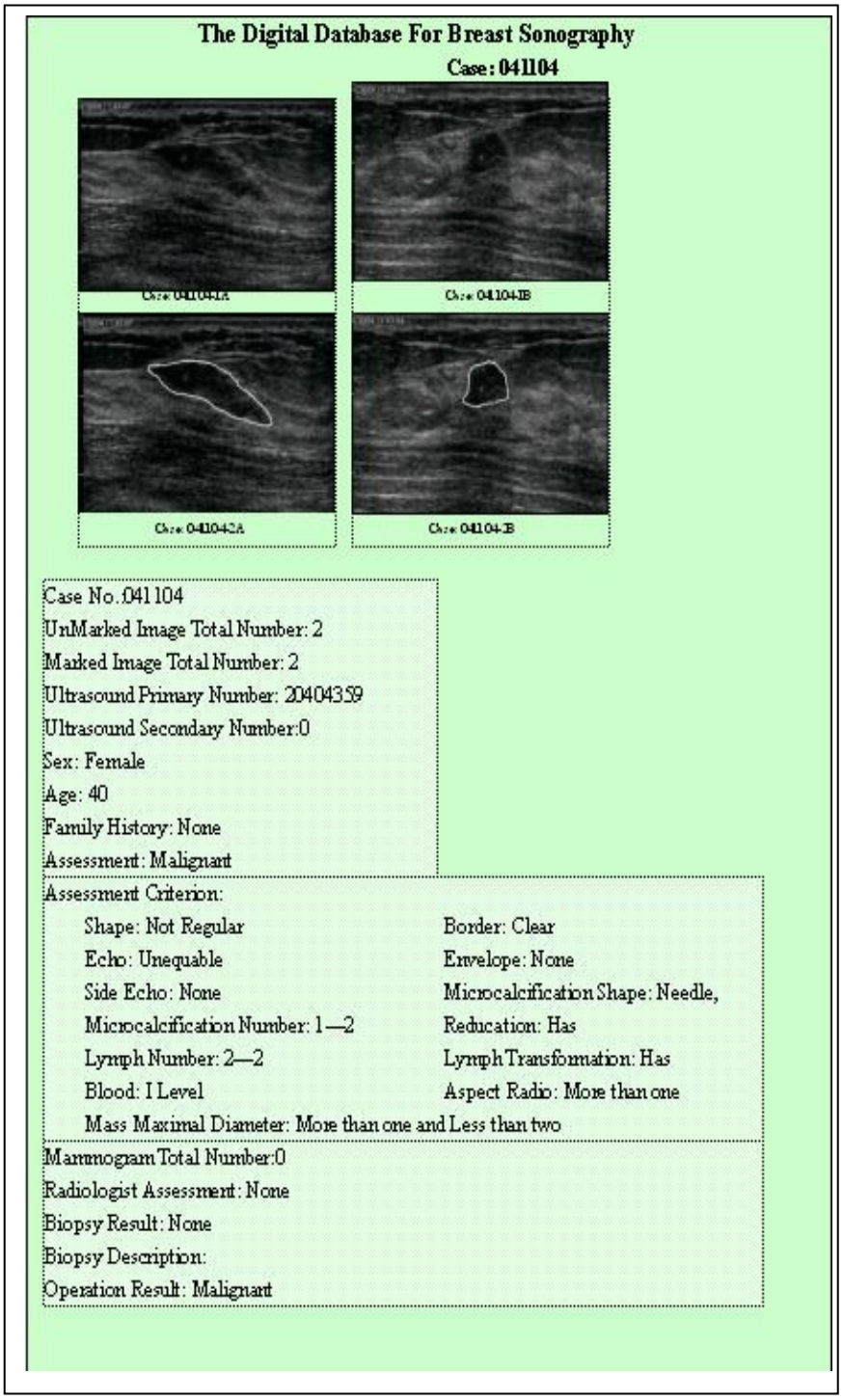


Figure. 1. The thumbnail and related information of the images for case 041104

Usually only abnormal findings are recorded. However, in our study, women with health breasts were also examined to provide normal images for the image base.

An important work for constructing the database is to select the proper parameters for digital images: small pixel images provide high spatial resolution but require lots of storage space and long computation time for image processing. The parameters used for the images in this database were appropriate for detecting breast cancer and for the evaluation of computer-aided diagnosis (CAD) algorithms and systems.

## 2.3. Categorization

The cases were divided into three classes (normal, benign, malignant) based on the findings. Normal class contains images from screening examinations that were diagonized as normal by at least 3 radiologists. Benign class contains images in which suspicious region was found and the patient was recalled for additional work-up that resulted in a benign finding. Each malignant case included at least two images, and lesions on images had outlined by radiologist.

Figure 1 is an example in the database. The top ones are the original images, and the bottom ones are the images with marks in solid lines surrounding the suspicious areas.

## 2.4. Preprocessing

Before inputting into the database, some patient information was removed from the record. The database contains text files including the following information of each patient: age, sex, final diagnosis, degree of subtlety, location of the mass, etc. [10-11]. The details are listed in Table 1

## 3. Software Tools

All images are saved in lossless jpg format. Data for an image measurement were stored in a file, which was written in ASCII to describe information about the sample, experiment, operation, etc. As shown in Figure 1, and each case had a ground truth data that depicted the patient's information and characters of lesions. The boundary of abnormality was marked manually by radiologists.

**Table1** Image Database Form

| Item | Option | | | |
|---|---|---|---|---|
| Image Number | | | | |
| Original Image | | | | |
| Marked Image | | | | |
| Patient Sex | Female () | | Male () | |
| Patient Age | | | | |
| Has Family History | Has () | | None () | |
| Assessment | Mass () | Micro-calcification () | Benign () | Malignant () |
| Type | Mass () | | Micro-calcification () | |
| Assessment Criterion | | | | |
| Has Mammogram | Has () | | None () | |
| Radiologist Assessment | Mass () | Micro-calcification () | Benign () | Malignant () |
| Biopsy | None () | Normal () | Benign () | Malignant () |
| Note | | | | |

The images can be downloaded from http://pr-ai.hit.edu.cn/DDBS.html. For any commercial usage, the users should get the permission by writing to gyh78@126.com. For non-commercial usage, the authors should write the acknowledgement: The images were provided by Harbin Institute of Technology and the Second Affiliated Hospital of Harbin Medical University.

## 4. Conclusions

An open digital breast ultrasound image base is established. A large set of digital breast ultrasound images is collected and can be used to evaluate and compare the performance of algorithms and systems. All information about the patients excepting names and the characteristics of lesions on ultrasound images is available and sorted in a file. The lesions biopsy or operation results are provided and the suspicious areas are outlined.

## References

[1] M. Heath, K. Bowyer, D. Kopans, R. Moore and P. Kegelmeyer Jr. "The Digital Database for Screen Mammography", *5th International Workshop on Digital Mammography*, ECCV, 2004

[2] http://bias.csr.unibo.it/fvc2004/
[3] http://www.neuroinformatik.ruhr-uni-bochum.de/ini/PEOPLE/rolf/messer-icpr.html
[4] http://www.ajronline.org/cgi/content/full/174/1/71
[5] A. T. Stavros, D. Thickman, C.L. Rapp, M. A. Dennis, S. H. Parker, G. A Sisney. "Solid Breast Nodules: Use of Sonography to Distinguish between Benign and Malignant Nodules. " *Radiology* 196, 123-134,1995
[6] H. D. Cheng, J. L. Wang and X. J. Shi, "Microcalcification Detection using Fuzzy Logic and Scale Space Approaches", *Pattern Recognition*, 37(2): 363-375, 2004.
[7] H. D. Cheng and Y. Cui, "Mass Lesion Detection with a Fuzzy Neural Network", *Pattern Recognition*, 37(6): 1189-1200, 2004.
[8] H. D. Cheng, X. P. Cai, X. W. Chen, L. M. Hu and X. L. Lou, "Computer-Aided Detection and Classification of Microcalcifications in Mammograms: a Survey", *Pattern Recognition* 36(12): 2967-2991, 2003
[9] H. D. Cheng, Y. G. Hu, D. L. Hung and C. Y. Wu, "Breast Cancer Classification Using Fuzzy Central Moments", *Fuzzy Logic in Medicine*, S. Barro and R. Marin, eds. , Springer-Velerg, 2002.
[10] American College of Radiology. American College of Radiology standards. *Reston (VA): American College of Radiology*; 2002.
[11] A. F. Miguel, A. F. Patricia, A. L. Luis, S. M. Jose, F. P. Rafael and T. P. Agustin, "Computer-Aided Measurement of Solid Breast Tumor Features on Ultrasound Images", *ECCV*, 2004.