DETECÇÃO DE MASSAS EM IMAGENS DA MAMA USANDO ÍNDICES DE DIVERSIDADE E ALGORITMOS DE SEGMENTAÇÃO EM GRAFO*

DETECTION OF MASSES IN BREAST IMAGES USING DIVERSITY INDEX AND GRAPH-BASED SEGMENTATION ALGORITHMS

DETECCIÓN DE MASAS EN IMÁGENES DE MAMA UTILIZANDO ÍNDICE DE DIVERSIDAD Y ALGORITMOS DE SEGMENTACIÓN EN GRAFO

André de Souza Moreira
Geraldo Braz Junior
Simara Vieira da Rocha
Aristófanes Correa Silva
Anselmo Cardoso Paiva

Resumo: O câncer de mama tem sido um dos tipos mais frequentes de câncer. Entre a população feminina, esta neoplasia é a principal causa da morte para indivíduos entre 35 e 55 anos de idade. Apesar de ainda não haver modos efetivos de prevenir o câncer de mama, o tratamento do câncer de mama em estágio inicial proporciona maiores chances de cura ao paciente, além de um tratamento menos agressivo. Por isso, a mamografia de rastreamento tem sido fundamental na detecção precoce desta neoplasia. Entretanto, alguns resultados destes exames são comprometidos por diversos fatores, entre eles a qualidade da imagem mamográfica. Neste cenário, a comunidade científica tem despendido esforços visando à construção de sistemas CAD e CADx a fim de dar suporte ao processo de detecção e diagnóstico do câncer de mama através de técnicas de processamento de imagens e visão computacional em imagens médicas. Este artigo apresenta uma proposta de metodologia para a construção de um sistema CAD/CADx que auxilie o processo de detecção e diagnóstico de massas em imagens da mama.
Palavras-chave: Câncer de mama. Mamografia. Detecção. Diagnóstico.

Abstract: Breast cancer has been one of the most frequent types of cancer. Among female population, this disease is the major cause of death for women between 35 and 55 years of age. Although there is still no effective ways to prevent breast cancer, the treatment of breast cancer at an early stage provides greater chances of cure for the patient, and less aggressive treatment. For this reason, screening mammography has been instrumental in the early detection of this malignancy. However, some results of these tests are compromised by several factors, including the quality of the mammographic image. In this scenario, the scientific community has made efforts aimed to building CAD/CADx systems to support the process of detection and diagnosis of breast cancer using techniques of image processing and computer vision in medical imaging. This article proposes a methodology for building a CAD/CADX to assist the process of detection and diagnosis of masses in breast imaging.
Keywords: Breast cancer. Mammography. Detection. Diagnosis.

Resumen: El cáncer de mama es uno de los tipos más frecuentes de cáncer. Dentro del grupo femenino, este tipo de cáncer es la principal causa de muerte en mujeres entre 35 y 55 años de edad. Aunque aún no hay formas efectivas de prevenir el cáncer de mama, el tratamiento del cáncer de mama en una etapa temprana proporciona mayores posibilidades de curación para el paciente, y menos tratamiento agresivo. Por esta razón, la mamografía ha sido fundamental en la detección precoz de esta neoplasia. Sin embargo, algunos de los resultados de estos ensayos se ve comprometida por varios factores, incluyendo la calidad de la imagen mamográfica. En este escenario, la comunidad científica ha realizado esfuerzos encaminados a la construcción de sistemas CAD / CADx para apoyar el proceso de detección y diagnóstico de cáncer de mama utilizando técnicas de procesamiento de imágenes y visión por computador en imágenes médicas. En este artículo se propone una metodología para la construcción de un sistema CAD / CADX para ayudar en el proceso de detección y diagnóstico de las masas en imágenes mamarias.
Palabras clave: Cáncer de mama. Mamografía. Detección. Diagnóstico.

1 INTRODUÇÃO

O termo câncer refere-se a um grande conjunto de doenças que afetam qualquer parte do corpo, tendo como principal característica o crescimento anormal da célula e que pode invadir outros órgãos. Tal processo é conhecido como metástase e é a principal causa de

Trabalho premiado durante o XXIV Encontro do SEMIC, realizado na UFMA entre os dias 05 a 08 de novembro de 2012.
*Artigo recebido em dezembro 2012
Aprovado em fevereiro 2013

morte por câncer (WORLD HEALTH ORGANIZATION, 2012).

No ano de 2008, o câncer levou 7.6 milhões de pessoas ao óbito no mundo todo, equivalente a 13% de todas as mortes ocorridas naquele ano, sendo que os principais tipos de câncer registrados naquele ano são o de pulmão, estômago, fígado, colo e o de mama. Cerca de 30% das mortes por câncer são devidos aos cinco principais riscos comportamentais e alimentares: índice de massa corporal elevado; baixa ingestão de frutas, legumes e verduras; a falta de atividade física; tabagismo e uso de álcool (WORLD HEALTH ORGANIZATION, 2012).

No Brasil, estimativas apontam que 518.510 novos casos de câncer serão registrados no ano de 2012, onde 257.870 casos surgirão no sexo masculino e 260.640 no grupo feminino. Para o grupo feminino, o câncer de mama será o predominante, representando 52.680 novos casos, 27,9% de todos os novos casos de câncer no grupo feminino (INSTITUTO NACIONAL DO CÂNCER,2012). Ainda segundo o Instituto Nacional do Câncer (2012), o estado do Maranhão encerrará o ano de 2012 com 460 novos casos de câncer de mama, sendo que 190 casos surgirão na capital, São Luís, representando uma taxa bruta de 35,65 de incidência para 100 mil habitantes.

A melhor forma de combater o câncer de mama ainda reside na detecção precoce do mesmo, daí a importância da realização da mamografia de rastreamento, principalmente para as mulheres acima dos quarenta anos de idade, dentro do grupo de risco (AZEVEDO; PEIXOTO, 1993). Quando encontrado em estágios iniciais, possibilita um tratamento menos agressivo e com maiores chances de cura ao paciente. Visando promover a saúde pública de qualidade, o governo brasileiro criou a Lei 11.664/2008 que entrou em vigor em 29 de abril de 2009 a qual dispõe sobre a efetivação de ações de saúde que assegurem a prevenção, a detecção, o tratamento e o seguimento dos cânceres do colo uterino e de mama, no âmbito do Sistema Único de Saúde - SUS. Além disso, diversas campanhas foram feitas a fim de conscientizar as mulheres, principalmente as que estão dentro do grupo de risco, da importância da mamografia.

Tais campanhas levaram ao recorde de 2.139.238 de mamografias realizadas em 2012 pelo Sistema Único de Saúde (SUS), representando um aumento de 41% no número de mamografias entre as mulheres na faixa prioritária (50 a 69 anos) se comparado ao mesmo período de 2010. No Maranhão, o Sistema Único de Saúde (SUS) realizou 28.984 mamografias no ano de 2012, representando, até então, um aumento de 10% se comparado com o mesmo período de 2011.

1.1 O problema

Apesar de a mamografia ser grande aliada na detecção precoce do câncer de mama, vários fatores têm interferido na análise da mesma, resultando em falhas nos laudos emitidos por radiologistas que variam entre 10% a 30% (BIRD; WALLACE; YANKASKAS, 1992). Em entrevista recente, o secretário de Atenção à Saúde, Helvécio Magalhães, do Ministério da Saúde, afirmou que algumas cidades do Brasil chegam a ter até 50% de dificuldade de leitura e de diagnóstico fazendo uso da mamografia.

Tais ocorrências levaram o Ministério da Saúde a instituir o Programa Nacional de Qualidade em Mamografia (PNQM) com o objetivo de minimizar os riscos associados ao uso dos raios-X, assim como garantir a qualidade no diagnóstico.

Diversos fatores contribuem para uma possível falha na detecção ou caracterização do câncer de mama, em geral podemos citar o uso de técnicas radiográficas inadequadas, lesão com características sutis ou incomuns, erro de interpretação e, principalmente, a restrição de resolução e contraste da imagem da mamografia que resultam na diminuição da visualização de tumores de mama e microcalcificações em pacientes com tecido fibroglandular denso (HUYNH; JAROLIMEK; DAYE, 1998).

1.2 Sistemas CAD/CADx

Diante deste cenário, técnicas de processamento de imagens e visão computacional têm sido adotadas para dar suporte aos especialistas da área, a fim de minimizar os erros envolvidos no diagnóstico médico em busca de lesões no parênquima mamário através do uso de médicas da mama. Tais técnicas têm servido como suporte para a construção de sistemas de detecção auxiliado por computador (CAD - Computer-Aided Detection) e sistemas de diagnóstico auxiliado por computador (CADx - Computer-Aided Diagnosis).

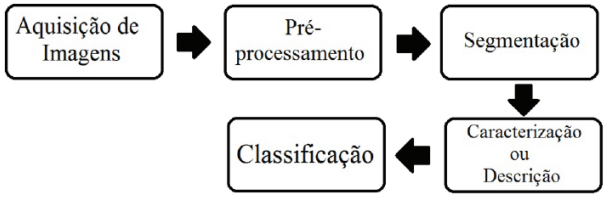
Esses sistemas têm sido de grande importância para diminuição dos erros dos diagnósticos, atuando como uma segunda opinião médica onde o especialista corrobora ou não com o laudo dado por estes sistemas. Pesquisas demonstram que os casos de acertos no laudo de câncer de mama poderiam ser aumentados de 5% a 15% caso sejam utilizados sistemas CAD (FREER; ULISSEY, 2001).

De forma geral, as metodologias utilizadas por sistemas CAD/CADx baseiam-se no fluxo das etapas do processamento de imagens, representado pela Figura 1. O processamento digital de imagens envolve processos cujas entradas e saídas são imagens e, além disso, envolve processos de extração de atributos de imagens até – e inclusive – o reconhecimento de objetos individuais.

Vários trabalhos têm sido propostos a fim de auxiliar a detecção precoce desta neoplasia. Em Giuliato, Barcellos e Dias, (2008) é utilizado um filtro de suavização anisotrópica via Equações Diferenciais Parciais na fase de pré-processamento para a detecção de regiões suspeitas em mamografias. Em Zhang (2011)

é proposto um algoritmo automático para detecção de tumores em imagens mamográficas através do uso da lógica de fuzzy.

Figura 1 - Representação das etapas do processamento de imagem



Fonte: Gonzalez e Woods (2000).

Este artigo propõe uma nova metodologia capaz de dar suporte ao diagnóstico médico na tarefa de detecção de anomalias na mama através do uso de imagem médica da mama. Para tal, é sugerido o uso de técnicas de realce da imagem analisada e o uso de algoritmos de segmentação em grafo para a extração de regiões de interesses. Por fim, algumas abordagens são utilizadas a fim de minimizar o número de regiões que serão analisadas e as remanescentes são caracterizadas através de índices de diversidades para classificá-las com o uso de máquinas de vetores de suporte.

Este artigo apresenta, primeiramente, conceitos teóricos acerca das técnicas empregadas na construção da metodologia proposta. Tais conceitos englobam o CLAHE e Mean-Shift para o realce da imagem, o uso de algoritmo de segmentação baseado em grafo, índices de diversidade de espécie com matriz de co-ocorrência e run-length para caracterização das regiões e o uso de máquinas de vetores de suporte para classificar as regiões sob análise. Na segunda parte, são apresentadas as etapas seguidas na metodologia utilizando os conceitos abordados previamente. Por fim, são exibidos os resultados obtidos nos testes realizados e uma discussão final sobre o assunto.

2 FUNDAMENTAÇÃO TEÓRICA

Esta seção descreve os principais conceitos teóricos empregados na construção da metodologia proposta neste trabalho.

2.1 Realce de imagem

Como discutido anteriormente, grande parte das mamografias realizadas apresentam baixa qualidade na imagem produzida, o que torna o processo de detecção de neoplasias mais difícil. Portanto, é importante para o sucesso da metodologia que todas as estruturas do parênquima mamário apresentem um bom contraste em relação aos outros elementos. Diante dessa tarefa, o CLAHE e o Mean-Shift têm se demonstrado eficazes para o a realce dessas estruturas e ambos são explicitados a seguir.

O Contrast Limited Adaptive Histogram Equalization, CLAHE (PISANO, 1998), objetiva realçar a imagem através da análise de diversos histogramas, onde cada um corresponde a uma região da imagem e então cada histograma é equalizado a fim de redistribuir os valores de luz da imagem. Tal técnica tem a vantagem de proporcionar um bom contraste na imagem, mesmo que a distribuição dos valores dos pixels não seja similar em torno de toda a imagem.

Já o Mean-Shift (COMANICIU; MEER, 2002) é uma técnica, um procedimento interativo que localiza os máximos de uma função de densidade a partir dos dados discretos amostrados dessa função. Para cada ponto, o algoritmo delimita uma região onde será calculado o centro de massa do mesmo. Encontrado o centro de massa, essa região é deslocada de tal forma que o centro dela fique sobre o local onde o centro de massa foi calculado.

Ambas as técnicas podem ser combinadas de tal forma que o resultado produzido consiste em estruturas bem definidas e com ausência de ruídos.

2.2 Segmentação baseada em grafo

A segmentação de uma imagem visa delimitar regiões de interesses presentes na imagem. Vários algoritmos e técnicas de segmentação têm sido propostos, cada um apresentando um comportamento e resultado peculiar. Em geral, as técnicas mais utilizadas por algoritmos de segmentação baseiam-se na análise de descontinuidade, presença de bordas e disposição da intensidade dos pixels na imagem.

Uma técnica que tem sido amplamente utilizada para segmentar imagens é o uso do particionamento de grafos. Dada uma imagem I, pode-se construir um grafo G = (V,E,W), com os pixels de I modelados como nós do grafo (V), e pixels com distância ≤ Gr são conectados por uma aresta do grafo (E). Um valor de peso W(i,j) mensura a probabilidade dos pixels i e j pertencerem à mesma região. A partição deste grafo provê a segmentação de regiões da imagem.

O particionamento de um grafo corresponde ao corte do mesmo originando subgrafos, em que e tal corte é associado a um valor algébrico que representa o grau de semelhança entre as componentes disjuntas geradas. Este valor pode ser calculado através da Equação 1.

Equação 1

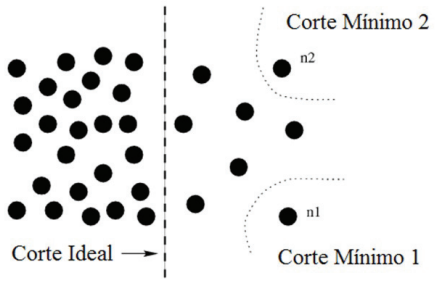
$$Corte(A,B) = \sum_{i \in A, j \in B} p(i,j)$$

onde p(i,j) corresponde ao peso da aresta que conectava os dois subgrafos gerados.

De modo geral, as técnicas de segmentação baseada em grafo tentam minimizar o

custo do corte, porém tal prática pode levar à baixa qualidade de segmentação, tendo em vista que o cálculo do custo não leva em consideração o número de vértices presentes em cada um subgrafo, podendo acontecer, assim, segmentações triviais como mostrados na Figura 2.

Figura 2 – Imagem representando segmentação de baixa qualidade. A ponderação das arestas é inversamente proporcional à distância entre eles. Portanto, uma abordagem de segmentação baseada apenas no corte mínimo nos sugere o corte mínimo 1 e 2, enquanto que o corte desejável é representado pela linha tracejada.



Fonte: Shi e Malik (1997)

Como pode ser observado na imagem anterior, pontos isolados têm preferência no corte devido a minimização no valor produzido por este corte. Visando resolver este problema, Shi e Malik (2000) propõem um novo cálculo no custo do corte que gera subgrafos com quantidade de nós mais balanceados. Tal cálculo é obtido pela Equação 2.

Equação 2

$$NCut(A,B) = \frac{Cut(A,B)}{Volume\ (A)\ x\ Volume\ (B)}$$

Esta função pode ser reescrita usando uma função indicadora de grupo binário $X_l \in \{0,1\}^N$, com $X_l(i) = 1$ se o pixel i pertence ao segmento l.

Seja $X = [X1,X2]$ e D uma matriz diagonal onde $D(i,j) = \sum_j W(i,j)$. O critério utilizado para a quantidade da segmentação é definido como:

Equação 3

$$maximize\ \epsilon(X) = \frac{1}{2}\sum_{l=1}^{2}\frac{X_l^T\, WX_l}{X_l^T\, DX_l}$$

Uma função de custo *k-way* generalizada pode ser definida como X = [X1, ... ,XK]. Encontrar a partição Ncut ótima do grafo é uma tarefa NP-complexo. A técnica de partição de um grafo de espectro permite solucionar este problema usando uma solução em um espaço contínuo pela computação do autovetor K correspondente ao maior autovalor K em:

Equação 4

$$WV = \lambda DV$$

A qualidade global da segmentação depende da afinidade par a par dos pixels no grafo, quesito que é levado em conta na ponderação do grafo. Duas simples, mas eficientes, medidas de afinidade são: intensidade e contorno.

Para a intensidade, é levado em conta o fato de que pixels que estão próximos e possuem intensidades similares, provavelmente pertencem ao mesmo grupo. Portanto, a ponderação do grafo pode ser seguida através da Equação 5.

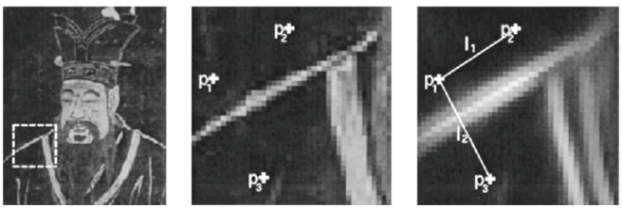
Equação 5

$$W_I(i,j) = e^{-\frac{\|x_i - x_j\|}{\sigma_x} - \frac{\|I_i - I_j\|}{\sigma_I}}$$

onde X e I denotam a localização e intensidade do pixel, respectivamente. Conectar pixels pela intensidade é útil para juntar partes disjuntas do objeto. Porém, devido à heterogeneidade da textura, levar em conta somente a intensidade não garante boas segmentações.

Outra característica de uma imagem que pode ser levada em conta na ponderação do grafo é a presença de bordas, pois, de modo geral, as bordas da imagem constituem uma provável fronteira para os objetos presentes na mesma. Tal quesito é particularmente útil quando o fundo da imagem possui valores de intensidades semelhantes ao corpo do objeto, como mostrado na Figura 3.

Figura 3 – Imagem extraída de Malik et al. (2001). À esquerda, imagem original com bounding box demonstrando a região a ser analisada. Pixels P1, P2 e P3 com intensidades similares, porém uma borda separa P1 e P3, mas não P2 o que sugere que P1 e P2 pertencem a um mesmo objeto, enquanto P3 pertence a outro



Fonte: Malik et al. (2001)

A afinidade entre dois pixels pode ser mensurada através da magnitude das arestas da imagem entre ambos, conforme a Equação 6.

Equação 6

$$W_C(i,j) = e^{-\frac{max_{x\ \in line(i,j)}\|Edge(x)\|^2}{\sigma_C}}$$

onde line(i,j) é uma linha reta ligando os dois pixels i e j; e Edge(x) é o peso da aresta na localização x.

2.3 Índices de diversidade

Os índices de diversidade de espécies são índices que são utilizados geralmente em investigações demográficas e de biodiversidade, tais índices mensuram o quão heterogêneo é a amostra em análise. Estes índices podem ser aplicados ao processamento de imagem fazendo uma analogia dos elementos da imagem com elementos da natureza. Cada ROI (Region of Interest) da imagem passa a ser considerada uma amostra do estudo e cada elemento dessa amostra, neste caso os pixels, pode ser considerado uma espécie.

a) Índice de Simpson

Índice de Simpson é um dos parâmetros que permitem medir a riqueza de organismos. Para isso, ele utiliza o número de espécies em seu habitat e abundância relativa de cada espécie. O índice de Simpson representa a probabilidade de que dois indivíduos dentro de um mesmo habitat sejam escolhidos aleatoriamente e ambos pertençam à mesma espécie. O índice é dado pela seguinte equação:

Equação 7

$$S = \frac{\sum_{i=1}^{S} n_i(n_i - 1)}{N\,(N-1)}$$

onde:
- S é o índice de Simpson;
- N é o número total de todos os indivíduos;
- n é o número de indivíduos por espécie.

b) Índice de Shannon

O índice de Shannon, ou índice Shannon-Weaver ou índice do Shannon-Wiener é um índice de diversidade usado para mensurar a entropia dos dados. Tal índice pode ser calculado pela seguinte fórmula:

Equação 8

$$H' = -\sum_{i=1}^{S} p_i \ln p_i$$

onde:
- H' é o índice de Shannon
- S O número de espécies. Chamado também de riqueza.
- N O número total de todos os indivíduos.
- P_i é abundância relativa de cada espécie, calculada pela proporção dos indivíduos de uma espécie pelo número total dos indivíduos na comunidade.

c) Índice de McIntosh

O índice de McIntosh é outro índice de diversidade ecológica que pode ser calculado pela seguinte expressão:

Equação 9

$$D = \frac{N - U}{N - \sqrt{N}}$$

onde:
- D é o índice de McIntosh;
- N é o número total de indivíduos da amostra;
- U é calculado por:

Equação 10

$$U = \sqrt{\sum N_i^2}$$

onde:
- N_i é o número de indivíduos pertencentes a i-ésima espécie.

2.4 Matriz de co-ocorrência e Run-Length

Durante a descrição da região de interesse, é importante que além da distribuição intensidade de cinza, também seja levada em consideração a disposição espacial de cada elemento, permitindo uma descrição mais precisa da região.

Portanto, o uso de um simples histograma de uma dimensão não possibilita uma descrição apropriada da região, visto que uma região em preto e branco disposta da forma de um tabuleiro de xadrez seria descrita da mesma forma que uma região do mesmo tamanho e com metade preta e a outra metade branca. Neste caso, uma melhor descrição da região é obtida com o uso da matriz de co-ocorrência.

A matriz de co-ocorrência de uma imagem I quantizada em N níveis de cinzas para dado um ângulo Θ e uma distância d é representada por uma matriz M_{NxN} onde $M(p,q)$ indica a quantidade de ocorrências onde um pixel de intensidade q é vizinho de um pixel de intensidade p a uma distância d e formando um ângulo Θ com o mesmo.

Já a matriz Run-Length codifica a textura levando em consideração a quantidade de vezes que um mesmo nível de cinza aparece em uma sequência de um determinado tamanho e uma determinada direção.

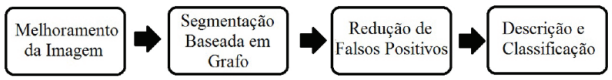
3 METODOLOGIA PROPOSTA

Este trabalho propõe uma metodologia para auxiliar a tarefa de detecção de neoplasias através do uso de imagens digitalizadas de mamografias.

Não entraremos em detalhes sobre o processo de aquisição das imagens, primeira etapa do processamento de imagens, por estar fora do objetivo deste trabalho, já que as imagens

utilizadas são provenientes da base de dados que é colocada à disposição da comunidade científica, mais conhecida como MIAS (Suckling, 1994).

Figura 4 - Etapas da metodologia desenvolvida



Fonte: Elaborada pelo autor

Nas seções seguintes, serão descritas as abordagens adotadas em cada etapa da metodologia, descrita na Figura 4.

3.1 Melhoramento da imagem

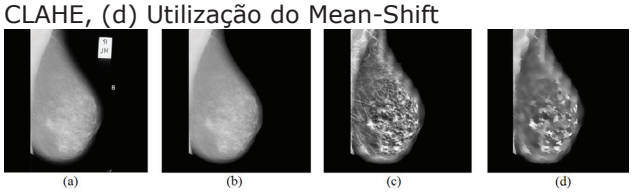
O uso de imagens mamográficas que apresentam boa qualidade torna-se um fator decisivo para o sucesso dos sistemas de auxílio à detecção. Por este motivo, fez-se necessário suprimir os ruídos e realçar as estruturas internas da mama através de procedimentos aqui descritos.

Grande parte dos ruídos está presente no fundo da imagem mamográfica, por isso a primeira etapa desta metodologia visa extraí-lo através da aplicação da função gaussiana de tamanho 12, seguida de uma binarização com limiar 128.

Posteriormente, é utilizado o CLAHE com contraste igual a 0,018, para obter uma imagem em que as estruturas de alta densidade sejam destacadas e que preserva associações locais de intensidades.

Em seguida, o resultado do CLAHE é combinado com o MeanShift proporcionando um contraste das estruturas internas da mama e diminuindo os ruídos presentes nesta região. Ao fim dessas etapas, a imagem já está pronta para ser segmentada. A figura abaixo apresenta os resultados obtidos na etapa de melhoramento.

Figura 5 - Sequência seguida na etapa de melhoramento da imagem. (a) Imagem Original. (b) Imagem após a extração do fundo, (c) Utilização do CLAHE, (d) Utilização do Mean-Shift



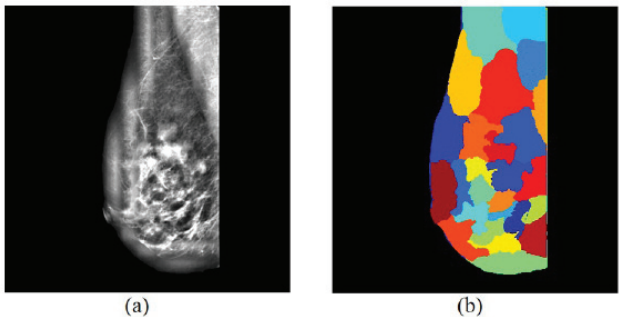
Fonte: Elaborado pelo autor

3.2 Segmentação baseada em grafo

Após a etapa de melhoramento da imagem, é realizada a segmentação, responsável por dividir uma imagem em múltiplas regiões ou segmentos de interesse.

A segmentação baseada em grafo tem demonstrado resultados satisfatórios quando aplicado em imagens mamográficas que apresentam o parênquima mamário realçado e com o fundo previamente extraído. Ao fim da realização dos testes, cada imagem foi segmentada três vezes, variando o número de segmentos em 30, 35 e 40. A mudança no número de segmentos em cada segmentação realizada não ocasionou alterações significativas na qualidade da segmentação obtida. A Figura 6 representa o resultado obtido com o uso da técnica de segmentação que fora apresentada.

Figura 6 - Imagem mdb265, MIAS. a) Imagem com histograma equalizado e fundo extraído, b) resultado da segmentação com 30 grupos

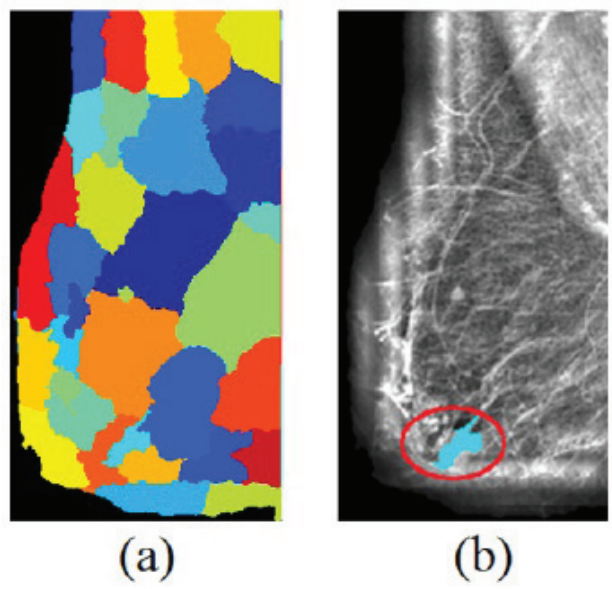


Fonte: Elaborado pelo autor

3.3 Redução de falsos positivos

A saída da etapa de segmentação de imagem é formada por múltiplos grupos onde todo pixel da imagem de entrada pertence exatamente a um grupo. Contudo, como pode ser visto na Figura (a), alguns desses grupos podem ser descartados facilmente por representarem regiões que apresentam características bem diferentes de uma região correspondente a uma massa, por exemplo, o fundo da imagem mamográfica pode ser removido fazendo uma análise do formato geométrico do mesmo. A remoção desses grupos confere uma redução nas taxas de falso positivo, além de diminuir o custo computacional empregado para a execução desta metodologia. As características que são analisadas nessa fase incluem a dimensão do grupo, número de pixels presentes no grupo, análise da forma geométrica, homogeneidade e média da intensidade.

Figura 7 – Comparativo do resultado da fase de redução de falsos positivos. (a) Imagem mamográfica segmentada, (b) Um único grupo restante após a fase de redução de falsos positivos. Cada grupo gerado pela segmentação está representado por uma cor diferente na imagem e a demarcação em vermelho em (b) representa a região que o especialista destacou como suspeita



Fonte: Elaborado pelo autor

Em geral, regiões de massas apresentam alta densidade, portanto, altos valores de intensidade na imagem. Em razão dessa característica, um dos critérios adotados para descarte dos grupos é a verificação da média da intensidade de cada grupo, sendo utilizado um limiar que avalia se essa região é suspeita ou não de apresentar uma massa.

Outro aspecto característico de regiões de massa é a homogeneidade da textura que esta região apresenta. Portanto, com o cálculo do desvio padrão das intensidades dos pixels presentes no grupo, alguns grupos que apresentam desvio padrão elevado podem ser descartados.

Por fim, é feita uma análise sobre o formato geométrico do grupo analisado, pois uma região de massa possui normalmente um comportamento geométrico circular (RANGAYYAN et al., 1997). Para esta análise é utilizado o template matching com um modelo circular e espiculado, visto que na maioria das vezes o câncer de mama apresenta contorno irregular, impreciso e espiculado.

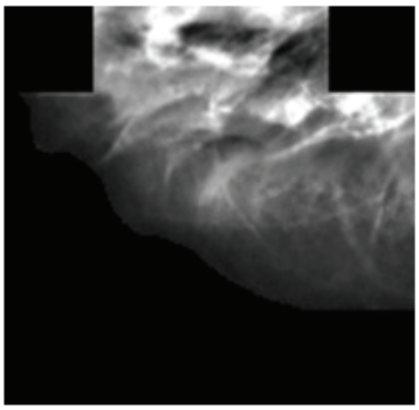
3.4 Descrição e classificação

Na etapa de descrição é realizada a extração de atributos dos grupos oriundos da etapa anterior para que posteriormente esses dados possam ser analisados e classificados, transformando, assim, informação em conhecimento.

Para tal, é importante a análise não somente dos grupos em si, mas também das regiões vizinhas ao grupo de tal forma que seja realizada uma descrição mais rica e precisa em detalhes da região. Portanto, cada análise é feita em relação as seguintes extrações dos grupos:

1. Cruz: A região é extraída em um formato de cruz, para isso, a região central da cruz corresponde ao menor quadrado envolvente à região e as quatro pontas da cruz são as regiões vizinhas da região central que distam 30 pixels da ROI.

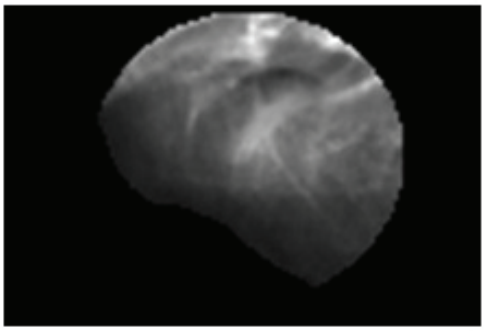
Figura 8 - Região extraída com formato de cruz



Fonte: Elaborado pelo autor

2. Elipse: Região referente à menor elipse que envolve toda a região da ROI.

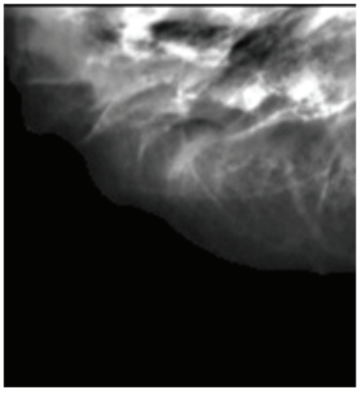
Figura 9 - Região extraída com o formato da menor elipse envolvente



Fonte: Elaborado pelo autor

3. GAP: A região extraída compreende um quadrado onde cada um dos seus lados possui um tamanho de 60 pixels maior em relação ao menor quadrado que envolve a região.

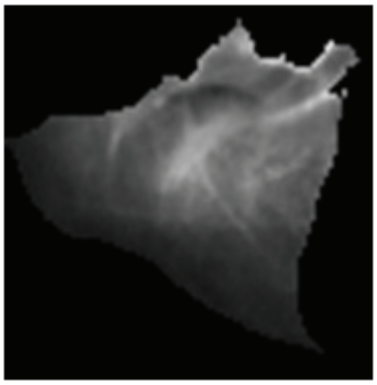
Figura 10 - Região extraída com folga (GAP)



Fonte: Elaborado pelo autor

4. Somente ROI: A região extraída é exatamente a região segmentada, ou seja, o menor quadrado envolvente sem o fundo.

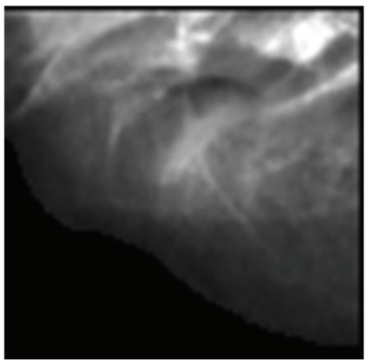
Figura 11 - Extração da região igualmente como ela foi segmentada



Fonte: Elaborado pelo autor

5. Mínimo Quadrado: A região extraída corresponde ao menor quadrado envolvente.

Figura 12 - Região extraída equivalente ao menor quadrado envolvente da região.



Fonte: Elaborado pelo autor

Para descrever as características das regiões, são utilizados índices de diversidade de espécies e, posteriormente, para cada índice é calculado a matriz de co-ocorrência e run-length objetivando a descrição da disposição desses índices nessas regiões.

3.5 Avaliação dos resultados

Duas medidas importantes para avaliar a metodologia são: a sensibilidade e a acurácia. A sensibilidade é a medida de acerto da metodologia levando em conta somente as regiões que apresentam neoplasia, calculada por:

Equação 11

$$S = VP/(VP + FN)$$

onde:

- **S** é a sensibilidade;
- **VP** é a quantidade de regiões que foram indicadas pela metodologia como uma região que apresenta neoplasia e de fato é uma região com neoplasia;
- **FN** é a quantidade de regiões que foram indicadas pela metodologia como uma região que não apresenta neoplasia, mas, na verdade, esta região apresenta neoplasia.

Já a acurácia representa a taxa de acerto global da metodologia. Tal medida pode ser calculada por:

Equação 12

$$A = \frac{VP + VN}{VP + FP + VN + FN}$$

onde:

- **A** é a acurácia;
- **VP** é a quantidade de regiões que foram indicadas pela metodologia como uma região que apresenta neoplasia e de fato é uma região com neoplasia;
- **FN** é a quantidade de regiões que foram indicadas pela metodologia como uma região que não apresenta neoplasia, mas, na verdade, esta região apresenta neoplasia.
- **VN** é a quantidade de regiões que foram indicadas pela metodologia como uma região que não apresenta neoplasia e de fato é uma região sem neoplasia;
- **FP** é a quantidade de regiões que foram indicadas pela metodologia como uma região que apresenta neoplasia, mas, na verdade, esta região não apresenta neoplasia.

4 RESULTADOS

Ao todo, foram utilizadas 74 imagens da base MIAS (SUCKLING et al., 1994) para a realização dos testes, sendo que cada imagem utilizada apresenta pelo menos uma região de massa. Como o número de segmentos pouco interferiu na qualidade de segmentação, foram utilizadas somente as imagens segmentadas em 40 grupos.

Para o cálculo dos limiares utilizados na etapa de redução de falsos positivos, foram utilizadas oito imagens que mais representam todas as imagens presentes na base para extrair as ROIs referentes à região com nódulo e então foi utilizada a média da intensidade e do desvio padrão encontrado nessas regiões como limiares. Os valores obtidos para estes limiares foram 180 para a média e 37,8 para o desvio padrão. O limiar utilizado para o índice de circularidade foi 0,45, sendo estimado empiricamente. Após a fase de redução de falsos positivos, a metodologia apresentou 85% de sensibilidade e uma taxa média de falso positivos de 6,67.

Após a redução de falsos positivos e extração das características das regiões resultantes, foram utilizados 200 casos de grupos que continham alguma região de massas e 200 regiões que não correspondem a uma massa para o treinamento do SVM. Na fase de predição, foram utilizadas 395 regiões de massas e 1750 de regiões normais para cada codificação de textura realizada.

A Tabela 1 apresenta as cinco melhores abordagens levando em consideração as menores taxas da média de falsos positivos por imagem, enquanto que a Tabela 2 exibe os cinco melhores resultados obtidos, levando-se em consideração as maiores taxas de sensibilidade.

No geral, pode-se perceber que as técnicas que resultam em menores taxas de falsos positivos por imagem não necessariamente implicam em uma boa taxa de sensibilidade, ou seja, em acertos dos casos onde há uma anomalia. Este fato pode ser comprovado comparando a Tabela 1 com a Tabela 2, onde nenhuma das cinco metodologias que apresentaram as menores taxas média de falsos positivos por imagem não fazem parte das cinco melhores taxas de sensibilidade.

O índice de Simpson foi o índice que mais resultou em acertos nas regiões com alguma displasia, além disso, é válido notar que o cálculo do índice levando-se em consideração o formato geométrico de cruz gerou tanto a maior sensibilidade quanto a menor taxa da média de falsos positivos por imagem.

5 CONSIDERAÇÕES FINAIS

Diante dos problemas enfrentados diariamente por profissionais da saúde na tarefa de identificar neoplasias presentes na mama através das imagens radiográficas da mesma, os sistemas computacionais de auxílio à detecção e diagnósticos têm demonstrado ser de grande utilidade para realização dessas tarefas, proporcionando uma redução nos casos de falsos negativos e, consequentemente, elevando o grau de confiança no diagnóstico médico.

Por outro lado, o emprego desses sistemas tem evitado que alguns procedimentos invasivos tenham que ser realizados para constatar a opinião médica sobre o caso analisado. Além disso, as técnicas computacionais apresentadas têm sido cada vez mais úteis na tentativa de melhorar a qualidade da imagem em análise, visto que, em grande parte, a qualidade da imagem adquirida está diretamente relacionada com a tecnologia empregada nos mamógrafos e a aquisição de mamógrafos modernos implica em alto custo financeiro.

É válido salientar que apesar das grandes contribuições dadas por sistemas de apoio à decisão, nenhum desses sistemas visa substituir ou invalidar o diagnóstico médico, mas sim corroborar e garantir mais confiabilidade tanto para o médico quanto ao paciente.

Em relação à metodologia proposta, percebe-se que os índices de diversidade configuram-se uma boa forma de descrever a textura das regiões analisadas e que o formato geométrico utilizado para o cálculo desses índices é uma variável que deve ser levada em conta, visto que as regiões vizinhas a uma região de massa também apresentam características peculiares.

A metodologia apresentada neste trabalho mostrou-se promissora no sentido de identificar e classificar anormalidades presentes no parênquima mamário, apresentando resulta-

Tabela 1 - Cinco melhores resultados em relação à média de falsos positivos por imagem

Região	Técnica	Média FP/ Imagem	Sensibilidade	Acurácia
Cruz	Shannon e McIntosh com Matriz de Co-ocorrência	7,22	0,81	0,72
Cruz	Shannon com Matriz de Co-ocorrência	7,84	0,79	0,69
Gap	Shannon e McIntosh	8,11	0,64	0,65
Elipse	Shannon e McIntosh com Matriz de Co-ocorrência	8,38	0,82	0,68
Elipse	McIntosh	8,70	0,70	0,64

Fonte: Elaborado pelo autor

Tabela 2 - Cinco melhores resultados em relação a taxa de Sensibilidade

Região	Técnica	Média FP/Imagem	Sensibilidade	Acurácia
Cruz	Simpson	17,91	0,90	0,36
Elipse	Simpson com Matriz de Co-ocorrência	15,31	0,85	0,44
MINQUAD	Simpson com Matriz Run-Length	16,58	0,83	0,40
Elipse	McIntosh com Matriz de Co-ocorrência	14,61	0,82	0,46
Elipse	Shannon e McIntosh com Matriz de Co-ocorrência	8,38	0,82	0,68

Fonte: Elaborado pelo autor

dos satisfatórios nos testes realizados. Além disso, é possível obter melhores resultados restringindo ainda mais as características que as regiões segmentadas devem apresentar para serem consideradas suspeitas, configurando-se como uma proposta para trabalho futuro, aliado a uma pré-extração do músculo peitoral, conferindo um menor custo computacional e uma significativa redução de falsos positivos, uma vez que grande parte dos grupos que correspondem ao músculo peitoral não serem descartados na etapa de redução de falsos positivos por apresentarem características semelhantes a uma região de massa, como alta intensidade na imagem e por ser uma região homogênea.

REFERÊNCIAS

AZEVEDO, C. M; PEIXOTO, J. E. *Falando sobre mamografia*. Rio de Janeiro: INCA, 1993.

BIRD, R. E.; WALLACE, T. W.; YANKASKAS, B. C. `Analysis of cancers missed at screening mammography. *Radiology,* v. 184, p. 613–617, 1992.

COMANICIU, Dorin; MEER, Peter. Mean shift: a robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 24, p. 603-619, 2002.

FREER, T. W.; ULISSEY, M. J. screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. *Radiology,* v. 220, n. 3, p. 781-786, 2001.

GIULIATO, Denise; BARCELIOS, Celia A.; DIAS, Walter D. O uso de equações de difusão no processo de detecção de regiões suspeitas em mamografias. In: Workshop de Informática Média, Belém-PA, 2008, p. 121-130.

GONZALES, Rafael C.; WOODS, Richard E. *Digital Image Processing*, 2.ed. [S.l.]: Prentice Hall, 2002.

HUYNH, P.T.; JAROLIMEK, A. M.; DAYE, S. The false-negative mammogram. *RadioGraphics*, v. 18, p. 1137-1154. Sept./Oct. 1998.

INSTITUTO NACIONAL DO CÂNCER. Disponível em: <http://www.inca.gov.br>. Acesso em: 16 dez. 2012.

KERLIKOWSKE, K. et al. Performance of screening mammography among women with and without a first-degree relative with breast cancer. *Annals Int. Med*. v. 133, p. 855–863, 2000.

KOLB, T. M.; LICHY, J.; NEWHOUSE, J. H. Comparison of the performance of screening mammography, physical examination, and breast US and evaluation of factors that influence them: an analysis of 27,825 patient evaluations. *Radiology,* v. 225, p. 165–175, 2002.

MALIK, J. et al. Contour and texture analysis for image segmentation. Int. *J. of Computer Vision*, v. 43, n. 1, p. 7–27, 2001.

PISANO, E. Contrast limited adaptive histogram equalization image processing to improve the detection of simulated spiculations in dense mammograms. *Journal of Digital Imaging*, p. 193-200, 1998.

PORTAL DA SAÚDE. Disponível em: <http://portalsaude.saude.gov.br>. Acesso em: 16 dez. 2012.

RANGAYYAN, R.M. et al. Measures of acutance and shape for classification of breast tumors. In: *Medical Imaging, IEEE Transactions*, v. 16 p. 799-810, 1997.

SHI, Jianbo; MALIK, Jiltendra. Normalized cuts and image segmentation. *IEEE Conference on Computer Vision and Pattern Recognition*, p 731–737, 1997.

SUCKLING J. et al. The mammographic images analysis society digital mammogram database. *Exerpta Medica. International Congress Series,* v. 1069, p. 375-378, 1994.

WORLD HEALTH ORGANIZATION. Disponível em: <http://www.who.int>. Acesso em: 17 dez. 2012.

ZHANG, L. A novel automatic tumor detection for breast cancer ultrasound Images. *IEEE Transactions*, v. 1, p. 401-404, 2011.