Compressão de Imagens Mamográficas Utilizando Segmentação e o Algoritmo PPM

José R. T. Marques[1], Glauber M. Pires[2], Leonardo V. Batista[3], JanKees v. d. Poel[4]

[1, 2, 3]Departamento de Informática, Universidade Federal da Paraíba – Brasil
[3]Programa de Pós-graduação em Informática, Universidade Federal da Paraíba – Brasil
[3, 4]Programa de Pós-graduação em Engenharia Mecânica, Universidade Federal da Paraíba – Brasil

Resumo – Este trabalho descreve um método de compressão de imagens mamográficas que utiliza segmentação e o algoritmo de compressão de dados *Prediction by Partial Matching* (PPM), juntamente com decomposição da imagem em planos de bits. O compressor será futuramente incluído em um sistema de Diagnóstico Auxiliado por Computador ainda em desenvolvimento. Os resultados obtidos mostram que o método atinge taxas de compressão competitivas em relação à de outros compressores avançados, e traz algumas vantagens adicionais, discutidas neste trabalho.

Palavras-chave: Mamografia, Compressão, *Prediction by Partial Matching* (PPM), Planos de Bits.

Abstract – This work describes a mammographic image compression method that uses segmentation and the prediction by partial matching (PPM) data compression algorithm, together with a bit plane image decomposition. The compressor will be included in a Computer-Aided Diagnosis System which is still being developed. Results show that the method achieves competitive compression rates in comparison with other advanced compressors, and brings some additional benefits which will be discussed in this work..

Key-words: Mammography, Compression, *Prediction by Partial Matching* (PPM), Bit Planes.

1. Introdução e Motivação

Grave problema de saúde pública mundial, o câncer de mama é a principal causa de óbitos por câncer na população feminina brasileira. A Organização Mundial de Saúde (OMS) recomenda o rastreamento em massa para enfermidades que constituam problemas sérios de saúde pública, desde que sua detecção precoce leve à redução da morbidade e da mortalidade, como é o caso do câncer de mama.

O diagnóstico auxiliado por computador – *Computer-Aided Diagnosis* (CAD) [1][2] pode facilitar o trabalho do radiologista e até emitir uma segunda opinião, aumentando as chances de um diagnóstico precoce.

A utilização de bancos de dados de imagens mamográficas em formato digital e as práticas de telemedicina exigem armazenar e transmitir grandes quantidades de dados. A título de ilustração, a digitalização das imagens de um único exame mamográfico com duas incidências (médio-lateral oblíqua e crânio-caudal) por mama e utilizando uma resolução adequada ao diagnóstico pode ocupar até 120 Mbytes de espaço em disco. O problema torna-se ainda mais relevante quando se tem em mente que uma única clínica de porte médio pode efetuar diariamente dezenas de exames mamográficos.

Devido a este fato, técnicas eficientes de compressão de dados são necessárias para reduzir custos de armazenamento e transmissão. Entretanto, tais técnicas não devem gerar perdas significativas de informação na imagem, pois isto poderia comprometer o diagnóstico médico e a análise das mamografias por sistemas CAD [3].

Escarpinati e Schiabel [3] avaliaram cinco técnicas de compressão sem perdas sobre um conjunto de 16 imagens mamográficas de mamas densas, digitalizadas em um scanner Laser Lumiscan 50 com uma resolução de contraste de 12 bits por pixel e resolução espacial de 0,15 mm. As técnicas avaliadas foram: RLE (*Run-Length Encoding*) [4] com decomposição em planos de bits; DAC (DAdos com Compressão) [5]; codificação baseada em dicionário LZW (Lempel-Ziv-Welch) [6]; padrão JPEG sem perdas (lossless JPEG, LJPEG) [7]; algoritmo de Huffman com modelo semi-adaptativo não-contextual [8]; LOCO-I (LOw COmplexity LOssless COmpression for Images) [9]; e CALIC (Context-based, Adaptive, Lossless Image Coding) [10]. Os melhores resultados foram obtidos pelo LJPEG, que em média reduziu as imagens para aproximadamente 18% do tamanho original. No outro extremo, o RLE com decomposição por plano de bits não conseguiu comprimir as imagens - ao contrário, ampliou-as 9,8%, em média. O desempenho insatisfatório desta última técnica deve-se provavelmente à utilização do RLE, altamente ineficiente quando

longas seqüências de símbolos repetidos não são prováveis.

Este trabalho propõe um método de compressão de imagens mamográficas que utiliza segmentação e o algoritmo *Prediction by Partial Matching* (PPM), juntamente com decomposição da imagem em planos de bits. Pretende-se mostrar que a decomposição em planos de bits em associação com um esquema de modelagem avançado produz um esquema de compressão eficaz e traz uma série de benefícios adicionais.

2. Materiais e Métodos

O compressor de dados *Prediction by Partial Matching* (PPM) [11][12] é uma técnica de codificação por entropia baseada na modelagem estatística adaptativa e na predição por contexto. O PPM é considerado um dos compressores de propósito genérico mais eficazes da atualidade.

O modelo PPM utiliza um conjunto de no máximo K símbolos precedentes como contexto para estimar a distribuição de probabilidades condicionais para o próximo símbolo da mensagem. O modelo alimenta um codificador aritmético, que atribui a cada símbolo um número de bits igual a sua informação condicional, que por sua vez depende da probabilidade de ocorrência do símbolo condicionada ao contexto [13]. Assim, o esquema de codificação aritmética é capaz de igualar a entropia da fonte em todos os casos, atingindo compressão máxima para o modelo utilizado [11][13].

A decomposição em plano de bits decompõe uma imagem S de n bits por pixel (ou seja, $2n$ níveis de cinza) em n imagens binárias, ou plano de bits, s_0, s_1, …, s_{n-1}. Um pixel no plano s_i equivale ao i-ésimo bit do pixel na mesma posição em S. A Figura 1 ilustra a decomposição em planos de bits. Nesse exemplo, cada pixel na imagem original possui três bits; assim, a decomposição gera três imagens binárias [4]. O primeiro plano de bits é formado pelos bits mais significativos de cada pixel, e assim sucessivamente. A Figura 2 mostra uma imagem mamográfica e dois dos seus planos de bits: a imagem original, o plano de bits s_8 e o plano de bits s_{10}.

O método de compressão proposto no presente trabalho envolve decompor em planos de bits a imagem mamográfica para, em seguida, comprimir separadamente os planos de bits pelo PPM, adaptado para alfabeto binário, o que reduz drasticamente os requisitos computacionais do compressor.

Com o intuito de melhorar a razão de compressão (RC), o background foi separado da região da mama através de segmentação por limiarização. Seja $S(x,y)$ o valor do pixel na posição (x,y) e L um valor de limiar escolhido de acordo com as características do *background* da imagem mamográfica:

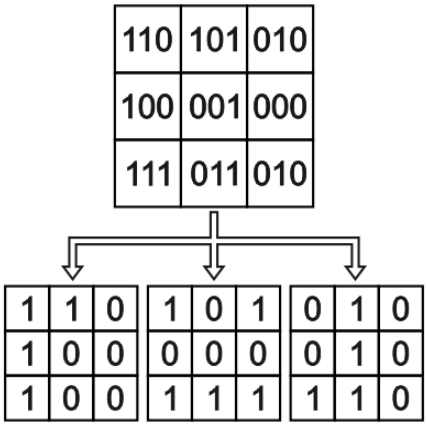

Figura 1: Decomposição de uma imagem em planos de bits.

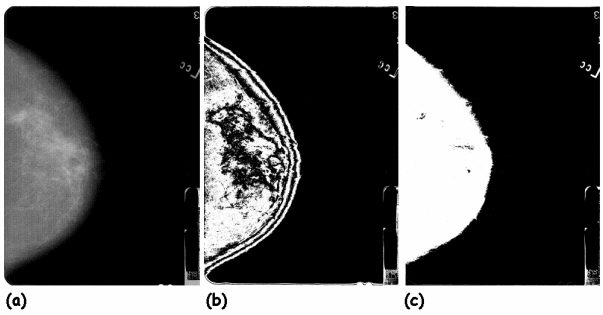

Figura 2: Imagem mamográfica e dois dos seus planos de bits: (a) imagem original; (b) plano de bits s_8; (c) plano de bits s_{10}.

O processo de segmentação baseia-se no fato de que o *background*, em geral, é a região mais escura da mamografia, cujos pixels normalmente possuem nível de cinza próximos a zero. Serão criados dois modelos estatísticos diferentes, um para cada região presente na mamografia, o que tende a melhorar a compressão, pois os modelos serão mais específicos para cada região.

Uma imagem binária adicional, com bits de valor '1' indicando região de mama e de valor '0' indicando região de background é anexada ao arquivo comprimido, para permitir a descompressão com os modelos corretos para cada pixel.

O banco de dados de imagens utilizado para testar o compressor é o *Digital Database for Screening Mammography* (DDSM) [14]. Esta base de dados vem sendo largamente utilizada como *benchmark* em vários artigos na área de mamografia [15], por ser de utilização gratuita e conter uma grande quantidade e diversidade de casos, compostos por imagens e informações técnicas e clínicas correspondentes.

Para os testes foram escolhidas de forma aleatória doze imagens dentre aquelas presentes no DDSM [14]. Todas as imagens escolhidas foram digitalizadas por um scanner Lumisys 200 de 12 níveis/pixel e 50 microns.

Nas imagens de teste provenientes do DDSM há 2949 níveis de cinza diferentes, codificados em 16 bits de tal forma que assumem valores inteiros entre 0 e 65355. Uma vez que 12 bits seriam suficientes para codificar cada pixel, antes da decomposição em planos de bits os níveis são mapeados para valores inteiros entre 0 e 2948, e representados em 12 bits, gerando 12 planos de bits. O grau de compressão é aferido considerando que as imagens originais têm 12 bits/pixel, e não 16 bits/pixel.

Para mensurar o grau de compressão, utiliza-se neste trabalho a taxa de compressão percentual, *p*, definida como:

$$p = \frac{T_c}{T_o} \times 100\% \qquad (1)$$

Na Equação (1), T_c e T_o são o tamanho da imagem comprimida e o tamanho da imagem original, respectivamente. Em T_c são incluídas todas as informações necessárias à descompressão das imagens.

O compressor foi escrito em linguagem Java e dividido em módulos independentes. A arquitetura proposta para o compressor/ descompressor está mostrada na Figura 3.

Figura 3: Arquitetura do Compressor/ Descompressor Proposto.

Durante a implementação do sistema, foram testadas algumas variações:

- Ausência da segmentação de background e mama;
- Ausência do mapeamento dos níveis de cinza; e
- Utilização de um único modelo para todos os planos de bits.

Todas as combinações dessas variações foram testadas até chegar à variação que apresentou os melhores resultados.

4. Resultados Experimentais

O Gráfico 1 permite avaliar o ganho de compressão que se obteve com a segmentação das áreas de mama e background. O limiar de segmentação utilizado foi *L = 0*, o que significa que pixels com valor zero são considerados background, e os demais são considerados pixels de mama.

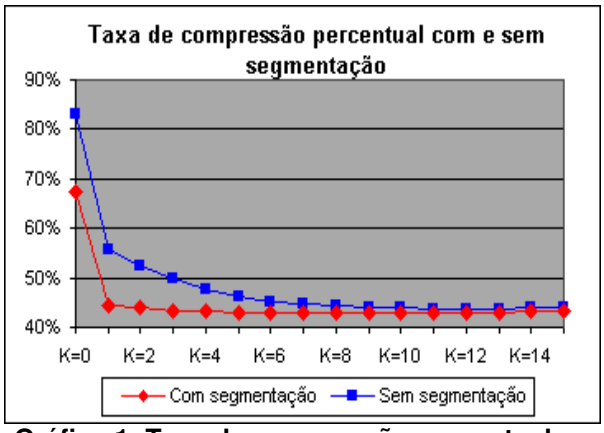

Gráfico 1: Taxa de compressão percentual, *p*, com e sem segmentação.

A Tabela 1 e a Tabela 2 mostram as taxas de compressão percentual média, *p*, com segmentação, na região da mama e nas imagens completas (mama e background), para cada plano de bits e tamanho de contexto máximo (parâmetro K do PPM) entre 0 e 12. Todas as 12 imagens de teste foram utilizadas para a construção da Tabela 1.

Tabela 1: Taxa de compressões percentual média para cada plano de bits e tamanho de contexto K, K = 0, 1, ..., 6

	K=0	K=1	K=2	K=3	K=4	K=5	K=6
Bit 0	**91,7**	100,0	100,0	100,0	100,0	100,0	100,0
Bit 1	**91,7**	100,0	100,0	100,0	100,0	100,0	100,0
Bit 2	**91,7**	100,0	100,0	100,0	100,0	100,0	100,0
Bit 3	**91,7**	100,0	100,0	100,0	100,0	100,0	100,0
Bit 4	**91,6**	100,0	100,0	100,0	100,0	100,0	100,0
Bit 5	**91,6**	99,7	99,6	99,5	99,5	99,4	99,3
Bit 6	91,3	92,0	91,1	90,6	90,3	90,0	89,9
Bit 7	90,2	65,5	62,6	60,9	60,1	59,6	59,4
Bit 8	85,0	39,3	35,8	33,9	32,8	32,2	31,9
Bit 9	72,2	20,7	18,5	17,1	16,3	15,9	15,7
Bit 10	67,6	10,1	8,9	8,2	7,8	7,6	7,4
Bit 11	23,4	3,3	3,0	2,8	2,7	2,6	2,6
Mama	81,6	69,2	68,3	67,8	67,5	67,3	67,2
Total	50,7	38,4	37,6	37,2	37,0	36,9	36,8

Tabela 2: Taxa de compressões percentual média para cada plano de bits e tamanho de contexto K, K = 7, 8, ..., 12

	K=7	K=8	K=9	K=10	K=11	K=12
Bit 0	100,0	100,0	100,0	100,1	100,1	100,3
Bit 1	100,0	100,0	100,0	100,1	100,1	100,3
Bit 2	100,0	100,0	100,0	100,1	100,1	100,3
Bit 3	100,0	100,0	100,0	100,1	100,1	100,3
Bit 4	100,0	100,0	100,0	100,0	100,1	100,2
Bit 5	99,3	99,2	99,2	99,2	99,3	99,4
Bit 6	89,8	89,7	89,6	**89,6**	89,6	89,7
Bit 7	59,2	59,1	59,1	**59,0**	59,1	59,1
Bit 8	31,7	31,5	31,4	31,4	**31,3**	31,4
Bit 9	15,5	15,3	15,3	15,2	15,2	**15,2**
Bit 10	7,3	7,2	7,2	7,2	7,1	**7,1**
Bit 11	2,6	2,5	2,5	2,5	**2,5**	2,5
Mama	67,1	67,1	67,0	**67,0**	67,1	67,1
Total	36,8	36,7	36,7	**36,7**	36,7	36,8

A Tabela 3 apresenta uma comparação entre o sistema proposto, com segmentação e tamanho máximo de contexto K = 10, e diversos esquemas de compressão comerciais, utilizando 6 imagens de teste.

Tabela 3: Comparativo entre o compressor proposto e esquemas de compressão comerciais.

Técnica de Compressão	*p*
PNG 16 Bits	27%
WINZIP 10.0 com PPMd	33%
Sistema Proposto	35%
7ZIP	36%
RAR (WinRar)	39%
ZIP (WinRar)	51%

Os testes dos compressores foram feitos em um microcomputador com processador AMD Duron™ de 1,2GHz, 256 MBytes de memória DIMM, disco rígido PATA com 40 GBytes de memória, sistema operacional Windows XP com Service Pack 2 e Máquina Virtual Java versão 1.5.0_06-b05. Uma imagem de 30MBbytes é comprimida/descomprimida em aproximadamente 4,3 minutos, aí inclusos o tempo de leitura e escrita em disco rígido. Este tempo não inclui as etapas de segmentação e mapeamento, pois estas operações são muito rápidas quando comparadas às operações de leitura e escrita em disco.

5. Discussão e Conclusões

Este trabalho apresentou um método de compressão de imagens mamográficas utilizando segmentação e o algoritmo PPM, juntamente com decomposição em planos de bits.

Como se pode verificar nas Tabelas 1 e 2, a melhor compressão foi obtida para tamanho máximo de contexto K = 10 e K = 11, resultando em uma taxa de compressão percentual 36,7%. Isso significa que o compresso proposto foi capaz de reduzir as imagens a 36,7% do seu tamanho original, sem perda de informação.

Pretende-se mostrar que a decomposição em planos de bits em associação com um esquema de modelagem avançado produz um esquema de compressão eficaz e traz uma série de benefícios adicionais.

Uma comparação direta com os resultados apresentados por Escarpinati e Schiabel [3] não pode ser feita, pois as bases de imagens utilizadas para testes são diferentes. Diferenças no scanner, no nível de ruído, na resolução espacial e de contraste, no nível de contraste e na densidade radiológica das mamas (em [3] foram utilizadas exclusivamente mamas densas) afetam sobremaneira os resultados.

O método proposto foi comparado com compressores de propósito geral avançados e muito utilizados atualmente. A Tabela 3 resume os resultados. O WinZip 10.0 utilizando o compressor PPM tipo D, e o WinRar utilizando compressores RAR e Zip foram testados em modo de compressão máxima, e os demais em modo padrão.

Pode-se perceber que o método apresenta desempenho similar ao WinZip-PPM-D em modo de compressão máxima, ao 7Zip em modo padrão e ao WinRar-RAR em modo de compressão máxima. O PNG de 16 bits apresenta desempenho um pouco superior ao do compressor proposto que, por sua vez, é bem superior ao WinRar-ZIP em modo de compressão máxima.

O Gráfico 1 indica que a segmentação de mama e background melhora substancialmente a eficiência de compressão para contextos pequenos. Para K = 1, a segmentação leva praticamente à melhor compressão (obtida com K = 10). Sem segmentação, por outro lado, uma compressão semelhante só é atingida para K = 7. Considerando-se que o tempo de processamento e a utilização de memória pelo PPM aumentam com o tamanho do contexto, a segmentação é altamente recomendável.

O compressor proposto apresenta uma série de vantagens em relação à utilização de compressores de propósito genérico ou a formatos gráficos genéricos, como o PNG, relacionadas principalmente à decomposição por plano de bits:

- Redução drástica nos requisitos computacionais do PPM
- Extensão imediata para compressão *lossy-to-lossless* em fluxo único, característica importante para telemedicina. A transmissão progressiva dos planos de bits mais significativos seria muito eficiente. De acordo com a Tabela 2, o plano de bits s_{11} se reduz

para aproximadamente 2,5% do seu tamanho original; o plano 10 para 7,1% e assim por diante.

- Possibilidade de analisar o ruído presente na mamografia. A Figura 4 mostra uma ampliação do plano de bits s_0, de aparência ruidosa. Esta observação se repete para os planos s_0 a s_4. Apenas a partir do plano s_6 se pode observar alguma estrutura nos padrões. As Tabelas 1 e 2 confirmam a forte presença de ruído nestes planos. Observe-se que apenas para contexto de tamanho máximo K = 0, ou seja, apenas com o PPM em modo não contextual, se obteve alguma compressão para os planos s_0 a s_4. A medida em que se aumentou o tamanho do contexto, deixou de haver compressão e, para contextos 10, 11 e 12 passou a ocorrer expansão desses planos. Bell, Cleary e Witten [11] descrevem um comportamento semelhante quando o PPM tenta comprimir seqüências pseudo-aleatórias, e o justificam afirmando que nestes casos o PPM procura uma estrutura onde não há nenhuma. Essa observação corrobora os indícios de que os planos menos significativos das mamografias da base de testes utilizada estão fortemente contaminados por ruído. A eliminação destes planos poderia talvez até melhorar a qualidade diagnóstica das imagens, ou o desempenho de sistemas CAD. Essa afirmação será verificada futuramente em conexão com o sistema CAD em desenvolvimento pelos autores do presente trabalho, e em avaliações por radiologistas.

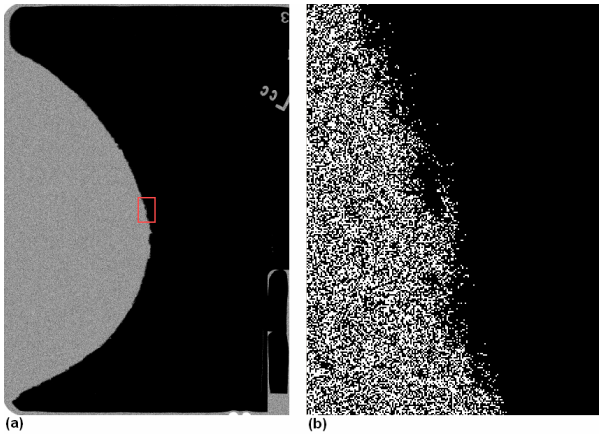

Figura 4: Ruído presente no plano de bits s_0: (a) Plano de bits s_0 da imagem mamografia completa; (b) Zoom da área destacada em vermelho.

Testes futuros com o compressor deverão avaliar a influência dos padrões de densidade mamária, de acordo com a classificação BI-RADS™ [16], sobre as taxas de compressão.

No estágio atual, o sistema possui limitações em relação ao tempo de processamento. Entretanto, vale salientar que o sistema ainda está em fase de aperfeiçoamento, apresentando um código sem muitas otimizações.

Uma variação simples poderia melhorar a compressão: utilizar para cada plano um tamanho de contexto máximo ótimo. Assim, de acordo com as indicações de compressão máxima em negrito nas Tabelas 1 e 2, para os planos s_0 a s_5, seria utilizado K = 0; para os planos s_6 e s_7, K = 10, e assim por diante. Essa e outras variações já estão sendo pesquisadas e desenvolvidas e serão divulgadas em breve.

6. Agradecimentos

Ao CNPq, pelo apoio na forma de bolsas de pesquisa.

7. Referências

[1] Almeida, C. W. D.; Poel, J.; Batista, L. V.; Amorim, H. L. E (2005), "Análise de Formas Baseada no Método da *Curvature Scale Space* Para Tumores de Câncer de Mama". In: *Anais do V Workshop de Informática Médica*, Porto Alegre, v. 1. p. 20-24.

[2] Astley, S., Gilbert, F. (2004), "Computer-aided Detection in Mammography", *Clinical Radiol.*, n. 59, pp. 390-399.

[3] Escarpinati, M. C., Schiabel, H. (2002), "Avaliação de Técnicas de Compressão sem Perdas Aplicadas a Imagens Mamográficas Digitais de Mamas Densas". In: Anais do XVIII Congresso Brasileiro de Engenharia Biomédica (CBEB'2002), São José dos Campos, SP, Brasil v. 1, pp. 195-198.

[4] Gonzalez, R. C., Woods, R. E. (2002), "Digital Image Processing" (2nd Ed). Prentice Hall, Inc., Upper Saddle River, New Jersey, USA.

[5] Martins, J. M. (1995), "Desenvolvimento de um Tomógrafo de Ressonância Magnética: Integração e Otimização", Tese (Doutorado) – Instituto de Física de São Carlos, Universidade de São Paulo, São Paulo, Brasil.

[6] Welch, T., (1984), "A Technique for High Performance Data Compression", IEEE Computer, v. 17, n. 6, p. 8-19.

[7] Wallace, G. K. (1991), "The JPEG Still Picture Compression Standard," Communications of the ACM, v. 34, n. 4, p. 30-44.

[8] Huffman, D. A. (1952), "A method for the construction of minimum-redundancy codes". *Proceedings of the I.R.E.*, pp. 1098-1102.

[9] Weinberger, M. J.; Seroussi, G.; and Sapiro, G. (1996), "LOCO-I: A Low Complexity, Context-Based, Lossless Image Compression Algorithm." Proceedings of the IEEE Data Compression Conference, Snowbird, p. 140-149, March, 1996.

[10] Wu, X., Memon, N. (1997), ”Context-Based, Adaptive, Lossless Image Coding”, IEEE transactions on Communications, vol.45, no.4, pp 437-444, April.
[11] Bell, T., Cleary, J., Witten, I. (1984), “Data compression using adaptive coding and partial string matching”. *IEEE Transactions on Communications*, v. 32, n. 4, pp. 396-402.
[12] Moffat, A. (1990), “Implementing the PPM data compression scheme”. *IEEE Transactions on Communications*, v. 38, n. 11, pp. 1917-1921.
[13] Shannon, C. E. “A Mathematical Theory of Communication.” Bell Syst. Tech. J., v. 27, p. 379-423, 1948.
[14] Heat, M., Bowyer, K., Kopans, D., Moore, R., Kegelmeyer, P. (2000), “The digital database for screening mammography”. In: *Digital Mammography – Proceedings of the 5th International Workshop on Digital Mammography (IWDM2000)*, Yaffe, M. J. (Ed.), Toronto, pp. 212-218.
[15] Heat, M., Bowyer, K. (2000), “Mass detection by relative image intensity Digital Mammography”. In: *Digital Mammography – Proceedings of the 5th International Workshop on Digital Mammography (IWDM2000)*, Yaffe, M. J. (Ed.), Toronto, pp. 219-225.
[16] American College of Radiology (2003). *Breast Imaging Reporting and Data System® (BI-RADS®) Atlas*. 5th Edition. Reston, VA.

8. Contato

JOSÉ RAPHAEL TEIXEIRA MARQUES
Endereço profissional:
Universidade Federal da Paraíba
Departamento de Informática
Cidade Universitária – Campus I
CEP: 58.059-900 – João Pessoa – PB
Endereço Pessoal:
Av. Cel. Augusto F. Maia, 206 – José Américo – João Pessoa – PB – Brasil – CEP 58.073-000
Fone: (83) 8828-3314
E-mail: jose_raphael_marques@yahoo.com.br