USP

Campus de São Carlos

ELABORAÇÃO DE BANCO DE IMAGENS MAMOGRÁFICAS DIGITALIZADAS

Autor: Rodrigo Henrique Benatti

Orientador: Prof. Dr. Homero Schiabel

UNIVERSIDADE DE SÃO PAULO

ESCOLA DE ENGENHARIA DE SÃO CARLOS

Serviço de Pós-Gradução EESC/USP

EXEMPLAR REVISADO

Data de entrada no Serviço: 01/07/03

Ass.:

ELABORAÇÃO DE BANCO DE IMAGENS MAMOGRÁFICAS DIGITALIZADAS

Rodrigo Henrique Benatti

EESC - USP Biblioteca CAMPUS DE S. CARLOS

Dissertação apresentada à Escola de Engenharia de São Carlos, da Universidade De São Paulo, como requisito para obtenção do Título de Mestre em Engenharia Elétrica.

DEDALUS - Acervo - EESC

31100043558

Orientador: Prof. Associado HOMERO SCHIABEL

São Carlos

2003

Class. TESE
Cutt. 0699
Tombo T226/03
Sysno 1326762

Ficha catalográfica preparada pela Seção de Tratamento
da Informação do Serviço de Biblioteca – EESC/USP

B456e

Benatti, Rodrigo Henrique
Elaboração de banco de imagens mamográficas digitalizadas / Rodrigo Henrique Benatti. -- São Carlos, 2003.

Dissertação (Mestrado) -- Escola de Engenharia de São Carlos-Universidade de São Paulo, 2003.
Área: Engenharia Elétrica.
Orientador: Prof. Assoc. Homero Schiabel.

1. Banco de dados. 2. Mamografia. 3. Digitalização de imagens. I. Título.

FOLHA DE JULGAMENTO

Candidato: Engenheiro **RODRIGO HENRIQUE BENATTI**

Dissertação defendida e julgada em 30-04-2003 perante a Comissão Julgadora:

APROVADO

Prof. Assoc. **HOMERO SCHIABEL (Orientador)**
(Escola de Engenharia de São Carlos/USP)

APROVADO

Prof. Dr. **ADERSON TADEU BEREZOWSKI**
(Faculdade de Medicina de Ribeirão Preto/USP)

Aprovado

Prof. Dr. **LUÍS GUSTAVO NONATO**
(Instituto de Ciências Matemáticas e de Computação/USP)

Prof. Assoc. **MURILO ARAUJO ROMERO**
Coordenador do Programa de Pós-Graduação
em Engenharia Elétrica

Profa. Assoc. **MARIA DO CARMO CALIJURI**
Presidente da Comissão de Pós-Graduação

Agradecimentos

Ao Prof. Homero Schiabel, que há muito tempo deixou de ser "Professor Homero" para se tornar simplesmente "Homero", pela orientação, amizade e compreensão nos momentos de indecisão (e ele sabe que não foram poucos), desde os tempos de iniciação científica até o final deste trabalho. Lá se vão 7 anos!! Obrigado.

Ao meu pai e minha mãe, que mesmo geograficamente distantes sempre me apoiaram e incentivaram a levar o trabalho adiante, sempre dando suporte às minhas escolhas. MUITO obrigado mesmo !!

À Fátima, agora professora Dra., por ser minha segunda orientadora. Obrigado pela ajuda técnica, pelo apoio, por me ajudar a cumprir as metas de acordo com o planejado, e até mesmo pelos "puxões de orelha".

Ao Guilherme Correa Stamato, o "Dr. Haule" pelos anos de convivência na República da Bruxa® e por ser, muitas vezes, o ajudante-orientador-controlador-republicano e, obviamente, companheiro de festas e etc. "Valeu, fio".

Aos amigos da República da Bruxa® (atuais e antigos moradores), por suportarem a ala "geriátrica" da casa, muitas vezes sem entender os motivos, mas também pela ajuda no longo período de convivência.

À grande amiga Ana "Embrapa" Claudia Martinez, pela ajuda não só técnica-computacional, mas também pela ajuda "extra-laboratório", que você sabe foi muito grande. Aliás, continua até agora....

Ao Claudio pela ajuda na resoluçao dos problemas computacionais, sempre dando dicas e indicando os melhores caminhos para se atingir os objetivos da melhor forma possível, sempre com bom humor. Agradeço também às indicações visando aumentar minha coleção de CD's.

À todo o pessoal do LADI-SEL que me ajudou a dar continuidade ao projeto de mestrado e que hoje podem ver o trabalho finalizado. Obrigado meu povo !!

À CAPES e ao Departamento de Engenharia Elérica (SEL-EESC- USP) pelo apoio financeiro durante a elaboração do trabalho.

SUMÁRIO

LISTA DE FIGURAS iii
LISTA DE ABREVIATURAS E SIGLAS iv
RESUMO v
ABSTRACT vi
Capítulo 1 1
1. Introdução 1
1.1 Exame mamográfico 1
1.2 Sistemas de apoio ao diagnóstico 3
1.3 Objetivos 4
1.4 Disposição do trabalho 5
Capítulo 2 7
2. Digitalização de imagens 7
2.1 Digitalização e processamento de imagens mamográficas 7
2.2 Aquisição das imagens 7
Capítulo 3 13
3. Banco de Dados de Imagens Médicas 13
3.1 Conceitos básicos 13
3.2 Modelos de dados 14
3.3 Processamento de imagens 20
3.4 Banco de imagens existentes 22
3.4.1 MIAS (Mammographic Image Analysis Society) 22
3.4.2 DDSM: "Digital Database for Screening Mammography" 23
3.4.3 UCSF/LLNL "Digital Mammogram Library" 25
3.4.4 The CALMa project. 26
3.4.5 "Nijmegem Digital Mammogram Database" 27
Capítulo 4 28
4. Banco de imagens em sistemas de auxílio ao diagnóstico (CAD's). 28
4.1 Conjuntos de imagens 28
4.2 Influência nos testes de processamento 29

4.3 Critérios de qualidade da imagem 33
4.4 Interface 35
Capítulo 5 37
5. Metodologia 37
5.1 Aquisição das imagens 37
5.2 Modelagem da base de imagens 40
5.3 Buscas e consultas ao banco de imagens 42
5.4 Desenvolvimento da interface 44
5.5 Armazenamento das imagens 46
Capítulo 6 48
6. Resultados e conclusões 48
6.1 Composição do banco de imagens 48
6.2 Testes de processamento 52
Capítulo 7 57
7. Discussão e conclusões 57
8. Referências 61

LISTA DE FIGURAS

FIGURA 1.1: ESQUEMA DE FORMAÇÃO DA IMAGEM MAMOGRÁFICA 2
FIGURA 1.2: EXEMPLO DAS DUAS VISÕES MAMOGRÁFICAS: (A) CRÂNIO-CAUDAL E (B) MÉDIO-LATERAL OBLÍQUA 3
FIGURA 2.1: DIFERENÇA ENTRE MÉTODOS DE DIGITALIZAÇÃO: (A) REFLEXÃO E (B) TRANSMISSÃO. 9
FIGURA 2.2: IMAGENS DIGITALIZADAS COM PIXELS DE 0,5 MM E 0,1 MM, RESPECTIVAMENTE. 10
FIGURA 3.1: DIAGRAMA ENTIDADE-RELACIONAMENTO 16
FIGURA 3.2: EXEMPLO DA INTERFACE DE BUSCA DO DDSM 25
FIGURA 4.1: ESTRUTURAS DE INTERESSE DA MAMA: (A) MICROCALCIFICAÇÕES, (B) CALCIFICAÇÕES E (C) NÓDULO 29
FIGURA 5.1: DIAGRAMA ENTIDADE-RELACIONAMENTO SIMPLIFICADO RELATIVO AO BANCO DE IMAGENS 42
FIGURA 5.2: EXEMPLO DE CONSULTAS AO BANCO DE IMAGENS (A) POR CARACTERÍSTICAS E (B) POR REGIÕES 45
FIGURA 6.1 : DISTRIBUIÇÃO DAS IMAGENS DO BANCO, DE ACORDO COM AS ESTRUTURAS DE INTERESSE 49
FIGURA 6.2: DISTRIBUIÇÃO DAS IMAGENS DE ACORDO COM O INTERVALO DE IDADES. 50
FIGURA 6.3: EXEMPLO DA TELA PRINCIPAL DO PROGRAMA 52
FIGURA 6.4: CURVA ROC PARA O PROCESSAMENTO DE REALCE DE CONTRASTE ([NUNES 2002]) 53
FIGURA 6.5: EXEMPLO DA INTERFACE DO "SOFTWARE"DE SEGMENTAÇÃO DE NÓDULOS ([SANTOS2002]). 54
FIGURA 6.6: CURVA ROC PARA A SEGMENTAÇÃO DE NÓDULOS ([SANTOS 2002]) 55
FIGURA 6.7: CURVA ROC PARA O PROCESSAMENTO DE CLASSIFICAÇÃO DE NÓDULOS ([SANTAELLA 2002]). 56

LISTA DE ABREVIATURAS E SIGLAS

CAD: Computer-Aided Diagnosis

CALMa: Computer Assisted Library for Mammography

DDSM: Digital Database for Screening Mammography

FTM: Função de Transferência de modulação

HC-FMRP/USP: Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto / Universidade de São Paulo

JPEG: Joint Photographic Experts Group

Mb: Megabyte

MIAS: Mammographic Image Analysis Society

PGM: Portable Gray Map

ROC: Receiver Operating Characteristic

SQL: Structured Query Language

TAR: Tape Archived

TIFF: Tagged Image File Format

UCSF/LLNL: University of California at San Francisco / Lawrence Livermore National Laboratory

Unifesp: Universidade Federal de São Paulo

RESUMO

Este trabalho teve como objetivo o desenvolvimento de um banco de imagens mamográficas digitalizadas para oferecer suporte a treinamento e testes de processamentos em esquemas de auxílio ao diagnóstico (CADs).

Juntamente com o banco de dados associado, desenvolveu-se ainda uma interface gráfica que permitiu a recuperação das imagens e que auxiliou no acesso às imagens de interesse.

Foram incluídas imagens de vários hospitais, com características diferentes quanto aos mamógrafos utilizados. Além disso, o banco de dados também incorpou dados clínicos das pacientes analisadas e laudos radiológicos e patológicos emitidos por médicos.

O banco de imagens mostrou-se abrangente em relação às questões dos parâmetros de digitalização utilizados, às características das pacientes e contemplou os diversos tipos de patologias mamárias existentes.

Realizaram-se testes de processamento de imagens, mostrando a viabilidade e a facilidade do uso do banco de imagens em situações práticas do desenvolvimento de novas técnicas e algoritmos computacionais relacionados à área de mamografia.

ABSTRACT

This work had the main purpose of develop a digital mammographic image database to provide support to training and testing procedures of image processing techniques commonly used in Computer-Aided Diagnosis schemes.

Also, a graphic interface was developed in order to ease the retrieval and access to images of interest accordingly to the processing technique in test.

Images from several hospitals were included in the database, with different mammographic equipments features. Besides, the database was also provided with patient's clinical information and, when available, radiological reports issued by expert physicians.

The image database was developed so a wide range of pathologies could be expressed and different digitization and acquiring parameters could be used in the images included.

Image processing tests using images from the database was done, and results showed that the database was useful and easy to use in practical situations that demand tests of new image processing techniques development and computerized algorithms related to mammographic research area.

Capítulo 1

1. Introdução

1.1 Exame mamográfico

A mamografia é reconhecida como sendo o exame mais indicado no diagnóstico de doenças mamárias. Esse fato deve-se principalmente a sua boa resolução, facilidade de obtenção (inclusive quanto à questão econômica) e a sua boa resposta de contraste.

O exame mamográfico é obtido através da imagem gerada pela absorção de fótons de raios-X pelo filme mamográfico. De acordo com a quantidade de fótons que impressionam o filme, este se torna mais ou menos escuro após a revelação. Isso permite diferenciar estruturas mamárias com diferentes densidades uma vez que os raios-X atravessam tecidos menos densos com maior facilidade causando maior escurecimento do filme.

A Figura 1.1 mostra esquematicamente como a imagem mamográfica é obtida:

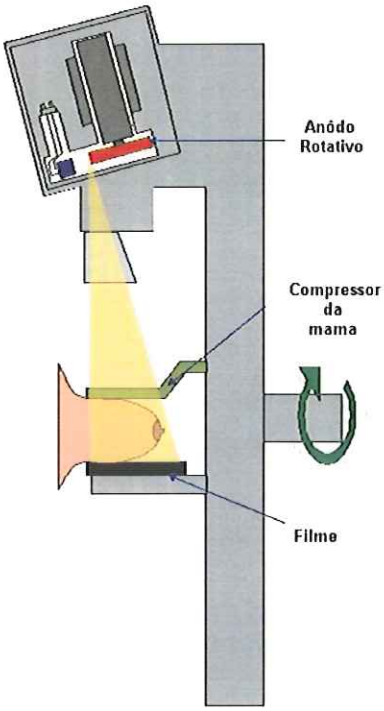
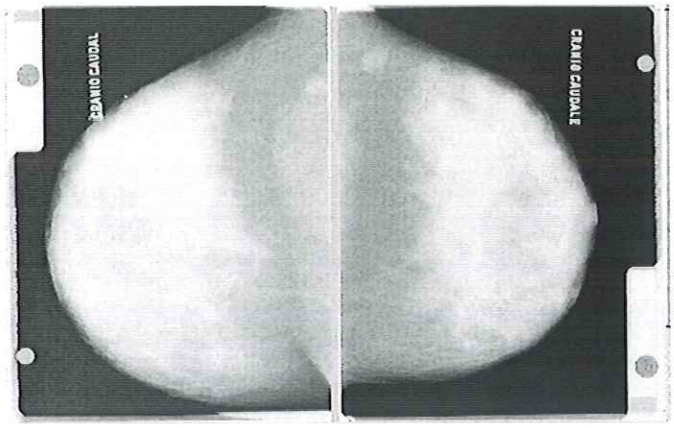


Figura 1.1: Esquema de formação da imagem mamográfica

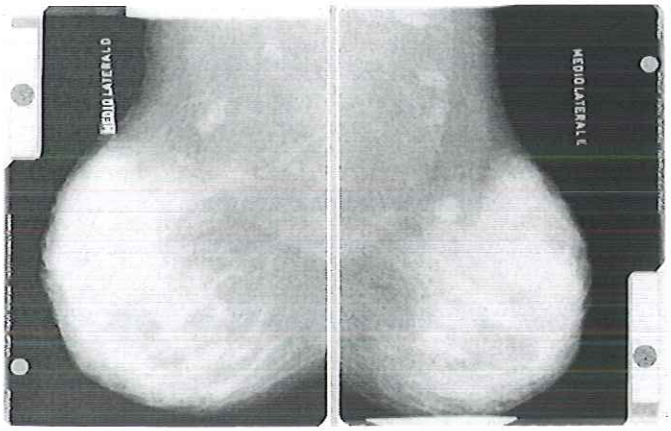
Esta característica que a mamografia possui de ser impressa em filme limita a visualização espacial da mama como um todo, impossibilitando a precisão necessária para a localização de estruturas patológicas no seu interior com a utilização de apenas uma imagem. Frente a este problema, o exame mamográfico é geralmente constituído de quatro imagens, sendo duas visões para cada mama, chamadas de Crânio-Caudal e Médio-Lateral Oblíqua.

Com este artifício torna-se possível a visualização total da mama, incluindo-se a área linfática axilar, que muitas vezes pode ajudar no diagnóstico de doenças mamárias. Além disso, nos casos em que é necessária a extração de tecido para biópsia ou remoção de partes da mama, a localização espacial exata da estrutura tem que ser conhecida, o que é facilitado com a combinação das duas visões disponíveis.

A Figura 1.2 mostra um exemplo das quatro imagens que geralmente são obtidas em um exame mamográfico:



(a)



(b)

Figura 1.2: Exemplo das duas visões mamográficas: (a) Crânio-Caudal e (b) Médio-Lateral Oblíqua

1.2 Sistemas de apoio ao diagnóstico

Esquemas CAD (de "Computer-Aided Diagnosis") são esquemas computacionais que têm como principal objetivo auxiliar o diagnóstico médico servindo como uma segunda opinião à pessoa que irá fornecer o laudo final. O público alvo de sistemas CAD em mamografia é composto por radiologistas ou ginecologistas.

Frente a isso, o sistema proposto deve atender às diversas demandas, tanto da área médica quanto da área computacional. Portanto, devem ser

incluídas no CAD rotinas que possibilitem ao médico ter liberdade de escolha sobre o procedimento a ser tomado. Por outro lado, deve possuir a capacidade de fornecer respostas confiáveis e úteis ao diagnóstico.

Para que o esquema CAD apresente a confiabilidade desejada, novos algoritmos e procedimentos de processamento de imagens devem ser testados de modo a propiciar detecções e classificações úteis como segunda opinião ao diagnóstico. É nesta etapa que o uso de banco de imagens digitalizadas desempenha papel fundamental, uma vez que os testes demandam grande quantidade de imagens para que possam ser implementados. Além disso, a diversidade e abrangência do banco também são características essenciais às imagens que são utilizadas nos testes de novas rotinas.

Outro fator que evidencia a importância na utilização de banco de imagens diz respeito à facilidade com que elas são recuperadas e apresentadas ao usuário, de forma a agilizar o processo de busca das imagens de interesse, de acordo com o processamento a ser empregado.

Possíveis diferenças nos conjuntos de imagens utilizados nos testes de processamento também são fatores que podem alterar a avaliação do desempenho da técnica em teste, uma vez que as imagens podem a presentar características discrepantes quanto ao contraste utilizado, qualidade do mamógrafo e relação de contraste entre o fundo da imagem e a estrutura patológica em análise.

Dessa forma, o desenvolvimento de esquemas CAD deve vir sempre acompanhado de uma base de imagens que possa ser utilizada de forma confiável ao processamento. Nos casos em que isso não ocorre, deve-se incluir informações quanto às imagens utilizadas durante os testes na avaliação final dos resultados do processamento.

1.3 Objetivos

Assim, este trabalho tem como objetivo a elaboração de um banco de imagens que inclua informações clínicas e de histórico das pacientes cadastradas, laudos radiológicos, características de diferentes aparelhos mamográficos e de

diferentes digitalizadores, uma vez que subsidiará a utilização de imagens mamográficas digitalizadas para posterior processamento e testes de algoritmos. O banco de imagens servirá como base de consultas para treinamento e desenvolvimento de sistemas de processamento a serem integrados a um esquema CAD em mamografia.

Além disso, o banco de imagens deve dar suporte ao desenvolvimento de "softwares" dedicados ao processamento de imagens e deve ser suficientemente abrangente para que as principais patologias e estruturas de interesse estejam representadas, de modo a validar os testes de processamento e a proporcionar buscas por imagens com determinadas características. Opções de seleção por característica de pacientes também devem estar integradas ao sistema.

Uma interface gráfica também deve ser desenvolvida com o objetivo de facilitar o acesso às imagens e agilizar a busca e recuperação de imagens ou regiões de interesse, de acordo com características da paciente ou da mama.

1.4 Disposição do trabalho

Além desta introdução, o presente trabalho divide-se em mais 6 capítulos, sendo que os capítulos 2, 3 e 4 apresentam revisões bibliográficas quanto ao processo de digitalização de imagens, os conceitos e a elaboração de banco de dados e sua aplicação em imagens médicas e uma breve apresentação dos bancos de imagens que foram desenvolvidos até agora e são apresentados na literatura.

No capítulo 5 será apresentada a metodologia do trabalho, onde serão abordados os aspectos considerados durante a elaboração do presente banco de imagens e detalhes sobre sua implementação.

O capítulo 6 refere-se aos resultados alcançados com a construção da base de dados, destacando seu papel no processamento de imagens mamográficas, assim como os desempenhos obtidos pelas técnicas de processamento testadas com a utilização dessa base de imagens.

Finalmente, no capítulo 7 será feita uma breve discussão sobre as dificuldades encontradas na obtenção das imagens e no processo de digitalização

e armazenamento das mesmas, assim como aspectos ligados à usabilidade prática do banco de imagens desenvolvido.

Capítulo 2

2. Digitalização de imagens

2.1 Digitalização e processamento de imagens mamográficas

O processamento de imagens médicas vem sendo largamente utilizado com o objetivo de auxiliar no diagnóstico, sempre tendo a função de fornecer uma segunda opinião ao médico que irá emitir o laudo final. Para isso, são desenvolvidos algoritmos de pré-processamento, detecção e classificação aplicados para cada tipo de imagem e de acordo com as estruturas de interesse de cada caso.

Como já mencionado, no presente trabalho será abordada a imagem mamográfica. As principais estruturas de interesse nessas imagens são microcalcificações, calcificações e nódulos pois podem indicar o início de patologias da mama e sua detecção é fundamental no diagnóstico precoce destas patologias. Devido ao tamanho e às características de absorção de radiação X dessas estruturas, elas podem ser de difícil visualização a olho nu, sendo muitas vezes necessário o uso de lentes de aumento para a boa identificação e diagnóstico.

2.2 Aquisição das imagens

Com a digitalização das imagens, todo esse procedimento pode ser feito computacionalmente com o auxílio de "softwares" específicos para esse tipo de aplicação.

Para que os esquemas computacionais tenham sucesso, a digitalização deve seguir alguns critérios mínimos quanto à qualidade da imagem e do

dispositivo digitalizador utilizado. Somente assim, a imagem digital pode representar com razoável fidelidade as informações presentes no filme original.

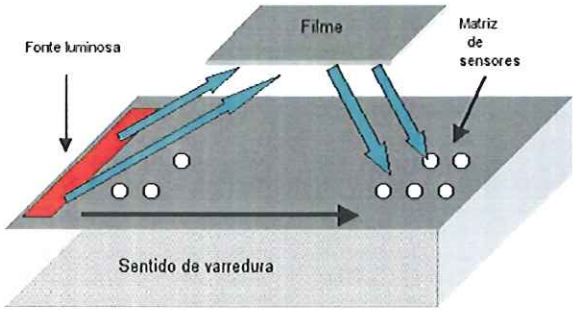
Os principais fatores que afetam a qualidade da imagem digitalizada e que são extremamente importantes para esquemas de processamento são os seguintes:

- Método de digitalização utilizado;
- Resolução espacial;
- Resolução de contraste;
- Nível de ruído presente na imagem.

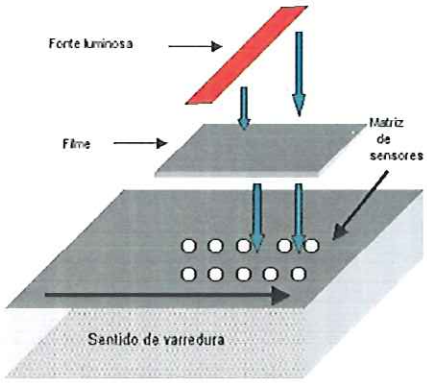
Cada aspecto/parâmetro citado será analisado com maior detalhamento para que fique claro o papel que desempenham na aquisição e no processamento de imagens médicas.

Quanto ao método de digitalização, existem duas maneiras de se fazer a aquisição da imagem digital considerando-se a fonte luminosa e os sensores utilizados: por reflexão e por transmissão. Na digitalização por reflexão utiliza-se a intensidade luminosa que é refletida pela imagem e medida pelos sensores para se fazer a conversão das densidades ópticas (graus de escurecimento) presentes no filme em níveis de cinza. Este método não é utilizado na geração de imagens médicas utilizadas para processamento, pois apresenta problemas de borramento e perda de precisão de bordas causados pelo espalhamento da luz inerente ao fenômeno de reflexão. Já no processo de digitalização por transmissão, a luz incidente atravessa o filme e é então captada por sensores. Este processo reduz muito a limitação quanto ao borramento das imagens e melhora muito a definição de eventuais bordas presentes na imagem. Este método é o mais utilizado na digitalização de imagens médicas, principalmente em imagens radiográficas, pois neste caso os pequenos detalhes presentes na imagem desempenham papel fundamental no diagnóstico final.

A Figura 2.1 mostra esquematicamente a diferença entre os métodos de digitalização por reflexão e por transmissão:



(a)



(b)

Figura 2.1: Diferença entre métodos de digitalização: (a) reflexão e (b) transmissão.

Atualmente, com o objetivo de eliminar completamente o borramento, os digitalizadores mais utilizados para fins médicos empregam tecnologia laser, que utiliza este tipo de feixe como fonte luminosa. Este efeito é obtido pelo digitalizador laser devido à coerência de seu feixe, que apresenta quase nenhum espalhamento, diferentemente da luz branca utilizada na maioria dos digitalizadores comerciais.

Outros dois fatores muito importantes que devem ser levados em conta quando analisamos imagens, principalmente as empregadas em medicina, são as resoluções, espacial e de contraste. A resolução espacial pode ser entendida como sendo a informação disponível na imagem digitalizada por unidade de área da imagem original. Sendo assim, em imagens médicas é desejado que se utilize a maior resolução espacial possível. Esse fato implica a digitalização de filmes

com tamanhos de pixel (unidade de informação na imagem) cada vez menores, justificando, assim, o uso de "scanners" laser para esse fim.

A Figura 2.2 mostra a mesma região de um filme mamográfico digitalizada com tamanhos de pixels (resolução espacial) diferentes:

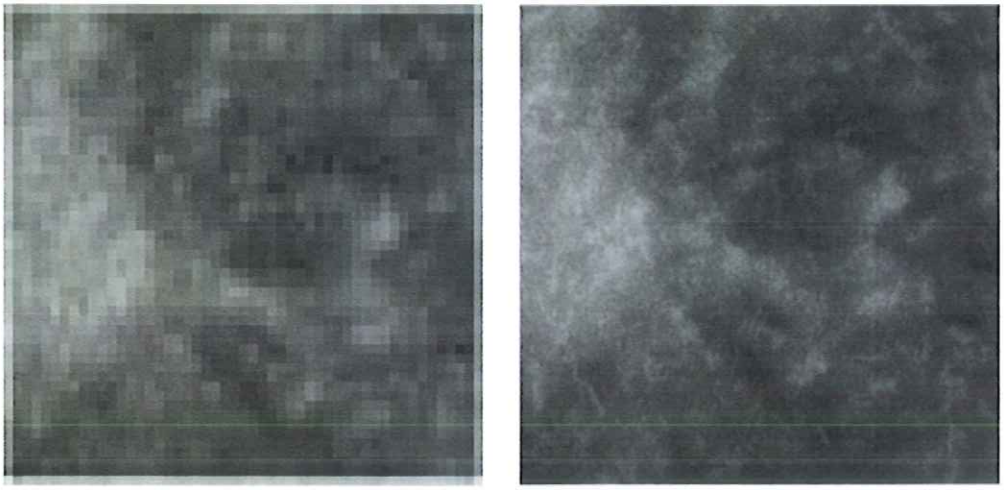


Figura 2.2: Imagens digitalizadas com pixels de 0,5 mm e 0,1 mm, respectivamente.

O estabelecimento de um limite no tamanho mínimo do pixel também apresenta vantagens consideráveis para o armazenamento das imagens, uma vez que os arquivos gerados por imagens digitais geralmente são de grande tamanho. Para casos de aplicações em banco de imagens, onde grande quantidade de imagens deve ser acessada ou manipulada, esta limitação desempenha papel fundamental no seu desempenho. Outro ponto que deve ser destacado diz respeito à velocidade de processamento em imagens com grande resolução espacial, já que o aumento da resolução espacial implica aumento no número de pixels a serem lidos e interpretados pelo sistema computacional desenvolvido.

A segunda resolução que deve ser levada em conta é a chamada resolução de contraste, que pode ser definida como a quantidade de níveis de quantização que podem ser expressos na imagem. Para imagens de raios-X, estes níveis de quantização representam a quantidade de bits disponível na imagem digital para se expressar a escala dos níveis de cinza entre pixels puramente brancos e pixels totalmente pretos. Dessa forma pode-se afirmar que quanto maior for a resolução

de contraste, maior será a diferenciação entre os níveis de cinza dos pixels da imagem, alterando o desempenho de esquemas de processamento.

Esta diferença no desempenho do processamento pode ser melhor observada e avaliada nos casos em que são utilizadas imagens de baixo contraste entre o que é considerado fundo da imagem (background) e as estruturas de interesse.

Em imagens mamográficas, o baixo contraste se apresenta principalmente quando são analisadas imagens de mamas densas, uma vez que nestes casos o tecido mamário é composto em sua maioria por tecido fibroglandular, que absorve mais raios-X, causando um menor escurecimento no filme. Este tipo de comportamento é semelhante ao de estruturas como microcalcificações ou nódulos. Sendo assim, pode haver um possível mascaramento de eventuais estruturas de interesse presentes na mama, dificultando a precisão do diagnóstico.

Embora o olho humano não consiga diferenciar mais do que 32 níveis de cinza em uma imagem, para a análise digital pixel a pixel o incremento na resolução de contraste pode ser determinante na segmentação e detecção de estruturas que apresentem níveis de cinza próximos.

Atualmente, a maioria das imagens médicas utilizadas em processamento é digitalizada com resolução de contraste de 10 ou 12 bits, embora os formatos digitais nos quais as imagens são armazenadas apresentem suporte para arquivos com até 16 bits/pixel de níveis de cinza.

No entanto, estudos anteriores (NUNES 2001) mostraram que a partir de um certo limite, o aumento na resolução espacial da imagem acrescenta pouco ao desempenho do processamento, apresentando ainda a desvantagem de acrescentar ruídos indesejáveis à imagem. No caso de algoritmos de processamento relacionados à detecção de microcalcificações mamárias ou de quaisquer outras estruturas de pequeno tamanho, a presença de ruído eleva consideravelmente a taxa de detecções falsas-positivas, deteriorando o desempenho do esquema de processamento.

O ruído pode ser definido como qualquer sinal/evento que prejudique a fidelidade e a qualidade da transmissão de dados/informações entre a entrada e a

saída de determinado sistema. No caso particular de imagens digitalizadas, este fenômeno se faz presente na forma de pequenos pontos presentes na imagem, resultante de falhas no processo de aquisição da imagem ou dos parâmetros utilizados. Eventuais borramentos ou perda de precisão de bordas na imagem final também podem ser interpretados como ruídos pois prejudicam a correta avaliação da imagem obtida.

Capítulo 3

3. Banco de Dados de Imagens Médicas

3.1 Conceitos básicos

Bancos de dados são ferramentas muito importantes em uma série de processos. Estes processos nem sempre estão ligados à área computacional e em muitos casos acabam sendo encobertos pelo objetivo principal da aplicação.

Um exemplo clássico é o de bancos de dados utilizados em instituições financeiras, onde o foco principal do processo diz respeito a aplicações, resgates e movimentação de fundos e investimentos. Este caso apresenta um enorme banco de dados integrado ao sistema principal, que deve gerenciar fonte e destino das aplicações, montante envolvido e o aspecto mais importante, que diz respeito ao tempo de atualização das operações e da disponibilização dos novos dados para que sejam utilizados em outras operações.

O conceito de banco de dados para uso computacional surgiu na década de 50 com o advento dos primeiros computadores, mas obviamente encontrava-se muito limitado pois o espaço disponível para armazenamento de dados e a velocidade de processamento ainda eram insuficientes para se obter o desempenho desejado.

Anteriormente, bancos de dados eram basicamente uma coleção de dados ordenados que eram armazenados na forma de arquivos (físicos) ou catálogos que podiam ser consultados por qualquer pessoa. Pode-se notar que esse método não era prático e, além de tudo, apresentava o grave risco de ser desorganizado por pessoas não-familiarizadas com os arquivos.

O emprego de banco de dados em aplicações computacionais apresenta grandes vantagens frente a outros métodos de armazenamento, entre eles

independência e consistência dos dados e recursos de preservação das informações ("backup"). (LYMAN 2000)

A independência dos dados é obtida através do recurso da criação de visões que os bancos de dados oferecem. Estas visões permitem a criação de aplicações onde cada usuário tenha acesso somente às informações que podem ser a ele disponibilizadas.

Dessa forma, assegura-se que informações confidenciais ou que exijam maior grau de segurança não sejam visualizadas por qualquer usuário. Outro aspecto importante que deve ser enfatizado na criação destas visões diz respeito a modificações realizadas no conjunto de informações do banco de dados. Assim, as visões ajudam a evitar possíveis manipulações indesejadas.

Outra característica importante da independência dos dados diz respeito ao local onde estão armazenadas as informações. Para aplicações em banco de dados, somente os dados e sua forma de acesso são importantes, não se fazendo distinção quanto ao local exato do armazenamento. Isso possibilita o desenvolvimento de aplicações multi-usuário com acesso remoto aos dados, além de fornecer suporte para "backup" das informações em lugares diferentes daquele onde o sistema principal está sendo utilizado.

Quanto à consistência dos dados, pode-se citar a validade, a completeza e a consistência das informações. Estas características podem ser obtidas de diversas formas tais como: obtenção de dados válidos para cada domínio disponível (caracteres, inteiro, real, data, etc) e a disponibilização de todos os dados necessários mesmo em casos onde as informações encontrem-se armazenadas em locais diferentes.

3.2 Modelos de dados

A limitação de desempenho dos computadores da década de 50 teve papel importante no desenvolvimento de banco de dados, uma vez que foi necessário o desenvolvimento de algoritmos e estratégias de acesso e armazenamento otimizadas, pois soluções para a parte de "hardware" estavam limitadas pela tecnologia da época.

Já na década de 70, surgiu o conceito de modelo de dados. Esses modelos são estruturas de armazenamento e acesso aos dados do mundo real que podem ser interpretadas por qualquer pessoa e que são utilizadas no desenvolvimento e na manutenção de sistemas. Entre os modelos existentes, os primeiros que foram desenvolvidos e que são utilizados até hoje são o Modelo Relacional apresentado por Codd em 1970 (CODD 1970) e o Modelo Entidade-Relacionamento (CHEN 1976). O sucesso obtido por esses modelos deve-se principalmente à facilidade de aprendizagem e por desempenharem papel fundamental no desenvolvimento e principalmente na manutenção de sistemas previamente desenvolvidos.

O Modelo Entidade-Relacionamento pertence à classe dos modelos conceituais. Esta classe tem como principal objetivo a estruturação do sistema, sem ênfase na posterior implementação. Sendo assim, o Modelo Entidade-Relacionamento é mais intuitivo se comparado ao Modelo Relacional. Além disso, este modelo é essencialmente gráfico, facilitando a visualização geral e o entendimento do sistema, inclusive nos casos em que são necessárias atualizações e manutenções futuras. (STOREY 2001)

O Modelo Entidade-Relacionamento apresenta 4 construtores que são utilizados na representação do sistema:

- Conjunto de Entidades;
- Conjunto de Relacionamentos;
- Atributos de Entidades;
- Atributos de Relacionamento.

O conjunto de Entidades é definido como sendo o conjunto de dados obtidos no mundo real que pertencem a uma mesma classe: pessoas, alunos, usuários, imagens, etc. O modelo não trata destas entidades isoladamente mas sim do conjunto aos quais elas pertencem. Estes conjuntos são simbolizados no diagrama Entidade-Relacionamento por retângulos com seus respectivos nomes. As entidades possuem atributos que são associados separadamente para cada conjunto, caracterizando unicamente cada um deles.

Já os conjuntos de Relacionamentos são os construtores que associam os diferentes conjuntos de entidades. São simbolizados por um losango e estão sempre ligados aos conjuntos de entidades. Importante salientar que embora os relacionamentos tenham a principal função de conectar entidades, eles também podem possuir atributos próprios. Este caso ocorre em situações nas quais o atributo não pode ser incluído em nenhuma das entidades envolvidas, sendo inerente ao próprio processo indicado pelo relacionamento.

Todos os atributos, independentemente de pertencerem às entidades ou aos relacionamentos, devem ser indicados no diagrama por uma elipse juntamente com o nome do atributo. Deve ainda estar ligado à entidade/relacionamento ao qual pertence.

Existe ainda no Modelo Entidade-Relacionamento uma sintaxe mínima para ser usada na representação que deve indicar, além das entidades, relacionamentos e atributos, as chaves existentes nas entidades e relacionamentos, os papéis dos relacionamentos e a cardinalidade de cada um deles.

Um pequeno exemplo de um diagrama Entidade-Relacionamento é mostrado na Figura 3.1:

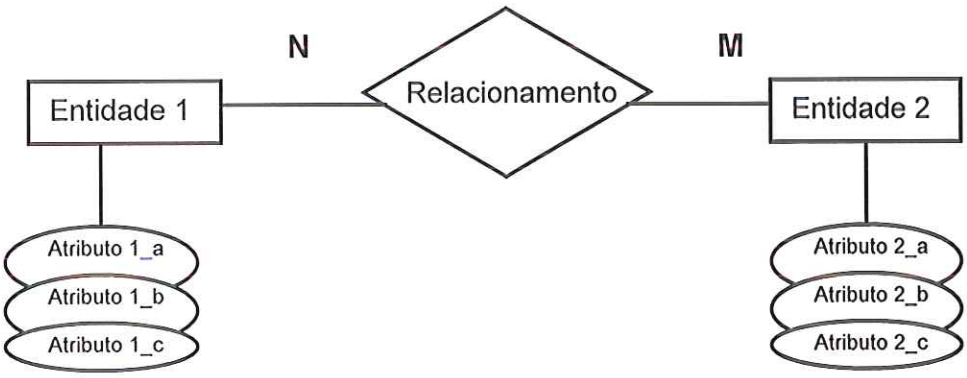


Figura 3.1: Diagrama Entidade-Relacionamento

Já o Modelo Relacional pertence à classe dos modelos chamados de "físicos", pois é voltado para a implementação do banco de dados propriamente dito e é baseado em relações. Neste modelo, todos os dados devem ser

agrupados em tabelas inter-relacionadas de forma lógica de acordo com os parâmetros de projeto especificados para a aplicação a ser desenvolvida. Deve-se levar sempre em consideração o futuro acesso que será feito a estas tabelas, de modo a otimizar o procedimento de busca e recuperação dos dados de interesse.

De acordo com o modelo proposto, as tabelas criadas para a implementação do banco de dados devem obedecer a algumas regras para que se possa obter o correto funcionamento do sistema e o desempenho desejado. Estas regras são basicamente:

- Unicidade da chave da relação;
- Integridade de entidade;
- Integridade referencial.

Para melhor entendimento das regras citadas anteriormente é necessário o conhecimento de alguns conceitos empregados no Modelo Relacional que serão abordados a seguir.

O primeiro conceito diz respeito à própria constituição das tabelas. As relações são compostas por atributos, definidos como os campos presentes na tabela, e valores, que são os dados obtidos no mundo real e armazenados de acordo com o atributo adequado.

A partir daí pode-se definir os conceitos de chave, chave-primária e chave-estrangeira:

- Chave: Conjunto de atributos cujos valores identificam, de forma única, cada registro existente na tabela.
- Chave-primária: Menor conjunto de atributos pertencente à chave que é utilizado como identificador de um registro.
- Chave-estrangeira: Conjunto de atributos de uma determinada tabela, que não são chave nesta tabela, mas que são parte da chave de outra tabela e que são compatíveis em domínio entre si. Tem a finalidade de permitir o relacionamento entre tabelas.

Neste ponto, pode-se definir melhor as regras necessárias para a utilização do Modelo Relacional.

A regra de unicidade da chave impõe ao sistema que os valores dos atributos-chave não se repitam na mesma tabela. É importante destacar o fato de que a chave pode possuir mais de um atributo, sendo que alguns valores destes atributos podem se repetir. A restrição só se aplica para os casos em que todos os valores dos atributos componentes da chave coincidem.

A segunda regra diz respeito à integridade da entidade, que é uma restrição ao valor dos atributos-chave e impõe que nenhum atributo-chave pode apresentar valor nulo. Esta condição pode ser explicada frente ao fato de que os atributos-chave são utilizados como "referência" de determinada tabela.

A restrição de integridade referencial impõe ao sistema que os valores dos atributos chave-estrangeira de determinada tabela sejam nulos ou iguais aos valores destes mesmos atributos na tabela em que não fazem parte da chave.

Para a implementação do banco de dados e as respectivas tabelas do modelo relacional, faz-se necessário um mapeamento completo do sistema exibido no Diagrama Entidade-Relacionamento.

O mapeamento deve obedecer a uma série de regras para que a implementação final retrate fielmente o sistema. Entre essas regras, podemos destacar duas principais, que dizem respeito ao mapeamento dos conjuntos de entidades e relacionamentos.

Conjuntos de Entidades devem ser sempre mapeados como uma nova relação, tendo-se como atributos-chave os atributos-chave mostrados no diagrama Entidade-Relacionamento.

Os conjuntos de relacionamentos têm mapeamentos diversos, de acordo com a sua cardinalidade, conforme descrito a seguir:

- Relacionamentos com cardinalidade 1:1 - Não são mapeados em relações próprias, sendo que seus atributos são adicionados a uma das relações geradas pelas entidades ligadas ao relacionamento. Além

disso, devem ser também incluídos nessa relação, os atributos-chave da relação que não foi alterada.

- Relacionamentos com cardinalidade 1:N - Seguem a mesma regra anterior, pois não são mapeados em relação própria. No entanto, os atributos do relacionamento devem, necessariamente, ser incluídos na relação gerada pela entidade de cardinalidade N, também respeitando a inclusão dos atributos-chave da relação de cardinalidade 1.
- Finalmente, relacionamentos com cardinalidade N:M devem ser mapeados em uma nova relação contendo os atributos do próprio relacionamento acrescidos dos atributos-chave de ambas as relações ligadas pelo relacionamento.

A organização dos dados em tabelas de acordo com as regras do modelo relacional apresentadas anteriormente facilita bastante o acesso, a busca e a disponibilização das informações. Para esta finalidade, foi criada uma linguagem voltada ao suporte dos bancos de dados chamada SQL (do inglês "Structured Query Language").

A linguagem SQL foi criada pela IBM, pois esta foi a primeira companhia a usar banco de dados de forma avançada e a disponibilizar o seu sistema para outras aplicações. Posteriormente foi modificada devido à necessidade de normalização frente aos bancos de dados em desenvolvimento, sendo que a versão mais utilizada é a chamada SQL2 (ou SQL92). Esta versão foi adaptada em 1992 para suportar inovações presentes em aplicações recentes e foi aprovada pela ANSI (American National Standards Institute), tornando-se o padrão utilizado em todas os sistemas que fazem uso da linguagem SQL. (DATE 1996)

O padrão SQL apresenta a vantagem de integrar todos os comandos necessários à criação e manipulação de banco de dados relacionais. Além disso, engloba todas as funções das "sub-linguagens" mostradas a seguir:

- DDL: Linguagem de Definição de Dados;
- DML: Linguagem de Manipulação de Dados;

- DCL: Linguagem de Controle de Dados.

3.3 Processamento de imagens

O processamento de imagens, principalmente quando se considera o uso de imagens médicas e, como conseqüência, a elaboração de diagnósticos através de imagens, envolve a realização de testes extensivos para a validação dos algoritmos utilizados.

Na realização desses testes, grande quantidade de imagens faz-se necessária, uma vez que as diversas possibilidades de processamento como parâmetros, classes de imagens, informações presentes em cada tipo de imagem e características de sua obtenção (equipamento, digitalizador, etc) devem ser abordadas. Questões ligadas ao desempenho dos algoritmos de processamento e do sistema como um todo, também podem estar relacionadas à quantidade e à qualidade do conjunto de imagens utilizados durante os testes (SCHIABEL 2001)

Frente a isso, um conjunto de imagens que seja consistente, confiável e suficientemente amplo desempenha papel importante dentro dos esquemas de processamento. Além disso, o banco de dados deve ser o mais uniforme possível (incluindo-se as imagens) para que não exerça influência no processamento e, mais ainda, para que possa ser utilizado com diferentes sistemas operacionais, diferentes plataformas de trabalho e por diferentes grupos de pesquisa.

A busca por um banco de dados único, que possa ser utilizado como padrão de imagens mamográficas, tem sido realizada com maior ênfase desde a década de 90 (KARSSEMEIJER 1993), (NISHIKAWA 1993) sem obter resultados concretos satisfatórios até o momento. Isto se deve principalmente às limitações de formato das imagens utilizadas e ao próprio espaço computacional que as imagens digitalizadas demandam.

Como será citado adiante, o único banco de imagens que se aproxima de um "padrão" é o desenvolvido pelo MIAS (Mammographic Image Analysis Society), sediado no Reino Unido. Mesmo tendo sido utilizado por diversos grupos de pesquisa ao redor do mundo, este banco de imagens apresenta a desvantagem de possuir imagens digitalizadas com 8 bits de resolução de contraste, quando o consenso no meio científico atual é de que imagens de certas

classes, como as mamográficas, devem possuir no mínimo 10 bits para representar os níveis de cinza presentes na imagem de modo satisfatório (DANCE 1993).

A quantidade de imagens necessárias para a composição de um banco que seja representativo ainda é uma questão aberta, mas a literatura aponta na direção de que o total de imagens disponíveis deve representar, estatisticamente, a realidade clínica de ocorrência quando se consideram os diversos tipos de lesões ou anomalias que podem ocorrer e que são detectados através da mamografia.

Sendo assim, pode-se esperar que um grande número de casos incluídos no banco de imagens esteja associado a exames que não apresentam qualquer estrutura de interesse para o processamento, uma vez que muitas pacientes realizam somente exames de rotina, sem apresentar nenhum tipo de suspeita anterior. Além disso, a inclusão de casos não-patológicos no banco de imagens possibilita a realização de testes de detecção de falso-positivos.

Outro fato que deve ser levado em conta na elaboração de bancos de imagens diz respeito aos percentuais de imagens que apresentam estruturas diagnosticadas como sendo benignas e malignas. Clinicamente, a ocorrência de casos benignos supera muito a de casos considerados malignos, e esta proporção, quando possível, também deve ser respeitada nos bancos de imagens (KIMME-SMITH 1993).

Além disso, dados referentes à paciente e seus respectivos exames não podem deixar de ser incluídos no banco de dados , pois essas informações devem ser a principal forma de busca por imagens e estruturas de interesse. Quando possível, deve-se acrescentar também imagens de uma mesma paciente que foram obtidas em datas diferentes, com o objetivo de se fazer comparações quanto ao desenvolvimento de patologias ou de se fazer acompanhamento clínico.

3.4 Banco de imagens existentes

Atualmente, alguns bancos de dados estão sendo desenvolvidos por diferentes grupos de pesquisas ao redor do mundo com o principal objetivo de fornecer suporte a aplicações em processamento de imagens e na validação de algoritmos de processamento.

No caso das imagens médicas, que é o objeto de estudo deste trabalho, os bancos de imagens já desenvolvidos ou em processo de elaboração apresentam grandes diferenças entre si. Esta diferenciação deve-se principalmente à grande quantidade de diagnósticos médicos que levam em consideração imagens geradas pelos diversos exames existentes. Dessa forma, os bancos são, geralmente, compartimentados de acordo com o exame e a natureza da imagem de interesse a ser posteriormente processada e/ou visualizada.

No caso das imagens mamográficas, que são geradas por aparelhos de raios-X, existem atualmente 5 grandes bancos de dados já completos ou em desenvolvimento, sendo que 3 deles estão disponíveis para uso de outros grupos de pesquisa e 2 ainda continuam restritos para tal tipo de aplicação.

A seguir é apresentada uma breve discussão sobre as características de cada um desses 5 bancos:

3.4.1 MIAS (Mammographic Image Analysis Society)

O banco de imagens proposto pelo MIAS (SUCKLING 1994) é o mais utilizado hoje em dia por grupos de pesquisa espalhados pelo mundo que realizam os mais diversos testes de processamento em imagens mamográficas. Esse fato deve-se principalmente por ter sido o primeiro banco de imagens a ser efetivamente concluído (1994) e também por estar disponível através da Internet já há um longo tempo. De acordo com os próprios criadores, o banco de imagens do MIAS possui atualmente 57 cópias completas sendo utilizadas em 11 países.

As imagens incluídas neste banco foram obtidas através do Programa Britânico de Acompanhamento Médico (United Kingdom National Screening Programme) e conta com imagens de pacientes com idades entre 50 e 65 anos.

Todas as imagens disponíveis são da vista médio-lateral e foram digitalizadas em um "scanner" da marca JOYCE-LOEBL (SCANDIG3) que fornece imagens com tamanho de pixel de até 50 μm (resolução espacial) e resolução de contraste de 8 bits/pixel (256 níveis de cinza). Segundo o fabricante deste "scanner", sua resposta é linear para o intervalo de densidades óticas compreendido entre 0 e 3,2, que é a faixa de maior interesse na digitalização de imagens mamográficas.

O banco conta com 320 filmes digitalizados (aproximadamente 160 pacientes) e está armazenado no formato de compressão .TAR, utilizado pelo sistema operacional UNIX. Existe ainda uma segunda versão deste banco, chamada de "MIAS MiniDatabase", onde são disponibilizados 322 filmes digitalizados com 8 bits/pixel e armazenado no formado .PGM. Esta versão apresenta a grande desvantagem de ter sido elaborada com tamanho de pixel de 200 μm, o que limita sua aplicação prática em testes de processamento.

É disponibilizado ainda, juntamente com esta versão, informações sobre regiões de interesse para eventuais processamentos na forma de diagnóstico acompanhado de coordenadas do centro da região e do raio da mesma.

3.4.2 DDSM: "Digital Database for Screening Mammography"

Este banco de imagens foi desenvolvido pela University of South Florida (Estados Unidos) e foi encerrado em 1999 (HEATH 2000). Conta com 2620 filmes digitalizados, sendo 4 filmes referentes a cada exame (4 vistas mamográficas usuais), totalizando aproximadamente 650 pacientes.

As imagens foram obtidas junto a 4 hospitais (Massachussets General Hospital, Wake Forest Universtiy School of Medicine Hospital, Sacred Heart Hospital e Washington University of St. Louis School of Medicine Hospital) e digitalizadas em 4 diferentes "scanners", cada qual com características próprias. As características dos "scanners" são mostradas na Tabela 3.1:

Tabela 3.1: Características de digitalizadores do DDSM

CARACTERÍSTICA / SCANNER	RESOLUÇÃO ESPACIAL	RESOLUÇÃO DE CONTRASTE
DBA M2100 ImageClear	42 μm	16 bits/pixel
HOWTEK 960	43 μm	12 bits/pixel
LUMYSIS 200 LASER	50 μm	12 bits/pixel
HOWTEK MULTIRAD850	43 μm	12 bits/pixel

Juntamente com as imagens presentes neste banco, constam também informações consideradas relevantes para o acompanhamento da paciente e para eventuais processamentos futuros, tais como: idade da paciente, data da digitalização, data da realização do exame e informação referente à densidade da mama fornecida por radiologista, segundo critério adotado pelo ACR (American College of Radiology).

Com o objetivo de preservar a identidade da paciente, as imagens mostradas ao usuário final tiveram as partes em que constavam informações sobre ela removidas. Também para facilitar a visualização, o fundo da imagem foi removido de forma automática, evitando a presença de artefatos indesejados na imagem. O armazenamento das imagens foi feito utilizando-se algoritmo de compressão JPEG sem perdas.

O banco de imagens inclui também um "software" para detecção de massas mamográficas. Outro pacote é utilizado para a conversão de formatos de imagens em 8 e 16 bits, redimensionamento em outras resoluções espaciais (diferentes da aquisição original) e gravação em 2 formatos: PGM (Portable Gray Map) e TIFF (Tagged Image File Format).

Um exemplo do formulário utilizado na interface final e na busca de imagens é mostrado na Figura 3.2:

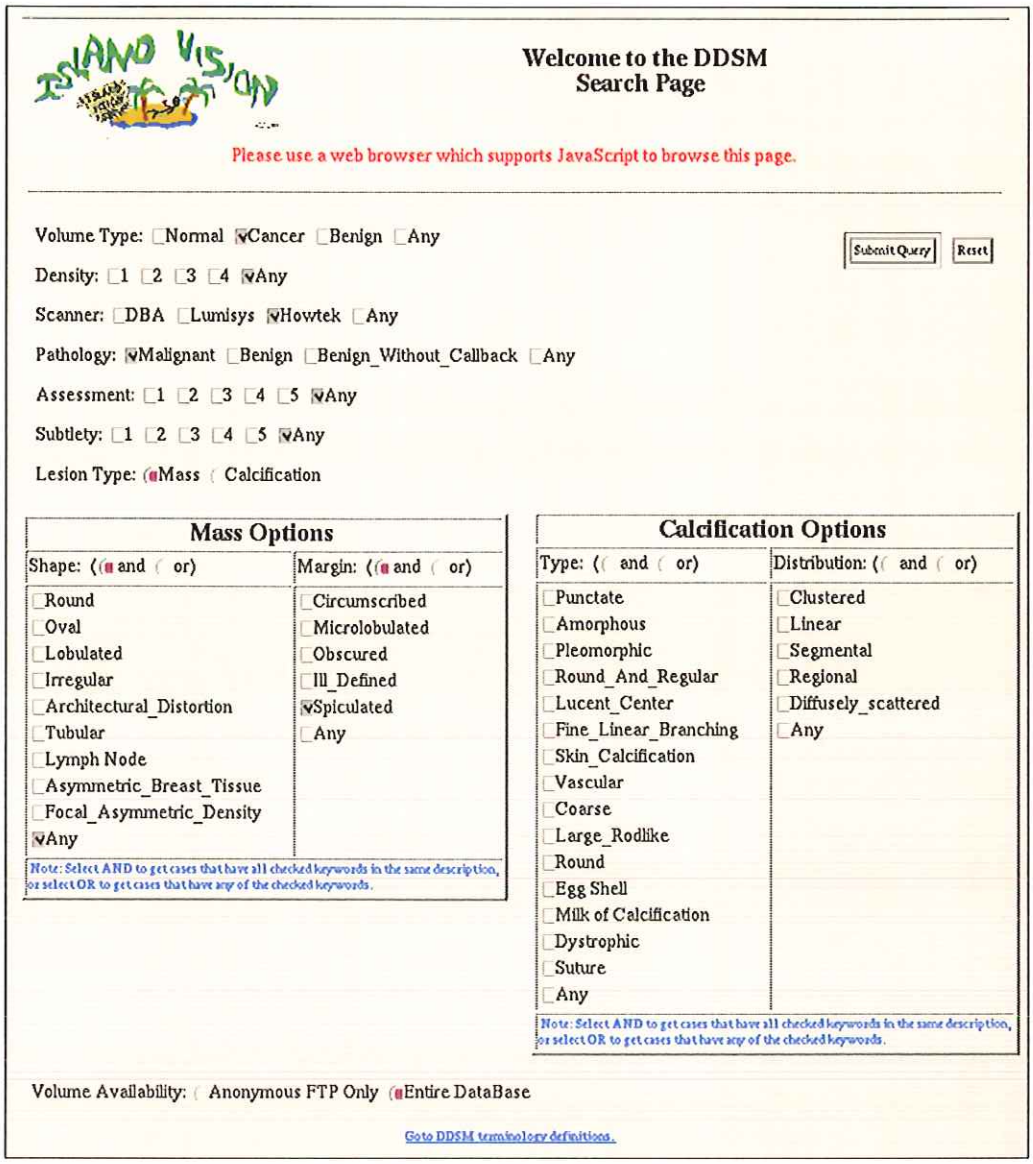


Figura 3.2: Exemplo da interface de busca do DDSM

3.4.3 UCSF/LLNL "Digital Mammogram Library"

O banco de dados criado pela parceria entre a University of California at San Francisco e o Lawrence Livermore National Laboratory (MASCIO 1996) vem sendo largamente utilizado, principalmente dentro dos Estados Unidos, devido a duas características muito importantes: está totalmente disponível em CD-ROM, ocupando pouco espaço em disco, e utiliza compressão sem perdas com redução

entre 10 e 30 vezes no tamanho do arquivo original, o que facilita a transmissão entre pontos remotos através de redes de comunicação.

Um aspecto desvantajoso desta base de imagens frente às duas citadas anteriormente é o baixo número de imagens (e conseqüentemente pacientes) disponível. Até o momento, estão cadastradas 50 pacientes do programa de acompanhamento da University of California (UCSF "Screening Program") que deram origem a 198 filmes digitalizados, sendo quatro filmes de cada uma das 49 pacientes correspondendo às quatro vistas mamográficas e mais duas imagens de uma paciente que foi submetida a mastectomia total.

As imagens foram digitalizadas utilizando-se um "scanner" DuPont (Wilmington, DE) que gera imagens com 12 bits de resolução de contraste e tamanho de pixel de 35 μm. Sendo assim, o banco necessita de aproximadamente 6 GB de espaço em disco, tornando o armazenamento na forma de CD-ROM uma necessidade real quando é considerada a transmissão das imagens para outros grupos de pesquisas.

3.4.4 The CALMa project

O projeto CALMa ("Computer Assisted Library for Mammography") é um projeto de colaboração entre o Instituto Nazionale di Física Nucleare (Itália) e alguns centros médicos italianos (Bari, Bologna, Livorno, Trento e Udine) com o objetivo de desenvolver um esquema CAD ("Computer-Aided Diagnosis") baseado em redes neurais (AMENDOLIA 2001).

Integrada a este projeto foi desenvolvida também uma base de imagens para ser utilizada nos testes de processamento e na validação/treinamento das redes neurais que deveriam ser utilizadas. As imagens obtidas junto aos hospitais encontram-se hoje armazenadas em 240 CD-ROM's, mas ainda não se apresentam disponíveis para uso externo.

Todas as 3000 imagens, referentes à aproximadamente 1400 pacientes, foram digitalizadas com resolução de contraste de 12 bits/pixel e 85 μm de resolução espacial. Do total de imagens, 1/3 são de casos considerados

patológicos, o que possibilita o treinamento e o teste das redes neurais com uma grande diversidade de opções.

3.4.5 "Nijmegem Digital Mammogram Database"

Esta base de imagens mamográficas é a menos utilizada atualmente devido ao seu reduzido número de casos/pacientes e ao formato utilizado no armazenamento. Ainda é utilizada em alguns estudos, mas somente com o propósito de comparações entre desempenhos de esquemas de processamento.

A base foi desenvolvida por um grupo de pesquisa sediado na Holanda e conta com apenas 40 imagens provenientes de 21 pacientes. As imagens foram digitalizadas com 12 bits/pixel e resolução espacial de 100 μm.

Capítulo 4

4. Banco de imagens em sistemas de auxílio ao diagnóstico (CAD's)

4.1 Conjuntos de imagens

Como já mencionado anteriormente, um banco de dados direcionado para esquemas CAD deve ser o mais amplo e completo possível. Este objetivo deve sempre ser buscado levando-se em consideração as estatísticas de ocorrências clínicas relacionadas a cada patologia ou estrutura de interesse, de modo a retratar de maneira eficiente o universo em que o banco de imagens está inserido.

A grande importância da existência de um conjunto de imagens que retrate realisticamente a ocorrência clínica das patologias mamárias para o desenvolvimento de esquemas computacionais de auxílio ao diagnóstico (CAD's) já foi discutida. No entanto, até este ponto, nada foi dito a respeito do papel que diferentes conjuntos de imagens podem ter na performance propriamente dita do CAD, no caso em que estas imagens são utilizadas para testes de algoritmos de processamento.

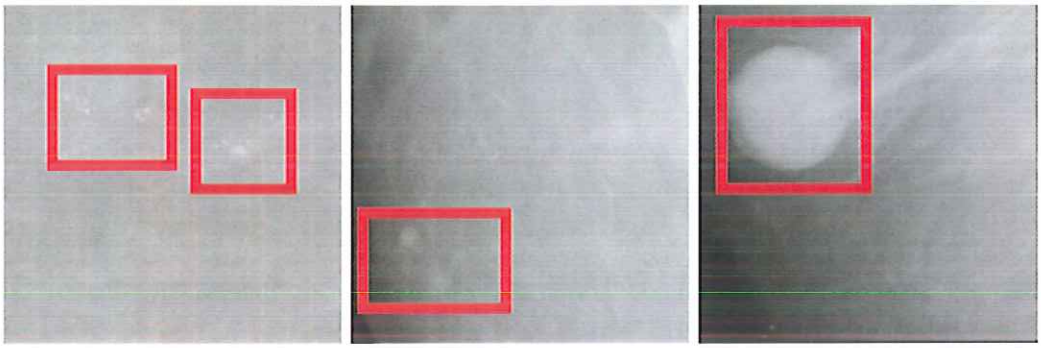
Neste trabalho o interesse está restrito a imagens mamográficas digitalizadas e suas estruturas de interesse para o processamento computacional, que são principalmente: microcalcificações, calcificações, nódulos e massas nodulares. Estas estruturas podem representar, de acordo com a classe médica, indícios do desenvolvimento de diversas patologias da mama, entre elas o câncer.

É neste estágio que o processamento de imagens digitalizadas torna-se importante, uma vez que o desenvolvimento de algoritmos computacionais específicos pode detectar estruturas de tamanho reduzido, como é o caso das microcalcificações e calcificações, muitas vezes não percebidas pelos

radiologistas. Com a detecção destas estruturas de pequeno tamanho, pode-se acompanhar o desenvolvimento das mesmas e, em muitos casos, pode-se fazer o diagnóstico precoce das patologias mamárias, aumentando a chance de recuperação total da paciente.

No caso dos nódulos e massas não identificadas, o processamento de imagens pode desempenhar papel importante na etapa de classificação da estrutura, distinguindo diferentes formatos de nódulos. Esta classificação pode fornecer subsídios para que o diagnóstico entre estruturas consideradas benignas e malignas seja feito com maior segurança.

Na Figura 4.1 é mostrado um exemplo das três principais estruturas citadas acima. Por questões de clareza e visualização, apenas parte da mama está mostrada, na forma de regiões de interesse (ROIs):



(a) (b) (c)

Figura 4.1: Estruturas de interesse da mama: (a) Microcalcificações, (b) Calcificações e (c) Nódulo

4.2 Influência nos testes de processamento

A escolha dos casos que são utilizados em cada teste pode influenciar decisivamente na análise de desempenho do CAD em questão. Dessa forma, quaisquer comparações envolvendo diferentes processamentos são questionáveis se nenhuma informação adicional for fornecida, uma vez que a performance depende da "dificuldade" associada aos casos analisados. Conjuntos que possuam imagens com estruturas de interesse evidenciadas irão ocasionar

processamentos mais precisos e, obviamente, taxas menores de erros. (NISHIKAWA 1994)

Atualmente, diferentes grupos de pesquisa desenvolvem novas técnicas de processamento e algoritmos de otimização de desempenho de esquemas de auxílio ao diagnóstico e para isso desenvolvem também bancos de imagens para a realização de seus testes. (NISHIKAWA 1998)

Desta forma, a validação dos resultados obtidos com essas novas técnicas é feita basicamente empregando-se as imagens disponíveis dentro do próprio grupo. Devido ao fato de que estas imagens são obtidas em mamógrafos diversos e digitalizadas em diferentes equipamentos, as mesmas podem possuir características diferentes quanto à resolução espacial, resolução de contraste, formato de armazenamento do arquivo digital e ao tamanho do ponto focal do mamógrafo utilizado, gerando imagens com qualidades diferentes entre si. Assim, estas técnicas de processamento poderiam ser melhor avaliadas e rapidamente descartadas caso apresentassem baixo desempenho

Além disso, deve-se considerar que a distribuição de patologias mamárias em determinada região pode diferir muito da distribuição observada em outros lugares, sendo que o percentual de imagens consideradas normais e patológicas presentes no banco de imagens também pode, e deve, variar, de acordo com o local onde está sendo desenvolvido.

Este procedimento não invalida o algoritmo utilizado, mas o índice de acertos/erros obtido deve ser analisado considerando os fatores mencionados anteriormente e também a qualidade das imagens processadas.

Estudos recentes mostram que, no caso de imagens mamográficas digitalizadas para fins de processamento, deve ser utilizada resolução de contraste de, no mínimo, 10 bits (DANCE 1993), (RANGAYYAN 1993). Assim, é possível diferenciar de forma adequada o maior intervalo possível de densidades óticas (graus de escurecimento) presentes no filme que, depois de digitalizadas, são expressas pelos valores dos níveis de cinza da imagem.

Atualmente existe um consenso também no que diz respeito à resolução espacial que deve ser utilizada no caso de imagens digitalizadas empregadas em

processamento. Para imagens que serão empregadas em testes de detecção de calcificações/microcalcificações deve ser usado tamanho de pixel de 0,1 mm (100 μm). (KARSSEMEIJER 1993)

Este valor foi determinado após a realização de testes comprovando o fato de que resoluções com pixels menores do que 0,1 mm não contribuem significativamente para o aumento das taxas de detecções verdadeiro-positivas, mas aumentam de forma inconveniente as detecções falso-positivas (NUNES 2001).

O fato de se aumentar o número de falso-positivos contribui para a deterioração do desempenho dos esquemas computacionais de processamento e podem, inclusive, falsear diagnósticos que eventualmente sejam feitos de forma visual.

Estudos anteriores mostram que mesmo pequenas alterações, da ordem de 10%, no conjunto de imagens empregado podem causar uma perda de desempenho de até 25% na sensibilidade do sistema, o que significa redução significativa na taxa de detecções verdadeiras-positivas. Além de rotinas de detecção, foram também testados processamentos para classificação de estruturas mamárias, sendo que mudanças na base de imagens também causaram perdas sensíveis no desempenho geral do sistema (NISHIKAWA 1998), (SCHIABEL 2000), (SCHIABEL 2001).

Uma possível solução para esse problema seria a utilização do mesmo conjunto de imagens na avaliação dos diversos CADs e das novas técnicas desenvolvidas. Isso demandaria a criação de uma base de imagens comum, de uso público, que pudesse ser utilizada sempre que necessário. Entretanto, a criação desta nova base também apresentaria desvantagens, como, por exemplo, a necessidade de reavaliação das técnicas de processamento sempre que a base de imagens fosse alterada, seja por inclusão de novos casos, o que alteraria a proporção de casos normais/patológicos, seja por avanços tecnológicos relacionados às características de aquisição do aparelho de raios-X ou dos digitalizadores.

Além disso, durante a criação de uma base de imagens comum seria necessária a realização de estudos quanto às taxas de ocorrência das patologias mamárias de acordo com a região considerada, uma vez que estas patologias estão relacionadas à população submetida ao exame. Somente assim, a nova base poderia ser considerada representativa da realidade clínica existente, justificando seu uso comum durante os testes de processamento.(NISHIKAWA 1998)

Outra desvantagem relacionada à base comum diz respeito ao próprio desenvolvimento das técnicas de processamento. Alguns CADs em desenvolvimento poderiam ter seu desempenho otimizado para este conjunto de imagens e apresentar resultados satisfatórios, mas ainda assim não apresentarem a mesma eficiência quando implementados em casos clínicos reais, que é o objetivo final de qualquer sistema de apoio ao diagnóstico.

Uma segunda alternativa que amenizaria o efeito do conjunto de imagens sobre o processamento seria a realização de intercâmbio entre as entidades de pesquisa para a troca de imagens para que os sistemas em desenvolvimento pudessem ser treinados e testados com uma abrangência maior de casos e originadas em mamógrafos e digitalizadores diferentes. (NISHIKAWA 1994)

A troca de imagens apresenta o principal inconveniente de que as imagens teriam que sofrer um processo de uniformização no que diz respeito aos formatos de arquivos usados para o armazenamento, a plataforma de trabalho empregada e as características dos digitalizadores (tamanho de pixel, quantidade de níveis de cinza utilizados, curva característica de calibração do digitalizador, modelo do dispositivo, etc).

A terceira solução a longo prazo apresentado na literatura (NISHIKAWA 1994) de processamento de imagens diz respeito à criação de critérios objetivos para a medida da qualidade da imagem utilizada para testes. Não só as características de aquisição da imagem devem ser levadas em conta nestas medidas, mas também informações que quantifiquem, de alguma forma, o "grau de dificuldade" associado a estruturas presentes na imagem. Imagens com estruturas evidenciadas apresentariam menor "grau de dificuldade", uma vez que

contribuiriam para o aumento de detecções e classificações verdadeiras-positivas durante os testes do CAD. Por outro lado, imagens com baixo contraste ou com estruturas de interesse de difícil detecção seriam consideradas como possuindo maior grau de dificuldade, por apresentarem maiores chances de levar o sistema a resultados falsos-negativos, degradando sua performance geral.(NISHIKAWA 1994)

No cálculo destes critérios de qualidade associados às imagens, também os níveis de cinza e o tamanho de pixel obtido devem ser levados em consideração para a correta classificação da imagem.

4.3 Critérios de qualidade da imagem

Nenhuma das soluções apresentadas anteriormente foram de fato implementadas até o momento, sendo que a recomendação atual para que desempenhos de CADs diferentes possam ser comparados de modo razoável é de que os percentuais de acertos/erros e o desempenho geral do sistema desenvolvido sejam acompanhados de informações adicionais. Exemplos destas informações adicionais são os métodos utilizados na seleção dos casos (exames consecutivos, casos típicos de determinada estrutura de interesse, imagens consideradas "difíceis", imagens aleatórias, etc) e características físicas das imagens, através da apresentação de histogramas mostrando o intervalo de tamanho no qual as estruturas de interesse se encontram, características de contraste geral da imagem.

Além disso, curvas ROC ("Receiver Operating Characteristic") também são de grande utilidade na avaliação do desempenho de sistemas computacionais dedicados, sendo que estas curvas também podem sofrer influência do conjunto de imagens presentes no estudo.

A principal vantagem apresentada pelas curvas ROC está no fato de que estas curvas levam em consideração não só o percentual de acertos do sistema, mas também os erros nas detecções. Isto propicia uma avaliação global e completa do esquema de processamento em avaliação.

Outro aspecto que deve ser considerado quando utiliza-se curvas ROC diz respeito à característica "visual" da mesma, que pode transmitir informações de modo mais intuitivo ao observador quando comparado com tabelas e percentuais de acertos/erros.

Trabalhos anteriores mostram que, em geral, imagens com pouco contraste (consideradas de difícil interpretação), experimentos que utilizam pequena ou insuficiente quantidade de imagens e ainda imagens com baixos índices de qualidade acarretam piores resultados gerais quando comparados com outros esquemas que utilizam as mesmas técnicas, mas com conjuntos de imagens diferentes. (NISHIKAWA 1998)

Para imagens mamográficas, as estruturas de interesse (microcalcificações, calcificações e nódulos) apresentam maior densidade quando comparados com o tecido mamário adjacente, por tratarem-se de estruturas que contêm Cálcio em sua composição. Nos nódulos (que não contêm Cálcio), a maior densidade é causada pela diferente composição dos tecidos nodulares em comparação com os tecidos considerados normais. Isso ocasiona a visualização destas imagens com menor grau de escurecimento no filme.

Para que o banco de imagens fique mais abrangente e completo, outros dados além das imagens também devem ser armazenados com o objetivo de proporcionar ao usuário final a visualização da maior quantidade possível de informações e fornecer mais subsídios no auxílio ao diagnóstico.

A incorporação de novas informações deve permitir ainda que testes de processamento sejam feitos de acordo com diferentes características. Estas características não precisam, necessariamente, levar em conta apenas aspectos de aquisição e digitalização das imagens, mas podem também considerar dados como histórico das pacientes, idade das pacientes analisadas e até mesmo características clínicas como a densidade associada aos tecidos que compõem a mama.

Com esse objetivo, deve-se usar os laudos fornecidos pelos radiologistas como uma ferramenta de apoio ao desenvolvimento do sistema. Através destes

laudos pode-se organizar os procedimentos de consulta ao banco e, ao mesmo tempo, utilizá-los na avaliação dos algoritmos de processamento.

Além disso, acompanham o laudo informações clínicas da paciente, observadas geralmente pelo próprio ginecologista. Neste conjunto de informações também estão disponíveis dados tais como: número de gestações da paciente, período de lactação para cada filho, antecedentes de câncer ou outras doenças presentes na família, uso de anticoncepcional, simetria entre as mamas, etc. Todos estes dados devem ser incluídos no banco visando à sua abrangência. Além disso, devem ser disponibilizados de forma organizada.

A utilização do diagnóstico ou das suspeitas fornecidas pelo laudo radiológico apresenta a vantagem de possibilitar comparações entre exames de uma mesma paciente que tenham sido realizados em datas diferentes. Assim, pode-se acompanhar o histórico da evolução de eventuais doenças da mama, o tratamento ministrado no período (quando necessário) e também laudos anátomo-patológicos e citológicos anteriores. Esta outra modalidade de laudo é realizada através da análise patológica laboratorial e tem como objetivo identificar a classe e a composição celular do tecido sob análise.

No caso de imagens da mama, as amostras de tecido geralmente são compostas de nódulos ou partes da mama extraídas por conter estruturas suspeitas. Nos laudos emitidos pelo patologista podem-se identificar possíveis células com formação neoplásica, indicativas da presença de câncer ou, por outro lado, eliminar a necessidade de controles mais rígidos quando o exame laboratorial indica somente a presença de células consideradas normais, ausentes de qualquer agente carcinogênico ou patológico.

4.4 Interface

Uma das características de maior importância na construção de sistemas computacionais é o programa aplicativo que será a interface entre os dados armazenados e o usuário.

É desejável que se desenvolvam interfaces que sejam suficientemente fáceis e intuitivas para que possam ser utilizadas por pessoas sem grande experiência no uso de computadores. Sendo assim, o sistema final deve mostrar-

se robusto e transparente quanto à capacidade de processamento da máquina usada na visualização, uma vez que, atualmente, a diversidade quanto à velocidade de processadores, placas de vídeo e disponibilidade de memória é considerável.

Apesar da simplicidade necessária citada anteriormente, o sistema deve possibilitar consultas abrangentes e diversificadas, de modo a apresentar atrativos quanto à sua utilização. Outra característica desejável no que diz respeito à interface é a apresentação e a distribuição das informações na tela, que deve ser completa e amigável.

Por motivos éticos e de preservação da identidade das pacientes, quaisquer informações presentes na imagem além da própria mama e das estruturas de interesse devem ser removidas do arquivo antes do mesmo ser armazenado e mostrado ao usuário final, de modo a eliminar quaisquer sinais que possam identificar a origem da imagem, tais como: nome da paciente, número de registro no hospital correspondente, etc.

Na elaboração do banco de imagens também é de interesse o intervalo de idade nos quais as pacientes situam-se quando da obtenção das imagens. Esta estatística possibilita a realização de estudos/processamentos relacionados às características das mamas de acordo com a idade da paciente.

De acordo com estudos anteriores, a idade pode influenciar na densidade das mamas, alterando assim os parâmetros que serão utilizados no processamento. Neste ponto, a existência dos laudos associados também tem importância fundamental no processamento e na interpretação das imagens, uma vez que fatores externos também podem alterar a composição do tecido mamário. Entre estes fatores, pode-se citar eventuais tratamentos de reposição hormonal, uso de anticoncepcionais e quantidade de gestações da paciente. Todos estes dados estão presentes no prontuário clínico da paciente e devem também ser incluídos na base de dados para que as consultas possam ser realizadas.

Capítulo 5

5. Metodologia

5.1 Aquisição das imagens

Durante o desenvolvimento do banco de imagens procurou-se obter a maior quantidade possível de mamogramas durante o período do trabalho, de forma a incluir imagens provenientes de diferentes hospitais com os quais o nosso grupo de pesquisa mantém convênio de colaboração: Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto (HCFMRP-USP), Santa Casa de Misericórdia de São Carlos e Hospital São Paulo. Também tivemos acesso a algumas imagens do Hospital Amaral Carvalho, da cidade de Jaú, que é considerado referência no tratamento e combate ao câncer dentro do Estado de São Paulo.

A abrangência quanto aos hospitais nos quais as imagens foram obtidas apresentou também como importante característica a diversidade de patologias mamárias retratadas nos exames. Este fato é de grande relevância em qualquer banco de imagens a ser desenvolvido, uma vez que oferece a possibilidade para que grupos de pesquisa que realizam estudos considerando-se diversas estruturas de interesse utilizem o banco nos testes de processamento e no treinamento de pessoal ligado a esta área.

Além do fato de obtermos grande número de imagens, deve-se acrescentar ainda que os aparelhos mamográficos utilizados também apresentavam características diferentes entre si, principalmente em relação às imagens do Hospital das Clínicas de Ribeirão Preto, uma vez que os mamogramas obtidos correspondiam a um longo período de tempo e os mamógrafos foram substituídos neste período.

Uma importante diferença entre os mamógrafos que subsidiaram esse projeto está no tamanho do ponto focal apresentado por cada aparelho. O ponto focal é considerado fator decisivo na qualidade da imagem quando aspectos de resolução espacial real são analisados. Pontos focais de pequeno tamanho permitem detectar detalhes na imagem mamográfica, o que é fundamental para o posterior processamento de microcalcificações e calcificações. Já aparelhos com pontos focais de dimensões maiores podem acarretar borramento na imagem, prejudicando dessa forma a correta identificação destas estruturas de interesse que podem ser decisivas no diagnóstico.

A Figura 5.1 mostra as características dos pontos focais de cada um dos mamógrafos utilizados na aquisição das imagens incluídas no presente banco. Os valores correspondem ao tamanho do ponto focal medido no centro do campo de radiação. Este valor é utilizado como referência nos cálculos para estimação do tamanho efetivo do ponto focal em outras regiões do campo.

Figura 5.1: Características dos mamógrafos utilizados

Mamógrafo	Local	Tamanho do ponto focal (mm)
Philips Mammodiagnost*	HC-FMRP/USP	1,20 x 0,75
CGR Senographe 500 T**	HC-FMRP/USP	1,50 x 0,90
GE Senographe DMR ***	HC-FMRP/USP	0,45 x 0,37
CGR Senographe 600 T	Hosp. S. Paulo/UNIFESP	0,68 x 0,54
Philips Mammodiagnost MD 4000	Hosp. S. Paulo/UNIFESP	0,58 x 0,40
Lorad M III	Sta. Casa S. Carlos	0,60 x 0,40

Outro fator que altera a qualidade da imagem e que pode influenciar de forma decisiva no processamento diz respeito ao "scanner" utilizado no processo de digitalização das imagens. Como citado anteriormente, esta influência pode ser causada tanto pelas diferenças na resolução espacial (tamanho de pixel, conforme

* Equipamento utilizado rotineiramente até final dos anos 80.

** Equipamento utilizado rotineiramente até 1996.

*** Equipamento atualmente utilizado na rotina dos exames mamográficos naquele hospital.

mostrado anteriormente na Figura 2.2) utilizada quanto na resolução de contraste (níveis de cinza) proporcionada pelo digitalizador.

No presente trabalho, utilizamos dois digitalizadores a laser, ambos da marca Lumisys (Lumiscan 50 e Lumiscan 75) e que, segundo o fabricante, possibilitam obter imagens com até 12 bits de resolução de contraste (4096 níveis de cinza). Testes realizados utilizando-se medições de densidades óticas em ambos os aparelhos mostraram que a resolução de contraste real alcançada é de aproximadamente 3600 níveis de cinza [ESCARPINATI 2002].

Foi comprovado, no entanto, que estes 3600 níveis de cinza são suficientes para um bom desempenho do processamento. Este fato pode ser melhor entendido quando se nota que as principais estruturas de interesse para o processamento de imagens mamográficas têm característica de maior densidade, apresentando, portanto, imagens mais claras. Os níveis de cinza representados por estas estruturas estão localizados em sua maior parte nos valores mais próximos ao início da escala, não tendo sua visibilidade alterada significativamente pela diferença de amostragem citada.

A principal diferença entre os digitalizadores utilizados diz respeito à questão da resolução espacial. Enquanto o modelo Lumiscan 50 tem a capacidade de digitalizar as imagens com tamanho de pixel de 0,150 mm, o modelo Lumiscan 75 pode fornecer imagens com pixels de até 0,075 mm.

Esta diferença no tamanho dos pixels pode alterar o processamento, conforme citado anteriormente, mas não afeta a visualização normal da imagem no banco de imagens, uma vez que a maior limitação, neste caso, está na resolução máxima que pode ser expressa pelo "display" utilizado, no caso, o monitor de computador. A única alteração significativa que pode ser notada quando as imagens são comparadas é o efeito de aumento no tamanho dos arquivos gerados por ambos os digitalizadores. Deve-se salientar que uma diminuição de 50% no tamanho de cada pixel da imagem acarreta a quadruplicação na quantidade de pontos expressos na imagem por unidade de área. Desse modo, o tamanho do arquivo também é quadruplicado, acentuando

ainda mais o problema da escolha da melhor forma de armazenamento das imagens.

5.2 Modelagem da base de imagens

A modelagem do banco foi realizada com base nas informações colhidas nos laudos e em contatos com médicos radiologistas e ginecologistas, que guiaram a definição dos fatores que deviam fazer parte do conjunto de dados incorporados à base. Além disso, as opções de busca/consulta foram implementadas de modo a atender tanto aos requisitos médicos como aos interesses que possam existir em futuros processamentos de imagens.

O modelo de dados utilizado na elaboração do banco de imagens incluiu três tabelas: uma com os atributos das pacientes, outra com os atributos de cada conjunto (composto por quatro imagens) que compõem o exame, e ainda uma terceira, que contém as informações referentes às regiões de interesse previamente selecionadas.

A tabela de pacientes inclui dados da paciente que podem ser utilizados na busca e são utilizados também para identificação junto ao hospital, de modo a possibilitar possível acompanhamento médico no caso de patologias ou controle nos casos em que nada é detectado.

Cada exame é identificado pela data em que foi realizado, e possui seu próprio conjunto de imagens (único), que é recuperado dos arquivos de imagens através dos dados incluídos na tabela. Pelo fato de que cada paciente só realiza um exame de cada vez, a identificação da imagem é feita de forma única e a recuperação/armazenamento pode ser facilmente realizada.

Na tabela de exames também estão armazenados dados como a idade da paciente na data da realização do exame, a existência das três principais estruturas de interesse para processamentos (microcalcificações, calcificações e nódulos), o laudo emitido pelo radiologista e ainda um atributo que indica se existem regiões de interesse previamente selecionadas nas imagens deste exame.

A presença dos laudos radiológico, anátomo-patológico e citológico (quando realizado/disponível) apresenta a importante vantagem de indicar, de forma clara, a existência de eventuais estruturas patológicas nas mamas e, ainda, sua localização exata. Estas informações têm grande importância quando o interesse do processamento é específico, focado em uma única estrutura a ser estudada. O atributo "região" da tabela de exames indica a existência ou não de regiões pré-selecionadas em cada exame. Através deste campo é que as pesquisas por região são realizadas.

As regiões de interesse não são armazenadas diretamente no banco de imagens, mas a tabela de regiões funciona como referência para que as mesmas possam ser visualizadas, de forma que o usuário pode salvar cada uma das regiões de acordo com o seu interesse particular. A tabela de regiões também oferece o recurso de mostrar a região pré-selecionada diretamente na imagem do mamograma completo, propiciando assim uma visão global da localização de cada estrutura na mama da paciente. Os principais atributos desta tabela são as coordenadas (X,Y) dos pontos localizados no extremo superior esquerdo e no extremo inferior direito de cada região. Assim, cada região fica determinada de forma única e pode ser exibida na tela e mostrada juntamente com o mamograma completo.

A cardinalidade dos relacionamentos entre essas três tabelas é 1:N, já que uma mesma paciente pode ter realizado exames em datas diferentes. No caso do segundo relacionamento, esta cardinalidade fica evidente, uma vez que uma mesma imagem pode apresentar mais de uma região de interesse a ser selecionada.

A Figura 5.2 mostra o diagrama Entidade Relacionamento relativo ao banco de imagens desenvolvido:

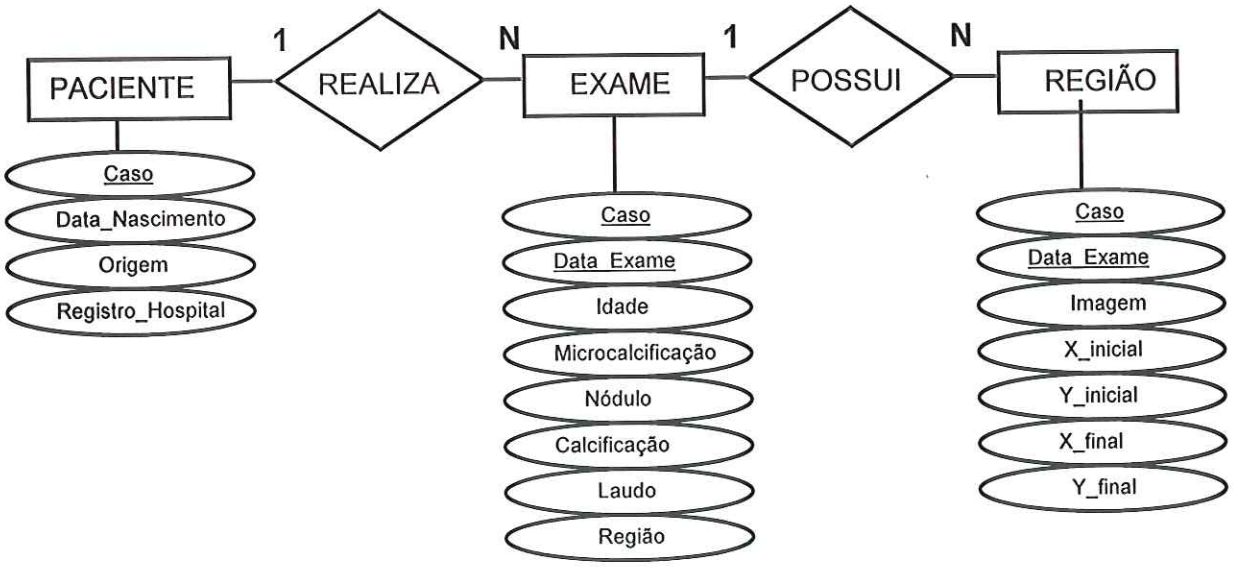


Figura 5.2: Diagrama Entidade-Relacionamento simplificado relativo ao banco de imagens

5.3 Buscas e consultas ao banco de imagens

As opções de consulta estão agrupadas de modo a permitir a seleção dos casos/imagens de interesse de acordo com o hospital de origem, o intervalo de idade da paciente e estrutura de interesse. A única restrição à busca de imagens diz respeito à seleção do hospital de origem. O sistema exibe mensagem de aviso quando nenhum dos hospitais disponíveis for selecionado, exigindo que o usuário faça a escolha do hospital de interesse. Todas as demais opções de busca podem ser feitas de forma conjunta ou exclusivas, conforme seja necessário.

Também foi incluída a opção de se fazer a busca de acordo com as regiões de interesse. Através dessa modalidade de busca, a etapa de recuperação e recorte manual dos mamogramas foi eliminada do processo de seleção de imagens para processamento.

Esta nova rotina de busca propiciou considerável ganho de tempo na busca por imagens, que é um fator importante quando um grande volume de imagens é necessário. As consultas podem ser feitas de forma conjunta ou separadamente, possibilitando-se restringir a quantidade e as características dos casos disponíveis que serão apresentados como resultado.

O resultado da seleção traz as quatro imagens normalmente associadas a cada paciente/exame, o hospital onde foram obtidas, o mamógrafo utilizado, a idade da paciente e a resolução espacial empregada no processo de digitalização. Além disso, as informações disponíveis nos laudos (radiológicos e anátomo-patológicos) também são disponibilizadas em um campo específico. A exibição destas informações oferece o recurso de se fazer comparações quanto à evolução dos exames de uma mesma paciente em várias datas.

Devido à resolução utilizada na digitalização das imagens e às diferentes resoluções empregadas nos monitores atuais, as quatro imagens de cada caso são exibidas simultaneamente em um tamanho menor do que o original e com menor quantidade de níveis de cinza, possibilitando a visão completa e total das imagens de interesse selecionadas através da consulta.

No entanto, as regiões pré-selecionadas ou recortadas pelo próprio usuário são mostradas separadamente, em seu tamanho original. Além disso, o programa oferece a opção de se aplicar "zoom" na região exibida.

A visualização das imagens é feita através da conversão dos arquivos de imagens, originalmente no formato TIFF (12 bits), em imagens com 256 níveis de cinza (8 bits/pixel), pois este é o padrão para visualização de imagens em monitores convencionais. A conversão é feita com o uso de rotina específica desenvolvida anteriormente dentro do grupo. Esta conversão, apesar de aparentemente degradar a qualidade da imagem, não altera o arquivo gerado na digitalização, sendo apenas e tão somente uma ferramenta de visualização.[ESCARPINATI 2002]

Vale salientar que o usuário sempre tem a possibilidade de acesso às imagens em seu formato original TIFF, com resolução espacial original de 0,15 mm ou 0,075 mm e 12 bits de resolução de contraste, que podem ser visualizadas em qualquer "software" que forneça suporte a este formato. Dessa forma, pode-se também selecionar as imagens para processamento utilizando-se todos os recursos oferecidos por este formato.

Com o recurso da seleção por regiões pré-definidas, estas também podem ser salvas pelo usuário com a finalidade de processamento. O mamograma

completo também pode ser salvo em seu formato original, de acordo com o interesse do processamento.

5.4 Desenvolvimento da interface

O banco de imagens foi desenvolvido utilizando-se linguagem visual *Delphi*, sendo que as imagens e laudos foram incorporados com o intuito de tornar ainda mais abrangente o universo de imagens disponíveis para consulta.

A escolha pelo *Delphi* apresentou ainda a vantagem de ter sido desenvolvido para uso sob o sistema operacional *Windows* (*Microsoft*). Vantagem adicional pode ser atribuída ao fato do *Delphi* ser uma linguagem compilada, gerando, portanto, códigos em linguagem de máquina (arquivos executáveis) em tempo de desenvolvimento ("Design-Time"), não necessitando estar instalado na máquina final para que os aplicativos sejam executados de forma correta.

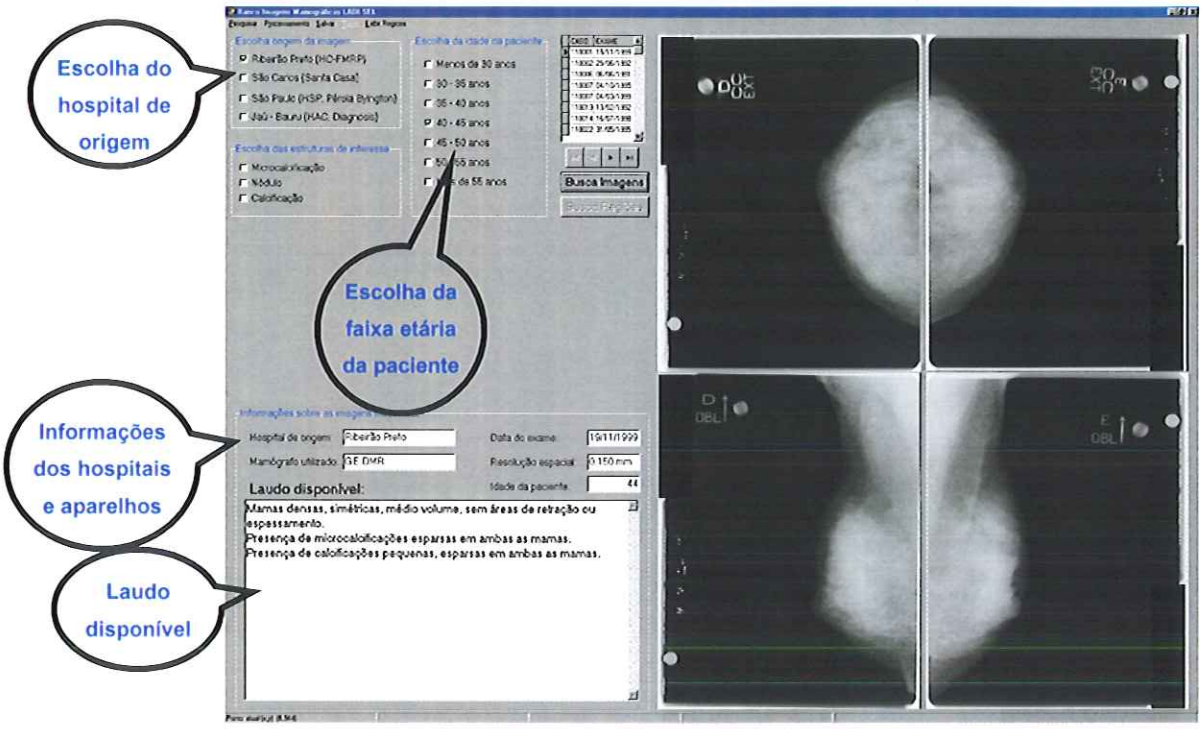
A interface também foi desenvolvida usando-se o *Delphi* por questões de integração com "softwares" previamente desenvolvidos dentro do grupo, pela facilidade de implementação e, principalmente, por oferecer suporte a banco de dados (Paradox) através do uso da linguagem SQL, já discutida anteriormente.

A estrutura utilizada possibilitou a apresentação das diversas opções de buscas na mesma tela, na forma de "menus" que são habilitados/desabilitados conforme as funções do programa são usadas. Este recurso impede que o programa pare de funcionar (travamento) por uso indevido, uma vez que o usuário só tem disponível as opções das operações permitidas.

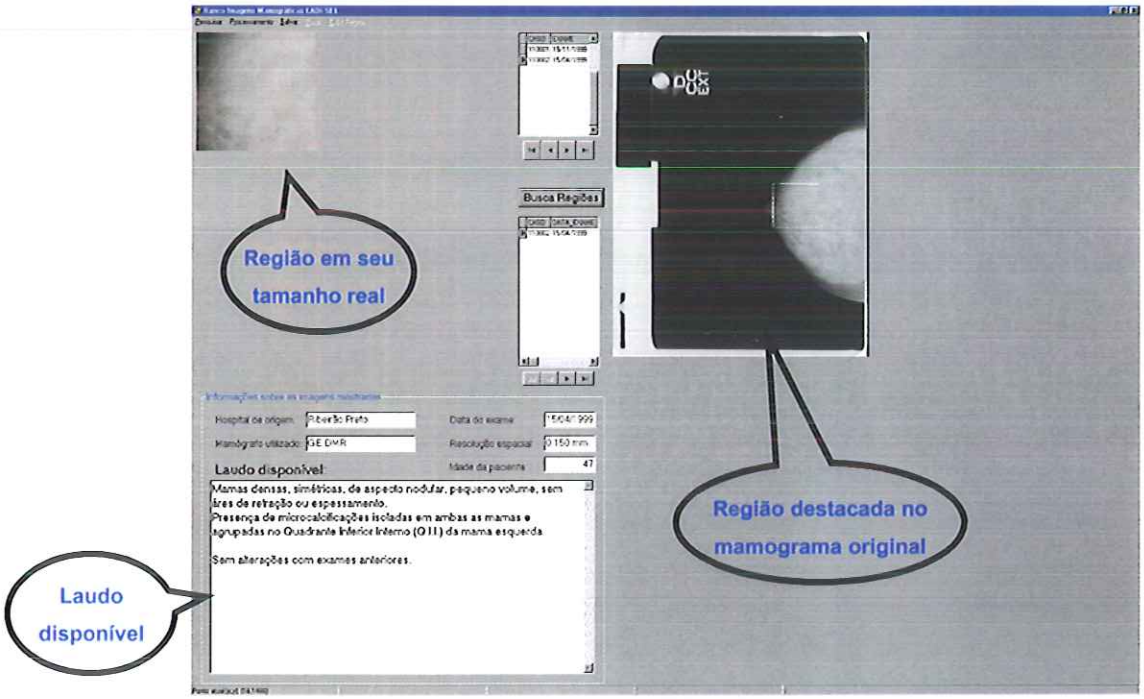
As imagens apresentadas na saída do sistema foram modificadas de modo a não exibirem qualquer informação que possibilite a identificação da paciente ou do hospital em que foram obtidas. O recurso de ocultar este tipo de informação tem por objetivo a preservação da identidade da paciente, por questões éticas e também para possibilitar eventuais distribuições destas imagens (ou versões das mesmas) para outros grupos de pesquisa.

Apesar disso, uma versão de "backup" foi feita, com o intuito de permitir futuras identificações junto ao hospital de origem quanto à incorporação de novas imagens e laudos de uma mesma paciente. No entanto, esta versão será de acesso restrito, pelas razões de ética já citadas anteriormente.

A Figura 5.3 mostra dois exemplos de consultas ao banco de imagens:



(a)



(b)

Figura 5.3: Exemplo de consultas ao banco de imagens (a) por características e (b) por regiões

5.5 Armazenamento das imagens

Um dos principais problemas enfrentados durante a elaboração do banco de imagens foi o armazenamento e o acesso às imagens, devido ao grande espaço computacional ocupado pelos exames (compostos por 4 imagens cada). Cada arquivo de imagem gerado pelo digitalizador Lumiscan 50 ocupa, aproximadamente, 3 Mb de espaço em disco. Quanto ao modelo Lumiscan 75 este espaço aumenta para aproximadamente 11 Mb, limitando consideravelmente a utilização total do banco desenvolvido em sistemas que não possuam alta capacidade de armazenamento/processamento.

A solução alternativa utilizada na resolução deste problema de armazenamento em disco foi o uso de CD's, para que o banco de imagens pudesse se tornar o mais "portátil" possível e para que não prejudicasse o desempenho da máquina que está sendo utilizada para visualização. Além dos arquivos com as imagens, os arquivos necessários para a execução do banco de dados propriamente dito e suas respectivas tabelas também encontram-se gravadas em CD's.

Gerou-se também um programa de instalação visando à utilização do banco de imagens em qualquer computador que utilize plataforma Windows, independente da configuração ou dos recursos disponíveis na máquina. Este programa gerencia a gravação das tabelas e do aplicativo em locais específicos da máquina "hospedeira" de modo a permitir o correto acesso à todos os recursos oferecidos pelo banco de imagens. O único requisito quanto ao computador utilizado é a presença de um leitor de CD-ROM, uma vez que esta é a forma de acesso às imagens disponíveis.

O espaço computacional ocupado atualmente pelas 3300 imagens que fazem parte do banco é de aproximadamente 20 Gb, sendo que se encontra armazenado em 30 CD's.

O sistema desenvolvido reconhece, via código, a localização de cada exame de acordo com a paciente e data da realização do exame, indicando ao usuário em qual CD o exame se encontra armazenado. Quando se aplicar, o programa exibe mensagem de aviso para a troca do CD em uso para que o

gerenciador do banco de dados possa fazer a recuperação da imagem de interesse.

Capítulo 6

6. Resultados e conclusões

6.1 Composição do banco de imagens

Quanto ao número de imagens, para este trabalho foram totalizadas 3300 imagens dos diversos casos disponíveis. A maior parte destas imagens (90%) apresenta os respectivos laudos radiológicos acompanhados dos laudos anátomo-patológicos e citológicos (quando realizados). O restante das imagens (aproximadamente 10%) encontra-se sem laudo devido a razões tais como perda/extravio do laudo dentro do próprio hospital ou por óbito da paciente, o que impossibilitou o acesso a essas informações.

O atual banco de imagens encontra-se em constante expansão, uma vez que os convênios de cooperação continuam em curso, através de projetos conjuntos com o Laboratório de Análise e Digitalização de Imagens (LADI-SEL), onde foi desenvolvido este trabalho.

O banco conta com aproximadamente 830 exames obtidos junto a 680 pacientes dos quatro hospitais com os quais mantemos convênio. Quanto aos hospitais de origem, o total de imagens está assim distribuído:

- 1440 imagens (243 pacientes) obtidas junto ao Hospital das Clínicas da Faculdade de Medicina de Ribeirão Preto (HC-FMRP);
- 1008 imagens (252 pacientes) obtidas junto ao Hospital São Paulo;
- 656 imagens (164 pacientes) obtidas junto à Santa Casa de Misericórdia de São Carlos;
- 196 imagens (49 pacientes) obtidas junto ao Hospital Amaral Carvalho de Jaú.

As imagens que fazem parte do banco abrangem uma vasta gama de estruturas de interesse e de doenças da mama. Também foram incluídas imagens que não apresentavam qualquer tipo de anomalia, pelos motivos estudados na literatura e expostos anteriormente.

A Figura 6.1 mostra de forma gráfica a distribuição das imagens de acordo com as estruturas patológicas visualizadas na imagem:

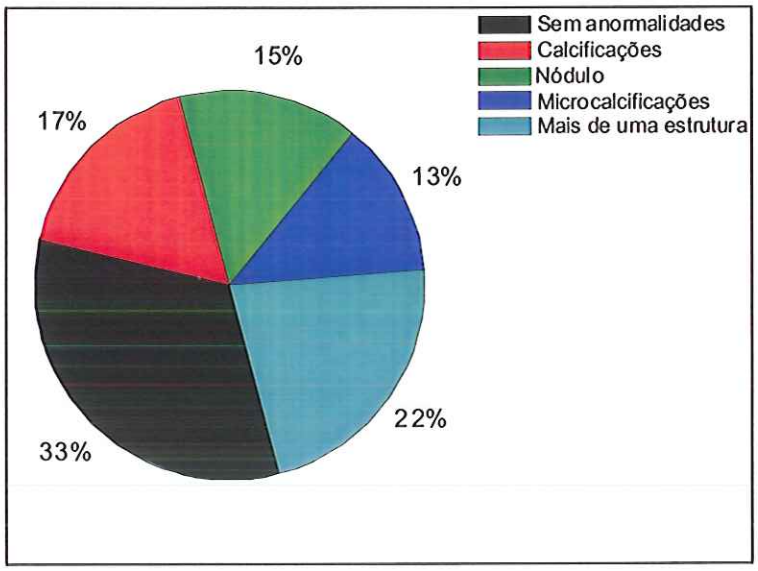


Figura 6.1 : Distribuição das imagens do banco, de acordo com as estruturas de interesse

O gráfico com a distribuição das patologias mostra que o banco de imagens desenvolvido retrata com satisfatória fidelidade a realidade clínica observada nos hospitais estudados, uma vez que praticamente um terço das imagens incorporadas apresentam laudo sem qualquer tipo de anormalidade. Estes exames correspondem à grande parte da população feminina que se submete à mamografia preventiva ou mesmo de acompanhamento e controle clínico. Portanto, é de se esperar que sejam retratados em maior número dentro deste estudo. Além disso, as imagens consideradas "não-patológicas" têm grande utilidade na avaliação das detecções falsas-positivas dos diversos algoritmos de processamento de imagens.

Dentre as imagens com algum tipo de patologia, pode-se notar uma distribuição praticamente uniforme (13%, 15% e 17%) entre as três principais doenças de mama. Este fato também retrata a realidade clínica analisada e apresenta a vantagem de contemplar de forma igualitária as estruturas de interesse para o processamento. Isto mostra que o banco de imagens encontra-se preparado para ser utilizado na busca/recuperação de imagens independentemente de qual seja o algoritmo em teste.

Outro fato importante que deve ser levado em conta no processamento de imagens mamográficas diz respeito ao intervalo de idade das pacientes estudadas. A Figura 6.2 apresenta graficamente a distribuição das imagens de acordo com a idade das pacientes. Neste gráfico estão representadas todas as imagens, não importando se os laudos estão ou não cadastrados no banco de dados. Isto é possível porque as idades são cadastradas independentemente dos laudos radiológicos e dos prontuários clínicos.

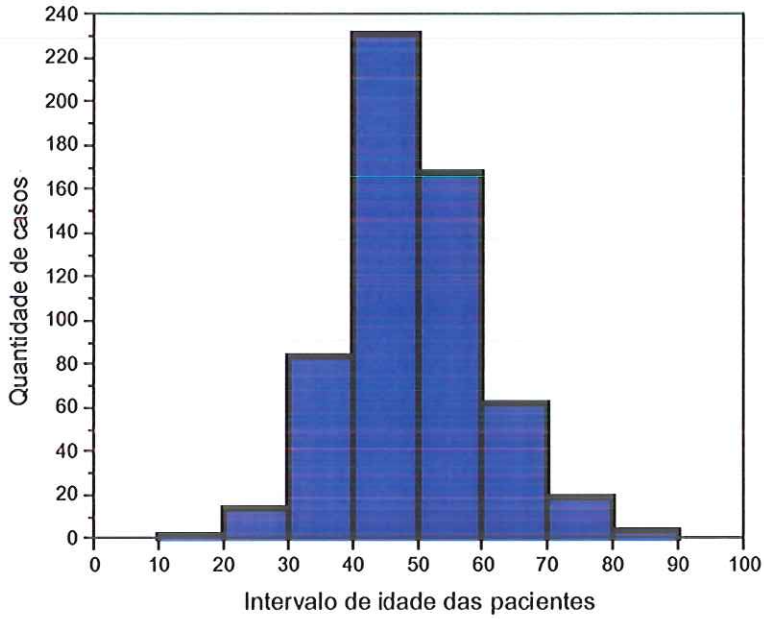


Figura 6.2: Distribuição das imagens de acordo com o intervalo de idades.

O gráfico apresentado mostra de forma clara que grande parte dos exames (aproximadamente 48%) concentra-se em pacientes com idades entre 40 e 60

EESC - USP
Biblioteca
CAMPUS DE S. CARLOS

anos. Este fato é conseqüência direta de fatores como a recomendação de se realizar mamografias de base em pacientes com idades a partir de 40 anos. Outro fator que leva a esta distribuição é a crescente busca por tratamentos de reposição hormonal, o que leva a realização de mamografias de controle e que geralmente é feito também a partir dos 40 anos.

Um terceiro fator que leva a esta concentração de exames nesta faixa etária é o nível de risco de desenvolvimento de doenças da mama neste período. Pacientes de menos idade geralmente apresentam mamas com maior densidade, e possíveis estruturas patológicas serão detectadas, com maior probabilidade, através de exames de ultra-sonografia e não por mamografias.

Além disso, estatísticas médicas mostram que pacientes com mais de 60 anos que não apresentam histórico de doenças mamárias, dificilmente irão desenvolver qualquer tipo de patologia da mama após esta idade, diminuindo-se assim a necessidade da realização de mamografias em pacientes nesta faixa etária.

Um exemplo da tela principal do programa é mostrado na Figura 6.3:

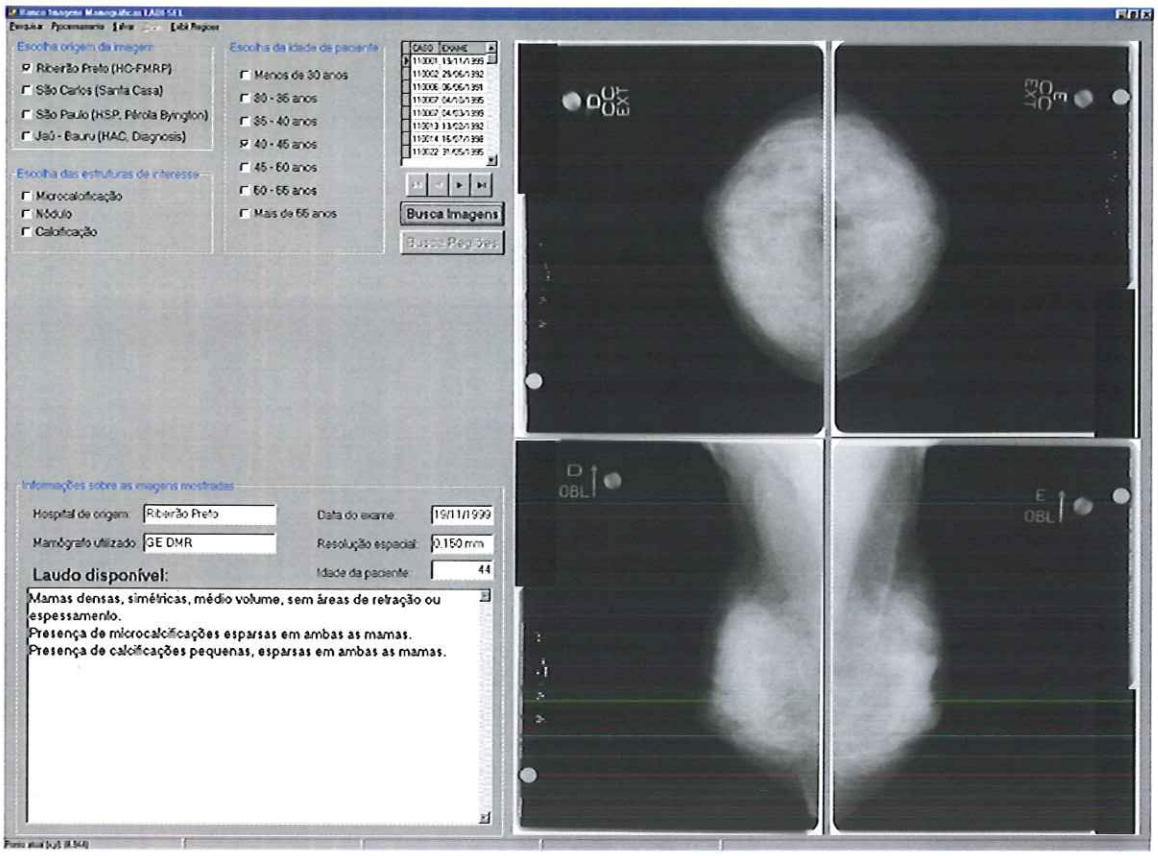


Figura 6.3: Exemplo da tela principal do programa

6.2 Testes de processamento

Com o objetivo de testar a viabilidade do uso prático do banco de imagens, foram realizados testes de processamento levando-se em conta a idade da paciente e a densidade associada às mamas. Este estudo mostrou que o banco tem a abrangência desejável, e necessária, para ser utilizado nos testes que necessitam de grande quantidade de imagens, como é o caso dos processamentos realizados para validação de novos algoritmos computacionais.

Além disso, o banco foi também utilizado como suporte para testes em processamento de detecção e classificação de estruturas nodulares, com bons resultados, mostrando que a diversidade desejada foi alcançada. O banco mostrou-se robusto quanto à interface, sendo utilizada com sucesso por várias pessoas dentro do grupo.

Nos testes realizados para detecção de microcalcificações, foram selecionadas 75 regiões onde o laudo radiológico informava presença de

agrupamento de microcalcificações. Estas imagens foram submetidas, primeiramente, à técnica de realce de contraste por FTM (Função de Transferência de Modulação) e então o esquema já desenvolvido de detecção pôde ser normalmente utilizado. ([NUNES 2002])

As regiões de interesse utilizadas nos testes foram extraídas do banco através da opção de consulta/busca por hospitais, uma vez que o interesse principal residia em aplicar correções via FTM, sendo que para isto, o aparelho mamográfico deve ser conhecido.

Neste caso, a base de imagens mostrou-se particularmente útil no que diz respeito ao recorte e seleção das regiões de interesse, que foi feita de forma direta, através das ferramentas presentes na interface implementada.

A Figura 6.4 mostra a curva ROC que expressa o desempenho do processamento de detecção de agrupamento de microcalcificações com a aplicação da técnica de realce de contraste.

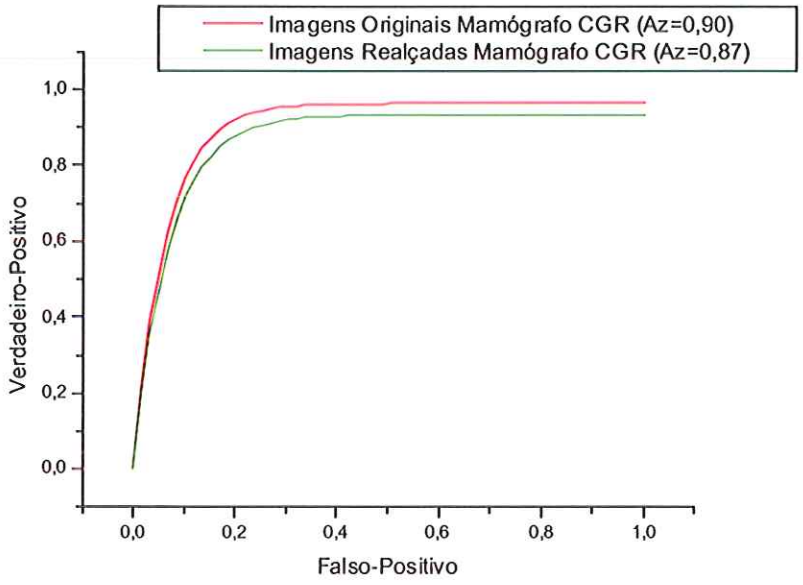


Figura 6.4: Curva ROC para o processamento de realce de contraste ([NUNES 2002])

No estudo realizado para a detecção de nódulos mamários, foi utilizada uma técnica de segmentação conhecida como Watershed, aplicada e

aperfeiçoada dentro do grupo e que mostrou bons resultados quando aplicado às imagens que fazem parte do banco de imagens. ([SANTOS 2002])

A região selecionada para o processamento dos nódulos foi submetida também ao processamento de equalização antes da aplicação da técnica de Watershed, sendo este procedimento de fundamental importância na comparação dos desempenhos. A Figura 6.5 mostra o esquema de processamento aplicado à estas imagens:

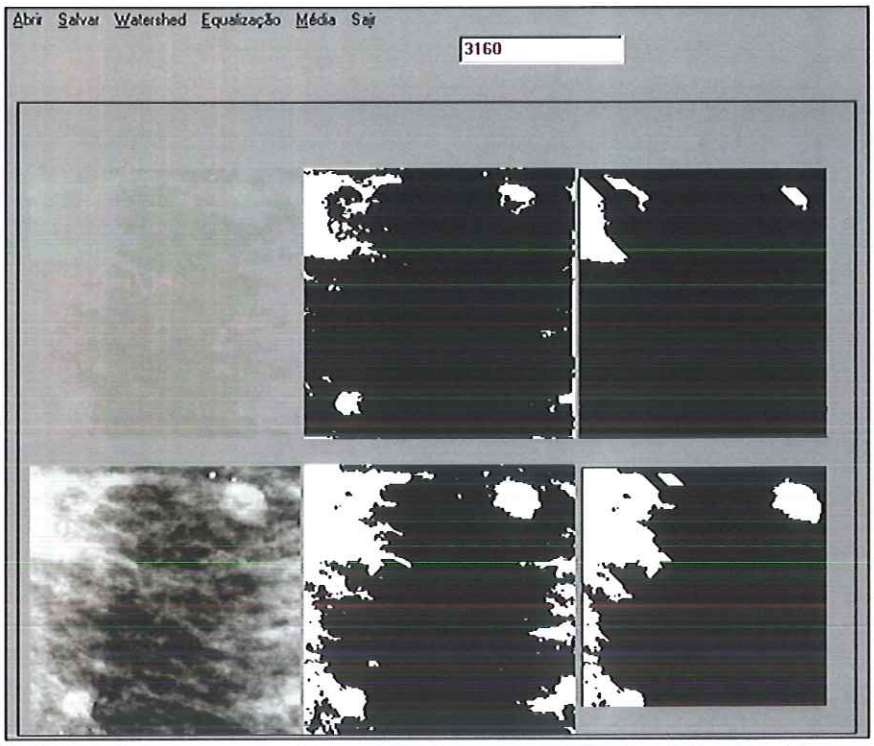


Figura 6.5: Exemplo da interface do "software"de segmentação de nódulos ([SANTOS2002])

Neste caso específico foram processadas 320 imagens (regiões de interesse), incluindo-se neste conjunto imagens que possuíam nódulos indicados pelos laudos radiológicos e também imagens que não apresentavam qualquer estrutura nodular, com o objetivo de testar o desempenho da técnica desenvolvida.

As imagens selecionadas tiveram como característica comum a presença de alta densidade nas mamas, que é um fator complicador na detecção de estruturas nodulares patológicas, uma vez que estas estruturas podem ser interpretadas como densificação usual da mama em estudo.

A Figura 6.6 mostra a curva ROC que analisa o desempenho do processamento realizado:

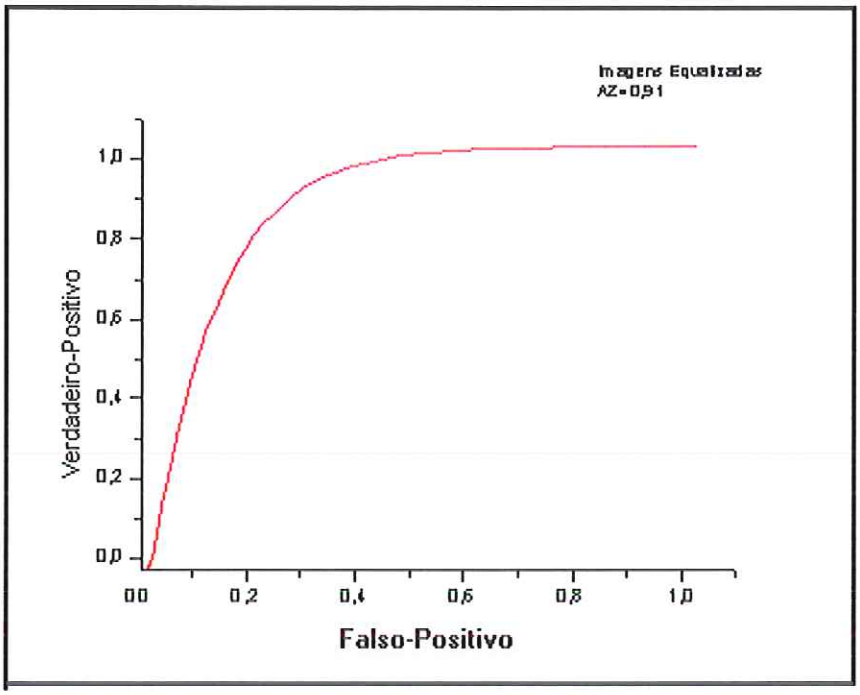


Figura 6.6: Curva ROC para a segmentação de nódulos ([SANTOS 2002])

Em outro estudo envolvendo nódulos de mama, foi desenvolvido sistema para a classificação destas estruturas de acordo com a sua forma, uma vez que esta é, geralmente, a principal característica observada pelos radiologistas na elaboração dos diagnósticos.([SANTAELLA 2002])

Neste caso, 60 imagens contendo nódulos de diversas formas (circular, lobular e espiculado) foram selecionadas para os testes, demonstrando mais uma vez a versatilidade obtida no desenvolvimento do banco de imagens.

A técnica utilizada para a classificação das imagens foi baseada na aplicação da transformada Wavelet nas imagens segmentadas obtidas de nódulos

mamários. A transformada empregada decompõe a imagem em componentes de freqüência, possibilitando a sua aplicação na forma de filtros.

No caso em questão, a aplicação sucessiva desta transformada e a quantificação da contribuição das componentes de baixas e altas freqüências para a composição final da imagem (quando a borda do nódulo é analisada) resultou em informações sobre a suavidade ou não da periferia do nódulo. Uma vez analisada esta informação, a classificação em "classes de circularidade" foi possível, tornando capaz a diferenciação entre nódulos circulares, lobulados e espiculados, de acordo com a característica de sua borda.

Este procedimento resultou em desempenhos satisfatórios de classificação, conforme expresso na curva ROC mostrada na Figura 6.7, que expressa a diferenciação entre a clase dos nódulos espiculados e a classe conjunta dos nódulos circulares e lobulados:

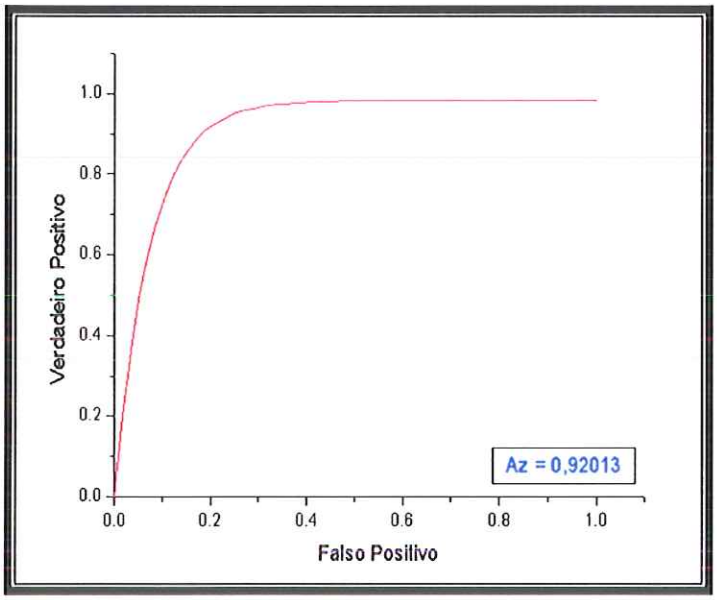


Figura 6.7: Curva ROC para o processamento de classificação de nódulos ([SANTAELLA 2002]).

A busca automática das imagens diminuiu, de forma decisiva, o tempo necessário para a recuperação das imagens de interesse, agilizando assim as rotinas de processamento e possibilitando, ao mesmo tempo, que mais imagens fossem selecionadas nos testes de novos algoritmos.

Capítulo 7

7. Discussão e conclusões

Os resultados mostraram que o banco de imagens pode ser utilizado de forma prática e rápida em testes de processamento, atingindo seu objetivo principal.

O ponto de destaque do banco implementado reside na sua forma intuitiva de ser utilizado, através da interface voltada à plataforma Windows e por apresentar "menus" de opções de fácil entendimento.

Além disso, a instalação do sistema completo não requer recursos especiais quanto ao "hardware" ou aos "softwares" previamente instalados. O arquivo de configuração ("setup") apresentou tamanho razoável (10 Mb), possibilitando que o mesmo seja gravado juntamente com o CD-ROM de armazenamento de imagens. Este fato contribuiu para a portabilidade do sistema, sendo que o único pré-requisito necessário ao sistema "hospedeiro" é a existência do leitor de CD-ROM, uma vez que as imagens são recuperadas e mostradas através deste tipo de mídia.

As imagens foram armazenadas em 30 CD's, totalizando aproximadamente 20 Gb de espaço em disco, o que reporta a abrangência da base de dados desenvolvida. Além disso, a gravação em CD-ROM apresentou-se como solução viável para a utilização e eventual distribuição do banco de imagens, uma vez que os arquivos envolvidos são de grande tamanho e, sendo assim, demandariam grande espaço livre em disco para a sua correta implementação. Dessa forma, a limitação de espaço foi solucionada.

Duas das principais dificuldades encontradas durante o desenvolvimento do banco foi a obtenção e a digitalização dos mamogramas. Quanto às imagens originais, o principal problema enfrentado foi a disponibilização dos exames pelos hospitais citados anteriormente. Muitas vezes as imagens não puderam ser encontradas pois se encontravam em poder da paciente ou seriam utilizadas em eventuais retornos ao hospital.

No que diz respeito à digitalização, o fator limitante foi o tempo disponível para que as imagens fossem armazenadas no banco, uma vez que os filmes originais não podiam permanecer por longo tempo fora dos arquivos hospitalares. Este problema foi contornado através da digitalização contínua dos mamogramas, de modo a não comprometer o processo de armazenamento das imagens e nem os possíveis interessados em consultar estas mesmas imagens através dos arquivos/prontuários das pacientes diretamente no hospital onde os exames foram realizados.

A obtenção dos laudos radiológicos também representou, em alguns casos, certa limitação, uma vez que muitos laudos se apresentavam incompletos ou com informações insuficientes ao cadastramento das imagens/exames. Além disso, como as imagens presentes no banco abrangem longo período de tempo (principalmente no que se refere ao HC-FMRP), alguns prontuários/laudos se encontravam em arquivos semi-ativos ou em arquivos-mortos, o que limitou o acesso a estas informações. Houve casos, também, de falecimento da paciente no decorrer do tratamento, o que inviabilizou totalmente o acesso aos dados, junto ao hospital, referentes aos exames realizados anteriormente.

Foi elaborado também um quadro comparativo entre os bancos de imagens já existentes/disponíveis e o banco de imagens desenvolvido neste trabalho, no qual pode-se observar sua compatibilidade e proximidade com os demais sistemas desenvolvidos por outros grupos de pesquisas com a mesma área de interesse. O quadro é mostrado a seguir:

Tabela 7.1: Quadro comparativo dos bancos de imagens existentes

Características	Bancos de imagens					
	MIAS	DDSM	UCSF/LLNL	CALMa	Nijmegem	LADI-SEL
Origem	Grã-Bretanha	Estados Unidos	Estados Unidos	Itália	Holanda	Brasil
Quantidade de imagens	320	2620	198	3000	40	3300
Resolução de contraste (bits/pixel)	8	12/16	12	12	12	12
Resolução espacial (μm)	50	42/50	35	85	100	100
Quantidade de hospitais	_	4	_	5	_	3
Formato da imagem armazenada	PGM	JPEG sem perdas	_	_	_	TIFF

Sugestões para a continuidade do banco já desenvolvido incluem a disponibilização destas imagens (juntamente com o programa para acesso às imagens) para outros grupos que tenham interesse no processamento de imagens mamográficas, além da continuação de sua utilização com futuras técnicas de processamento que venham a ser desenvolvidas dentro do próprio LADI-SEL. Pode-se citar ainda como projeto futuro, a disponibilização remota (via Internet) dos dados atualmente incorporados ao banco de imagens.

Por todo o exposto, pode-se afirmar que o banco desenvolvido encontra-se totalmente preparado para oferecer suporte aos testes de processamento com imagens mamográficas e que apresenta abrangência e facilidade de uso, atingindo, dessa forma, seu objetivo maior, que é o de contribuir para o desenvolvimento de esquemas CAD e, conseqüentemente, para o diagnóstico precoce do câncer de mama.

Deve-se salientar, por fim, que o banco continua em expansão, através da incorporação de novas imagens/laudos provenientes dos hospitais com os quais mantemos convênio ou que venham a fazê-lo futuramente.

8. Referências

AMENDOLIA, S.R.; BISOGNI, M.G.; BOTTIGLI, U.; CECCOPIERI, A.; DELOGU, P.; DIPASQUALE, G.; FANTACCI, M.E.; LORENZINI, E.; MARCHI, A.; MARZULLI, V.M.; OLIVA, P.; PALMIERO, R.; REGGIANI, M.; ROSSO, V.; STEFANINI, A. ; STUMBO, S.; TANGARO, S.; VENIER, O. (2001) – "The CALMA Project" - Nuclear Instruments & Methods in Physics Research Section A-Accelerators Spectrometers Detectors and Associated Equipment, Vol. 461 (1-3) pag. 428-429, 2001

CHEN, P. P-S. (1976) "The Entity-Relationship model – Toward a Unified view of data", Association for Computing Machinery Transactions on Database Systems, Vol. 1, No. 1, pags. 9-36, March 1976].

CODD, E.F. (1970) - "A Relational Model of data for large shared data banks", Communications of the Association for Computing Machinery, Vol. 13; No. 6, pags 377-387, June 1970

DANCE, D. R. (1993) - "Design of a common database for research in mammogram image analysis" - SPIE 93: Biomedical Image Processing and Biomedical Visualization, v. 1905, p. 538-539 - 1993.

DATE, C. J.; DARWEN, H. (1996)– "A guide to SQL standard" – Addison Wesley Profeesional, 1996

ESCARPINATI, M.C. (2002)– "Investigação de formatos e compressão de imagens digitais para processamento de imagens mamográficas de mamas densas" – Dissertação de Mestrado – Março/2002

HEATH, M.; BOWYER K.; KOPANS D.; MOORE R.; KEGELMEYER, JR., P. (2000)- "The Digital Database for Screening Mammography", Proceedings of the 5th International Workshop on Digital Mammography (Toronto, Canada, June 2000), Medical Physics Publishing

KARSSEMEIJER, N. (1993)– "A common database for research in mammographic image analysis" - SPIE 93: Biomedical Image Processing and Biomedical Visualization, v. 1905, p. 542-543 - 1993

KIMME-SMITH, C. (1993)– "Clinical considerations for a mammography database" - SPIE 93: Biomedical Image Processing and Biomedical Visualization, v. 1905, p. 546-547 - 1993

LYMAN J. A., HERSH W.R., SPACKMAN K.A. (2000)- "Representing Clinical Information in an Internal Medicine Teaching Image Database". Proceedings of the American Medical Informatics Association Fall Symposium 2000

MASCIO, L. N.; FRANKEL, S. D.; HERNANDEZ, J. M.; LOGAN, C.M.(1996)- "Building the LLNL/UCSF Digital Mammogram Library with Image Groundtruth" - Proceedings of the 3rd International Workshop on Digital Mammography, Chicago, June 1996 (Doi K ed.) Elsevier Science BV: Amsterdam 1996.

NISHIKAWA, R. M. (1993)– "Design of a common database for research in mammogram image analysis" - SPIE 93: Biomedical Image Processing and Biomedical Visualization, v. 1905, p. 548-549 – 1993

NISHIKAWA. R. M.; GIGER, M. L.; DOI, K.; METZ, C. E..; YIN, F-F.; VYBORNY, C.L.; SCHMIDT, R. A. (1994)– "Effect of case selection on the

performance of computer-aided detection schemes". Medical Physics, v. 21, p. 265-269, 1994

NISHIKAWA, R. M.; YARUSSO, L. M. (1998)– "Variations in measured performance of CAD schemes due to database composition and scoring protocol" - SPIE 99: Medical Imaging, v. 3338, p. 840-844 – 1998

NUNES, F. L. S.; SCHIABEL, H.; ESCARPINATI, M. C.; GOES, C. E. (2001)– "Comparisons of different contrast resolutions effects on a computer-aided detection system intended to cluster microcalcifications detection in dense breasts images" – Journal of Digital Imaging, v. 14, n. 2, suppl. 1, p. 217-219, 2001

NUNES, F.L.S; SCHIABEL, H; BENATTI, R.H.(2002) – Contrast enhancement in dense breast images using the modulation transfer function – Medical Physics, Vol. 29, No. 12, p. 2925-2936, 2002

RANGAYYAN, R.; PARANJAPE, R.; SHEN, L.; DESAUTELS, J. E. L. (1993)– "A database for mammographic image research"- SPIE 93: Biomedical Image Processing and Biomedical Visualization, v. 1905, p. 550-551 – 1993

SANTAELLA, C. H. M. (2002)– "Classificação de nódulos em Imagens Mamográficas digitais por Transformada "Wavelet" " - Dissertação de Mestrado – Escola de Engenharia de São Carlos – (EESC-USP), Setembro/2002

SANTOS, V.T. (2002)– "Segmentação de Imagens Mamográficas para detecção de nódulos em mamas densas" - Dissertação de Mestrado – Escola de Engenharia de São Carlos – (EESC-USP), Novembro/2002

SCHIABEL, H.; NUNES, F. L. S.; ESCARPINATI, M. C.; BENATTI, R. H. (2000)– "Performance of a processing scheme for clustered microcalcifications detection with different images database" – Proceedings of the World Congress on Medical Physics and Biomedical Engineering – Chicago (USA), Jul 23-28, 2000.

SCHIABEL, H.; NUNES, F. L. S.; ESCARPINATI, M. C.; BENATTI, R. H. (2001) – "Investigations on the effect of different characteristics of images sets on the performance of a processing scheme for microcalcifications detection in digital mammograms" – Journal of Digital Imaging, v. 14, n. 2, suppl. 1, p. 224-225, 2001.

STOREY, V.C., (2001) "Understanding and representing relationship semantics in database design", Natural Language Processing and Informations Systems, Vol. 1959, pgs 79-90, 2001.

SUCKLING, J.; PARKER, J.; DANCE D.; ASTLEY, S.; HUTT, I.; BOGGIS, C.; RICKETTS, I.; STAMATAKIS, E.; CERNEAZ, N.; KOK, S.; TAYLOR, P.; BETAL, D.; SAVAGE, J.; (1994)"The mammographic images analysis society mammogram database" - Exerpta Medica. International Congress Series 1069: 375-378 (1994) mias@sv1.smb.man.ac.uk]