RECUPERAÇÃO DE IMAGENS MAMOGRÁFICAS COM BASE NO CONTEÚDO VISUAL UTILIZANDO UMA BASE DE DADOS DE REFERÊNCIA

Júlia Epischina Engrácia de Oliveira

Júlia Epischina Engrácia de Oliveira

RECUPERAÇÃO DE IMAGENS MAMOGRÁFICAS COM BASE NO CONTEÚDO VISUAL UTILIZANDO UMA BASE DE DADOS DE REFERÊNCIA

Tese de Doutorado apresentada
ao Doutorado em Bioinformática
da Universidade Federal de Minas Gerais
como requisito para a obtenção do grau de Doutor.

Orientador:
Arnaldo de Albuquerque Araújo (UFMG)
Co-orientador:
Thomas M. Deserno (RWTH Aachen)

Belo Horizonte
Doutorado em Bioinformática – UFMG
2009

À minha mãe Nina e ao meu pai Júlio (*in memorian*).

Agradecimentos

À CAPES e CNPq pelo apoio financeiro.

Ao meu orientador, Arnaldo de Albuquerque Araújo, por suas sugestões e críticas que me ajudaram no desenvolvimento do trabalho.

Ao meu co-orientador, Thomas M. Deserno, por sua disponibilidade durante a minha estadia na Alemanha e por suas sugestões, críticas e incentivo durante o final desse trabalho.

Ao professor Alexei Manso Correa Machado, por sua disponibilidade e acompanhamento no final desse trabalho.

Aos colegas do laboratório NPDI, em especial ao Guillermo, Ana e Thatyene por toda a ajuda.

Aos colegas alemães Ben e Mark, por toda a ajuda durante a minha estadia na Alemanha.

Aos amigos que fiz durante o doutorado e estadia em Belo Horizonte: Renata, Sandra, Natália, Rodrigo, Dayane, Ana, Marcelo, André, Mariane, Khalil e Luciana e aos amigos de longe, Lisa, Vívian, Mírian e Rodrigo.

À Zilda, Paulinho e Guilherme, por me receberem com carinho em Belo Horizonte.

Ao Iouri e Gália, amigos queridos e família do coração, sem eles eu não teria a tranquilidade de estar longe quando foi necessário. Agradeço também pelo apoio, disponibilidade, críticas e sugestões que foram importantes para o desenvolvimento desse trabalho.

Aos meus sobrinhos, Júlia e Luiz Eduardo, e à Bellinha, por alegrarem e iluminarem a minha vida.

Ao meu pai Júlio, que tão cedo e repentinamente nos deixou, pelas conversas, pelo apoio incondicional, pelo incentivo, pelos conselhos e por sempre acreditar em mim. A saudade é eterna.

À minha mãe Nina, minha amiga e companheira, sem seu apoio, força e incentivo eu não teria chegado até aqui.

Resumo

Os sistemas de recuperação de imagens, como ferramenta de auxílio ao diagnóstico, podem ajudar o radiologista na sua tomada de decisão através da apresentação de um conjunto de imagens similares a uma imagem específica de busca. Lesões da mama são indicativas do câncer de mama e em certos tipos de tecidos da mama essas lesões podem se ocultar, visto que tecido e lesão aparecem como áreas mais brancas nas imagens mamográficas. Utilizando a resposta do computador como referência para o auxílio ao diagnóstico, este trabalho visa o desenvolvimento de um sistema de recuperação de imagens mamográficas que usa imagens da base de dados do projeto IRMA, que contém imagens mamográficas classificadas e verificadas por um experiente radiologista. Dois casos de estudo são propostos. O primeiro caso de estudo, o MammoSys, utiliza como padrão de busca os tipos de tecido da mama, de acordo com as quatro categorias BI-RADS propostas pelo Colégio Americano de Radiologia. A técnica análise dos componentes principais em duas dimensões é usada para caracterizar a diferença da textura dos tecidos da mama, a fim de representá-la apropriadamente e permitir também a redução de dimensionalidade do vetor de características obtido. Máquina de vetores de suporte é utilizada para o processo de recuperação das imagens. Valores médios de precisão estão entre 77,83% e 81,11% considerando um conjunto de 800 imagens mamográficas. O segundo caso de estudo, o MammoSysLesion, utiliza como padrão de busca, em conjunto, os tipos de tecido da mama e a existência ou não de uma lesão mamográfica e sua classificação, de acordo com as categorias BI-RADS. A técnica análise dos componentes principais em duas dimensões é utilizada para a caracterização dos tecidos e lesões da mama e a máquina de vetores de suporte é usada para o processo de recuperação das imagens. Valores médios de precisão estão entre 70,95% e 80,64% considerando um conjunto de 1.392 imagens mamográficas.

Palavras-chave: base de dados de imagens médicas, tecido da mama, lesão mamográfica, sistema de recuperação de imagens, análise dos componentes principais em duas dimensões, máquina de vetores de suporte.

Abstract

As a tool of aid of diagnosis, content-based image retrieval systems can help radiologists in reducing the variability of their analysis through the presentation of a similar set of images according to a specific query image. Breast lesions are indicative of breast cancer and some types of breast tissue can hide these lesions, as both lesion and tissue appear as white areas in mammographies. Using the computer answer as a reference for the aid of diagnosis, this work aims at developing a content-based image retrieval system that uses images of the IRMA database which contains mammographies classified and verified by an experienced radiologist. Two case studies are proposed. The first case study, MammoSys, uses breast density as a pattern of retrieval, according to the four BI-RADS categories proposed by the American College of Radiology. Two dimensional principal component analysis is used for the characterization of breast density texture, that allows for feature extraction at the same time that dimensionality reduction is performed. Support vector machine is used for the image retrieval task. Average precision rates are in the range from 77.83% to 81.11% considering a set of 800 mammographies. The second case study, MammoSysLesion, uses breast density together with the existence of a breast lesion and its classification, according to the BI-RADS categories. Two dimensional principal component analysis is used for breast density and lesions characterization, and support vector machine is used for the image retrieval task. Average precision rates are in the range of 70.95% and 80.64% considering a set of 1,392 mammographies.

Key words: medical images database, breast lesion, content based image retrieval system, two dimensional principal component analysis, support vector machine.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviações

2DPCA *Two Dimensional Principal Component Analysis*

ACR *American College of Radiology*

AMD *Assembled Matrix Distance*

BI-RADS *Breast Imaging Reporting and Data System*

CAD *Computer Aided Diagnosis*

CBIR *Content Based Image Retrieval*

CC Crânio-caudal

DDSM *Digital Database for Screening Mammography*

D-LDA *Direct Linear Discriminant Analysis*

GRNN *General Regression Neural Network*

ICS *Image Cytometry Standard*

IRMA *Image Retrieval in Medical Applications*

JPEG *Joint Pictures Expert Group*

k-NN *k-nearest neighbor*

LDA *Linear Discriminant Analysis*

LLNL *Lawrence Livermore National Laboratory*

MIAS *The Mammographic Image Analysis Society Digital Mammogram Database*

ML Médio-lateral

MNN *Modular Neural Network*

NGTDM *Neighborhood graytone difference matrix*

PCA *Principal Component Analysis*

PGM *Portable Grey Map*

PNG *Portable Network Graphics*

RBF *Radial-Basis Function*

ROC *Receiver Operator Characteristic Curve*

ROI *Region of Interest*

RWTH *Rheinisch-Westfälische Technische Hochschule*

SVD *Singular Value Decomposition*

SVM *Support Vector Machine*

Capítulo 1

Introdução

Neste capítulo, apresentam-se o contexto, a motivação e os desafios que deram origem ao desenvolvimento deste trabalho. Os principais objetivos são discutidos e as contribuições pretendidas são apresentadas, finalizando com a apresentação de como este documento está organizado.

1.1 Considerações Iniciais

As imagens médicas são importantes para o propósito de diagnóstico por estarem relacionadas ao histórico médico do paciente e a sua patologia. O câncer de mama representa uma das causas principais de morte entre mulheres no mundo [INCA, 2009, Xue and Michels, 2007], e a detecção precoce do câncer é o método mais eficaz de reduzir a mortalidade. A imagem mamográfica é o principal modo de investigação dessa doença e é obtida através da mamografia, que constitui uma forma particular de radiografia que utiliza níveis de radiação mais baixos que os utilizados em radiografia convencional, e destina-se a registrar imagens das mamas a fim de se diagnosticar a eventual presença de estruturas indicativas de doenças.

Wolfe [1976] em seu estudo sugeriu que existe uma relação entre os padrões de composição mamária, isto é, o tecido ou densidade mamária, e o risco do desenvolvimento do câncer de mama, visto que uma alta densidade da mama pode ocultar as lesões dificultando a detecção precoce do câncer.

De maneira a padronizar os relatórios sobre o diagnóstico de uma imagem mamográfica, melhorar a comunicação entre os radiologistas e auxiliar as pesquisas científicas foi desenvolvido em 1993 pelo Colégio Americano de Radiologia[1] (ACR - *American*

[1]http://www.acr.org

College of Radiology) o código padrão BI-RADS (*Breast Imaging Reporting and Data System*). Para a avaliação do tecido ou densidade da mama, o código é definido como:

- BI-RADS I: a mama é extremamente gordurosa.
- BI-RADS II: a mama é gordurosa e existe algum tecido fibroglandular.
- BI-RADS III: a mama é, de forma heterogênea, densa.
- BI-RADS IV: a mama é extremamente densa.

Um outro tipo de avaliação é relacionada a existência ou não de uma lesão e sua classificação:

- BI-RADS 0: necessidade de imagens adicionais ou é necessária a comparação com exames prévios.
- BI-RADS 1: achados mamográficos negativos – mamografia normal.
- BI-RADS 2: achados mamográficos benignos.
- BI-RADS 3: achados mamográficos provavelmente benignos.
- BI-RADS 4: achados mamográficos suspeitos – necessidade de avaliação adicional.
- BI-RADS 5: achados mamográficos altamente suspeitos e sugestivos de malignidade.
- BI-RADS 6: achados mamográficos com biópsia provando o câncer de mama.

Os radiologistas avaliam e relatam a densidade da mama através de uma análise visual da imagem mamográfica. Nesse cenário, sistemas computadorizados de auxílio ao diagnóstico (CAD - *Computer Aided Diagnosis*) [del Bimbo, 1999, Baeza-Yates and Neto, 1999, Lehmann et al., 2005, Doi, 2007] e sistemas de recuperação de imagens por conteúdo (CBIR - *Content Based Image Retrieval*) [Salton and McGill, 1983, Müller et al., 2004, Rahman et al., 2004, Oliveira et al., 2007] surgem como uma possibilidade de auxiliar os radiologistas na redução da variabilidade de sua análise e de melhorar a sua acurácia na interpretação da imagem mamográfica, mediante o uso da resposta do computador como referência.

Os sistemas CBIR, que fazem parte dos sistemas CAD, utilizam informações visuais extraídas das imagens para recuperar imagens similares a uma imagem específica de busca. Um sistema CBIR não necessita fornecer informações diagnósticas sobre as

imagens recuperadas, mas apenas apresentar imagens similares a um padrão específico. O esquema de um sistema CBIR pode ser visualizado na Figura 1.1. Para cada imagem da base de dados e da imagem de consulta, é extraída uma região de interesse (ROI - *Region of Interest*) que é representada através de atributos de cor, forma ou textura. Os atributos numéricos de cada uma das ROIs são extraídos para a definição do vetor de características que tem então sua dimensionalidade reduzida. Os mesmos atributos são também extraídos da ROI da imagem de consulta e comparados com os atributos armazenados na base de dados. É calculado um índice de similaridade entre as imagens, e aquelas mais próximas à imagem de consulta são recuperadas da base de dados e apresentadas ao usuário, em resposta à sua consulta.

Um sistema CAD ou CBIR efetivo, isto é, um sistema que forneça informações diagnósticas de uma imagem ou um sistema que apresente realmente imagens similares de acordo com um certo padrão, deve ser avaliado utilizando um grande número de imagens de referência com diagnóstico já aprovado por experientes radiologistas (padrão de ouro). Apesar de grandes bases de dados para imagens mamográficas, como por exemplo a DDSM (*Digital Database for Screening Mammography*)[2] [Heath et al., 1998] com aproximadamente 9.000 imagens, estarem publicamente disponíveis, os problemas para os pesquisadores que necessitam de dados de referência são muitos, já que os pesquisadores precisam enfrentar o problema de selecionar um número de casos apropriados suficientes para os procedimentos de desenvolvimento dos sistemas CAD ou CBIR.

Além disso, o grande volume de imagens médicas produzidas em hospitais e centros médicos vem crescendo rapidamente, tornando-se necessário um armazenamento seguro dessas imagens, bem como seu gerenciamento eficiente. O projeto IRMA (*Image Retrieval in Medical Applications*)[3] lida com esses tipos de problema, já que visa o desenvolvimento e implementação de métodos de alto nível para a recuperação de imagens médicas radiológicas baseada em conteúdo. A base de dados do projeto IRMA contém mais de 20.000 imagens radiológicas e, atualmente, mais de 10.000 imagens mamográficas, todas elas disponíveis com suas informações diagnósticas, oferecendo um incalculável suporte para o desenvolvimento de sistemas CBIR e CAD.

[2]http://marathon.csee.usf.edu/mammography/database.html
[3]http://www.irma-project.org

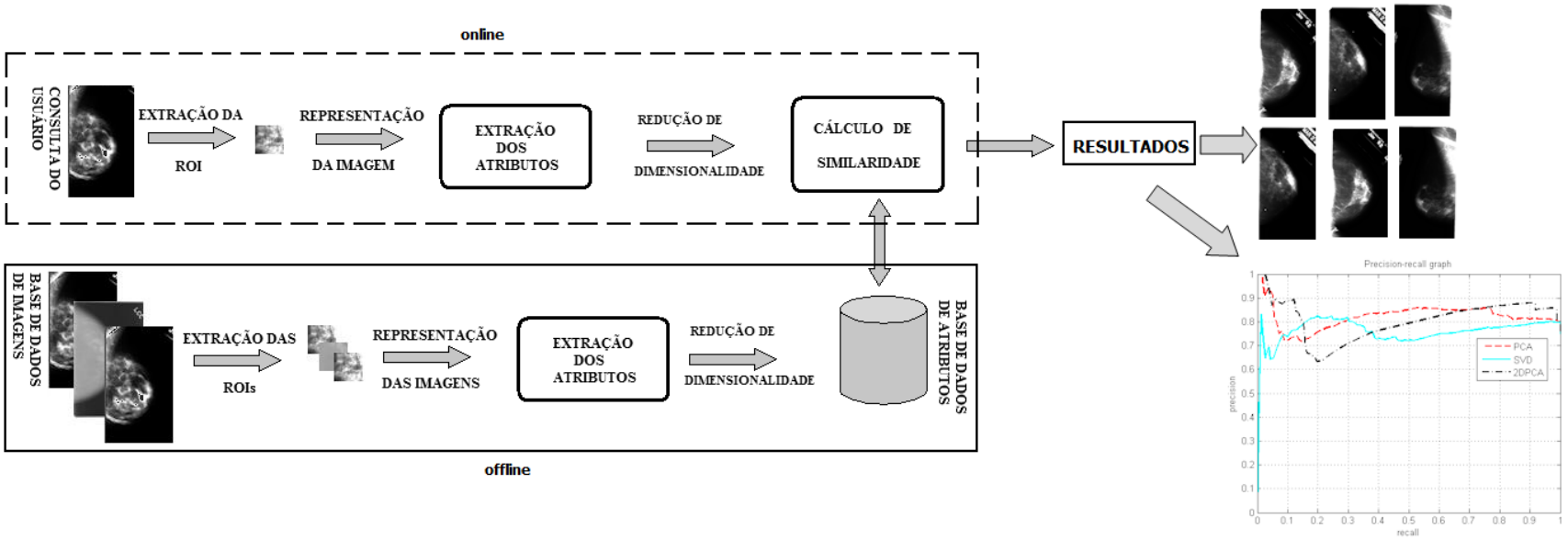

Figura 1.1. Sistema CBIR. Das imagens da base de dados e da imagem de consulta são extraídos os atributos que as representam, sendo calculado um índice de similaridade entre essas imagens que irá indicar as imagens mais relevantes à imagem de consulta para serem recuperadas da base de dados e apresentadas ao usuário.

1.2 Motivação

A consulta por similaridade de imagens, ao ser avaliada, deve corresponder ao esperado pelo usuário. Os critérios envolvidos na percepção humana devem ser considerados nessa avaliação, porém muitas vezes os próprios médicos divergem entre si quanto ao ponto de vista semântico referente à análise das imagens médicas. Um sistema CBIR pode auxiliar o médico em seu diagnóstico, fornecendo a ele um conjunto de imagens semelhantes ao de um paciente em questão.

O câncer de mama pode se ocultar em mamas que possuem uma alta densidade. Visto que os radiologistas possuem uma dificuldade em obterem um consenso na diferenciação dos tipos de tecidos da mama, através de um sistema CBIR eles podem analisar comparativamente as imagens mamográficas e obter um conjunto de imagens similares para o seu auxílio.

Considerando um sistema CBIR com base no tipo de tecido da mama, do ponto de vista clínico, esse sistema pode guiar os radiologistas para a detecção de uma lesão e sua posterior classificação. Do ponto de vista técnico, esse sistema é o primeiro e importante passo para o desenvolvimento de um sistema CAD.

Nos sistemas CBIR, as imagens são descritas por atributos ou vetores de características, que as representam sob um determinado aspecto. É um desafio a escolha de um conjunto de atributos que capture a essência das imagens, descrevendo-a por meio de uma quantidade pequena de valores. Também, o armazenamento e acesso rápido a essas imagens podem ser vistos como um problema.

1.3 Objetivos e Contribuições

O objetivo geral deste trabalho é apresentar um sistema de recuperação de imagens mamográficas utilizando para base de testes a base de dados de imagens mamográficas do projeto IRMA. As imagens são caracterizadas através de um atributo de textura e recuperadas utilizando-se um classificador, que indica a relevância de cada imagem recuperada para uma determinada imagem de busca. Assim sendo, as principais contribuições são apresentadas:

- Utilização do tecido da mama, de acordo com as quatro categorias BI-RADS, como padrão de busca para o caso de estudo denominado MammoSys, que introduz a análise dos componentes principais em duas dimensões (2DPCA - *Two Dimensional Principal Component Analysis*) para a caracterização da textura dos tecidos da mama, de maneira que as características do tecido da mama são

extraídas ao mesmo tempo que a redução da dimensionalidade dos vetores de características é realizada. A técnica 2DPCA é comparada com a decomposição em valores singulares (SVD - *Singular Value Decomposition*) e com a análise dos componentes principais (PCA - *Principal Component Analysis*) para a caracterização do tecido da mama. O classificador máquina de vetores de suporte (SVM - *Support Vector Machine*) é utilizado no processo de recuperação de imagens.

- Utilização do tecido da mama juntamente com a existência de uma lesão mamográfica e sua classificação, de acordo com as categorias BI-RADS, como padrão de busca para o caso de estudo denominado MammoSysLesion. As técnicas 2DPCA, SVD e PCA são comparadas para a caracterização do tecido da mama e a lesão. O classificador SVM é utilizado no processo de recuperação de imagens.

- Integração de quatro bases de dados já existentes ao projeto IRMA, de forma que todas as imagens possuam informações diagnósticas previamente avaliadas por um radiologista experiente. Essa base de dados de imagens mamográficas do projeto IRMA será utilizada para a realização dos testes desse trabalho.

1.4 Organização do Trabalho

O conteúdo desse trabalho está estruturado da seguinte forma:

- Neste primeiro capítulo, foi apresentada uma visão e a importância do processo de recuperação de imagens mamográficas para auxiliar o radiologista em seu diagnóstico.

- No Capítulo 2, são apresentados e analisados alguns trabalhos relacionados à recuperação de imagens mamográficas e à classificação dos tecidos e lesão da mama.

- No Capítulo 3, as técnicas 2DPCA, PCA e SVD são definidas e analisadas detalhadamente.

- No Capítulo 4, é descrita a utilização do classificador SVM no processo de recuperação das imagens mamográficas.

- No Capítulo 5, os casos de estudo realizados para o desenvolvimento e avaliação do sistema CBIR proposto são apresentados e os resultados obtidos são descritos e analisados.

- No Capítulo 6, as conclusões desse trabalho são apresentadas.

Capítulo 2

Trabalhos Relacionados

As imagens médicas são importantes para propósitos de diagnóstico por estarem diretamente relacionadas à patologia do paciente e seu histórico médico. A mamografia constitui uma forma particular de radiografia que utiliza níveis de radiação mais baixos que os utilizados em radiografia convencional e destina-se a registrar imagens das mamas, a fim de diagnosticar a eventual presença de lesões indicativas de câncer. O objetivo final do exame radiológico é produzir imagens detalhadas das estruturas internas da mama e uma boa interpretação dos resultados da mamografia torna-se imprescindível para o estabelecimento de um bom diagnóstico e da tomada acertada de decisões pelo radiologista.

De maneira a auxiliar os radiologistas em suas tomadas de decisões, pesquisas em sistemas envolvendo a recuperação por conteúdo visual de imagens mamográficas têm sido desenvolvidas nos últimos anos, destacando-se como uma área de pesquisa ativa no campo da visão computacional e conseqüentemente em processamento de imagens.

Neste capítulo, trabalhos relacionados ao desenvolvimento de sistemas de recuperação de imagens mamográficas são apresentados. Especificamente, são relatadas as diversas maneiras de caracterizar os tecidos da mama e as lesões mamográficas, e de recuperar as imagens mamográficas juntamente com a avaliação dos sistemas CBIR.

2.1 Caracterização dos Tecidos da Mama

A densidade ou tipo de tecido da mama é um indicador de risco do desenvolvimento do câncer de mama e, visualmente, os tipos de tecidos são diferenciados através da intensidade de níveis de cinza nas imagens mamográficas. A definição de um conjunto de características capaz de descrever de maneira efetiva os tipos de tecidos da mama é um desafio para o desenvolvimento de sistemas CBIR.

Kinoshita *et al.* [2007] utilizaram 88 atributos de maneira a caracterizar o tecido da mama de 1.080 imagens mamográficas do Hospital das Clínicas da Faculdade de Medicina da Universidade de São Paulo[1], Ribeirão Preto – SP, que foram previamente categorizadas manualmente por um radiologista de acordo com as quatro categorias BI-RADS. Todas as imagens mamográficas passaram por um processamento para a remoção de ruídos, tais como a etiqueta de identificação do paciente e do exame, e também foi feita uma segmentação da região da mama. De cada imagem mamográfica e de sub-regiões dessas imagens, foram extraídos os seguintes atributos:

- forma: a medida de solidez do contorno da região da mama binarizada;
- textura: 14 medidas de Haralick [1973] representando a variação dos níveis de cinza e pares de níveis de cinza;
- momentos: sete momentos de Hu invariantes à posição, escala e rotação;
- histograma: as medidas média, variância, assimetria, curtose e entropia da distribuição dos níveis de cinza das imagens;
- Radon: as medidas média, variância, assimetria, curtose e entropia das funções da transformada Radon;
- granulometria: as medidas média, variância, assimetria, curtose e entropia do histograma de distribuição dos tamanhos dos objetos presentes.

Esses atributos foram utilizados individualmente ou em conjunto para comporem o vetor de características representante da imagem mamográfica, e a técnica PCA foi aplicada para a redução de dimensionalidade desse vetor de características.

Os atributos do histograma de níveis de cinza foram um dos que proveram um alto nível de precisão na avaliação do sistema CBIR, e segundo os autores, a inclusão de outros atributos, como por exemplo a forma, pode melhorar a performance do sistema, ainda que seja necessário realizar um teste de como melhor selecionar e combinar os diversos atributos propostos.

Oliver *et al.* [2009] propõem uma técnica estatística para realizar uma segmentação do tecido da mama, que foi dividido em somente duas classes, gorduroso e denso. O tecido da mama foi caracterizado utilizando-se duas estratégias, PCA e a análise discriminante linear (LDA - *Linear Discriminant Analysis*), em que cada *pixel* de uma nova imagem mamográfica foi classificado em gorduroso ou denso, levando-se também

[1]http://www.hcrp.fmrp.usp.br/gpxsites/hgxpp001

em conta sua vizinhança. Da base de dados Trueta[2], foram extraídas 54 ROIs de tamanho 50 × 50 *pixels* e da base de dados MIAS (*The Mammographic Image Analysis Society Digital Mammogram Database*[3]) [Suckling, 1994], foram usadas todas as 322 imagens disponíveis. Apesar dos autores esperarem que a estratégia LDA fornecesse os melhores resultados, PCA foi que se sobressaiu, mesmo que a diferença entre as duas estratégias não tenha sido estatisticamente relevante. PCA forneceu uma classificação mais compacta com aproximadamente 90% de acurácia.

A proposta de dois softwares para estimar o tecido da mama, um semi-automático e outro automático, foi feita por Tagliafico *et al.* [2009]. Foram utilizadas 160 imagens mamográficas da Universidade de Gênova[4] e todas as imagens tiveram os tecidos da mama avaliados de acordo com as quatro categorias BI-RADS por dois radiologistas experientes.

Para o software semi-automático, dois radiologistas criaram um valor limite e os valores dos *pixels* das imagens mamográficas acima desse limite correspondiam ao tecido denso, e os valores dos *pixels* abaixo do limite correspondiam ao tecido gorduroso. Para o software automático, o histograma de níveis de cinza das imagens e seu histograma cumulativo foram obtidos para todas as imagens.

Os resultados apontam que a estimativa do tecido da mama pelo software automático é possível, elimina a subjetividade e pode ser utilizada não somente para propósitos de pesquisa mas também na prática do dia a dia pelos radiologistas.

Oliver *et al.* [2008] propõem um sistema CAD para a classificação dos tecidos da mama, utilizando duas bases de dados de imagens mamográficas, a MIAS e a DDSM. Da primeira, foram utilizadas todas as 322 imagens mamográficas, que tiveram sua densidade categorizada por três radiologistas, utilizando o padrão internacional BI-RADS. Da base de dados DDSM, foram utilizadas 833 imagens mamográficas que já possuiam sua densidade anotada.

Para a caracterização dos tecidos da mama, os autores utilizaram atributos de textura e morfológicos:

- textura: da matriz de co-ocorrência, foram extraídas as medidas contraste, energia, entropia, correlação, soma da média, soma da entropia, diferença da média, diferença da entropia e homogeneidade;

- morfológicos: área relativa e quatro primeiros histogramas de momento (intensidade média, desvio padrão, assimetria e curtose).

[2]http://eia.udg.es/ aoliver/publications/tesi/node137.html
[3]http://peipa.essex.ac.uk/ipa/pix/mias
[4]http://www.unige.it/

Não foi apresentado o resultado do melhor atributo para a caracterização dos tecidos da mama.

Utilizando as características extraídas das imagens mamográficas, Castella *et al.* [2007] desenvolveram um método semi-automático para estimar a categoria BI-RADS do tecido da mama. Foram utilizadas 352 imagens mamográficas da *Clinique des Grangettes*[5] em Genebra, Suíça. De cada imagem mamográfica, extraíram-se manualmente quatro ROIs, de tamanho 256 x 256 *pixels*, com auxílio do radiologista, de forma a evitar a seleção de artefatos como, por exemplo, o músculo peitoral.

As seguintes características foram extraídas de cada uma das ROIs:

- histograma: desvio padrão, balanço, assimetria e curtose;
- textura: da matriz de co-ocorrência, as medidas energia, entropia, contraste e homogeneidade;
- primitivas: ênfase em primitivas curtas, ênfase em primitivas longas, uniformidade dos níveis de cinza e uniformidade do tamanho das primitivas;
- análise fractal: dimensão fractal;
- matriz da diferença nas proximidades dos tons de cinza (NGTDM – *Neighborhood Graytone Difference Matrix*): contraste, complexidade, força e aspereza.

O discriminante linear de *Fischer* foi utilizado para reduzir a dimensionalidade do vetor de características e para identificar a melhor combinação dos atributos para a caracterização. Os dois atributos que melhor caracterizaram o tecido da mama foram a textura com a medida homogeneidade e a matriz da diferença nas proximidades dos tons de cinza com a medida aspereza.

Sheshadri *et al.* [2006] extraíram de 60 imagens mamográficas da base de dados MIAS, para caracterizar o tecido da mama de acordo com o padrão BI-RADS, os seguintes descritores de textura baseados no histograma de intensidade: média, desvio padrão, suavidade, terceiro momento, uniformidade e entropia.

Por sua vez, Wang *et al.* [2003] utilizaram 195 imagens mamográficas do Centro Médico de Pittsburgh[6] para avaliar automaticamente a densidade da mama de acordo com o padrão BI-RADS. De forma a caracterizar as imagens, do histograma de níveis de cinza foram obtidos o menor valor de intensidade da imagem, a razão entre o menor e o maior valor da intensidade e a razão da distância entre os valores inicial e de pico

[5]http://www.grangettes.ch
[6]http://www.upmc.com/Services/Radiology/Pages/default.aspx

para o total da gama de distância. Segundo os autores, o histograma de níveis de cinza representa com mais precisão o tecido da mama.

Através da classificação dos tecidos da mama de acordo com o padrão BI-RADS, Bovis *et al.* [2002] propuseram aumentar a sensibilidade na detecção do câncer de mama. Em 377 imagens mamográficas da base de dados DDSM, primeiramente foi aplicado um processo de segmentação, de forma a obter-se somente a mama, sem o fundo preto da imagem. Para a caracterização das imagens, foram extraídos atributos de textura:

- da matriz de co-ocorrência, as medidas segundo momento angular, contraste, correlação, diferença do momento inverso, soma da média, soma da variância, soma da entropia, entropia, diferença da média, diferença da variância, diferença da entropia, informação da medida de correlação I, informação da medida de correlação II, inércia e variância;
- da transformada de Fourier, a energia espectral total;
- da máscara de textura de Law, a energia total;
- da transformada *wavelet*, quatro características (desvio padrão, média, assimetria e curtose) dos coeficientes *wavelet* para três níveis de decomposição.

Ainda, uma série de atributos estatísticos foram extraídos dos valores de cinza: entropia, média, desvio padrão, assimetria e curtose. Também, um atributo de forma circular e a dimensão fractal usando coeficiente de Hurst descrito por Russ [1990] foram extraídos. Para reduzir a dimensionalidade do vetor de características foi aplicada a técnica PCA, selecionando-se os 30 primeiros autovalores.

2.2 Caracterização das Lesões da Mama

Verma *et al.* [2008] propuseram um novo algoritmo para a classificação de lesões mamográficas, especificamente massas, em imagens mamográficas. Da base de dados DDSM, ROIs de 100 imagens de casos benignos e 100 imagens de casos malignos foram caracterizadas através de atributos do histograma de níveis de cinza: histograma médio, contraste, energia, entropia, desvio padrão e assimetria.

Eltonsy *et al.* [2007] apresentaram uma técnica para a detecção automática da lesão mamográfica massa utilizando também a base de dados DDSM. Características morfológicas foram extraídas de ROIs de 540 imagens contendo massas malignas, 270 imagens contendo massas benignas e 164 imagens mamográficas normais.

A transformada *wavelet* foi utilizada por Hamad *et al.* [2006] para caracterizar 10 imagens mamográficas contendo a lesão microcalcificação e em 10 imagens normais da base de dados DDSM com o objetivo de verificar qual transformada *wavelet* - *Daubechies*, *Coiflet*, *Symlet* e biortogonal - melhor detecta esse tipo de lesão. Os autores apontam que a detecção pode ser melhorada com o uso de transformadas *wavelets* que possuem funções similares ao formato das microcalcificações.

Com o mesmo objetivo de detectar a lesão microcalcificação em imagens mamográficas, Nakayama *et al.* [2006] utilizaram 610 imagens mamográficas da base de dados DDSM, sendo que em 239 imagens a lesão era maligna e em 371 imagens a lesão era benigna. ROIs de tamanho 115 $\times$ 115 *pixels* contendo a lesão foram extraídas das imagens e então foram caracterizadas através da aplicação da matriz de Hessian [Neudecker and Magnus, 1988].

A Tabela 2.1 resume os trabalhos apresentados baseados na caracterização do tecido da mama e das lesões mamográficas, em que histograma refere-se ao histograma dos níveis de cinza das imagens e morfológicas refere-se às operações morfológicas realizadas nas imagens.

Tabela 2.1. Resumo dos trabalhos baseados no tipo de tecido ou lesão da mama com sistema desenvolvido, número de imagens mamográficas e características utilizadas.

Autor e Ano	**Sistema**	**Objetivo**	**N° de imagens**	**Características**
[Kinoshita et al., 2007]	CBIR	tecido	1.080	forma, momentos, textura, histograma, granulometria, Radon
[Oliver et al., 2009]	CAD	tecido	376	PCA, LDA
[Tagliafico et al., 2009]	CAD	tecido	160	histograma
[Oliver et al., 2008]	CAD	tecido	1.155	textura, morfológicas
[Castella et al., 2007]	CAD	tecido	352	histograma, textura, primitivas análise fractal, NGTDM
[H.S.Sheshadri, 2006]	CAD	tecido	60	histograma
[Wang et al., 2003]	CAD	tecido	195	histograma
[Bovis and Singh, 2002]	CAD	tecido	377	textura, forma, análise fractal
[Verma, 2008]	CAD	lesão	200	histograma
[Eltonsy et al., 2007]	CAD	lesão	974	morfológicas
[Hamad and Taouil, 2006]	CAD	lesão	20	transformada *wavelet*
[Nakayama et al., 2006]	CAD	lesão	610	matriz de Hessian

2.3 Processo de Recuperação das Imagens Mamográficas

A recuperação de imagens tem o propósito de recuperar, de uma base de dados, imagens que sejam relevantes a uma consulta. Kinoshita *et al.* [2007], após caracterizarem os tipos de tecido da mama, utilizaram o mapa neural de *Kohonen* [Kohonen, 1990] para o processo de recuperação. Essa técnica é uma rede não supervisionada, ou seja, não são necessários exemplos rotulados para treinar o classificador. O método *leave-one-out* [Dudda et al., 2001] foi utilizado para criar os casos de treinamento e classificação. Medidas de precisão e revocação foram usadas para avaliar o sistema CBIR, com a aplicação da regra da Tabela 2.2 para a comparação da categoria BI-RADS entre a imagem mamográfica de consulta e as imagens mamográficas recuperadas, já que, segundo os autores, a avaliação dos diferentes tipos de densidades da mama pode variar consideravelmente.

Tabela 2.2. Regras usadas para a verificação da precisão da recuperação [Kinoshita et al., 2007].

Categoria BI-RADS da imagem de consulta	**Categorias BI-RADS aceitáveis das imagens recuperadas**
I	I e II
II	I, II e III
III	II, III e IV
IV	III e IV

Na obtenção dos resultados, foram considerados os valores da precisão para 25% e 50% de revocação. Um dos melhores resultados foi obtido utilizando o atributo histograma para o processo de caracterização das imagens, e a precisão obtida foi de 82,77% e 80,45% para 25% e 50% de revocação, respectivamente.

No sistema CBIR de imagens mamográficas proposto por Lamard *et al.* [2007], para medir a distância entre dois vetores de características, foi empregada uma medida de divergência, a distância de *Kullback-Leibler* [Kullback and Leibler, 1951]. Uma medida de divergência mede a diferença entre duas distribuições de probabilidade. Para avaliar o sistema proposto, foi calculada a precisão média, que foi de 69,4%. Também, como resultado, foram apresentadas cinco imagens mais similares à da consulta. Segundo os autores, este é o melhor número de imagens que os médicos podem dar um diagnóstico mais preciso, pois muitas imagens tornam a interpretação mais difícil e poucas imagens não representam bem a patologia.

Wei *et al.* [2007] apresentam uma classificação dirigida pela recuperação de imagens mamográficas utilizando a lesão microcalcificação como padrão de busca, ou seja, como casos similares recuperados podem ser usados como referência para melhorar o desempenho de um classificador numérico. Para o sistema de recuperação, foi utilizado o método de similaridade proposto em [Wei et al., 2006]. As imagens mamográficas são pontuadas através de um estudo de observação humana, para o treinamento da função de similaridade. No processo de classificação, usou-se o classificador SVM adaptado pelos autores. Nesse esquema, os autores ajustam a função do classificador SVM de acordo com sua atuação nos casos similares ao caso de consulta, ou seja, o limiar de decisão é ajustado de acordo com o caso a ser recuperado e classificado. Durante a classificação, foi adotado o procedimento *leave-one-out*.

Por ser um sistema de classificação, o método foi avaliado através de curvas ROC (*Receiver Operator Characteristic Curve*) [Fawcett, 2004]. Os resultados mostram uma melhora da acurácia utilizando o classificador SVM de 77,8% para 82,2% usando o método proposto de SVM adaptado.

A lesão massa foi usada como padrão de busca no sistema CBIR de imagens mamográficas proposto por Felipe *et al.* [2006]. Após a caracterização das imagens, foram aplicadas na base de dados consultas com o algoritmo k-vizinhos mais próximos (k-NN – *k-nearest neighbor*) [Russ, 2007]. O sistema foi avaliado através de curvas de precisão e revocação e o melhor resultado obtido foi o de 80% de precisão para 20% de revocação, utilizando 16 momentos de *Zernike* para caracterizar as imagens.

El-Naqa *et al.* [2002] consideram o cálculo de similaridade entre as imagens um grande desafio para a elaboração de sistemas CBIR. O diferencial do trabalho consiste em usar dois estágios para classificar as imagens mamográficas ao invés de um só. As imagens mamográficas contendo a lesão microcalcificação foram caracterizadas e então os vetores de características obtidos foram utilizados nos estágios de classificação e recuperação. O primeiro estágio de classificação serviu para desconsiderar as imagens mamográficas que não são similares à da consulta, através dos classificadores discriminante linear de Fisher [Gonzalez et al., 2003] e o SVM. Para o classificador SVM, foram usados um *kernel* linear, um *kernel* radial e uma junção dos dois. O segundo estágio, que funciona como um tipo de ajuste fino, seleciona somente as imagens cujos coeficientes de similaridade estão mais próximos do coeficiente da imagem de consulta. Nesse segundo estágio, foram utilizados os classificadores rede neural (GRNN – *General Regression Neural Network*) e o SVM.

Os resultados foram obtidos através de curvas de precisão e revocação, e foram comparados com testes que utilizaram somente um estágio de classificação. O método proposto, utilizando o classificador SVM e o *kernel* radial, foi o que obteve melhores

resultados com uma precisão média de 93,7%.

2.4 Conclusão

Neste capítulo, foram apresentados trabalhos relacionados à caracterização do tecido da mama e à caracterização das lesões mamográficas, uma vez que a caracterização é um passo crucial para o processo de recuperação de imagens. Os principais atributos utilizados para a caracterização, e que obtiveram melhores resultados no desenvolvimento de sistemas CBIR de imagens mamográficas, foram o histograma de níveis de cinza e textura [Kinoshita et al., 2007, Wang et al., 2003]. Combinações desses atributos também foram obtidas, sendo então necessária a aplicação de técnicas de redução de dimensionalidade dos vetores de características. A técnica mais utilizada foi o PCA. Além disso, foram apresentados trabalhos que utilizam redes neurais e SVM [Kinoshita et al., 2007, Wei et al., 2007, Naqa et al., 2002] para o processo de recuperação das imagens.

Em sistemas CBIR as imagens são descritas por vetores de características que as representam sob determinado aspecto, e a similaridade entre as imagens é obtida através de medidas de distância ou classificadores, que indicarão a relevância das imagens para determinada consulta. A escolha de um conjunto de atributos, que capturem a essência das imagens, por meio de uma quantidade sucinta de valores e de uma medida de similaridade capaz de prover imagens mais similares à uma imagem de consulta da forma mais próxima a da percepção humana, é um desafio.

Capítulo 3

Caracterização dos Tecidos e Lesões da Mama

Em sistemas de recuperação de imagens, o acesso à informação é executado através dos atributos visuais extraídos das imagens, juntamente com modelos de similaridade, e se necessário, mecanismos de indexação apropriados.

A definição de um conjunto de características capaz de descrever de maneira efetiva cada região contida em uma imagem é uma das tarefas mais complexas na análise de imagens e, além disso, esse processo de caracterização afeta todos os processos subseqüentes do sistema CBIR [Baeza-Yates and Neto, 1999, Pedrini and Schwartz, 2008]. A análise de imagens é, tipicamente, baseada na forma, na textura, nos níveis de cinza ou nas cores presentes nas imagens.

Uma imagem digital pode ser representada por meio de uma matriz bidimensional, na qual cada elemento da matriz corresponde a um *pixel* da imagem. Os valores dos *pixels* são formados por números inteiros correspondendo aos níveis de cinza. Um outro tipo de representação é feita através do vetor de características, que deve ressaltar aspectos da imagem para facilitar a percepção humana, ser invariante às transformações da imagem e reduzir a dimensionalidade dessa imagem [Castelli and Bergman, 2001, Dudda et al., 2001].

Visualmente, mamas com tipos de tecidos denso e gorduroso diferenciam-se através da intensidade dos níveis de cinza nas imagens mamográficas, como pode ser visto na Figura 3.1. Considerando-se em uma mesma imagem o tipo de tecido e tipo de classificação da lesão, a Figura 3.2 apresenta os tipos de tecidos da mama em suas quatro categorias BI-RADS e as lesões da mama nas categorias BI-RADS 2 (benigno) e BI-RADS 5 (maligno). Como pela análise visual essas imagens podem ser diferenciadas pela intensidade dos níveis de cinza, a representação do tecido e lesão da mama

pode ser realizada através do histograma ou do atributo textura. O histograma de uma imagem corresponde à distribuição dos níveis de cinza da imagem e o contraste de uma imagem pode ser avaliado observando-se o seu histograma. Várias medidas estatísticas podem ser obtidas a partir do histograma de uma imagem, tais como os valores mínimo e máximo, o valor médio, a variância e o desvio padrão dos níveis de cinza da imagem. Contudo, visto que a textura encontra-se entre as características empregadas pelo sistema visual humano e contém informações sobre a distribuição espacial e a variação de luminosidade [Pedrini and Schwartz, 2008, Gonzalez et al., 2003], sua utilização para a representação dos tecidos e lesões da mama também torna-se apropriada.

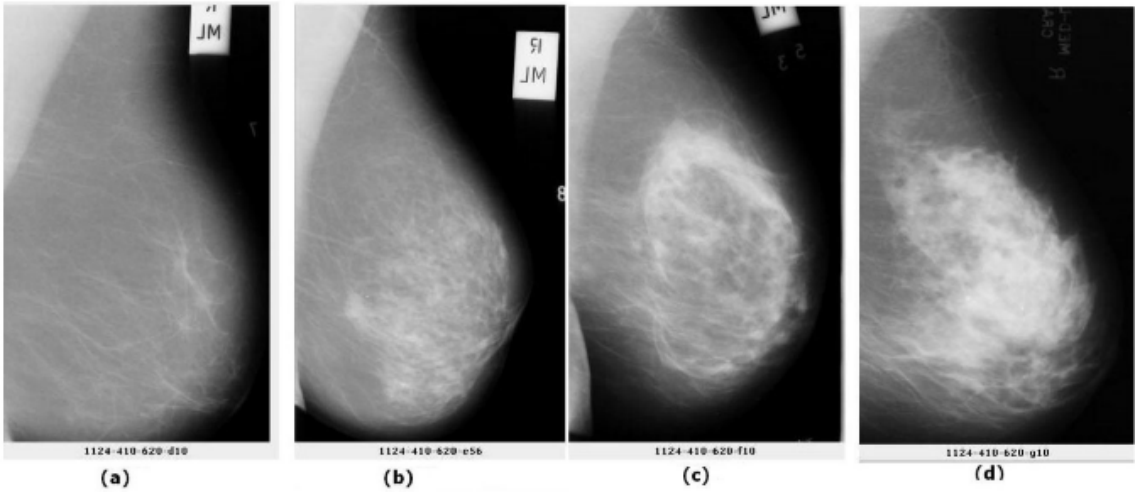

Figura 3.1. Imagens mamográficas de diferentes tipos de tecido: (a) Extremamente gordurosa, (b) Gordurosa com algum tecido fibroglandular, (c) Heterogeneamente densa, (d) Extremamente densa.

Os modelos de textura desenvolvidos na literatura podem ser divididos nas seguintes abordagens [Pedrini and Schwartz, 2008, Gonzalez et al., 2003, Acharya and Ray, 2005]:

- Abordagem estatística: a textura é representada indiretamente por propriedades que definem distribuições e relacionamentos entre os níveis de cinza dos *pixels* pertencentes a uma imagem, como por exemplo, através do cálculo da matriz de co-ocorrência e da função de auto-correlação.

- Abordagem geométrica: a textura é definida como sendo composta por "elementos de textura" ou primitivas e as características são extraídas através de medidas ou da posição espacial e relacionamento entre as primitivas. Por exemplo, através da unidade de textura e de sua simetria geométrica ou do grau de distribuição.

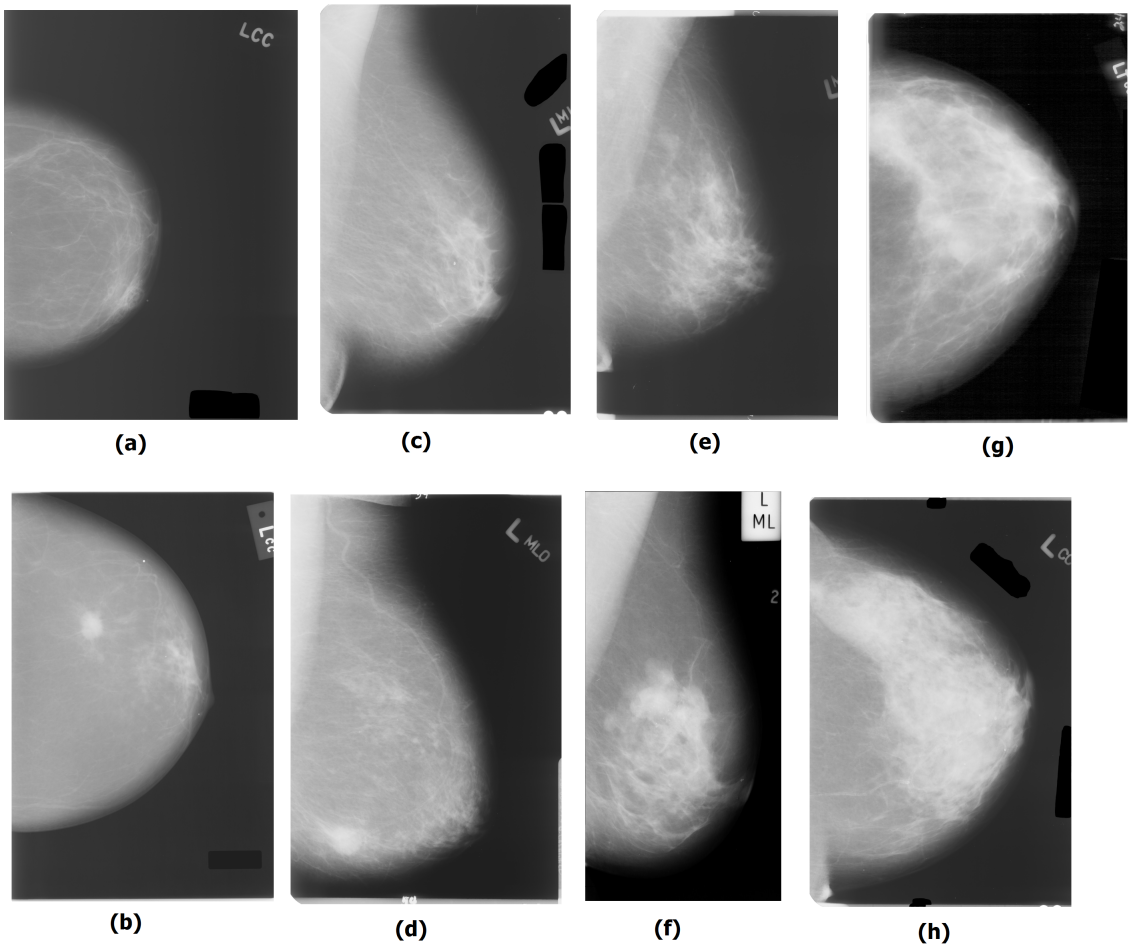

Figura 3.2. Imagens mamográficas de diferentes tipos de tecido e lesão: (a) Extremamente gordurosa com lesão benigna, (b) Extremamente gordurosa com lesão maligna, (c) Gordurosa com algum tecido fibroglandular com lesão benigna, (d) Gordurosa com algum tecido fibroglandular com lesão maligna, (e) Heterogeneamente densa com lesão benigna, (f) Heterogeneamente densa com lesão maligna, (g) Extremamente densa com lesão benigna e (h) Extremamente densa com lesão maligna.

- Abordagem baseada em modelos paramétricos: a textura é considerada uma amostra extraída de um processo definido por um conjunto de parâmetros que, servindo como modelo para textura, resumem suas características, como por exemplo, através dos campos aleatórios de Markov.

- Abordagem baseada em processamento de sinais: os descritores da textura são extraídos a partir da representação obtida após a aplicação de transformações na imagem de entrada. Exemplos dessa abordagem são a decomposição em valores singulares, espectro de Fourier e transformada *wavelet*.

Em razão de uma das dificuldades no uso do atributo de textura residir na alta dimensionalidade do vetor de características, torna-se necessária a redução da dimensão do espaço dos vetores de características, para tornar o algoritmo de recuperação de imagens computacionalmente tratável. Algumas das técnicas mais utilizadas para a redução de dimensionalidade são o PCA, análise fatorial e rede de Kohonen.

Neste capítulo, a análise dos componentes principais em duas dimensões (2DPCA) é introduzida para representar a textura do tecido da mama juntamente com as lesões mamográficas e reduzir a dimensão do vetor de características. Também, a decomposição em valores singulares é apresentada para a caracterização do tecido da mama e lesões mamográficas.

3.1 Análise dos Componentes Principais em Duas Dimensões

A técnica 2DPCA foi proposta por Yang *et al.* [2004] com a finalidade de representação de imagens, baseando-se nas imagens bidimensionais ao invés de nos vetores unidimensionais, como na técnica PCA [Dudda et al., 2001].

Alguns trabalhos empregaram a técnica 2DPCA para a representação de imagens de face e da palma da mão (*palmprint*). Por exemplo, Zuo *et al.* [2006] propuseram uma métrica (AMD - *Assembled Matrix Distance*) para medir a distância entre duas matrizes de características obtidas através da técnica 2DPCA. Primeiramente, eles utilizaram a base de dados de faces ORL (*Our Database of Faces*) (1992) para avaliar a técnica proposta em 400 imagens de tamanho 112 $\times$ 92 *pixels*. Apenas os primeiros quatro autovalores da matriz projetada 2DPCA foram escolhidos. Comparando com outros métodos de reconhecimento de imagens como o *Eigenfaces*, *Fisherfaces* e análise discriminante linear direta (D-LDA - *Direct Linear Discriminant Analysis*), a taxa de reconhecimento usando 2DPCA e AMD foi a maior obtida – 96,30%. Em segundo lugar, utilizando a base de dados de palma de mão PolyU (2004), eles utilizaram 600 sub-imagens de tamanho 128 $\times$ 128 *pixels* para testar a eficiência do método proposto. Os autores, nesse caso, escolheram manter os primeiros 18 autovalores, após fazerem testes com autovalores variando de um a 25. A comparação do método proposto foi feita com os três métodos de reconhecimento de imagens já citados, e novamente 2DPCA junto com o AMD obtiveram a maior taxa de reconhecimento – 97,67%.

Zhao *et al.* [2007], também, com o mesmo objetivo de reconhecimento de imagens, aplicaram a técnica 2DPCA na extração de características da palma da mão, removendo a informação de iluminação através da técnica w/o3 [Belhumeur et al., 1997]. Nessa

técnica, de maneira a não considerar a perturbação de diferentes condições de iluminação para as coleções de palma da mão e para direcionar a melhores resultados de reconhecimento, os primeiros três autovalores são removidos, já que representam essa informação de iluminação. Da base de dados PolyU foram escolhidas 600 imagens e de cada uma delas foi extraída a parte central da palma da mão. A técnica 2DPCA foi aplicada, os primeiros três autovalores foram descartados e a técnica PCA foi usada após a técnica 2DPCA para reduzir a dimensionalidade, em um processo chamado pelos autores de 2DPCA(w/o3)PCA. Sua performance foi comparada com outras técnicas de extração de características como o filtro 2DGabor, PCA, PCA(w/o3) e LDA. A técnica 2DPCA(w/o3)PCA consumiu menos tempo para a extração de características e também obteve a maior taxa de acurácia – 99,27% – utilizando um classificador proposto pelos autores, um classificado modular de rede neural modificado (MNN - *Modular Neural Network*).

PCA é uma técnica clássica de extração de atributos e representação de dados muito utilizada em áreas de reconhecimento de padrões e visão computacional e que recorre ao método de descorrelação de dados. Ela pode ser usada para a redução de dimensionalidade, retendo as características do conjunto de dados que contém os aspectos mais importante para sua identificação, pois ao se descorrelacionar os dados elimina-se parte da informação redundante em cada dimensão [Acharya and Ray, 2005].

O objetivo da técnica PCA é encontrar uma transformação mais representativa e compacta dos dados. Esse método transforma um vetor aleatório $\mathbf{x} \in \mathbb{R}^m$ em outro vetor $\mathbf{y} \in \mathbb{R}^n$ (para $n \leq m$), projetando $\mathbf{x}$ nas n direções ortogonais de maior variância – os componentes principais. Esses componentes são individualmente responsáveis pela variância dos dados e essa variação pode ser explicada por um número reduzido de componentes, sendo possível descartar os restantes sem grande perda de informação. A estimação dos componentes principais é relativamente simples, bastando utilizar a informação contida da matriz de covariância dos dados, pois a covariância é uma medida que descreve a variabilidade dos componentes das diferentes dimensões em relação com os restantes.

Seja Σ_x a matriz de covariância de um vetor aleatório e real $\mathbf{x}$. Se Σ_x for uma matriz não-singular[1], então Σ_x pode ser decomposta no seguinte produto matricial:

$$\Sigma_x = \Gamma_x \Lambda_x \Gamma_x^T$$

para

[1]Matriz não-singular é uma matriz invertível (Se a matriz $\mathbf{A}$ for não-singular, então o determinante $|\mathbf{A}| \neq 0$).

$$\Gamma_x = \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1n} \\ \vdots & \ddots & \vdots \\ \gamma_{n1} & \cdots & \gamma_{nn} \end{pmatrix} = [\gamma_1, \cdots, \gamma_n]$$

e

$$\Lambda_x = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & 0 & \vdots \\ \vdots & 0 & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_n \end{pmatrix}$$

em que $\lambda_1, \cdots, \lambda_n$ são os autovalores de Σ_x e estão em ordem descendente ($\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$). Já $\gamma_1, \cdots, \gamma_n$ são os autovetores correspondentes e cada coluna de Γ_x é um autovetor. Os autovetores de uma matriz são perpendiculares entre si, isto é, formam bases ortogonais.

Por sua vez, a idéia da técnica 2DPCA [Yang et al., 2004] é obter a matriz de covariância diretamente das matrizes bidimensionais das imagens ao invés de transformá-las em um vetor unidimensional, como na técnica PCA. A técnica 2DPCA, por ser baseada na matriz da imagem, é mais simples e é computacionalmente mais efetiva, ou seja, seu desempenho é melhor que a técnica PCA, e também pode melhorar significantemente a velocidade de extração de características [Yang et al., 2004].

Seja $\mathbf{X}$ um vetor coluna unitário t dimensional. A idéia da técnica 2DPCA é projetar uma imagem $\mathbf{A}$, que é uma matriz de tamanho $r \times s$, em $\mathbf{X}$ através da transformação linear:

$$\mathbf{Y} = \mathbf{AX} \tag{3.1}$$

Obtém-se um vetor projetado $\mathbf{Y}$, r dimensional, que é chamado vetor característica projetado da imagem $\mathbf{A}$.

Primeiramente, esse vetor $\mathbf{X}$ pode ser obtido através de um critério, o critério geral da dispersão total, que é caracterizado pelo traço da matriz de covariância:

$$J(\mathbf{X}) = tr(\mathbf{S}_x) \tag{3.2}$$

em que $\mathbf{S}_x$ denota a matriz de covariância dos vetores características projetados dos exemplos de treinamento e $tr(\mathbf{S}_x)$ denota o traço de $\mathbf{S}_x$. O significado de maximizar o critério em 3.2 é encontrar a direção da projeção $\mathbf{X}$, sobre a qual todos os exemplos são projetados. A matriz de covariância $\mathbf{S}_x$ pode ser denotada por:

$$\mathbf{S}_x = E[(\mathbf{Y} - E\mathbf{Y})(\mathbf{Y} - E\mathbf{Y})^T] =$$
$$E[\mathbf{AX} - E(\mathbf{AX})][\mathbf{AX} - E(\mathbf{AX})]^T =$$
$$E[(\mathbf{A} - E\mathbf{A})\mathbf{X}][(\mathbf{A} - E\mathbf{A})\mathbf{X}]^T$$

Então:

$$tr(\mathbf{S}_x) = \mathbf{X}^T[E(\mathbf{A} - E\mathbf{A})^T(\mathbf{A} - E\mathbf{A})]\mathbf{X} \tag{3.3}$$

A matriz $\mathbf{G}$, que é chamada de matriz imagem de covariância, pode então ser definida por:

$$\mathbf{G} = E[(\mathbf{A} - E\mathbf{A})^T(\mathbf{A} - E\mathbf{A})] \tag{3.4}$$

Essa matriz $\mathbf{G}$ é avaliada usando as imagens exemplo de treinamento. Considerando, por exemplo, a caracterização do tecido da mama para as quatro categorias BI-RADS. Existem n_i exemplos de treinamento para cada categoria. As imagens $\mathbf{A}_1, \mathbf{A}_2, \cdots, \mathbf{A}_Z$ ($Z = \sum_{i=1}^{Z} n_i$) denotam todos os exemplos de treinamento, onde $\mathbf{A}_i$ é uma matriz de tamanho $r \times s$.

A matriz de covariância $\mathbf{G}$ é a seguinte:

$$\mathbf{G} = \frac{1}{Z}\sum_{i=1}^{Z}(\mathbf{A}_i - \bar{\mathbf{A}})(\mathbf{A}_i - \bar{\mathbf{A}})^T \tag{3.5}$$

em que $\bar{\mathbf{A}} = \frac{1}{Z}\sum_{i=1}^{Z}\mathbf{A}_i$ é a matriz média de todos os exemplos de treinamento.

Alternadamente, o critério em 3.2 pode ser expresso por:

$$J(\mathbf{X}) = \mathbf{X}^T\mathbf{G}\mathbf{X} \tag{3.6}$$

O vetor $\mathbf{X}$ que maximiza esse critério é chamado de vetor de projeção ótima, e esses vetores $\mathbf{X}_{otm}$ são os vetores que maximizam $J(\mathbf{X})$, isto é, o autovetor de $\mathbf{G}$ correspondente ao maior autovalor. Os vetores de projeção ótima da técnica 2DPCA, $\mathbf{X}_1, \cdots, \mathbf{X}_d$, são os autovetores ortonormais de $\mathbf{G}$ que correspondem aos primeiros d maiores autovalores.

Esses vetores de projeção ótima são então usados para a extração de características. Para uma dada imagem $\mathbf{A}$:

$$\mathbf{Y}_t = \mathbf{AX}_t, \qquad t = 1, 2, \cdots, d \tag{3.7}$$

Uma família de características projetadas $\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_t$ é obtida, e é chamada de componentes principais (vetores) da imagem $\mathbf{A}$. Ao contrário da técnica PCA, em que o componente principal é um escalar, com a técnica 2DPCA cada componente principal é um vetor. Os vetores de componentes principais obtidos são usados para formar uma matriz $\mathbf{L} = [\mathbf{Y}_1, \mathbf{Y}_2, \cdots, \mathbf{Y}_t]$ de tamanho $r \times d$, que é chamada de matriz de características ou matriz imagem da imagem $\mathbf{A}$.

3.2 Decomposição em Valores Singulares

A decomposição em valores singulares (SVD) é uma das ferramentas mais úteis da álgebra linear e teve sua praticidade e utilidade demonstrada por Gene Golub [1983].

A utilização da técnica SVD na caracterização de imagens e redução de dimensionalidade foi empregada em diversos trabalhos, como por exemplo, por Chen *et al.* [2008]. Neste trabalho, a SVD foi utilizada para reduzir a matriz de características de imagens coloridas do estômago, de forma a ignorar dados redundantes e ruído, em um sistema médico CBIR. Os histogramas de cores de 1.345 imagens do estômago foram obtidos e a SVD foi realizada para compor uma nova matriz de características através da seleção dos k primeiros valores singulares. Essa seleção do valor k a ser utilizado é apontada como um desafio, visto que se esse valor for muito pequeno, pode-se perder informações importantes. A escolha desse valor foi decidida através de diversos experimentos. A recuperação das imagens foi realizada através da medida do cosseno. Foram testados valores de k variando de 5 a 200, e através de curvas de precisão e revocação foi visto que para valores de k acima de 100, não há variação dos resultados, e para valores de k entre 10 e 30 são obtidos os maiores valores de precisão. Os autores, em suas conclusões, apontam sobre a dificuldade da escolha de quantos valores singulares k são ideais para a melhora da performance do sistema CBIR, indicando a necessidade de maiores estudos nesse sentido.

No trabalho de Selvan *et al.* [2007], o principal objetivo da SVD é alcançar maiores taxas de reconhecimento em grandes bases de dados, requerendo menos computação. Os autores apresentaram uma nova proposta para a classificação de texturas em imagens, baseada na transformada *wavelet* e na SVD. Toda coleção de imagens da base de dados de textura Brodatz [Brodatz, 1966] foi usada para a classificação. Aplicou-se a transformada *wavelet* ortogonal com subamostragem em pares, e a função base de Daubechies foi a escolhida para realizar a decomposição e obtenção dos coeficientes. De forma a tornar os coeficientes da transformada *wavelet* menos sensíveis a variações locais foram aplicadas as operações de não linearidade (magnitude, alinha-

mento) e filtragem (filtro passa baixa retangular), e em seguida obteve-se a energia de cada coeficiente. Como o número de coeficientes da transformada *wavelet* obtido é grande, a demanda computacional para a estimação dos parâmetros é muito alta. Devido a isso, a SVD pode ser aplicada nesses coeficientes, visto que os valores singulares resultantes serão de número menor que os coeficientes. Os autores afirmam que a distribuição dos valores singulares dos coeficientes da transformada *wavelet* variam bastante de textura para textura. Isto implica que a distribuição dos valores singulares possui boas características de discriminação. A eliminação dos menores valores singulares também resulta em redução do ruído, facilitando a classificação efetiva de texturas na presença de ruído. A seleção do número k de valores singulares que são utilizados é muito importante para a obtenção de boas taxas de reconhecimento, porém é muito difícil a determinação exata desse valor, que é portanto obtido empiricamente. Os autores obtiveram o valor médio de k igual a 27. A distância de Kullback-Leibler foi utilizada como classificador, e os resultados do método proposto alcançam uma taxa de reconhecimento de 98,34%.

Wang *et al.* [2002], em seu trabalho, propuseram um método de classificação baseado em redes neurais, nas quais as características utilizadas são os valores singulares das imagens de faces. Para reconhecimento de imagens de faces, as propriedades técnicas relevantes da SVD são sua boa estabilidade (quando é inserida uma pequena perturbação na imagem da face não ocorre uma grande variação nos valores singulares) e a representação de propriedades algébricas e invariância de uma imagem pelos valores singulares. Segundo os autores, os valores singulares representam atributos importantes de uma matriz. Como as imagens podem ser observadas como matrizes, os valores singulares podem servir como características das imagens na avaliação de similaridade entre imagens. De dez diferentes imagens de 40 pessoas distintas foram extraídos os valores singulares, sem a indicação dos autores da quantidade k de valores singulares selecionados. A classificação foi realizada através da rede neural e comparada com a distância Euclidiana. Os resultados mostram uma taxa de reconhecimento de aproximadamente 80,9% utilizando a distância Euclidiana e de 92% utilizando a rede neural, enfatizando a idéia dos autores que essa classificação através de aprendizado supervisionado supera os problemas das técnicas existentes que utilizam a SVD e técnicas de aprendizado não supervisionado.

Diante do exposto, a SVD pode ser vista de três formas mutuamente compatíveis entre si. De um lado, a SVD pode ser visualizada como um método para a transformação de um conjunto de variáveis correlacionadas em um conjunto de variáveis não correlacionadas, ou seja, que expõe melhor as várias relações entre os dados originais. Ao mesmo tempo, a SVD é um método para a identificação e ordenação das dimensões

junto aos dados que apresentam as maiores variações. Isto se relaciona com a terceira forma de se visualizar a SVD, já que uma vez identificada onde está a maior variação dos dados, é possível encontrar a melhor aproximação dos dados originais usando uma menor dimensão. Por conseguinte, a SVD pode ser vista como um método para redução de dados.

A SVD é baseada em um teorema da álgebra linear que diz que uma matriz retangular pode ser separada no produto de outras três matrizes [Andrews and Patterson, 1976, Golub, 1983, Watkins, 1991, Strang, 1993, Élden, 2006, Pedrini and Schwartz, 2008]. Os elementos dessa matriz retangular podem ser compostos pela intensidade dos níveis de cinza dos *pixels* pertencentes a uma dada textura e os valores singulares obtidos como resultado dessa decomposição e suas distribuições provêem informações úteis sobre a textura de determinada imagem.

Considerando as definições básicas de álgebra linear descritas no Apêndice A, pode-se apresentar esse teorema através da seguinte notação: qualquer matriz $\mathbf{A}$ de tamanho $r \times s$, que pode representar uma certa imagem, pode ser fatorizada em:

$$\mathbf{A}_{r\times s} = \mathbf{U}_{r\times r}\mathbf{W}_{r\times s}\mathbf{V}^T_{s\times s} = [\mathbf{u}_1 \cdots \mathbf{u}_r][\sigma_1 \cdots \sigma_s][\mathbf{v}_1 \cdots \mathbf{v}_s]^T$$

A matriz $\mathbf{W}$ é diagonal e todos os σ são positivos e estão em ordem decrescente. As matrizes $\mathbf{U}$ e $\mathbf{V}$ devem ser matrizes ortogonais. Os seus vetores base devem ser ortonormais:

$$\mathbf{V}^T\mathbf{V} = \begin{pmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_i^T \end{pmatrix} \begin{pmatrix} \mathbf{v}_1 \cdots \mathbf{v}_i \end{pmatrix} = \begin{pmatrix} 1 \cdots 0 \\ \ddots \\ 0 \cdots 1 \end{pmatrix}$$

Então: $\mathbf{V}^T\mathbf{V} = \mathbf{I}$ e similarmente $\mathbf{U}^T\mathbf{U} = \mathbf{I}$.

Os valores singulares σ são obtidos extraindo a raiz quadrada dos autovalores diferentes de zero da matriz $\mathbf{A}^T\mathbf{A}$ ou da matriz $\mathbf{A}\mathbf{A}^T$:

$$\mathbf{A}^T\mathbf{A} = (\mathbf{U}\Sigma\mathbf{V}^T)^T(\mathbf{U}\mathbf{W}\mathbf{V}^T) = \mathbf{V}\mathbf{W}^T\mathbf{U}^T\mathbf{U}\mathbf{W}\mathbf{V}^T.$$

Como $\mathbf{U}^T\mathbf{U}$ é igual à matriz $\mathbf{I}$, então $\mathbf{W}^T$ é próximo a $\mathbf{W}$. Multiplicando essas matrizes diagonais obtém-se σ^2 e as colunas da matriz $\mathbf{V}$ são seus autovetores. É importante notar que os valores singulares σ não são os autovalores de $\mathbf{A}$. Na verdade, σ^2 é um autovalor de $\mathbf{A}^T\mathbf{A}$ e $\mathbf{A}\mathbf{A}^T$.

A SVD também pode ser utilizada para encontrar a melhor aproximação $\mathbf{A}_k$ com posto (*rank*) k para uma matriz de entrada $\mathbf{A}_{r\times s}$, ou seja, $\mathbf{A}_k = \mathbf{U}_k\mathbf{W}_k\mathbf{V}_k^T$ e posto $k < min(m, n)$ O armazenamento da aproximação de uma matriz resulta

em economias computacionais significantes sobre o armazenamento da matriz inteira. Com isso, reduz-se a dimensionalidade da matriz de entrada em questão. O grande desafio é encontrar o melhor posto k que irá melhorar a caracterização das imagens [Andrews and Patterson, 1976, Élden, 2006].

3.3 Conclusão

Um grande desafio para sistemas CBIR é a extração e seleção de características das imagens que as descrevam com precisão suficiente para a sua identificação. Para as imagens mamográficas, através de modelos de textura, essa representação precisa pode ser realizada. A escolha de técnicas que sejam capazes de caracterizar a imagem e reduzir a dimensionalidade dos vetores de características que as representem é desejável, pois o objetivo é fornecer ao usuário um sistema CBIR rápido e efetivo.

A técnica 2DPCA, através da seleção dos maiores autovalores e conseqüentemente, dos autovetores, reduz a dimensionalidade do vetor de características, além de melhorar o desempenho computacional do algoritmo de caracterização, visto que trabalha com a matriz das imagens ao invés de um vetor, como na técnica PCA. Já a técnica SVD utiliza os valores singulares para representar a variação de textura das imagens, bem como para reduzir a dimensionalidade do vetor de características ao se escolher os primeiros valores singulares de maior significância.

Capítulo 4

Processo de Recuperação com Base no Conteúdo Visual das Imagens Mamográficas

A recuperação com base no conteúdo visual de imagens tem o propósito de recuperar, de uma base de dados, imagens que são relevantes a uma consulta.

A imagem de consulta passa pelo processo de extração de características, explicado no capítulo anterior, no qual é gerado um vetor de características. Esse vetor de características é então submetido a uma busca por similaridade junto à estrutura que contém os vetores de características de todas as imagens armazenadas na base de imagens. Os identificadores das imagens resultantes da busca são utilizados para recuperar essas imagens da base de imagens, podendo as mesmas, assim, serem apresentadas ao usuário.

A classificação, com o objetivo de definição de similaridade entre as imagens, no processo de recuperação de imagens, serve para agrupar as imagens mais semelhantes e indicar a relevância das imagens para determinada imagem de consulta, de forma a tornar a recuperação mais rápida e efetiva [van Rijsbergen, 1979].

Neste trabalho, será explorado o uso da máquina de vetores de suporte (SVM - *Support Vector Machine*). Os resultados da aplicação dessa técnica são comparáveis e muitas vezes superiores ao obtidos por outros algoritmos [Naqa et al., 2004, Wong and Hsu, 2006], e além disso, exemplos de aplicações de sucesso podem ser encontrados em diversos domínios [Zhang et al., 2001, Yang et al., 2007, Akay, 2009].

4.1 Máquina de Vetores de Suporte

Máquina de vetores de suporte é uma estratégia de aprendizagem introduzida por Vapnik e colaboradores [1995], que foi desenvolvida para resolver problemas de classificação [Campanini et al., 2004, Arodz et al., 2005, Akay, 2009], mas que foi estendida para o uso em sistemas CBIR [Zhang et al., 2001, Rahman et al., 2005, Mumtaz et al., 2006, Wong and Hsu, 2006]. Especificamente para sistemas CBIR de imagens mamográficas, SVM foi usado nos trabalhos de El-Naqa *et al.* [2002, 2004], Wei *et al.* [2006] e Yang *et al.* [2007].

De acordo com a teoria de SVM, enquanto técnicas tradicionais para reconhecimento de padrões são baseadas na minimização do risco empírico – tenta-se otimizar o desempenho sobre o conjunto de treinamento, SVM minimiza o risco estrutural, isto é, a probabilidade de classificar de forma errada padrões ainda não vistos. Isto é chamado de princípio de indução, no qual obtêm-se conclusões genéricas a partir de um conjunto particular de exemplos.

O aprendizado indutivo pode ser dividido em dois tipos principais: supervisionado e não supervisionado [Dudda et al., 2001].

No aprendizado não supervisionado não existem exemplos rotulados. O algoritmo de aprendizado de máquina aprende a agrupar as entradas submetidas segundo uma medida de qualidade.

No aprendizado supervisionado, que é o caso do SVM, um agente externo é usado para indicar as respostas desejadas para os padrões de entrada. O classificador é treinado com um largo conjunto de dados rotulados. Neste caso, dado um conjunto de exemplos rotulados na forma (e_i, y_i), em que e_i representa um exemplo e y_i denota seu rótulo, deve-se produzir um classificador capaz de predizer precisamente o rótulo de novos dados. Esse processo de indução de um classificador a partir de uma amostra de dados é denominado treinamento. O classificador obtido também pode ser visto como uma função f, a qual recebe um dado e e uma predição ou rótulo y.

Os rótulos ou classes representam o fenômeno de interesse sobre o qual se deseja fazer previsões. Os rótulos podem assumir valores discretos $1, \cdots, p$. Um problema de classificação no qual $p = 2$ é denominado binário. Para $p > 2$ configura-se um problema multiclasse.

O treinamento do SVM consiste na resolução de um problema quadrático, que depende dos vetores de treinamento, de alguns parâmetros e da margem de separação. A solução desse problema fornece a informação necessária para se escolher, entre todos os dados de entrada, os vetores mais importantes, conhecidos como vetores de suporte, que definem o hiperplano de separação e são encontrados durante essa fase de

treinamento.

A vantagem do SVM para outros classificadores, como por exemplo, análise discriminante ou rede neural, é a generalização, ou seja, a classificação de novos dados através de uma fronteira mais distante dos dados de treinamento, como é visto na Figura 4.1.

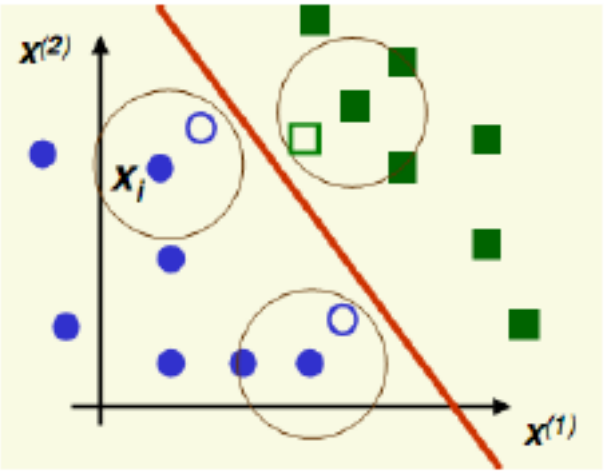

Figura 4.1. Generalização do método SVM. Círculos e quadrados cheios representam os dados de treinamento e círculos e quadrados vazios representam os novos dados a serem classificados.

No processo de classificação, o método SVM projeta os dados a serem classificados em um espaço de grande dimensão, onde certo critério é utilizado para a separação dos dados. Existem diferentes casos utilizados para a obtenção dessa fronteira e que são reportados a seguir.

4.1.1 Caso de separação linear

Em sua forma básica, SVMs são classificadores lineares que separam os dados em duas classes através de um hiperplano de separação.

Um hiperplano ótimo separa os dados com a máxima margem possível, que é definida pela soma das distâncias entre os pontos positivos (na Figura 4.2, representados pelos triângulos) e os pontos negativos (na Figura 4.2, representados pelos círculos) mais próximos do hiperplano. Esses pontos são chamados vetores de suporte (pontos circulados na Figura 4.2) e estão localizados nos planos $H1$ e $H2$. O hiperplano é construído com base em treinamento prévio em um conjunto finito de dados.

Supondo-se que exista um hiperplano que separa os exemplos negativos dos exemplos positivos. A equação de um hiperplano é apresentada na Equação 4.1, em que

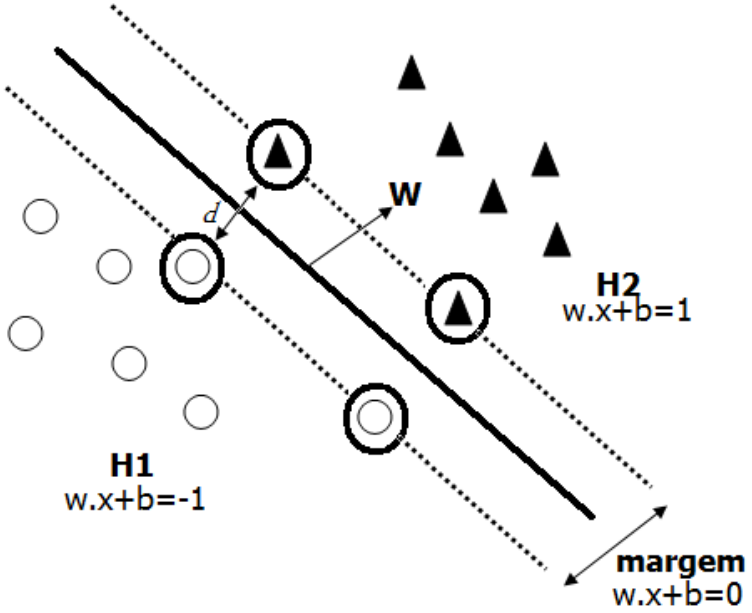

Figura 4.2. Classificação de um conjunto de dados utilizando SVM linear.

$\mathbf{w} \cdot \mathbf{x}$ é o produto escalar entre os vetores $\mathbf{w}$ e $\mathbf{x}$, $\mathbf{w} \in$ X é o vetor normal ao hiperplano e $\frac{|b|}{\|\mathbf{w}\|}$ é a distância perpendicular do hiperplano até a origem, com $b \in \mathbb{R}$.

$$f(x) = \mathbf{w} \cdot \mathbf{x} + b = 0 \tag{4.1}$$

Essa equação divide o espaço dos dados X em duas regiões: $\mathbf{w} \cdot \mathbf{x} + b > 0$ e $\mathbf{w} \cdot \mathbf{x} + b < 0$. Uma função sinal $g(x) = sgn(f(x))$ pode ser então empregada na obtenção das classificações, conforme a Equação 4.2:

$$g(\mathbf{x}) = sgn(f(\mathbf{x})) = \tag{4.2}$$

$$\begin{cases} +1, & \text{se } \mathbf{w} \cdot \mathbf{x} + b > 0; \\ -1, & \text{se } \mathbf{w} \cdot \mathbf{x} + b < 0. \end{cases}$$

Supõe-se que os dados de treinamento satisfazem as seguintes restrições:

$$x_i \cdot \mathbf{w} + b \geq +1 \quad para \;\; y_i = +1 \tag{4.3}$$

$$x_i \cdot \mathbf{w} + b \leq -1 \quad para \;\; y_i = -1 \tag{4.4}$$

E essas equações podem ser combinadas em

$$y_i(x_i \cdot \mathbf{w} + b) - 1 \geq 0, para\; i = 1, \cdots, n \tag{4.5}$$

em que n é o número de exemplos de treinamento.

Esse é um problema quadrático de otimização cuja solução possui uma ampla e estabelecida teoria matemática [Smola and Schölkopf, 2002]. Problemas desse tipo podem ser solucionados com a introdução de uma função Lagrangiana, que engloba as restrições à função objetivo, associados a parâmetros denominados multiplicadores de Lagrange α_i (Equação 4.6) [Smola and Schölkopf, 2002].

$$L(\mathbf{w}, b, \alpha) = \frac{1}{2} \parallel \mathbf{w} \parallel^2 - \sum_{i=1}^{n} \alpha_i (y_i (x_i \cdot \mathbf{w} + b) - 1) \tag{4.6}$$

A função Lagrangiana deve ser minimizada, o que implica em maximizar as variáveis α_i e minimizar $\mathbf{w}$ e b [Müller et al., 2001]. Tem-se então um ponto de sela, no qual:

$$\frac{\partial L}{\partial b} = 0 \quad e \quad \frac{\partial L}{\partial \mathbf{w}} = 0 \tag{4.7}$$

A resolução dessas equações leva aos resultados representados nas Equações 4.8 e 4.9.

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{4.8}$$

$$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i x_i \tag{4.9}$$

Substituindo as Equações 4.8 e 4.9 na Equação 4.6, obtém-se o seguinte problema de otimização (Equação 4.10):

$$Maximizar \ \ \alpha \equiv \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \tag{4.10}$$

Respeitando as restrições da equação linear:

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{4.11}$$

E as restrições da inequação:

$$\alpha_i \geq 0, \forall_i = 1, \cdots, n \tag{4.12}$$

Essa formulação é denominada forma *dual*, enquanto o problema original é referenciado como forma primal. É interessante observar que o problema *dual* é formulado

utilizando apenas os dados de treinamento e os seus rótulos.

4.1.2 Casos não lineares

Na maioria dos casos, a separação linear no espaço de entrada é uma hipótese restritiva para ser usada na prática, visto que existem muitos casos em que não é possível dividir satisfatoriamente os dados de treinamento através de um hiperplano [Vapnik, 1995].

O classificador SVM lida com problemas não lineares mapeando o conjunto de treinamento de seu espaço original, referenciado como de entradas, para um novo espaço de maior dimensão, denominado espaço de características [Hearst et al., 1998]. O teorema de Cover [Haykin, 1999] garante que um espaço de entrada com padrões não linearmente separáveis pode ser transformado em um novo espaço de características em que os padrões são linearmente separáveis, desde que duas condições sejam satisfeitas: a transformação seja não linear e a dimensão do espaço de características seja suficientemente grande. Assim, é possível construir um hiperplano ótimo nesse espaço de características.

Seja $\Phi : X \rightarrow \mathfrak{F}$ um mapeamento, em que X é o espaço de entradas e $\mathfrak{F}$ denota o espaço de características. A escolha apropriada de Φ faz com que o conjunto de treinamento mapeado em $\mathfrak{F}$ possa ser separado por um SVM linear, como é visto na Figura 4.3.

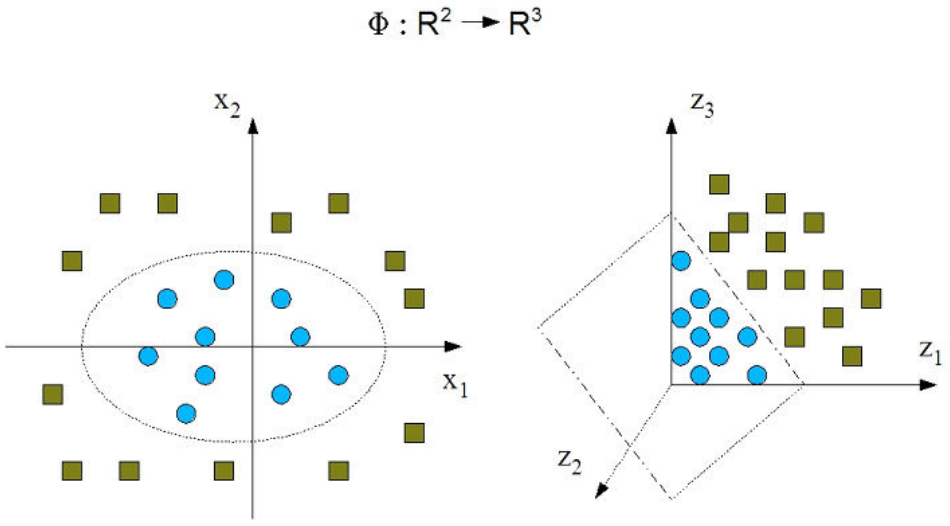

Figura 4.3. Dados de entrada mapeados em um espaço de características de maior dimensão.

Para realizar esse novo mapeamento, basta aplicar Φ aos exemplos presentes no problema de otimização representado na Equação 4.10, conforme ilustrado a seguir:

$$Maximizar\ \alpha \equiv \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j (\Phi(x_i) \cdot \Phi(x_j)) \tag{4.13}$$

De forma semelhante, o classificador extraído torna-se:

$$g(x) = sgn(f(x)) = sgn(\sum_{i=1}^{n} \alpha_i y_i \Phi(x_i) \cdot \Phi(x) + b) \tag{4.14}$$

Como $\mathfrak{F}$ pode ter uma dimensão muito alta, até mesmo infinita, a computação de Φ pode ser extremamente custosa ou inviável. Contudo, a formulação apresentada pela SVM não linear tem uma característica singular: um produto interno realizado no espaço de características. Esse produto interno pode ser usado para executar o mapeamento e é obtido com o uso de funções denominadas núcleos (*kernels*).

Um núcleo Q é uma função que recebe dois pontos x_i e x_j do espaço de entradas e computa o produto escalar desses dados no espaço de características:

$$Q(x_i, x_j) = \Phi(x_i) \cdot \Phi(x_j) \tag{4.15}$$

Alguns dos núcleos mais comumente utilizados na prática são o polinomial, gaussiano ou RBF (*Radial-Basis Function*) e sigmoidal, listados na Tabela 4.1. Cada um deles apresenta parâmetros que devem ser determinados pelo usuário, também indicados na Tabela 4.1.

Tabela 4.1. Exemplos de núcleos.

Tipo de núcleo	**Função** $Q(x_i, x_j)$	**Parâmetros**
Polinomial	$(\delta(x_i \cdot x_j) + \kappa)^{di}$	δ, κ, di
Gaussiano	$exp(-\sigma \parallel x_i - x_j \parallel)^2$	σ
Sigmoidal	$tanh(\delta(x_i \cdot x_j) + \kappa)$	δ, κ

4.1.3 Classificação em Múltiplas Classes

O classificador SVM foi originalmente desenvolvido para classificação binária, porém, no caso de classificação em p classes, $p > 2$, existem duas abordagens básicas [Crammer and Singer, 2000, Hsu and Lin, 2002], conforme a Figura 4.4. A primeira abordagem reduz o problema de múltiplas classes a um conjunto de problemas binários, através dos métodos decomposição um por classe (*one against all*) e separação das classes duas a duas (*one against one*). A segunda abordagem é a generalização de classificadores SVM binários para mais de duas classes e um dos métodos que utiliza essa abordagem é o método de Cramer e Singer.

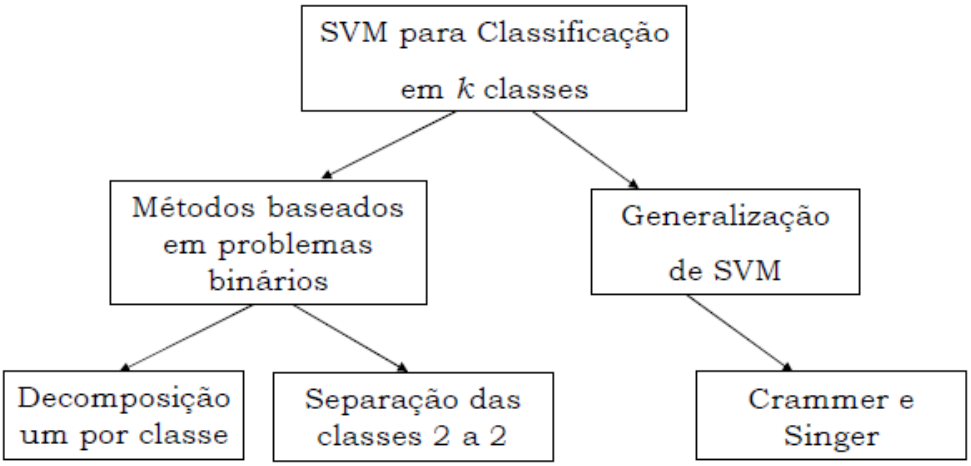

Figura 4.4. Métodos para classificação em múltiplas classes.

Cada conjunto de dados é treinada independentemente pelo classificador SVM. O conjunto de dados de treinamento (x_i, c_i) consiste em n exemplos pertencentes a M classes. O rótulo da classe $c_i \in 1, 2, \cdots, M$. Assume-se que o número de exemplos de cada classe é o mesmo, ou seja, n/M.

4.1.3.1 Decomposição um por classe

O método decomposição um por classe é provavelmente a primeira implementação da classificação em múltiplas classes através do SVM [Hsu and Lin, 2002].

Nesse método, SVM é aplicado para cada classe através da discriminação desta classe contra as classes restantes. Um dado de teste x é classificado usando uma estratégia de decisão, isto é, a classe com o valor máximo da função discriminante $f(x)$ é atribuída a esse dado. Todos os n exemplos de treinamento são utilizados na construção do vetor de suporte para uma classe. O vetor de suporte para uma classe p é construído utilizando o conjunto de exemplos de treinamento (x_i) e suas saídas desejadas (y_i).

A saída desejada y_i para um exemplo de treinamento x_i é definido por:

$$y_i = \begin{cases} +1, & \text{se } c_i = p; \\ -1, & \text{se } c_i \neq p. \end{cases}$$

Os exemplos com a saída desejada $y_i = +1$ são chamados exemplos positivos e os exemplos com a saída desejada $y_i = -1$ são chamados exemplos negativos. Um hiperplano ótimo é construído para separar n/M exemplos positivos de $n(M-1)$ exemplos negativos.

4.1.3.2 Separação das classes duas a duas

O método separação das classes duas a duas foi primeiro introduzido na técnica SVM como SVM em pares (*pairwise*).

Nesse método, SVM é utilizado em cada par de classes através do seu treinamento na discriminação de duas classes. Dessa forma, o número de vetores de suporte usado nesse método é $M(M-1)/2$. SVM para um par de classes (p, m) é construído utilizando exemplos de treinamento pertencentes a somente às duas classes.

A saída desejada y_i para o exemplo de treinamento x_i é dada por:

$$y_i = \begin{cases} +1, & \text{se } c_i = p; \\ -1, & \text{se } c_i = m. \end{cases}$$

A estratégia de utilizar um esquema de votos máximos é usada para determinar a classe de um dado de teste x. Se $f_{pm}(x)$, que é o valor da função discriminante do SVM para o par de classes (p, m), é positivo, então a classe p ganha um voto. Se não, a classe m ganha um voto. As saídas do classificador SVM são usadas para determinar o número de votos ganhos para cada classe. A classe com o maior número de votos é atribuída ao dado de teste. Quando existem múltiplas classes com um número máximo de votos, a classe com o valor máximo da magnitude total da função discriminante é atribuída ao dado. Esse valor é calculado por:

$$valor\ máximo_p = \sum_m |f_{pm}(x)|$$

onde o somatório ocorre sobre todos as m classes com as quais a classe p está pareada.

4.1.3.3 Método de Cramer e Singer

O método de Cramer e Singer [2000,2001] usa uma maneira mais natural de resolver o problema de classificação $p > 2$, que é construir uma função de decisão considerando todas as classes de uma vez. Nesse método, todos os exemplos de treinamento são usados ao mesmo tempo.

A abordagem desse método para o problema de multiclasses é a resolução através de um único problema de otimização, onde requer-se somente l variáveis oscilantes (*slack*):

$$min_{\mathbf{w},\xi} \frac{1}{2} \sum_{m=1}^{M} \mathbf{w}_m^T \mathbf{w}_m + C \sum_{i=1}^{l} \xi_i,$$

sujeito a:

$$\mathbf{w}_{y_i}^T \phi(x_i) - \mathbf{w}_m^T \phi(x_i) \geq 1 - \xi_i, \text{ para } i = 1, \cdots, l$$

A função de decisão resultante é:

$$f(x) = argmax_{m=1,\cdots,M}(\mathbf{w}_m^T \phi(x))$$

4.2 Conclusão

Neste capítulo, foi apresentado como o classificador SVM pode ser utilizado para a classificação de imagens através da obtenção de vetores de suporte e separação das imagens em diversas classes. O classificador SVM, para o processo de recuperação de imagens, agrupa as imagens semelhantes e o modelo gerado através do seu treinamento indica a relevância de cada imagem para uma imagem de consulta.

SVM caracteriza-se por apresentar uma boa capacidade de generalização, sendo robusto diante de dados de grande dimensão. Outra vantagem é o uso dos núcleos em casos não lineares, que torna o algoritmo do SVM eficiente, pois permite a construção de simples hiperplanos em um espaço de alta dimensão de forma tratável do ponto de vista computacional. Entre as principais limitações desse método encontram-se a determinação do melhor núcleo a ser utilizado diante do conjunto de dados.

Capítulo 5

Experimentos e Resultados

Neste capítulo, uma descrição do desenvolvimento do sistema CBIR proposto é apresentada. Em detalhes, são apresentados o caso de estudo que utiliza o tipo de tecido da mama de acordo com as quatro categorias BI-RADS como padrão de busca, chamado de MammoSys, e o caso de estudo que utiliza o tipo de tecido da mama juntamente com a lesão da mama e sua classificação, de acordo com as categorias BI-RADS, chamado de MammoSysLesion. A técnica 2DPCA é introduzida para a caracterização do tecido e da lesão da mama, e é comparada com as técnicas PCA e SVD, visto que essas são técnicas que também podem representar a textura e reduzir a dimensionalidade do vetor de características. Máquina de vetores de suporte é utilizada para o processo de recuperação das imagens mamográficas.

Também, é apresentada a integração de quatro bases de dados de imagens mamográficas já existentes ao projeto IRMA, e que são a base de testes desse trabalho.

5.1 Integração das bases de dados de imagens mamográficas ao Projeto IRMA

A base de dados de imagens mamográficas integrada ao projeto IRMA foi desenvolvida baseada na união das bases de dados DDSM, MIAS, LLNL (*Lawrence Livermore National Laboratory*) e imagens de rotina do hospital universitário da Universidade RWTH (*Rheinisch-Westfälische Technische Hochschule*) de Aachen, Alemanha. Os detalhes das bases de dados são dados a seguir.

DDSM. A base de dados DDSM [Heath et al., 1998] contém oficialmente 2.479 estudos (695 normais, 870 benignos, e 914 malignos casos). Cada estudo inclui duas imagens de cada mama, adquiridas nas direções crânio-caudal (CC) e médio-lateral

(ML) que foram escaneadas de imagens baseadas em filmes por quatro escaners diferentes com resolução entre 50 e 42 microns. Isso resulta em um total de 9.916 radiografias. As imagens são codificadas com um algoritmo de acordo com o padrão sem perdas JPEG (*Joint Pictures Expert Group*) e tiveram que ser convertidas em um arquivo de formato padrão utilizando um software fornecido pela página da internet da DDSM. Para todos os casos, existem arquivos de texto adicionais com informações do tipo de digitalizador e tipo de tecido de acordo com o padrão BI-RADS.

MIAS. A base de dados MIAS [Suckling, 1994] está disponível somente para propósitos de pesquisa científica e contém 322 imagens mamográficas, todas elas adquiridas na direção ML. Os arquivos das imagens estão disponíveis no formato PNG (*Portable Network Graphics*) e possuem anotações com os seguintes detalhes: um número de referência indicando mama esquerda ou direita, características do tipo de tecido, classe da lesão presente e coordenadas e tamanho dessas lesões.

LLNL. A base de dados LLNL[1] contém 197 imagens mamográficas, todas elas digitalizadas a 35 microns. As imagens estão armazenadas no formato ICS (*Image Cytometry Standard*) e tiveram que ser convertidas a um formato padrão utilizando um código fonte que converte imagens do formato ICS para o formato PGM (*Portable Grey Map*). Para 190 imagens mamográficas está disponível um texto contendo resultado da biópsia e padrão de ouro das imagens.

RWTH. O Departamento de Radiologia Diagnóstica do Hospital Universitário RWTH cedeu 170 casos em que as imagens foram adquiridas digitalmente através de um mamógrafo *General Electric Senographe* operando com um feixe baixo de energia de 26 a 32 kV e com um sistema de armazenamento da *Fuji/Philips* capaz de gravar 7 lp/mm. O cassete foi lido usando um *Philips PCR Eleva CosimaX*. Se disponível, um texto em alemão descrevendo o exame da mama, patologia, tipo de tecido e lesão estava incluído junto com a imagem digital.

De forma a integrar todas essas bases de dados no sistema IRMA, foi determinado um código IRMA [Lehmann et al., 2003] para as imagens mamográficas, visto que todas as imagens desse projeto são codificadas de acordo com um esquema multi-eixos, que fornece o padrão de ouro das imagens, e descrevem:

- técnica: modalidade de imageamento.
- direção: orientação da mama - CC ou ML.
- anatomia: região da mama examinada - mama direita ou mama esquerda.
- sistema biológico: tipo de tecido, estágio do tumor e tipo de lesão.

[1]Center For Health Care Technologies Livermore, Livermore, CA, USA

O código IRMA foi determinado automaticamente para a integração das quatro bases de dados ao sistema IRMA [de Oliveira et al., 2008].

Como resultado, estão disponíveis 10.509 imagens mamográficas que possuem o padrão de ouro estabelecido e verificado por um experiente radiologista, e de acordo com o padrão internacional BI-RADS. Mais detalhes da base de dados são encontrados no Apêndice B.

5.2 Metodologia aplicada aos casos de estudo MammoSys e MammoSysLesion

O sistema CBIR proposto foi implementado usando MatLab (*Matrix Laboratory*), utilizando as bibliotecas de processamento de imagens e matemática simbólica, e a biblioteca LIBSVM [Chang and Lin, 2001]. A extração de características foi executada em uma máquina equipada com IntelCore2Quad com processador 2,66 GHz, 8 GB de RAM e sistema operativo *Microsoft Windows* versão 64 bits. A recuperação de imagens foi realizada em uma máquina equipada com IntelCore2Duo com processador 2 GHz, 3 GB de RAM e sistema operativo *Microsoft Windows* versão 32 bits.

A Figura 5.1 apresenta a metodologia aplicada aos casos de estudo. O tamanho das imagens varia de 1.024 $\times$ 800 *pixels* a 1.024 $\times$ 340 *pixels* e para a aplicação da técnica 2DPCA é necessário que todas as imagens tenham um mesmo tamanho. Em vista disso, de cada imagem foi extraída uma ROI de tamanho 340 $\times$ 340 *pixels*, que permite a seleção somente da região de interesse da mama e exclui ruídos tais como anotações e rótulos de exame nas imagens mamográficas.

5.3 Caso de estudo MammoSys

As imagens mamográficas utilizadas para o desenvolvimento do caso de estudo proposto são da base de dados de imagens radiológicas do projeto IRMA e estão no formato PNG. De ambas projeções CC e ML foram selecionadas 800 imagens mamográficas sendo:

- 200 imagens mamográficas com tipo de tecido da categoria BI-RADS I;
- 200 imagens mamográficas com tipo de tecido da categoria BI-RADS II;
- 200 imagens mamográficas com tipo de tecido da categoria BI-RADS III;
- 200 imagens mamográficas com tipo de tecido da categoria BI-RADS IV.

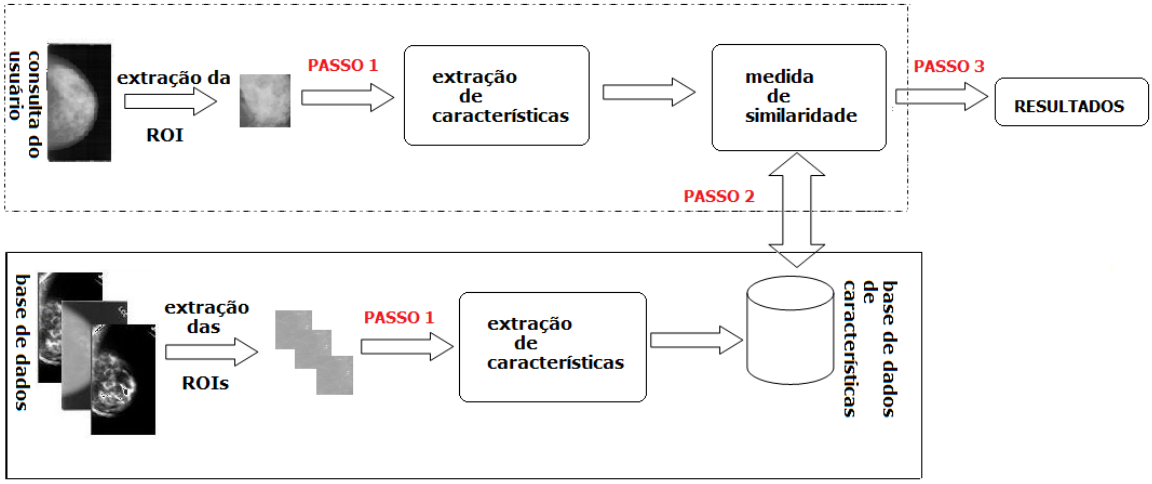

Figura 5.1. Metodologia aplicada aos casos de estudo.

Após a extração das ROIs, os passos seguidos para o desenvolvimento do sistema foram:

1° passo → Extração de características: As técnicas 2DPCA, PCA e SVD foram aplicadas em cada uma das 800 ROIs. Os seguintes componentes principais referentes aos primeiros d autovetores e que correspondem aos primeiros d maiores autovalores da matriz de covariância foram usados nos experimentos: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15 e 20, para formar o vetor de característica das imagens. Esses valores foram escolhidos empiricamente, como no trabalho de Zuo *et al.* [2006].

2° passo → Medida de similaridade entre as imagens: o classificador SVM foi usado para separar as quatro classes do conjunto de dados e determinar a relevância das imagens para uma determinada consulta. Usando a biblioteca LIBSVM, o conjunto de 800 vetores de características foi utilizado nos seguintes experimentos:

- Experimento 1: divisão do conjunto em 60% (120 de cada tipo de tecido) para treinamento e 40% (80 de cada tipo de tecido) para teste;
- Experimento 2: divisão do conjunto em 50% (100 de cada tipo de tecido) para treinamento e 50% (100 de cada tipo de tecido) para teste;
- Experimento 3: divisão do conjunto em 40% (80 de cada tipo de tecido) para treinamento e 60% (120 de cada tipo de tecido) para teste.

Além disso, os testes foram feitos para os núcleos polinomial e radial. O caso linear não foi considerado visto que na prática ele não se adequa a maior parte dos experimentos e também por ser difícil de lidar com dados que possuem ruídos [Burges, 1998].

3° passo → Avaliação do caso de estudo: medidas de precisão e revocação foram obtidas e todas as imagens mamográficas que não foram usadas para o treinamento do classificador SVM foram utilizadas como imagem de consulta. Para um conjunto de valores resultante de uma consulta a uma base de dados, precisão e revocação são definidas como:

$$precisão = \frac{TRO}{TO} \qquad\qquad revocação = \frac{TRO}{TR}$$

em que TRO significa o total de imagens relevantes obtidas no resultado, TO significa o total de imagens obtidas no resultado e TR significa total de imagens relevantes disponíveis na base de dados.

A construção de um gráfico apresentando diferentes valores de precisão × revocação produz uma curva onde, como regra de análise, o resultado é melhor à medida que a curva se aproxima do topo, ou seja, valores de precisão iguais a 1 ou 100%.

Foram considerados também valores de precisão para 10% de revocação, já que os radiologistas prestam mais atenção às primeiras imagens recuperadas pelo sistema. A relevância de cada imagem mamográfica recuperada para a imagem de consulta também foi obtida para as primeiras imagens recuperadas e apresentadas pelo sistema MammoSys.

O desempenho da técnica 2DPCA foi comparada com as técnicas SVD e PCA para a caracterização do tecido da mama e SVM para a tarefa de recuperação de imagens.

5.3.1 Resultados e Discussão

A Tabela 5.1 lista o tempo de execução, em minutos, da extração de características utilizando as técnicas 2DPCA, PCA e SVD para a caracterização do tecido da mama das 800 imagens mamográficas para os primeiros 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15 e 20 componentes principais. Nota-se, que para as técnicas 2DPCA e SVD, o tempo de extração de características foi semelhante e ambos foram mais rápidos que a técnica PCA, validando a teoria que diz que a técnica 2DPCA, por ser baseada na matriz da imagem, extrai as características mais rapidamente que a técnica PCA.

Na obtenção da precisão média, todas as imagens que não fizeram parte do conjunto de treinamento tiveram sua precisão calculada para então a média dessas precisões ser obtida, para os primeiros d componentes principais das técnicas 2DPCA, SVD e PCA, e também para os três experimentos usando SVM e o núcleo polinomial, que são apresentados na Tabela 5.2. Os testes aplicando SVM e o núcleo radial forneceram valores médios de precisão em torno de 30% a 35% para todas as técnicas, e portanto

Tabela 5.1. Tempo de execução, em minutos, da extração de características das 800 imagens mamográficas utilizando as técnicas 2DPCA, PCA e SVD, respectivamente, para todos os primeiros d componentes principais testados.

d	2DPCA	PCA	SVD
1	3,1	3,9	3,3
2	3,2	3,9	3,3
3	3,2	3,9	3,3
4	3,2	3,9	3,3
5	3,2	4	3,3
6	3,3	4	3,3
7	3,3	4,1	3,3
8	3,3	4,1	3,3
9	3,3	4,1	3,4
10	3,3	4,1	3,4
15	3,4	4,1	3,4
20	3,4	4,2	3,5

foi desconsiderado para a avaliação do sistema. Isso pode ter acontecido pelo fato do núcleo polinomial ser mais flexível e adaptável a poucas classes, sendo capaz de se adequar ao comportamento dos dados e separar as classes mais eficientemente do que o núcleo radial.

Pode-se observar na Tabela 5.2 que a técnica 2DPCA superou as técnicas PCA e SVD para todos os primeiros d componentes principais e para todos os três experimentos utilizando SVM. O maior valor de precisão média obtido foi de 81,11%, considerando-se os primeiros quatro componentes principais da técnica 2DPCA e treinando o classificador SVM com o núcleo polinomial com 60% dos vetores de características e testando com 40% dos vetores de características.

Para a técnica 2DPCA, o aumento da quantidade de componentes principais não decaiu o valor da precisão média, ao contrário do que ocorreu para as técnicas PCA e SVD. Para essas duas técnicas, em que os componentes principais são escalares, a matriz diagonal que contém os componentes principais possui d valores que são significantemente maiores que outros, e que estão ordenados do maior para o menor valor. Esse menores valores, que são próximos de zero, podem ser considerados como ruídos e podem não ser capazes de caracterizar apropriadamente uma imagem, explicando portanto os baixos valores da precisão média ao considerar-se um maior número de componentes principais. Além disso, para essas duas técnicas, o valor da precisão média foi maior considerando-se apenas um componente principal, já que, à medida que se aumenta o número de componentes principais, é possível aumentar a confusão do SVM em obter corretamente os vetores de suporte que irão separar as quatro categorias do

Tabela 5.2. Precisão média para os primeiros d componentes principais selecionados comparando as técnicas 2DPCA, PCA e SVD para a caracterização do tecido da mama e para os três experimentos utilizando SVM com o núcleo polinomial para o processo de recuperação.

d	Característica	Experimento 1	Experimento 2	Experimento 3
1	2DPCA	77,83%	76,02 %	75,47%
	PCA	71,09%	71,02%	71,02%
	SVD	71,92%	72,15%	72%
2	2DPCA	78,01%	77,99%	77,05%
	PCA	70,73%	70,64%	70,62%
	SVD	71,88%	72,05%	71,95%
3	2DPCA	80,06%	80,04%	78,96%
	PCA	70,55%	70,58%	70,26%
	SVD	74,99%	74,97%	74,52%
4	2DPCA	**81,11**%	80,08%	78,87%
	PCA	70,51%	70,39%	70,26%
	SVD	75,86%	75,63%	75,2%
5	2DPCA	80,16%	79,51%	78,51%
	PCA	70,61%	70,49%	70,07%
	SVD	75,75%	75,56%	75,21%
6	2DPCA	80,12%	79,94%	**80,26**%
	PCA	70,76%	70,25%	70,02%
	SVD	75,67%	75,51%	75,08%
7	2DPCA	80,38%	78,67%	78,88%
	PCA	70,37%	70,29%	69,85%
	SVD	75,48%	75,58%	75%
8	2DPCA	80%	79,27%	78,96%
	PCA	70,39%	70,12%	69,88%
	SVD	75,57%	75,4%	74,98%
9	2DPCA	80,07%	80,06%	78,75%
	PCA	70,32%	70,23%	69,93%
	SVD	75,48%	75,33%	74,9%
10	2DPCA	80,31%	79,01%	78,83%
	PCA	70,35%	70,05%	69,86%
	SVD	75,44%	75,11%	74,87%
15	2DPCA	80,06%	**80,13**%	79,67%
	PCA	70,51%	70,08%	69,85%
	SVD	74,99%	75,07%	74,7%
20	2DPCA	80,02%	80,04%	79,24%
	PCA	70,12%	70,02%	69,75%
	SVD	74,13%	75,01%	74,63%

tecido da mama e gerar o modelo que indica a relevância das imagens. Os resultados obtidos pela técnica 2DPCA mostraram-se estáveis para todos os d componentes prin-

cipais, indicando que a presença de ruídos nos dados é tão pequena que não influencia a caracterização dos tecidos da mama e que a escolha de menos valores de componentes principais alia a caracterização das imagens e a redução da dimensionalidade dos vetores de características.

Esses resultados poderiam ser melhorados utilizando-se juntamente, para a caracterização do tecido da mama com a técnica 2DPCA, outro tipo de atributo, como por exemplo, o histograma de níveis de cinza, que pode captar as diferenças entre os diferentes tipos de tecido, somando-se a caracterização da textura dos mesmos. Isto poderia ser realizado concatenando-se no vetor de características os dois tipos de atributos – histograma e textura – além de se verificar o peso de cada atributo para a representação dos tipos de tecidos da mama.

Também, comparando-se os três experimentos do classificador SVM para a divisão do conjunto de treinamento e teste, nota-se que são obtidos melhores resultados quando o SVM é treinado com um número maior de dados, que aumenta seu poder de generalização, ou seja, é obtida uma classificação mais correta dos dados de teste.

Ainda, de acordo com a Tabela 5.2, considerando-se o melhor resultado dos três experimentos com o classificador SVM e o núcleo polinomial, a Tabela 5.3 apresenta o tempo de execução, em segundos, do caso de estudo MammoSys, ou seja, de todo o processo de recuperação das imagens, paras todos os primeiros d componentes principais.

Tabela 5.3. Tempo de execução, em segundos, do caso de estudo MammoSys, comparando as técnicas 2DPCA, PCA e SVD para a caracterização do tecido da mama e SVM com o núcleo polinomial para o processo de recuperação, para todos os primeiros d componentes principais testados.

d	**2DPCA**	**PCA**	**SVD**
1	1.200,00	0,43	0,78
2	132,24	0,42	0,70
3	58,78	0,38	0,71
4	30,75	0,37	0,72
5	31,93	0,41	0,72
6	33,90	0,41	0,80
7	38,20	0,42	0,81
8	39,95	0,44	0,81
9	39,97	0,45	0,81
10	43,45	0,54	0,85
15	58,52	0,63	0,99
20	76,36	0,69	0,99

Apesar da técnica 2DPCA ter demorado mais tempo, para executar o processo

de recuperação das imagens mamográficas, com base no tipo de tecido da mama, isso era esperado, visto que cada componente principal para a técnica 2DPCA é um vetor, enquanto para as técnicas PCA e SVD cada componente principal é um escalar. Além disso, para os primeiros um, dois e três componentes principais, a técnica 2DPCA teve um tempo de execução maior que para os outros d valores. Algoritmos de aprendizado de máquina como o SVM são influenciados pelos dados, isto é, o número de características pode degradar o desempenho computacional. Se o número de características utilizado é muito pequeno ou não significante, pode fazer com que os vetores de suporte não sejam capazes de indicar corretamente a separação dos dados e indicar a relevância das imagens, não aprendendo a que classe esses dados pertencem, levando portanto mais tempo para executar a tarefa.

A Figura 5.2 mostra, considerando-se a precisão média, o gráfico de precisão $\times$ revocação usando os quatro primeiros valores dos componentes principais para as técnicas 2DPCA, PCA e SVD para a caracterização do tecido da mama e SVM com o núcleo polinomial para a tarefa de recuperação de imagens. A textura do tecido da mama foi melhor representada pelas características extraídas usando a técnica 2DPCA, que foi capaz de capturar a diferença entre as intensidades de níveis de cinza entre os diferentes tecidos da mama e caracterizá-los.

Com respeito à técnica 2DPCA e considerando-se uma imagem de consulta, para 10% de revocação, uma precisão de 90% significa que de 32 imagens mamográficas retornadas pelo sistema CBIR MammoSys, 28 imagens são relevantes à consulta do usuário.

A Figura 5.3 mostra um exemplo do sistema CBIR MammoSys. A imagem mamográfica de consulta pertence à categoria BI-RADS II (primeira imagem no alto e à esquerda). As imagens foram recuperadas com base na caracterização do tecido da mama com a técnica 2DPCA usando a seleção dos quatro primeiros valores dos componentes principais, e na recuperação utilizando SVM com o núcleo polinomial. Essa imagem de consulta foi sorteada entre todas as imagens do conjunto de testes. Nota-se que todas as imagens mamográficas recuperadas são da mesma categoria – BI-RADS II – da imagem de consulta.

Apresentando a interface do sistema MammoSys, a Figura 5.4 mostra um outro exemplo de recuperação de imagens mamográficas, também com base nos quatro primeiros componentes principais da técnica 2DPCA e SVM para o processo de recuperação. Todas as imagens mamográficas recuperadas pertencem à mesma categoria da imagem de consulta – BI-RADS IV – com exceção da sétima imagem mamográfica recuperada, que pertence à categoria BI-RADS III. Apesar dessa imagem mamográfica ser de um tipo de tecido denso, ela pode ter sido recuperada entre as primeiras pelo

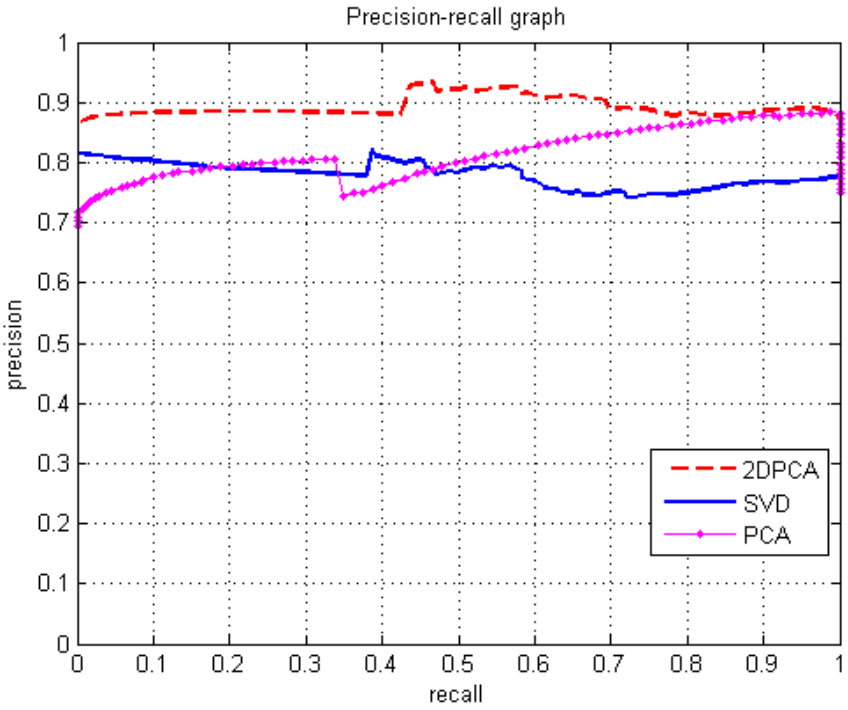

Figura 5.2. Curva precisão × revocação comparando as técnicas 2DPCA, PCA e SVD para a caracterização do tecido da mama e SVM para o processo de recuperação considerando-se a precisão média.

fato das imagens mamográficas da base de dados IRMA serem de diferentes bases de dados, e portanto terem sido obtidas de diferentes escaners e resoluções.

Esse mesmo exemplo da Figura 5.4 é apresentado na Figura 5.5 com a indicação da relevância de cada imagem mamográfica recuperada para a imagem mamográfica de consulta. A avaliação do sistema CBIR MammoSys através da indicação da relevância mostra que, considerando o valor médio da relevância das primeiras imagens recuperadas, essas imagens mamográficas recuperadas que estiverem abaixo desse valor médio, apesar de serem relevantes à consulta, não são tão similares à imagem de consulta e portanto não devem ser apresentadas ao usuário.

Nesse exemplo, considerando as oito primeiras imagens mamográficas recuperadas, o valor médio da relevância obtido pelo classificador SVM foi de 28,2806. Para a imagem mamográfica de consulta (primeira no alto e à esquerda), somente as quatro primeiras imagens retornadas pelo sistema MammoSys são relevantes à consulta e devem ser apresentadas ao usuário. As outras quatro imagens mamográficas são aqui mostradas apenas para a visualização das relevâncias, mas para o uso do sistema pelo radiologista não deveriam ser apresentadas. Esse sistema de avaliação excluiria a sétima imagem recuperada erroneamente pelo sistema, que é a única que não pertence à

mesma categoria BI-RADS da imagem de consulta.

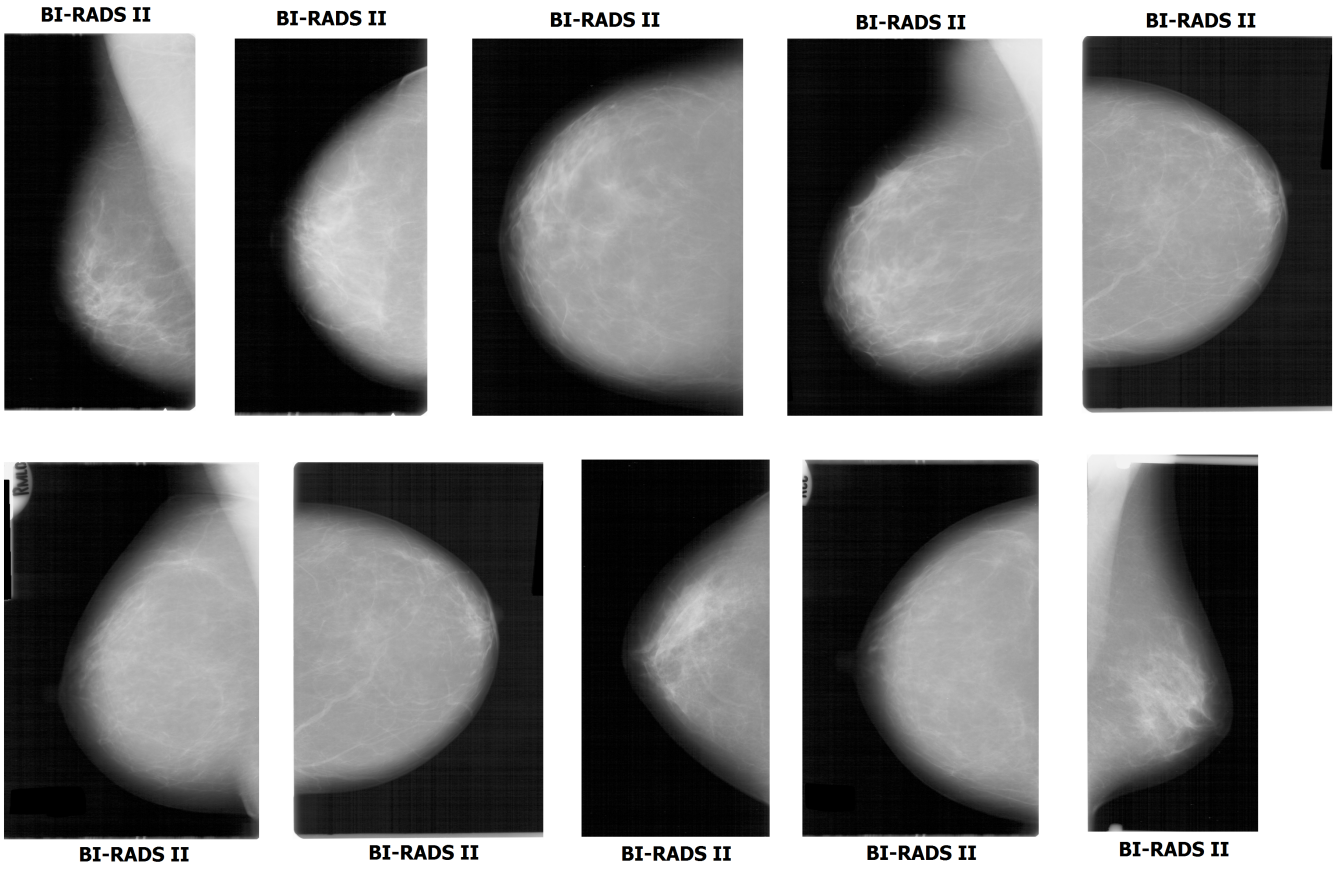

Figura 5.3. Exemplo do caso de estudo MammoSys de recuperação de imagens mamográficas com base no tipo de tecido da mama, utilizando os quatro primeiros componentes principais da técnica 2DPCA e SVM com o núcleo polinomial.

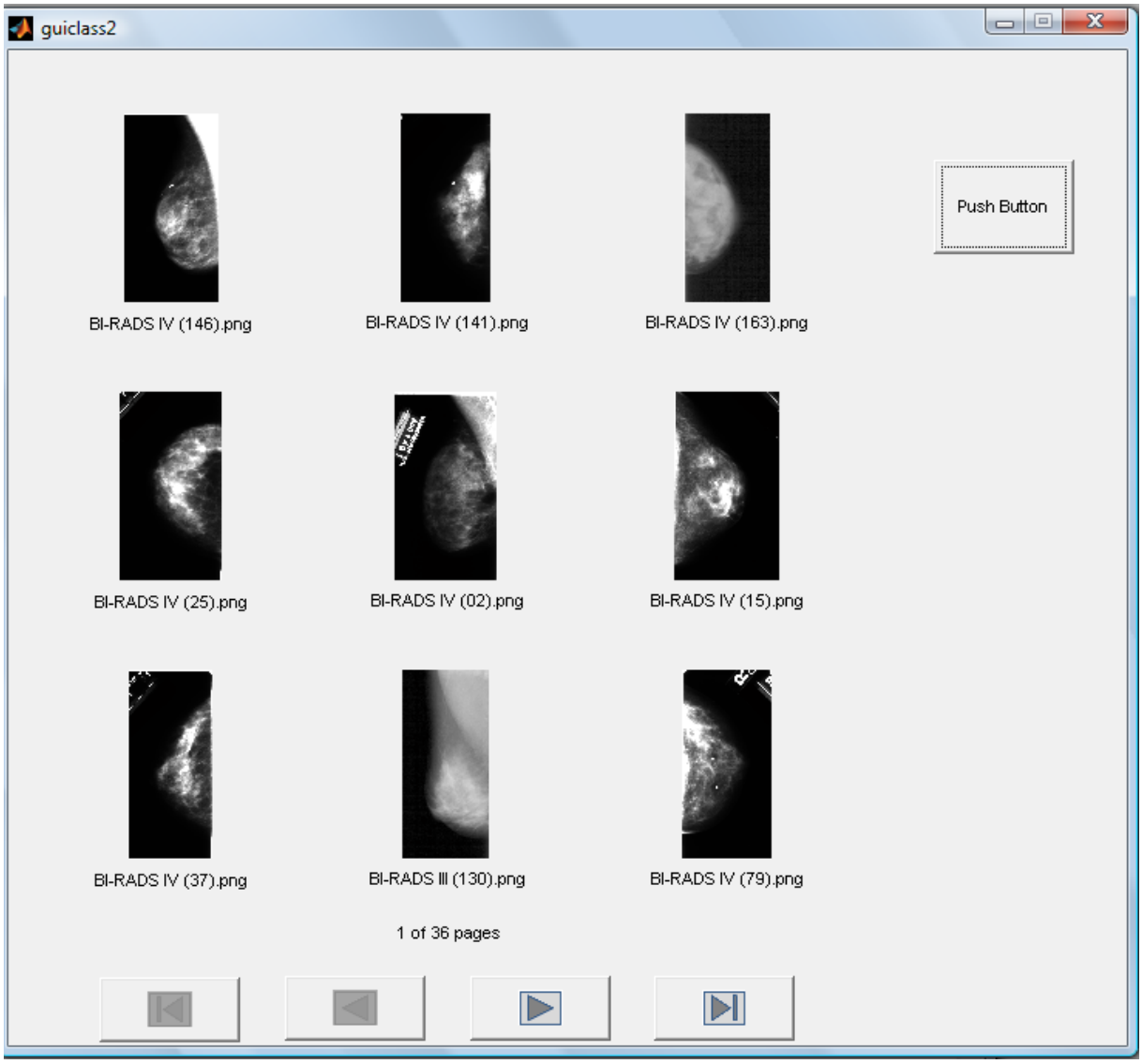

Figura 5.4. Exemplo da interface do caso de estudo MammoSys para a recuperação de imagens mamográficas com base no tipo de tecido da mama, utilizando os quatro primeiros componentes principais da técnica 2DPCA e SVM com o núcleo polinomial.

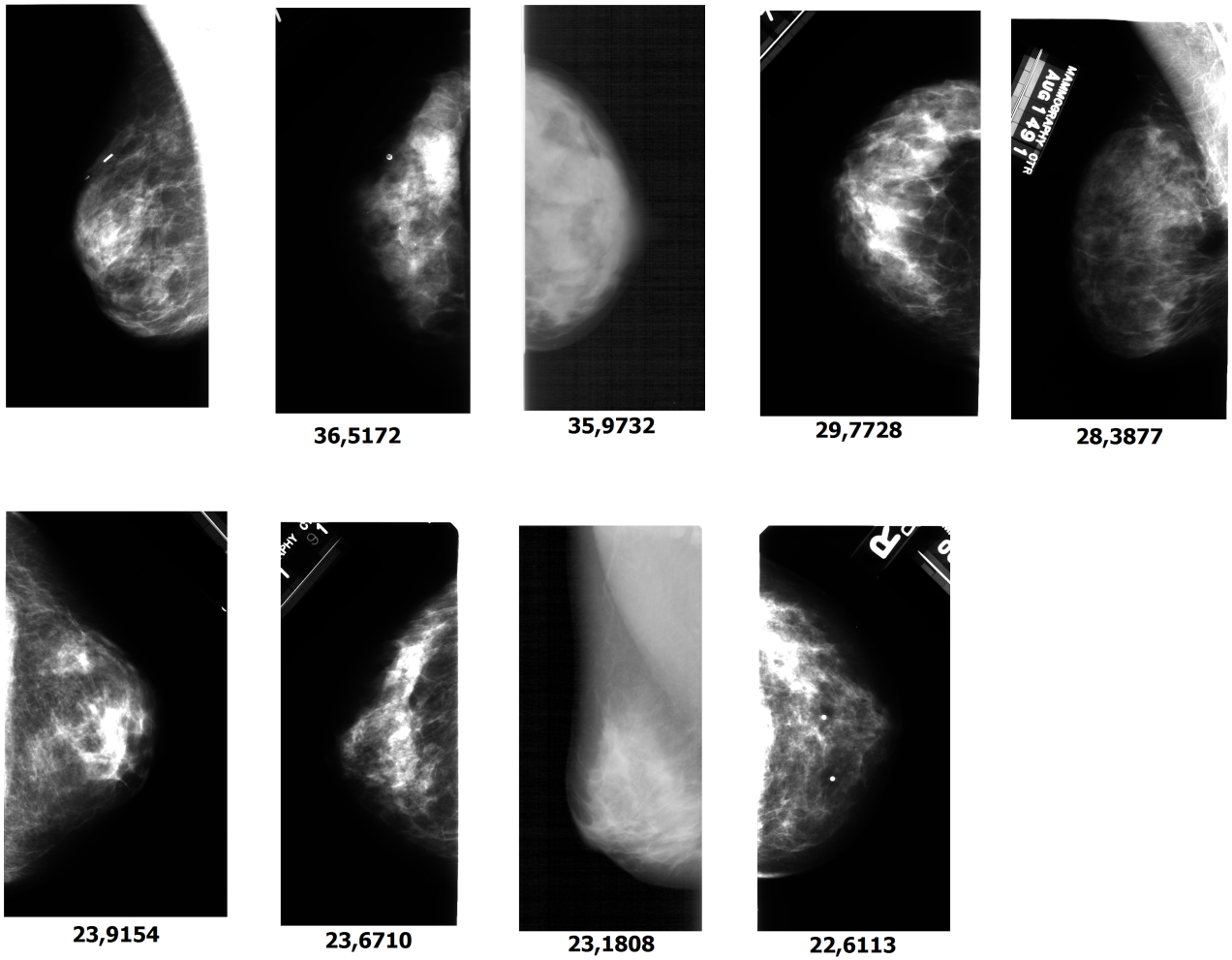

Figura 5.5. Segundo exemplo do caso de estudo MammoSys de recuperação de imagens mamográficas com base no tipo de tecido da mama, utilizando os quatro primeiros componentes principais da técnica 2DPCA e SVM com o núcleo polinomial, e com a indicação de relevância das imagens mamográficas recuperadas para a imagem de consulta (no alto e à esquerda).

5.4 Caso de estudo MammoSysLesion

As imagens mamográficas utilizadas para o desenvolvimento do caso de estudo proposto também são da base de dados de imagens radiológicas do projeto IRMA e de ambas projeções CC e ML foram selecionadas 1.392 imagens mamográficas sendo:

- 116 imagens mamográficas com tipo de tecido da categoria BI-RADS I e sem lesão mamográfica (categoria BI-RADS 1);
- 116 imagens mamográficas com tipo de tecido da categoria BI-RADS I e com lesão mamográfica benigna (categoria BI-RADS 2);
- 116 imagens mamográficas com tipo de tecido da categoria BI-RADS I e com lesão mamográfica maligna (categoria BI-RADS 5);
- 116 imagens mamográficas com tipo de tecido da categoria BI-RADS II e sem lesão mamográfica (categoria BI-RADS 1);
- 116 imagens mamográficas com tipo de tecido da categoria BI-RADS II e com lesão mamográfica benigna (categoria BI-RADS 2);
- 116 imagens mamográficas com tipo de tecido da categoria BI-RADS II e com lesão mamográfica maligna (categoria BI-RADS 5);
- 116 imagens mamográficas com tipo de tecido da categoria BI-RADS III e sem lesão mamográfica (categoria BI-RADS 1);
- 116 imagens mamográficas com tipo de tecido da categoria BI-RADS III e com lesão mamográfica benigna (categoria BI-RADS 2);
- 116 imagens mamográficas com tipo de tecido da categoria BI-RADS III e com lesão mamográfica maligna (categoria BI-RADS 5);
- 116 imagens mamográficas com tipo de tecido da categoria BI-RADS IV e sem lesão mamográfica (categoria BI-RADS 1);
- 116 imagens mamográficas com tipo de tecido da categoria BI-RADS IV e com lesão mamográfica benigna (categoria BI-RADS 2);
- 116 imagens mamográficas com tipo de tecido da categoria BI-RADS IV e com lesão mamográfica maligna (categoria BI-RADS 5).

Tabela 5.4. Número de imagens mamográficas utilizadas pertencentes às categorias BI-RADS para tecido de mama e lesão mamográfica.

Categorias	BI-RADS I	BI-RADS II	BI-RADS III	BI-RADS IV
BI-RADS 1	116	116	116	116
BI-RADS 2	116	116	116	116
BI-RADS 5	116	116	116	116
TOTAL	348	348	348	348

A Tabela 5.4 resume a quantidade de imagens mamográficas utilizadas.

As imagens mamográficas com lesões pertencentes às outras categorias BI-RADS – BI-RADS 0, BI-RADS 3, BI-RADS 4 – estão presentes nas bases de dados de imagens mamográficas em uma pequena quantidade. Após a integração das quatro bases de dados – DDSM, MIAS, LLNL e RWTH – ao sistema IRMA, existem:

- 8 imagens mamográficas da categoria BI-RADS 0;
- 12 imagens mamográficas da categoria BI-RADS 3;
- 6 imagens mamográficas da categoria BI-RADS 4.

Em vista disso, foram utilizadas para o caso de estudo MammoSysLesion apenas as imagens mamográficas das categorias citadas – BI-RADS 1, BI-RADS 2 e BI-RADS 5. Além disso, a categoria BI-RADS 6 refere-se a casos cujas lesões já foram identificadas na categoria BI-RADS 5 e que necessitam de acompanhamento e outro tipo de terapia ou imageamento.

Após a extração das ROIs, os passos seguidos para o desenvolvimento do sistema foram:

1° passo → Extração de características: As técnicas 2DPCA, PCA e SVD foram aplicadas em cada uma das 1.392 ROIs. Os seguintes componentes principais referentes aos primeiros d autovetores e que correspondem aos primeiros d maiores autovalores da matriz de covariância foram usados nos experimentos: 1, 2, 3, 4, 5, 6, 7, 8, 9 e 10, para formar o vetor de característica das imagens. Esses valores foram escolhidos empiricamente, como no trabalho de Zuo *et al.* [Zuo et al., 2006].

2° passo → Medida de similaridade entre as imagens: o classificador SVM foi usado para separar as 12 classes do conjunto de dados e determinar a relevância das imagens para uma determinada consulta. Usando a biblioteca LIBSVM, o conjunto de 1.392 vetores de características foi dividido em 60% (207 de cada tipo de tecido junto com a lesão) para treinamento e 40% (141 de cada tipo de tecido junto com a lesão) para teste, já que foi visto no caso de estudo anterior, o MammoSys, que SVM obtém

melhores resultados quando são usados mais dados para o seu treinamento. Além disso, os testes foram feitos para os núcleos polinomial e radial.

3° passo → Avaliação do caso de estudo: medidas de precisão e revocação foram obtidas e todas as 564 imagens mamográficas que não foram usadas para o treinamento do classificador SVM foram utilizadas como imagem de consulta.

Foram considerados também valores de precisão para 10% de revocação, já que os radiologistas prestam mais atenção às primeiras imagens recuperadas pelo sistema. A relevância de cada imagem mamográfica recuperada para a imagem de consulta também foi obtida para as primeiras imagens recuperadas e apresentadas pelo experimento MammoSysLesion.

O desempenho da técnica 2DPCA foi comparada com as técnicas SVD e PCA para a caracterização do tecido da mama e SVM para a tarefa de recuperação de imagens.

5.4.1 Resultados e Discussão

A Tabela 5.5 apresenta a precisão média para os primeiros d componentes principais, comparando a caracterização do tecido e lesão da mama utilizando as técnicas 2DPCA, PCA e SVD, e também comparando os núcleos polinomial e radial do classificador SVM. As 564 imagens do conjunto de teste tiveram sua precisão calculada e então a média dessas precisões foi obtida.

Pode-se observar na Tabela 5.5, que utilizando o núcleo polinomial, a técnica 2DPCA obteve os maiores valores de precisão média, mesmo que muito próximos dos valores obtidos pela técnica SVD, sendo que retendo os primeiros 10 componentes principais o valor médio da precisão foi de 80,38%. Para o núcleo radial, a técnica SVD só não superou a técnica 2DPCA para os primeiros 1 e 4 componentes principais, sendo o maior valor da precisão média obtido por essa técnica o de 78,87% considerando os primeiros 7 componentes principais. A técnica 2DPCA obteve o maior valor da precisão média, de 80,64% considerando os primeiros 4 componentes principais. A textura do tecido da mama e da lesão mamográfica foi melhor representada pelas características extraídas usando a técnica 2DPCA, que foi capaz de capturar a diferença entre as intensidades de níveis de cinza entre os diferentes tecidos da mama juntamente com a lesão mamográfica e caracterizá-los.

A eficácia do caso de estudo proposto poderia ser aumentada caracterizando-se separadamente a lesão da mama, através, por exemplo, de atributos de forma, que poderiam ser concatenados ao atributo de textura que caracteriza os tecidos da mama.

No caso desse caso de estudo, a inserção de mais componentes principais para a

Tabela 5.5. Precisão média para os primeiros d componentes principais selecionados comparando as técnicas 2DPCA, PCA e SVD para a caracterização do tecido e lesão da mama e comparando os dois núcleos, polinomial e radial, do classificador SVM.

d	Característica	Polinomial	Radial
1	2DPCA	78,14%	70,95%
	PCA	64,26%	67,01%
	SVD	63,04%	67,79%
2	2DPCA	78,54%	72,14%
	PCA	63,13%	67,92%
	SVD	74,83%	76,03%
3	2DPCA	78,28%	74,54 %
	PCA	65,45%	70,39%
	SVD	77,06%	77,75%
4	2DPCA	79,1%	**80,64 %**
	PCA	66,85%	68,54%
	SVD	77,94%	77,44%
5	2DPCA	79,43%	76,22%
	PCA	66,23%	69,63%
	SVD	77,51%	77,67%
6	2DPCA	79,26%	76,14%
	PCA	66,1%	70,23%
	SVD	78,92%	76,24%
7	2DPCA	78,86%	75,15%
	PCA	66,69%	67,25%
	SVD	78,28%	78,87%
8	2DPCA	80%	75,34%
	PCA	66,91%	69,72%
	SVD	78,09%	77,92%
9	2DPCA	80,12%	75,82%
	PCA	65,86%	69,19%
	SVD	79,03%	77,75%
10	2DPCA	**80,38%**	76,88%
	PCA	67,57%	71,17%
	SVD	78,79%	77,49%

caracterização do tecido e lesão da mama, no caso das técnicas PCA e SVD, não decaiu o valor da precisão média como no caso de estudo anterior, o MammoSys, indicando que apesar dos menores valores dos componentes principais normalmente serem ruído, eles ou foram considerados capazes de caracterizar cada uma das 12 classes para então o classificador SVM ser capaz de separá-las corretamente, tanto com a utilização do núcleo polinomial quanto com o núcleo radial, ou confundiram o classificador SVM, já

que com um número maior de classes, o limite de diferenciação entre elas é menor. Novamente, a técnica 2DPCA mostrou-se, em geral, numericamente invariante ao número de componentes principais, indicando a possibilidade de reduzir a dimensionalidade do vetor de características e ainda representar as características do tecido e lesão da mama retendo-se um menor número de componentes principais.

De acordo com a Tabela 5.6, que mostra o tempo de execução em segundos do processo de recuperação do caso de estudo MammoSysLesion, fazendo as comparações entre as três técnicas – 2DPCA, PCA e SVD – e entre os dois núcleos – polinomial e radial, é possível notar que ao utilizar-se a técnica 2DPCA e o núcleo radial o tempo de execução é muito menor que o tempo de execução usando a técnica 2DPCA e o núcleo polinomial, que por ser minutos torna-se inviável para a implementação de um sistema CBIR que possa ser usado pelos radiologistas, fornecendo as imagens recuperadas e relevantes em um tempo pequeno. Apesar do núcleo polinomial ter se adequado aos dados, a obtenção dos vetores de suporte foi custosa, ao contrário do núcleo radial que foi capaz de usar o número maior de classes para determinar o centro das funções da base radial, e com a obtenção dos vetores de suporte utilizar a capacidade de generalização do classificador SVM para agrupar corretamente os dados do conjunto de teste. Por isso, a escolha do núcleo radial torna-se mais apropriada.

A função núcleo aumenta o poder de classificação do SVM por flexibilizar a forma da superfície de separação dos dados, e também o classificador SVM depende da qualidade dos dados de treinamento para extrair bons modelos para a determinação da relevância das imagens. Nesse experimento, no caso da caracterização do tecido da mama juntamente com a lesão mamográfica, não houve a preocupação da identificação da localização da lesão para caracterizá-la separadamente. As lesões aparecem nas imagens, como é visto na Figura 5.6 (b), como regiões mais brancas que o tipo de tecido, sendo que em tipos de tecidos densos, a lesão pode se ocultar, já que esses tipos de tecidos aparecem como regiões mais brancas nas imagens. Os dados de treinamento do SVM gerados pelas características extraídas das images pelas técnicas 2DPCA, PCA e SVD foram capazes de definir os vetores de suporte necessários para a definição do hiperplano de separação dos dados nas diferentes classes, porém essa definição dos vetores de suporte levou mais tempo para o núcleo polinomial que para o núcleo radial. Também, os vetores de características contendo os componentes principais obtidos pela técnica 2DPCA, ou seja, vetores, foram capazes de caracterizar juntamente o tecido da mama e a lesão mamográfica tão bem quando os componentes principais obtidos pela técnica SVD, que são escalares. Com a inserção de outros atributos para caracterizar separadamente a lesão da mama, mais testes seriam necessários para determinar a superioridade ou não da técnica 2DPCA sobre a técnica SVD.

Tabela 5.6. Tempo de execução, em segundos, do sistema CBIR comparando as técnicas 2DPCA, PCA e SVD para a caracterização do tecido da mama e lesão juntamente com a SVM com o núcleo polinomial e radial para o processo de recuperação, para todos os primeiros d componentes principais testados.

d	Característica	Polinomial	Radial
1	2DPCA	8.000,00	3,6
	PCA	1,3	1,3
	SVD	1,7	1,6
2	2DPCA	7.600,00	4,2
	PCA	1,4	1,4
	SVD	1,1	1,1
3	2DPCA	6.800,00	4,3
	PCA	1,4	1,5
	SVD	1,2	1,2
4	2DPCA	6.200,00	4,3
	PCA	1,5	1,5
	SVD	1,2	1,1
5	2DPCA	4.800,00	4,4
	PCA	1,5	1,6
	SVD	1,2	1,1
6	2DPCA	4.300,00	4,4
	PCA	1,5	1,6
	SVD	1,2	1,2
7	2DPCA	2.500,00	4,3
	PCA	1,5	1,6
	SVD	1,2	1,2
8	2DPCA	1.700,00	4,5
	PCA	1,6	1,6
	SVD	1,2	1,2
9	2DPCA	845,00	4,6
	PCA	1,8	1,6
	SVD	1,2	1,2
10	2DPCA	601,00	4,9
	PCA	1,8	1,8
	SVD	1,4	1,3

A Figura 5.7 mostra, considerando-se a precisão média, o gráfico de precisão $\times$ revocação usando os quatro primeiros valores dos componentes principais para as técnicas 2DPCA, PCA e SVD para a caracterização do tecido da mama e SVM com o núcleo radial para a tarefa de recuperação de imagens.

Com respeito à técnica 2DPCA e considerando-se uma imagem de consulta, para 10% de revocação, uma precisão de 83% significa que de 56 imagens retornadas pelo sis-

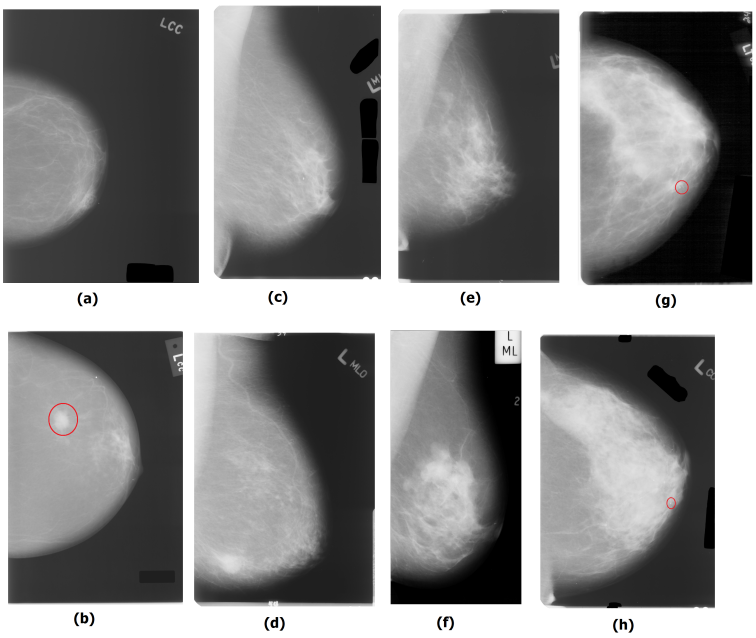

Figura 5.6. Imagens mamográficas de diferentes tipos de tecido e lesão: (a) Extremamente gordurosa com lesão benigna, (b) Extremamente gordurosa com lesão maligna, (c) Gordurosa com algum tecido fibroglandular com lesão benigna, (d) (c) Gordurosa com algum tecido fibroglandular com lesão maligna, (e) Heterogeneamente densa com lesão benigna, (f) Heterogeneamente densa com lesão maligna (g) Extremamente densa com lesão benigna e (h) Extremamente densa com lesão maligna.

tema CBIR MammoSysLesion, 47 imagens são relevantes à consulta do usuário. Como os tipos de tecidos densos podem ocultar certas lesões, é mais difícil a caracterização em conjunto do tipo de tecido e tipo de lesão da mama, visto que ambos aparecem como regiões mais brancas nas imagens, como pode ser visto nas Figuras 5.6 (g) e (h). Já considerando-se a técnica SVD e uma imagem de consulta, para 10% de revocação, uma precisão de 80% significa que de 56 imagens retornadas pelo sistema CBIR MammoSysLesion, 44 imagens são relevantes à consulta do usuário.

A Figura 5.8 mostra um exemplo do sistema MammoSysLesion. A imagem mamográfica de consulta possui o tecido da mama pertencente à categoria BI-RADS III – tecido denso heterogeneamente – e a lesão mamográfica pertencente à categoria BI-RADS 5 – lesão maligna – (primeira imagem no alto e à esquerda). As imagens foram recuperadas com base na caracterização do tecido e lesão da mama com a técnica 2DPCA e a seleção dos quatro primeiros valores dos componentes principais, e na recuperação utilizando SVM com o núcleo radial. Essa imagem de consulta foi sorteada entre todas as imagens do conjunto de testes.

Com exceção das imagens 6 (BI-RADS II e BI-RADS 1 – tecido da mama gor-

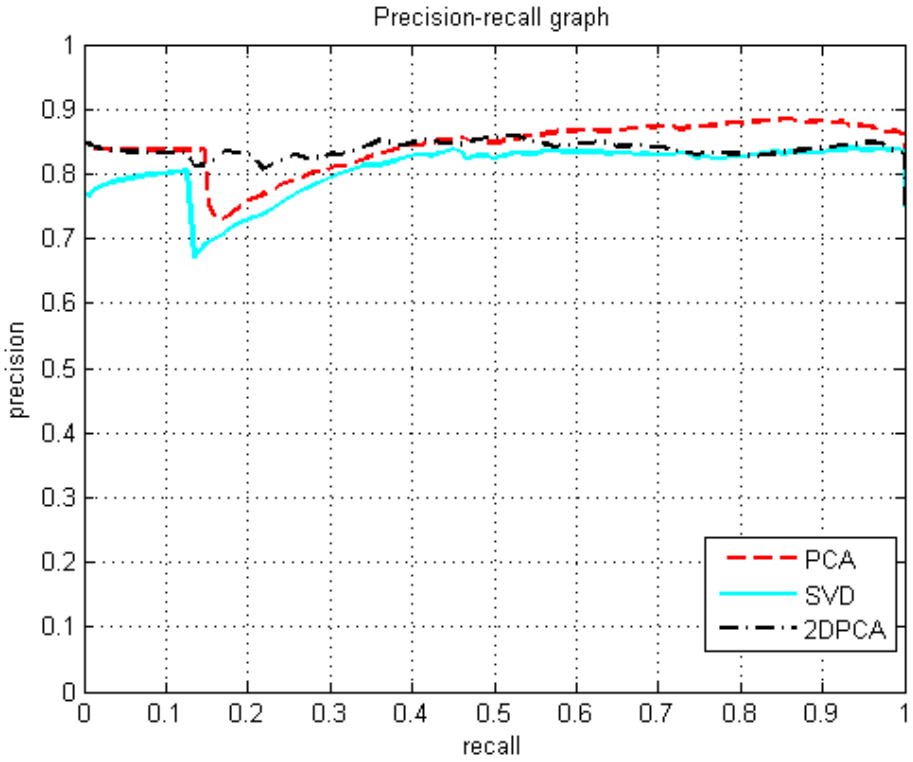

Figura 5.7. Curva precisão × revocação comparando as técnicas 2DPCA, PCA e SVD para a caracterização do tecido da mama e SVM para o processo de recuperação considerando-se a precisão média.

duroso com algum tecido fibroglandular sem lesão) e 8 (BI-RADS IV e BI-RADS 5 – tecido da mama extremamente denso com lesão maligna), todas as outras imagens recuperadas pertencem à mesma categoria do tipo de tecido da mama da imagem de consulta. Apesar de nem todas as imagens recuperadas possuírem a lesão pertencendo à mesma categoria da imagem de consulta, isso acontece pelo fato das lesões malignas aparecerem nas imagens com regiões contendo mais ramificações que as lesões benignas, e também ambas as lesões aparecerem como regiões mais brancas nas imagens. O descritor de textura não foi suficiente para captar as diferenças entre elas.

Esse mesmo exemplo da Figura 5.8 é apresentado na Figura 5.9 com a indicação da relevância de cada imagem mamográfica recuperada para a imagem mamográfica de consulta. A avaliação do sistema CBIR MammoSysLesion através da indicação da relevância mostra que, considerando o valor médio da relevância das primeiras imagens recuperadas, essas imagens mamográficas recuperadas que estiverem abaixo desse valor médio, apesar de serem relevantes à consulta, não são tão similares à imagem de consulta e portanto não devem ser apresentadas ao usuário. Nesse exemplo, considerando as oito primeiras imagens mamográficas recuperadas, o valor médio da relevância obtido

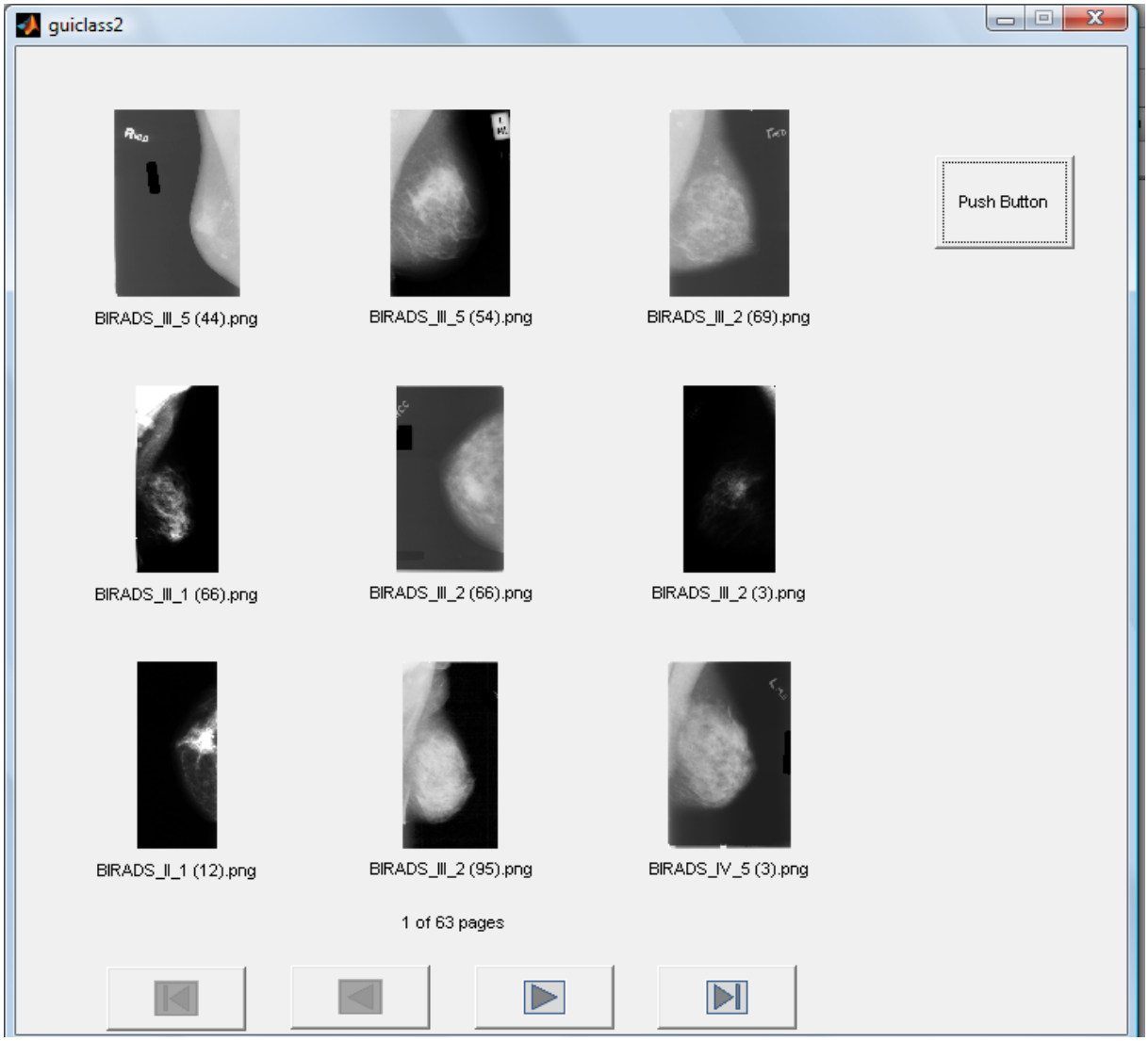

Figura 5.8. Exemplo do caso de estudo MammoSysLesion para a recuperação de imagens mamográficas com base no tipo de tecido e lesão da mama, utilizando os quatro primeiros componentes principais da técnica 2DPCA e SVM com o núcleo radial.

pelo classificador SVM foi de 1,0180. Considerando esse valor médio de relevância, seriam consideradas como imagens similares à imagem de consulta, as imagens recuperadas 1 – BI-RADS III e BI-RADS 5 (tecido denso heterogeneamente com lesão maligna), 2 – BI-RADS III e BI-RADS 2 (tecido denso heterogeneamente com lesão benigna) e 3 – BI-RADS III e BI-RADS 1 (tecido denso heterogeneamente sem lesão). Nota-se que, visualmente, todas as imagens são similares e que, apesar da terceira imagem recuperada não possuir lesão, seu tecido da mama aparentemente contém regiões mais brancas que as duas primeiras imagens recuperadas, que contêm lesões da mama. Nesse tipo de avaliação, as imagens recuperadas que não pertencem à mesma categoria BI-RADS,

para tecido da mama, da imagem de consulta, não seriam consideradas como imagens relevantes à consulta e não seriam apresentadas ao radiologista. Aqui, nesse exemplo, essas imagens foram mostradas apenas para exemplificar o experimento considerando a relevância das imagens.

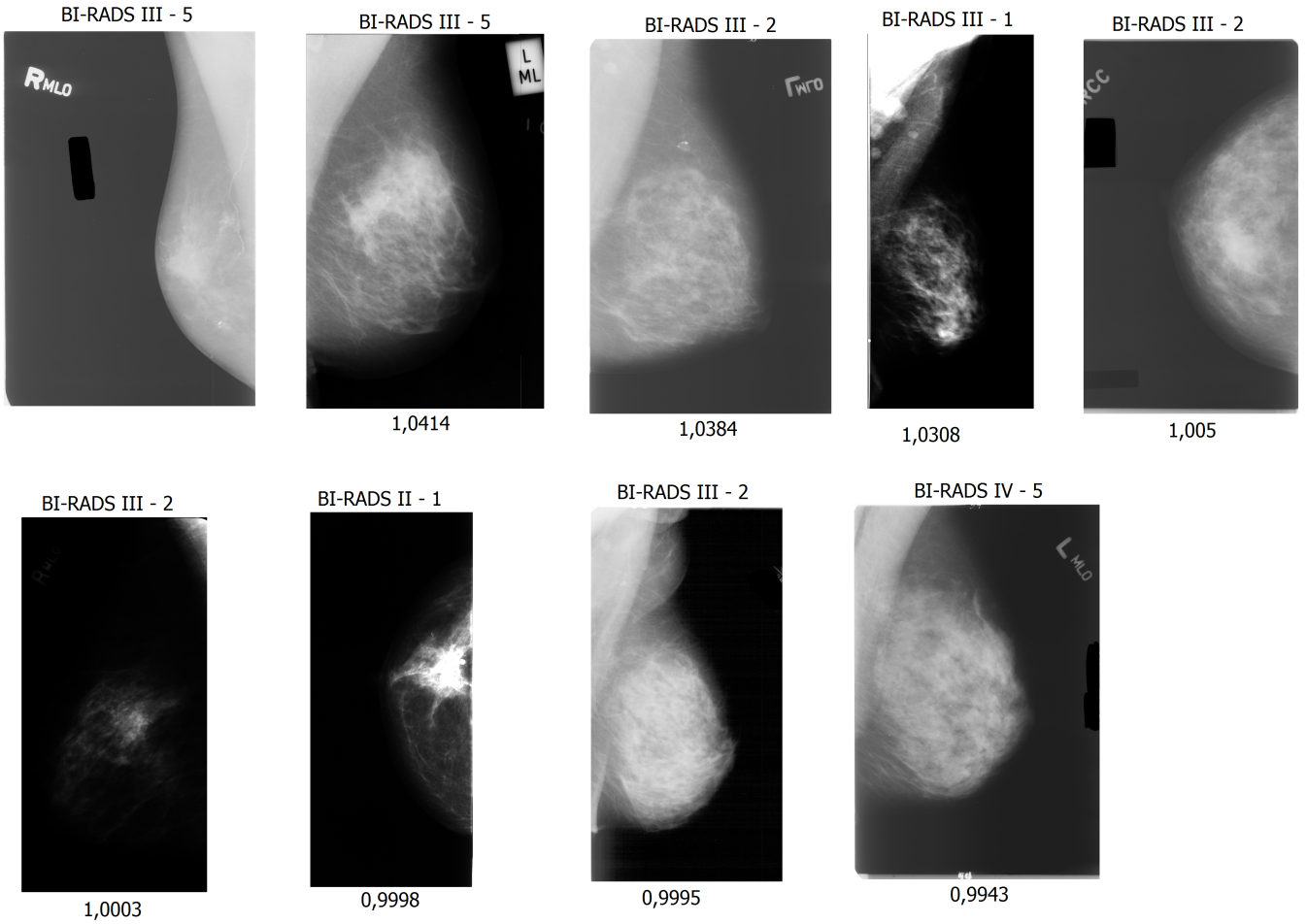

Figura 5.9. Exemplo do caso de estudo MammoSysLesion para a recuperação de imagens mamográficas com base no tipo de tecido e lesão da mama, utilizando os quatro primeiros componentes principais da técnica 2DPCA e SVM com o núcleo radial.

Capítulo 6

Conclusão

Neste capítulo, as principais contribuições e resultados obtidos com o projeto de doutorado são apresentadas. As publicações também são listadas. Algumas questões que surgiram ao longo do desenvolvimento do projeto são abordadas, assim como sugestões de direcionamento de trabalhos futuros.

6.1 Considerações Gerais

A proposta desse projeto foi contribuir para a área de recuperação com base no conteúdo visual de imagens mamográficas, fornecendo um sistema capaz de auxiliar radiologistas em seu diagnóstico ou um sistema para servir como um estágio de pré-processamento em aplicações de sistemas de auxílio ao diagnóstico para detecção de lesões mamográficas.

A extração de características de um conjunto de imagens que permita descrever de maneira efetiva cada região contida em uma imagem a partir de uma sequência pequena de valores numéricos é uma tarefa complexa e afeta os processos subseqüentes de um sistema CBIR. Nesse trabalho, a técnica análise dos componentes principais em duas dimensões foi introduzida para a caracterização da textura das imagens mamográficas, captando a diferença entre as intensidades dos níveis de cinza e também a distribuição espacial dos níveis de cinza com as variações de brilho.

Experimentos foram realizados de forma a determinar o número de componentes principais necessários para caracterizar a textura ao mesmo tempo em que a redução de dimensionalidade do vetor de características é realizada. Foram comparadas as técnicas 2DPCA, PCA e SVD para a caracterização da textura, sendo o classificador SVM utilizado para o processo de recuperação das imagens. Os resultados mostraram que retendo-se aproximadamente quatro componentes principais, é possível com a técnica

2DPCA caracterizar a textura e reduzir a dimensionalidade do vetor de características. Também, melhores resultados são obtidos ao treinar-se o classificador SVM com mais dados.

Uma importante característica na realização dos experimentos é o uso para os testes de imagens mamográficas que possuem o padrão de ouro estabelecido, já que todas as imagens contidas na base de dados IRMA foram previamente classificadas por experientes radiologistas, facilitando visualmente a avaliação dos dois experimentos.

6.2 Principais Contribuições e Resultados

Especificamente, dois casos de estudo para o desenvolvimento do sistema CBIR foram propostos, implementados e avaliados:

- **MammoSys**: nesse caso de estudo, os tecidos da mama foram caracterizados de acordo com as quatro categorias BI-RADS através da técnica 2DPCA, introduzindo essa técnica para a caracterização da textura dos tecidos. Ao mesmo tempo que essa técnica caracteriza a textura, reduz a dimensionalidade do vetor de características, permitindo que o classificador SVM indique a relevância das imagens para determinada consulta e separe corretamente as diferentes quatro classes: BI-RADS I, BI-RADS II, BI-RADS III e BI-RADS IV. A técnica 2DPCA foi comparada com as técnicas PCA e SVD. Os resultados obtidos mostram uma média de precisão de 81,11%, para os primeiros quatro componentes principais da técnica 2DPCA, treinando o classificador SVM com o núcleo polinomial com 60% dos vetores de características e testando com 40% dos vetores de características. O tempo de execução do processo de recuperação foi de 30,75 segundos.

- **MammoSysLesion**: nesse caso de estudo, os tecidos da mama foram caracterizados juntamente com as lesões mamográficas, utilizando as categorias do padrão internacional BI-RADS. A técnica 2DPCA foi introduzida para a caracterização em conjunto do tecido e lesão da mama, gerando vetores de características com dimensionalidade reduzida para o classificador SVM indicar a relevância das imagens para determinada consulta. Nesse caso, SVM com o núcleo radial forneceu os melhores resultados, 80,64% e 78,87% de média de precisão para as técnicas 2DPCA e SVD, respectivamente. O tempo de execução do sistema utilizando os quatro primeiros componentes principais da técnica 2DPCA foi de 4,3 segundos.

Além disso, a integração de quatro bases de dados já existentes ao projeto IRMA foi apresentada, fornecendo mais de 10.000 imagens separadas em diversas categorias

e com o padrão de ouro estabelecido e verificado por um radiologista experiente, permitindo o desenvolvimento e avaliação de sistemas CBIR.

6.3 Publicações

Este projeto de doutorado gerou as seguintes publicações:

- Júlia E. E. de Oliveira; Mark O. Gueld; Ilja Bezrukov; Bastian Ott; Thomas M. Deserno; Arnaldo de A. Araújo. Building a standard reference database for a computer-aided mammography diagnosis system. In: X-Meeting, 2007, São Paulo. Anais do X-Meeting, 1p, 2007.

- Júlia E. E. de Oliveira; Mark O. Gueld; Bastian Ott; Arnaldo de A. Araújo; Thomas M. Deserno. Towards a standard reference database for computer-aided mammography. In: SPIE Medical Imaging, 2008, San Diego, USA. Proceedings of SPIE, 9p, 2008.

- Júlia E. E. de Oliveira; Thomas M. Deserno; Arnaldo de A. Araújo. Breast lesions classification applied to a reference database. In: 2nd International Conference: E-Medical Systems (E-Medisys), 2008, Sfax. E-Medisys, 7p, 2008.

- Júlia E. E. de Oliveira; Ana Paula B. Lopes, Guillermo C. Chavez; Thomas M. Deserno; Arnaldo de A. Araújo. MammoSVD: a Content-Based Image Retrieval System Using a Reference Database of Mammographies. Proceedings of the 22nd IEEE International Symposium on Computer-Based Medical Systems, Albuquerque, USA, 4p, August 2009.

6.4 Proposta de Trabalhos Futuros

Novos trabalhos que venham a complementar ou dar continuidade a este são indicados a seguir.

- Utilização de outros atributos, como por exemplo, o histograma de níveis de cinza, em conjunto com o atributo de textura, para caracterizar os tecidos da mama. Os dois atributos podem ser concatenados para a formação do vetor de características das imagens e os pesos dos atributos, para a identificação da importância de cada um para a caracterização, podem ser obtidos.

- Caracterização, em separado, das lesões mamográficas, através de características morfológicas, nas quais o tamanho da lesão, forma e contorno são descritas.
- Verificação de outras funções de similaridade a fim de se identificar quais melhor se ajustam ao problema de recuperação de imagens mamográficas e aos vetores de características obtidos com as técnicas propostas.
- Recuperação das imagens mamográficas considerando-se a projeção utilizada no imageamento – crânio-caudal ou médio-lateral, visto que algumas lesões podem ser observadas ou estarem mais visíveis em apenas um tipo de projeção.

Referências Bibliográficas

[Acharya and Ray, 2005] Acharya, T. and Ray, A. K. (2005). *Image Processing: principles and applications*. John Wiley Sons.

[Akay, 2009] Akay, M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems with Applications*, 36:3240–3247.

[Andrews and Patterson, 1976] Andrews, H. C. and Patterson, C. L. (1976). Singular value decomposition and digital image processing. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-24(1):26–53.

[Arodz et al., 2005] Arodz, T., Kurdziel, M., Sevre, E. O. D., and Yuen, D. A. (2005). Pattern recognition techniques for automatic detection of suspicious-looking anomalies in mammograms. *Computer Methods and Programs in Biomedicine*, 79:135–149.

[Baeza-Yates and Neto, 1999] Baeza-Yates, R. and Neto, B. R. (1999). *Modern Information Retrieval*. Addison-Wesley Professinal.

[Belhumeur et al., 1997] Belhumeur, P. N., Hespanha, J. P., and Kriengman, D. J. (1997). Eigenfaces vs. fisherfaces: recognition using class specific linear projections. *IEEE Trans. PAMI*, 19(7):711–720.

[Bovis and Singh, 2002] Bovis, K. and Singh, S. (2002). Classification of mammographic breast density using a combined classifier paradigm. *Medical Image Understanding and Analysis (MIUA) Conference*.

[Brodatz, 1966] Brodatz, P. (1966). Textures: a photographic album for artists and designers. *New York, Dover*.

[Burges, 1998] Burges, C. J. C. (1998). A tutorial on support vector machines for pattern recognition. *Data Mining and Knowledge Discovery*, 2:121–167.

[Campanini et al., 2004] Campanini, R., Dongiovanni, D., Iampieri, E., Lanconelli, N., Masotti, M., Palermo, G., Riccardi, A., and Roffilli, M. (2004). A novel featureless approach to mass detection in digital mammograms based on support vector machines. *Physics in Medicine and Biology*, 49:961–975.

[Castella et al., 2007] Castella, C., Kinkel, K., Eckenstein, M. P., Sottas, P.-E., Verdun, F. R., and Bochud, F. O. (2007). Semiautomatic mammographic parenchymal patterns classification using multiple statistical features. *Academic Radiology*, 14:1486–1499.

[Castelli and Bergman, 2001] Castelli, V. and Bergman, L. D. (2001). *Image Databases - search and retrieval of digital imagery.* Wiley-Interscience, USA.

[Chang and Lin, 2001] Chang, C.-C. and Lin, C.-J. (2001). LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[Crammer and Singer, 2000] Crammer, K. and Singer, Y. (2000). On the learnability and design of output codes for multiclass problems. *Computer Learning Theory*, pages 35–46.

[de Oliveira et al., 2008] de Oliveira, J. E. E., Güld, M., de Albuquerque Araújo, A., Ott, B., and Deserno, T. (2008). Towards a standard reference database for computer-aided mammography. In *Proceedings of SPIE Medical Imaging*, volume 6915, page 69151Y, USA.

[del Bimbo, 1999] del Bimbo, A. (1999). *Visual Information Retrieval.* Morgan Kaufmann Publishers Inc., USA.

[Doi, 2007] Doi, K. (2007). Computer-aided diagnosis in medical imaging: historical review, current status and future potential. *Computerized Medical Imaging and Graphics*, 31:198–211.

[Dudda et al., 2001] Dudda, R. O., Hart, P. E., and Stork, D. G. (2001). *Pattern Classification.* John Wiley Sons.

[Eltonsy et al., 2007] Eltonsy, N. H., Tourassi, G. D., and Elmaghraby, A. S. (2007). A concentric morphology model for the detection of masses in mammography. *IEEE Transactions on Medical Imaging*, 26(6):880–889.

[Fawcett, 2004] Fawcett, T. (2004). *ROC graphs: notes and practical considerations for researchers.* Kluwer Academics.

[Golub, 1983] Golub, G. H. (1983). *Matrix computations*. Johns Hopkins series in the matematicals sciences.

[Gonzalez et al., 2003] Gonzalez, R. C., Woods, R. E., and Eddins, S. L. (2003). *Digital Image Processing using Matlab*. Prentice-Hall.

[Hamad and Taouil, 2006] Hamad, N. B. and Taouil, K. (2006). Exploring wavelets subband decomposition toward a computer aided detection of microcalcification in breast cancer. In *The 2nd International Conference on Distributed Frameworks for Multimedia Applications*, pages 1–8.

[Haykin, 1999] Haykin, S. (1999). *Neural Networks: a comprehensive foundation*. Prentice-Hall, 2nd edition.

[Hearst et al., 1998] Hearst, M., Schölkopf, B., Dumais, S., Osuna, E., and Platt, J. (1998). Trends and controversies – support vector machines. *IEEE Inteligent Systems*, 13(4):18–28.

[Heath et al., 1998] Heath, M., Bowyer, K., and et al., D. K. (1998). Current status of the digital database for screening mammography. *In: Digital Mammography, Kluwer Academic Publishers*, pages 457–460.

[H.S.Sheshadri, 2006] H.S.Sheshadri (2006). Breast tissue classification using statistical feature extraction of mammograms. *Medical Imaging and Information Sciences*, 23(3):105–107.

[Hsu and Lin, 2002] Hsu, C.-W. and Lin, C.-J. (2002). A comparison of methods for multiclass support vector machines. *IEEE Transactions on Neural Networks*, 13(2):415–425.

[INCA, 2009] INCA (2009). Instituto Nacional de Câncer. http://www.inca.gov.br.

[Kinoshita et al., 2007] Kinoshita, S. K., de Azevedo Marques, P. M., Jr, R. R. P., Rodrigues, J. A. H., and Rangayyan, R. M. (2007). Content-based retrieval of mammograms using visual features related to breast density patterns. *Journal of Digital Imaging*, 20(2):172–190.

[Kohonen, 1990] Kohonen, T. (1990). The self-organizing map. *Proceedings IEEE*, pages 1464–1480.

[Kullback and Leibler, 1951] Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86.

[Élden, 2006] Élden, L. (2006). Numerical linear algebra in data mining. *Acta Numerica*, pages 327–384.

[Lehmann et al., 2005] Lehmann, T. M., Güld, M. O., Deselaers, T., Keysers, D., Schubert, H., Spitzer, K., Ney, H., and Wein, B. (2005). Automatic categorization of medical images for content-based image retrieval and data mining. *Computerized Medical Imaging and Graphics*, 29(2):143–155.

[Lehmann et al., 2003] Lehmann, T. M., Schubert, H., Keysers, D., Kohnen, M., and Wein, B. (2003). The IRMA code for unique classification of medical images. In *Proceedings of SPIE*, volume 5033, pages 440–451.

[Müller et al., 2004] Müller, H., Michoux, N., Bandon, D., and Geissbuhler, A. (2004). A review of content-based image retrieval systems in medical applications - clinical benefits and future directions. *International Journal of Medical Informatics*, 73:1–23.

[Müller et al., 2001] Müller, K., Mika, S., Rätsch, G., Tsuda, K., and Schölkopf, B. (2001). An introduction to kernel-based learning algorithm. *IEEE Transactions on Neural Networks*, 12(2):181–201.

[Mumtaz et al., 2006] Mumtaz, A., Gilani, S. A. M., and Jameel, T. (2006). A novel texture image retrieval system based on dual tree complex wavelet transform and support vector machines. In *2nd International Conference on Emerging Technologies*, pages 108–114, Peshawar, Pakistan. IEEE - ICET 2006.

[Nakayama et al., 2006] Nakayama, R., Uchiyama, Y., Yamamoto, K., Watanabe, R., and Namba, K. (2006). Computer-aided diagnosis scheme using a filter bank for detection of microcalcification clusters in mammograms. *IEEE Transactions on Biomedical Engineering*, 53(2):273–283.

[Naqa et al., 2004] Naqa, I. E., Yang, Y., Galatsanos, N. P., Nishikawa, R. M., and Wernick, M. N. (2004). A similarity learning approach to content-based image retrieval: application to digital mammography. *IEEE Transactions on Medical Imaging*, 23(10):1233–1244.

[Naqa et al., 2002] Naqa, I. E., Yang, Y., Galatsanos, N. P., and Wernick, M. N. (2002). Content-based image retrieval for digital mammography. In *International Conference on Image Processing*, pages 141–144. Proceedings of the 2002.

[Neudecker and Magnus, 1988] Neudecker, H. and Magnus, J. R. H. . (1988). *Matrix diferential calculus with applications in statistics and econometrics*. John Wiley Sons, New York.

[Oliveira et al., 2007] Oliveira, M. C., Cirne, W., and de Azevedo Marques, P. M. (2007). Towards applying content-based image retrieval in the clinical routine. *Future Generation Computer Systems*, 23(3):466–474.

[Oliver et al., 2008] Oliver, A., Freixenet, J., Martí, R., Pont, J., Pérez, E., Denton, E. R., and Zwiggelaar, R. (2008). A novel breast tissue density classification methodology. *IEEE Transactions on Information Technology in Biomedicine*, 12(1):55–65.

[Oliver et al., 2009] Oliver, A., Lladó, X., Pérez, E., , Denton, E. R., Freixenet, J., and Martí, J. (2009). A statistical approach for breast density segmentation. *Journal of Digital Imaging*, Epub ahead of print - Published online.

[Pedrini and Schwartz, 2008] Pedrini, H. and Schwartz, W. R. (2008). *Analise de imagens digitais: princípios, algoritmos e aplicações*. Thomson Learning.

[Rahman et al., 2005] Rahman, M., Desai, B. C., and Bhattacharya, P. (2005). Supervised machine learning based medical image annotation and retrieval. In *Image CLEFmed*, pages 692–701.

[Rahman et al., 2004] Rahman, M., Wang, T., and Desai, B. C. (2004). Medical image retrieval and registration: towards computer assisted diagnostic approach. In *IDEAS Workshop on Medical Information Systems: the Digital Hospital (IDEAS-DH'04)*, pages 78–89, Washington DC. IEEE Computer Science.

[Russ, 2007] Russ, J. C. (2007). *The Image Processing Handbook*. CRC Taylor & Francis.

[Salton and McGill, 1983] Salton, G. and McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill Book Company, USA.

[Smola and Schölkopf, 2002] Smola, A. and Schölkopf, B. (2002). Learning with kernels. *The MIT Press*.

[Strang, 1993] Strang, G. (1993). *Introduction to Algebra Linear*. Wellesley-Cambridge Press, USA.

[Suckling, 1994] Suckling, J. (1994). The mammographic image analysis society digital datagram database. *Exerpta Medica International Congress Series*, 1069:375–378.

[Tagliafico et al., 2009] Tagliafico, A., Tagliafico, G., Tosto, S., Chiesa, F., Martinoli, C., Derchi, L. E., and Calabrese, M. (2009). Mammographic density estimation: comparison among BI-RADS categories, a semi-automated software and a fully automated one. *The Breast*, 18:35–40.

[van Rijsbergen, 1979] van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworth & Co, UK.

[Vapnik, 1995] Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer-Verlag, New York.

[Verma, 2008] Verma, B. (2008). Novel network architecture and learning algorithm for the classification of mass abnormalities in digitized mammograms. *Artificial Intelligence in Medicine*, 42:67–79.

[Wang et al., 2003] Wang, X. H., Good, W. F., Chapman, B. E., Chang, Y.-H., Poller, W. R., Chang, T. S., and Hardesty, L. A. (2003). Automated assestement of the composition of breast tissue revealed on tissue-thickness-corrected mammography. *American Journal of Roentgenology*, 180:257–262.

[Watkins, 1991] Watkins, D. S. (1991). *Fundamentals of matrix computations*. John Wileys Sons.

[Wei et al., 2007] Wei, C., Li, Y., and Li, C. (2007). Effective extraction of Gabor features for addaptative mammogram retrieval. In *International Conference on Multimedia and Expo*, pages 1503–1506.

[Wei et al., 2006] Wei, L., Yang, Y., Nishikawa, R. M., and Wernick, M. N. (2006). Mammogram retrieval by similarity learning from experts. In *IEEE International Conference on Image Processing*, pages 2517–2520. IEEE.

[Wong and Hsu, 2006] Wong, W.-T. and Hsu, S.-H. (2006). Application of SVM and ANN for image retrieval. *European Journal of Operational Research*, 173(3):938–950.

[Xue and Michels, 2007] Xue, F. and Michels, K. B. (2007). Intrauterine factors and risk of breast cancer: a systematic review and meta-analysis of current evidence. *Lancet Oncology*, 8:1088–1100.

[Yang et al., 2004] Yang, J., Zhang, D., Frangi, A. F., and yu Yang, J. (2004). Two-dimensional PCA: a new approach to apperance-based face representation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(1):131–137.

[Yang et al., 2007] Yang, Y., Wei, L., and Nishikawa, R. M. (2007). Microcalcification classification assisted by content-based image retrieval for breast cancer diagnosis. In *IEEE International Conference on Image Processing*, volume 5, pages 1–4. IEEE.

[Zhang et al., 2001] Zhang, L., Lin, F., and Zhang, B. (2001). Support vector machine learning for image retrieval. In *Proceeding of the 9th ACM International Multimedia Conference*, volume 9, pages 107–118, Canada.

[Zuo et al., 2006] Zuo, W., Zhang, D., and Wang, K. (2006). An assembled matrix distance metric for 2DPCA-based image recognition. *Pattern Recognition Letters*, 27:210–216.

Apêndice A

Revisão de Álgebra Linear

Uma matriz $\mathbf{A}$ $m \times n$ é uma tabela de mn números dispostos em m linhas e n colunas [Golub, 1983, Strang, 1993]:

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & \cdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

A i-ésima linha de $\mathbf{A}$ é:

$$[a_{i1}, a_{i2} \cdots a_{in}]$$

para $i = 1, \cdots, m$, e a j-ésima coluna de $\mathbf{A}$ é:

$$\begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{pmatrix}$$

para $j = 1, \cdots, n$. A notação $\mathbf{A} = (a_{ij})_{m \times n}$ também é usada. Pode-se dizer que a_{ij} é o elemento ou a entrada de posição i, j da matriz $\mathbf{A}$.

Se $m = n$, pode-se dizer que $\mathbf{A}$ é uma matriz quadrada de ordem n e os elementos $a_{11}, a_{22}, \cdots, a_{nn}$ formam a diagonal principal de $\mathbf{A}$.

A transposta da matriz $\mathbf{A} = (a_{ij})_{m \times n}$ é definida pela matriz nxm

$$\mathbf{B} = \mathbf{A}^T$$

e é obtida trocando-se as linhas com as colunas, ou seja,

$$b_{ij} = a_{ji}$$

para $i = 1, \cdots, n$ e $j = 1, \cdots, m$. Também é possível escrever $\mathbf{A}_{ij}^T = a_{ji}$.

Uma matriz identidade é uma matriz quadrada, que possui os valores 1 em sua diagonal e 0 no restante da matriz, e pode ser denotada por $\mathbf{I}_n$:

$$\mathbf{I}_n = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1 \end{pmatrix}$$

Uma matriz quadrada $\mathbf{A}$ é invertível ou não singular se existe uma matriz $\mathbf{B} = (b_{ij})_{n \times n}$ tal que:

$$\mathbf{AB} = \mathbf{BA} = \mathbf{I}_n$$

em que $\mathbf{I}_n$ é a matriz identidade. A matriz $\mathbf{B}$ é chamada inversa de $\mathbf{A}$. Se $\mathbf{A}$ não tem inversa, diz-se que $\mathbf{A}$ é não invertível ou singular.

O determinante é uma função da matriz quadrada que a reduz a um único número. O determinante de uma matriz $\mathbf{A}$ é denotado por $|\mathbf{A}|$ ou $det(\mathbf{A})$. Se $\mathbf{A}$ é uma matriz de tamanho 2×2, então:

$$|\mathbf{A}| = det(\mathbf{A}) = \begin{pmatrix} a & b \\ c & d \end{pmatrix} = ad - bc$$

Um vetor no espaço $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3)$ pode também ser escrito na notação matricial como uma matriz linha ou como uma matriz coluna:

$$\mathbf{V} = \begin{pmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \mathbf{v}_3 \end{pmatrix} \text{ ou } \mathbf{V} = [\mathbf{v}_1 \mathbf{v}_2 \mathbf{v}_3]$$

A norma de um vetor $\mathbf{V} = (\mathbf{v}_1, \cdots, \mathbf{v}_n)$ é dada por:

$$\|\mathbf{V}\| = \sqrt{\mathbf{v}_1^2 + \cdots + \mathbf{v}_n^2} = \sqrt{\sum_{i=1}^{n} \mathbf{v}_i^2}$$

Um vetor de norma igual a 1 é chamado de vetor unitário.

O produto escalar ou interno de dois vetores $\mathbf{V}$ e $\mathbf{W}$ é definido por

$$\mathbf{V} \cdot \mathbf{W} = \|\mathbf{V}\| \|\mathbf{W}\| \cos\theta$$

em que θ é o ângulo entre eles, ou 0 se $\mathbf{V}$ ou $\mathbf{W}$ é nulo.

Esses dois vetores $\mathbf{V}$ e $\mathbf{W}$ são ortogonais se $\mathbf{V} \cdot \mathbf{W} = 0$.

Para cada inteiro positivo n, o espaço vetorial $\mathbb{R}^n$ é definido pelo conjunto de todas as n-úplas ordenadas $X = (x_1, \cdots, x_n)$ de números reais. Um vetor $\mathbf{V} \in \mathbb{R}^n$ é uma combinação linear dos vetores $\mathbf{V}_1, \cdots, \mathbf{V}_k \in \mathbb{R}^n$, se existem escalares $x_1, \cdots, x_k$ que satisfaçam a equação:

$$x_1\mathbf{V}_1 + x_2\mathbf{V}_2 + \cdots + x_k\mathbf{V}_k = V$$

Diz-se que um conjunto $\mathcal{S} = \mathbf{V}_1, \cdots, \mathbf{V}_k$ de vetores do $\mathbb{R}^n$ é linearmente independente (L.I.) se a equação vetorial:

$$x_1\mathbf{V}_1 + x_2\mathbf{V}_2 + \cdots + x_k\mathbf{V}_k = \bar{0}$$

só possui a solução trivial, ou seja, a única forma de escrever o vetor nulo como combinação linear dos vetores $\mathbf{V}_1, \cdots, \mathbf{V}_k$ é aquela em que todos os escalares são iguais a zero. Caso contrário, isto é, se a equação vetorial possui solução não trivial, pode-se dizer que o conjunto $\mathcal{S}$ é linearmente dependente (L.D.).

Seja $\mathbb{W}$ um subespaço de $\mathbb{R}^n$. É dito que um subconjunto $V_1, \cdots, V_k$ de $\mathbb{W}$ é uma base de $\mathbb{W}$ se:

- $V_1, \cdots, V_k$ é um conjunto de geradores de $\mathbb{W}$, ou seja, todo vetor de $\mathbb{W}$ é combinação linear de $V_1, \cdots, V_k$ e,
- $V_1, \cdots, V_k$ é L.I.

O número de elementos de qualquer uma das bases de $\mathbb{W}$ é chamado de dimensão de $\mathbb{W}$.

Seja $\mathbf{A}$ uma matriz $m \times n$. O subespaço de $\mathbb{R}^n$ gerado pelas linhas de $\mathbf{A}$ é chamado espaço linha de $\mathbf{A}$, ou seja, o conjunto de todas as combinações lineares das linhas de $\mathbf{A}$. O subespaço de $\mathbb{R}^m$ gerado pelas colunas de $\mathbf{A}$ é chamado espaço coluna de $\mathbf{A}$, ou seja, o conjunto de todas as combinações lineares das colunas de $\mathbf{A}$.

Agora, seja $V_1, \cdots, V_k$ uma base do subespaço de $\mathbb{R}^n$. Pode-se dizer que $V_1, \cdots, V_k$ é uma base ortogonal, se $V_i \cdot V_j = 0$, para $i \neq j$, ou seja, se quaisquer dois vetores da base são ortogonais. E pode-se dizer que $V_1, \cdots, V_k$ é uma base ortonormal, se além de ser uma base ortogonal, $\| V_i \| = 1$, ou seja, o vetor V_i é unitário para $i = 1, \cdots, m$.

Uma função $\mathrm{T} : \mathbb{R}^n \to \mathbb{R}^m$ é uma transformação linear se

$$T(\alpha X) = \alpha T(X) \text{ e } T(X + Y) = T(X) + T(Y)$$

para todos $X, Y \in \mathbb{R}^n$ e todos escalares α.

Uma matriz $\mathbf{A}$ de tamanho $n \times n$ é diagonalizável se existem matrizes $\mathbf{P}$ e $\mathbf{D}$ tais que $\mathbf{A} = \mathbf{PDP}^{-1}$, ou equivalentemente $\mathbf{D} = \mathbf{P}^{-1}\mathbf{AP}$, em que $\mathbf{D}$ é uma matriz diagonal.

Seja $\mathbf{A}$ uma matriz de ordem n, denomina-se polinômio característico de $\mathbf{A}$ o polinômio $P(\lambda)$ obtido pelo cálculo de:

$$P(\lambda) = det(\mathbf{A} - \lambda\mathbf{I}).$$

A equação $P(\lambda) = 0$ é denominada equação característica de $\mathbf{A}$.

Os autovalores de uma matriz $\mathbf{A}$ são precisamente as soluções λ da equação característica.

Se $\lambda \in \lambda(\mathbf{A})$ então os vetores não-zero $x \in \mathbb{C}^n$ que satisfazem

$$\mathbf{A}x = \lambda x$$

são referidos como autovetores.

Além disso, se é definido o traço da matriz $\mathbf{A}$ (soma dos seus elementos diagonais) como:

$$\text{traço}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}, \quad A \in \mathbb{C}^{n \times n}$$

então:

$$\text{traço}(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i$$

Apêndice B

Base de Dados de Imagens Mamográficas integrada ao Projeto IRMA

A base de dados de imagens mamográficas do projeto IRMA foi desenvolvida a partir da união de outras quatro bases de dados de imagens mamográficas: DDSM (*Digital Database for Screening Mammography*), MIAS (Mammographic Image Analysis Society Digital Mammogram Database), LLNL (Lawrence Livermore National Laboratory) e imagens de rotina do Departamento de Radiologia Diagnóstica da Universidade Tecnológica de Aachen, Alemanha (RWTH – *Rheinische-Westfäliche Technische Hochschule*).

B.1 O sistema IRMA

Todas as imagens na base de dados do projeto IRMA são codificadas de acordo com um esquema uni-hierárquico e multi-eixos [Lehmann et al., 2003]. Os quatro eixos, cada um deles contendo de três a quatro posições hierárquicas, descrevem:

- *técnica*: modalidade de imageamento;
- *direção*: orientação do corpo;
- *anatomia*: região do corpo examinada e;
- *biossistema*: sistema biológico examinado,

resultando numa única sequência: TTTT–DDD-AAA–BBB.

Técnica. O código IRMA para técnica é TTTT = 11xx, onde 11 significa raio-x, radiografia simples, e as duas posições restantes são usadas para capturar a natureza das imagens (1 = imagem digital, 2 = imagem digitalizada) e sua resolução (exemplo, 42, 43.5, 50 ou 200 microns).

Direção. De acordo com a codificação, as direções para o imageamento da mama, isto é, crânio-caudal e médio-lateral, são denotadas por DDD = 310 (axial - crânio-caudal - não especificada) e DDD = 410 (outra orientação - oblíqua - não especificada), respectivamente.

Anatomia. O eixo para anatomia do código IRMA é utilizado para diferenciar direita de esquerda usando os códigos AAA = 610 (mama - lado direito - não especificado) e AAA = 620 (mama - lado esquerdo - não especificado), respectivamente.

Biossistema. O eixo para biossistema foi estendido de forma a capturar a densidade do tecido, patologia e descrição da lesão mamográfica. A primeira posição descreve o tipo de tecido de acordo com as classes ACR:

- 0: não especificada;
- 1-c: já em uso;
- d: a mama é praticamente toda gordurosa (*fat transparent*, ACR-1);
- e: densidade glandular com fibras dispersas (*fibroid glands*, ACR-2);
- f: densa de forma heterogênea (*heterogeneously dense*, ACR-3);
- g : extremamente densa (*extremely dense*, ACR-4);
- h : densa (ACR 3/4),

onde ACR 3/4 foi definida somente para a importação das imagens da base de dados MIAS.

A segunda posição captura a patologia, ou seja, o estágio do tumor, de acordo com a classificação BI-RADS:

- 0: necessita de avaliação adicional (não especificado, BI-RADS-0);
- 1: negativo (normal, BI-RADS-1);
- 2: benigno (BI-RADS-2);
- 3: provavelmente benigno (BI-RADS-3);

- 4: anormalidade suspeita (BI-RADS-4);
- 5: maligno (BI-RADS-5).

Finalmente, a terceira posição refere-se ao tipo de lesão e de acordo com o sistema BI-RADS são definidos oito tipos de lesões:

- 0: não especificado;
- 1: calcificação, não especificada;
- 2: microcalcificação;
- 3: macrocalcificação;
- 4: massa circunscrita;
- 5: massa espiculada;
- 6: outras massas;
- 7: distorção;
- 8: assimetria.

B.2 Integração das bases de dados de imagens mamográficas

Para todas as imagens mamográficas das bases de dados de imagens DDSM, MIAS, LLNL e RWTH, o código IRMA foi adaptado utilizando as descrições fornecidas pelas mesmas e uma visão geral destas bases de dados pode ser vista na Tabela B.1.

Tabela B.1. Resolução e tipo de imagem das bases de dados.

Base de dados	Resolução					Tipo de arquivo	
	x-min	x-max	y-min	y-max	bits	formato	padrão
DDSM	1411	5641	3256	7111	12	LJPEG	não
MIAS	334	1.000	802	1024	8	PNG	sim
LLNL	700	4494	2828	6874	12	ICS	não
RWTH	1582	4129	3382	5928	12	DICOM	sim

Os resultados da integração das quatro bases de dados ao sistema IRMA podem ser vistos nas Tabelas B.2 a B.4.

Tabela B.2. Estatística para a classe de tecidos após a integração.

Base de dados	Código IRMA BBB = xBB					Número de imagens
	d	e	f	g	h	total
DDSM	1.252	3.691	2.896	1.994	0	9.833
MIAS (antes da correção)	80(105)	84(104)	84(0)	74(0)	0(113)	322
LLNL	12	84	68	20	0	184
RWTH	48	78	42	2	0	170
IRMA	1.395	4.043	3.048	2.023	113	10.509

Tabela B.3. Estatística para a classe de patologia após a integração.

Base de dados	Código IRMA BBB = BxB						Número de imagens
	0	1	2	3	4	5	total
DDSM	0	6.181	1.848	0	0	1.804	9.833
MIAS (antes da correção)	0	206(209)	64(61)	0	0	52	322
LLNL	6	47	111	6	0	14	184
RWTH	2	69	87	6	6	0	170
IRMA	8	6.503	2.110	12	6	1.870	10.509

Tabela B.4. Estatística para a classe de lesão após a integração.

Base de dados	Código IRMA BBB = BBx									N^o de imagens
	0	1	2	3	4	5	6	7	8	total
DDSM	6.181	1.488	0	0	478	547	1.139	0	0	9.833
MIAS (antes da correção)	206(209)	23	0	3(0)	23	19	14	19	15	322
LLNL	65	112	4	0	0	0	0	0	3	184
RWTH	71	0	40	43	4	2	0	10	0	170
IRMA	6.523	1.623	44	46	505	568	1.153	29	18	10.509