

DESENVOLVIMENTO DE EXTRATORES DE CARACTERÍSTICAS PARA CLASSIFICAR TUMORES DE MAMA

LUCAS BATISTA LEITE DE SOUZA¹, DENISE GULIATO²

RESUMO

O câncer de mama é uma das principais causas de morte entre as mulheres. Sistemas de diagnóstico auxiliado por computador devem ser capazes de identificar de forma automática características relevantes para diagnóstico presentes na mamografia. Estas características devem ser discriminantes o suficiente para permitir classificar lesões como benignas ou malignas, ou recuperar imagens corretamente. Sendo assim, o sucesso de um sistema de apoio ao diagnóstico de câncer de mama está fortemente associado às características que representam os achados radiológicos em um mamograma. Este trabalho apresenta dois extratores de características baseados em forma (*Complexity Index* e *Elliptic Variance*), que extraem informações do contorno de uma dada lesão. Algumas destas informações mostraram-se relevantes para a classificação de uma lesão como maligna ou benigna.

PALAVRAS CHAVE: câncer de mama, extratores de características, processamento digital de imagens.

ABSTRACT

The breast cancer is one of the major reasons of death among women. Computer aided diagnostic systems must be able to automatically identify existent features in mammograms. These features must discriminate benign and malign tumors, or retrieve images correctly. The success of computer aided diagnostic system is strongly associated with the features that represent the radiological finding in a mammogram. This work presents two shape feature extractors (*Complexity Index* and *Elliptic Variance*), that extract information from the lesion. The

¹ Universidade Federal de Uberlândia, Faculdade de Computação, Avenida João Naves de Ávila 2121, Bloco 1B sala 1B121, Uberlândia, CEP: 38400-902, lucasbls1@comp.ufu.br .

² Universidade Federal de Uberlândia, Faculdade de Computação, Avenida João Naves de Ávila 2121, Bloco 1B sala 1B136, Uberlândia, CEP: 38400-902, guliato@ufu.br .

performance of these extractors was tested in a base of 111 contour lesions.

KEYWORDS: breast cancer, features extractors, digital image processing.

1. INTRODUÇÃO

Câncer de mama é o segundo tipo de câncer mais freqüente no mundo, com prevalência de aproximadamente um milhão de novos casos por ano [1]. É o tipo de câncer mais comum na Europa e nos Estados Unidos [2], [3]. No Brasil, de acordo com o Instituto Nacional de Câncer - Brasil, a estimativa de novos casos em 2009 é de 49 mil com um risco de 52 casos a cada 100 mil mulheres [4]. Apesar de ser considerado um câncer de bom prognóstico se diagnosticado e tratado precocemente, as taxas de mortalidade por câncer de mama continuam elevadas no Brasil, devido, provavelmente, ao diagnóstico tardio [4].

A mamografia por raio-X é ainda o exame mais adotado para detecção precoce de sinais de câncer de mama e tem um papel importante nas decisões terapêuticas a serem adotadas. A mamografia por raio-X pode revelar evidência de anormalidades como nódulos e calcificações, bem como sinais

sutis como assimetria bilateral e distorsão arquitetural [5].

Apesar dos avanços de qualidade alcançados na fabricação de equipamentos para aquisição de mamografias e pelas técnicas de filme de raio-X nos últimos anos, ainda hoje, em torno de 10% de tumores malignos não são detectados em mamografias de mulheres com idade acima de 50 anos e em torno de 25%, em mulheres com idade entre 40 e 49 anos [6], [7], [8]. Esses dados têm justificado o desenvolvimento de métodos de processamento digital de imagens para a detecção e análise de características obtidas a partir das mamografias, de tal maneira a auxiliar na redução de erros de diagnóstico, reduzir o uso de procedimentos auxiliares, reduzir mortalidade bem como os custos com a saúde.

Os sistemas de apoio ao diagnóstico envolvem extração de características a partir da mamografia. Tais características são usadas para

discriminar entre lesões benignas e malignas.

Neste artigo são apresentados dois extratores de características baseados na forma do contorno de uma dada lesão: *Complexity Index* e *Elliptic Variance*.

O artigo está estruturado da seguinte forma: a Seção 2 apresenta materiais e métodos; a Seção 3 apresenta o extrator *Complexity Index* (CI); a Seção 4 apresenta o extrator *Elliptic Variance* (Ev); a Seção 5 apresenta uma comparação entre os extratores CI e Ev com outros extratores existentes; finalmente a Seção 6 apresenta as conclusões do trabalho.

2. MATERIAIS E MÉTODOS

Os contornos de lesões utilizados para testar os extratores foram desenhados por um médico radiologista especializado na análise de mamogramas. No total são 111 contornos, contendo formas típicas e atípicas de 65 massas benignas e 46 tumores malignos.

O diagnóstico das lesões e classificação da base de dados foram feitos através de biópsia. Essa base de dados é a mesma utilizada por Rangayyan e Nguyen [9] e Guliato et al. [10]. Veja Guliato et al. [10] para mais detalhes.

Para medir a eficiência dos extratores na discriminação entre lesões benignas e malignas foi utilizada a análise ROC (*Receiver Operating Characteristic*) [11]. A análise ROC é uma ferramenta poderosa para medir e especificar problemas no desempenho do diagnóstico em medicina. Esta análise, por meio de um método gráfico simples e robusto, permite estudar a variação da *sensibilidade* e *especificidade*, para diferentes valores de corte. Na análise ROC, quanto mais próximo de 1.0 for a área abaixo da curva (Az), mais discriminante é o extrator. No caso específico dos extratores apresentados neste trabalho, o objetivo é a discriminação entre lesões malignas e benignas. Veja [11] para mais detalhes.

Os dois extratores de características apresentados neste artigo foram desenvolvidos na linguagem de programação C, em um computador com as seguintes características:

- Processador Intel Pentium 4, 2.8GHz;
- Memória Ram DDR2 1GB;
- Sistema Operacional GNU-Linux, distribuição Slackware 12, kernel 2.4;
- Compilador gcc 2.3.

3. O EXTRATOR *COMPLEXITY INDEX* (CI)

Tumores de mama malignos tipicamente aparecem em mamogramas com contornos rugosos, espiculados, microlobulados e com áreas côncavas, e lesões benignas têm contornos lisos, redondos, ovais, macrolobulados e convexos [10]. Sendo assim, os tumores malignos tendem a apresentar contornos mais complexos que as massas benignas. O descritor de características (CI) deve ser aplicado sobre a aproximação poligonal de um contorno. Essa aproximação poligonal foi gerada a partir do modelo proposto por Guliato et al. [10], que preserva os espículos e outros detalhes que são importantes para o diagnóstico do câncer de mama. A Figura 1 mostra um exemplo de um contorno (em linha pontilhada) e de sua aproximação poligonal (em linha contínua) usando essa técnica. Note que neste processo, artefatos e ruídos são eliminados do contorno.

A forma global de um contorno é um fator importante para determinar a complexidade de um contorno. Como exemplo, veja a Figura 2. O objeto 2(a) apresenta uma forma convexa, enquanto o objeto 2(b) apresenta uma forma mais espiculada.

O extrator CI leva em consideração a complexidade da forma de um dado contorno. As subseções 3.1, 3.2 e 3.3 mostram as métricas utilizadas para descrever o índice de complexidade de uma lesão. Na subseção 3.4 é apresentada a fórmula final utilizada pelo extrator *Complexity Index*.

3.1 A Frequência de Regiões Côncavas ou Índice de Espículos (Feq)

Seja N o número de vértices do contorno poligonal de uma lesão cujas arestas adjacentes formam ângulo interno maior do que 180° . Quanto maior o valor de N , maior a quantidade de regiões côncavas no contorno. Veja a Figura 3.

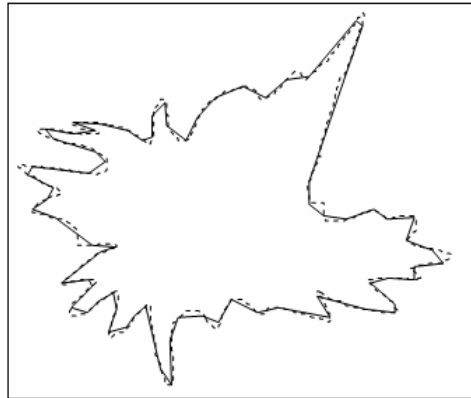


Figura 1: Polígono gerado a partir do contorno original de um tumor, com o objetivo de reduzir ruído, eliminar artefatos e preservar informações relevantes para diagnóstico.

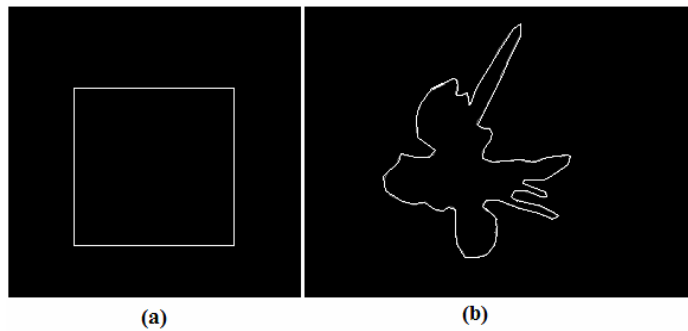


Figura 2: Exemplos de contornos com diferentes complexidades.

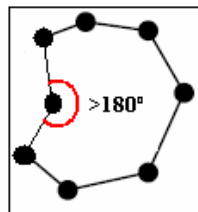


Figura 3: Contorno em que $N = 1$.

Seja V o número de vértices de um polígono. A seguinte propriedade pode ser verificada na Inequação 1.

$$N \leq V - 3 \quad (1)$$

Para normalizar N para o intervalo $[0,1]$ usa-se a Equação 2, onde N_{norm} é um valor normalizado para N .

$$N_{norm} = \frac{N}{V - 3} \quad (2)$$

Ainda deve-se definir uma terceira equação para fazer com que contornos com poucas regiões côncavas possuam valor próximo a 0, e contornos com muitas regiões côncavas possuam valor próximo a 1 [12]. A Equação 3 mede a

freqüência com que ocorrem regiões côncavas no contorno de lesão (Feq) [12].

$$Feq = 16 * (Nnorm - 0.5)^4 - 8 * (Nnorm - 0.5)^2 + 1 \quad (3)$$

3.2 A Amplitude das Regiões Côncavas (Ampl)

A Freqüência de espículos (Feq) não descreve nenhuma informação a respeito dos espículos formados. Com o objetivo de quantificar essa amplitude deve-se calcular a relação entre o perímetro da lesão e o perímetro de seu casco convexo. O casco ou envelope convexo de uma imagem é o menor contorno poligonal convexo contendo todos o contorno da lesão. A Figura 4 mostra o casco convexo (em azul) de um dado contorno (em vermelho).

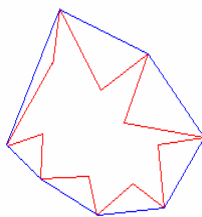


Figura 4: Casco convexo (em azul) de um dado contorno (em vermelho).

Para o cálculo do envelope convexo foi utilizado o algoritmo Graham

Scan [13], que possui custo $O(n \log n)$, onde n é o número de pontos do contorno.

Seja $Ampl$ a relação entre o perímetro da lesão e o perímetro de seu casco convexo, seja PL o perímetro do contorno e PC o perímetro do casco convexo. $Ampl$ é então definido pela Equação 4.

$$Ampl = \frac{PL - PC}{PL} \quad (4)$$

Observe nas Figuras 5 e 6, que quanto maior a diferença entre PL e PC , maior é o valor de $Ampl$, indicando que valores altos para $Ampl$ tendem a caracterizar contornos irregulares.

3.3 Desvio da Área em Relação ao Casco Convexo (Dev)

Uma outra medida que caracteriza a forma é uma modificação do descritor $Ampl$. Neste caso é calculada a relação entre a área do tumor e a área do seu casco convexo. Sejam AL e AC respectivamente a área do contorno da lesão e a área do casco convexo da lesão. Então, o desvio que ocorre entre a área do contorno e a área do seu casco convexo é dado pela Equação 5.

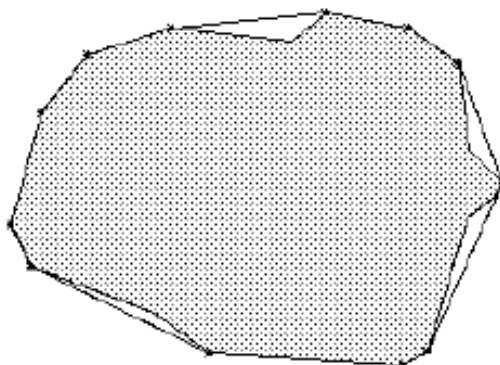


Figura 5: Contorno com $Ampl = 0.02$ e $Dev = 0.04$.

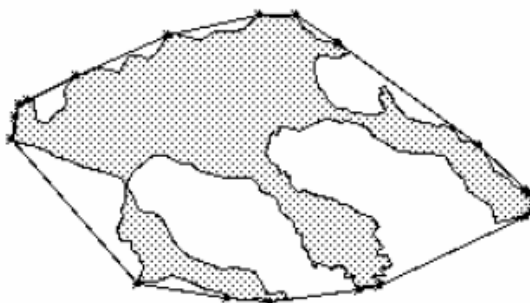


Figura 6: Contorno com $Ampl = 0.5$ e $Dev = 0.47$.

$$Dev = \frac{AC - AL}{AC} \quad (5)$$

Observe nas Figuras 5 e 6, que quanto maior a diferença entre AC e AL, maior é o valor de Dev , indicando que valores altos para Dev tendem a caracterizar lesões com formato mal definido e irregulares.

3.4 Cálculo Final do Índice de Complexidade (CI)

Os três parâmetros anteriormente descritos (Feq , $Ampl$ e Dev) vão ser usados para calcular um valor real (CI) que representa a medida do índice de complexidade de um contorno de lesão. A idéia é que os contornos mais complexos (lesões malignas) possuam um valor de complexidade mais próximo de 1, e os

contornos menos complexos (lesões benignas) apresentem um valor mais próximo de 0. É possível combinar estes parâmetros de várias maneiras, e com diferentes pesos para cada parâmetro, ou usá-los isoladamente. Várias dessas combinações foram testadas e os resultados podem ser vistos na Tabela 1.

A última combinação presente na Tabela 1 e que apresenta o melhor resultado foi proposta por Brinkhoff et al. [12], em que foi observado que o aumento relativo do perímetro (*Ampl*) tem uma correlação mais significativa com o cálculo da complexidade do contorno. Quanto mais rugosa a forma de uma lesão, maior é o valor de *Ampl*. Uma simples irregularidade em um contorno, como a presença de uma única região côncava, pode causar um grande aumento do perímetro (*Ampl*).

Sendo assim, deve-se combinar *Ampl* com *Feq*. Por isso Brinkhoff et al. [12] propuseram que esses dois parâmetros fossem multiplicados, chegando-se à Equação 6.

$$CI = Ampl * Feq + Dev \quad (6)$$

Tabela 1: Combinações dos parâmetros e seus resultados.

Combinação	Resultado (Az)
ampl	0.9090
feq	0.8676
dev	0.8679
ampl*feq	0.9070
ampl*dev	0.8890
feq*dev	0.8712
ampl*feq*dev	0.8893
ampl + feq + dev	0.8893
ampl*feq + dev	0.9102

A curva ROC correspondente à Equação 6 pode ser vista na Figura 7.

A Figura 8 ilustra alguns exemplos em que pode-se perceber que o descritor proposto diferencia bem lesões malignas (como o contorno 8(c)) e benignas (como os contornos 8(a) e 8(b)), pois as primeiras apresentam valores bem superiores em relação às segundas. O número dentro de cada contorno é o valor do índice de complexidade (CI) calculado de acordo com a Equação 6.

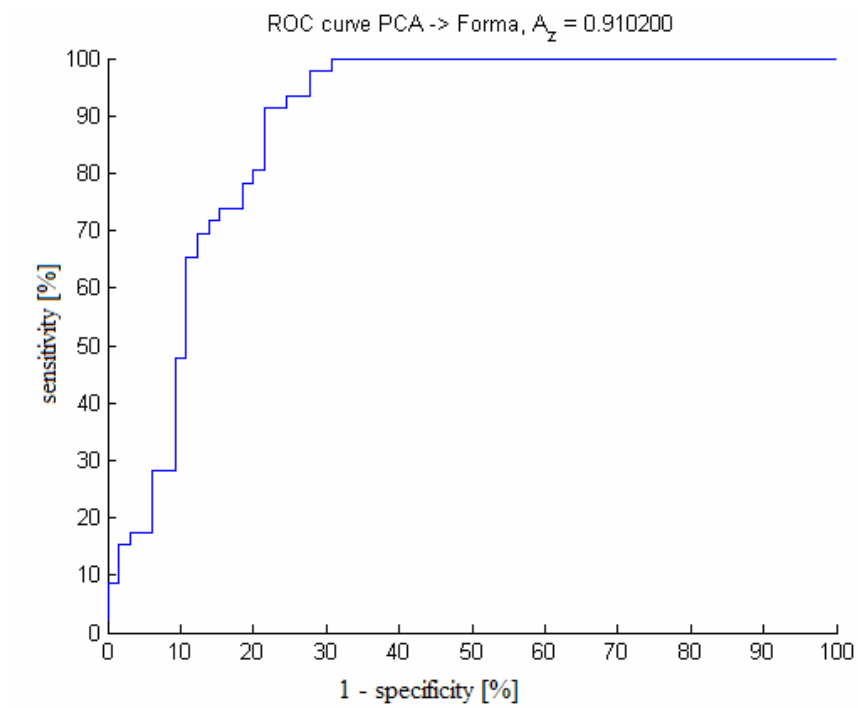


Figura 7: Curva ROC para $CI = \text{ampl} * \text{freq} + \text{conv}$

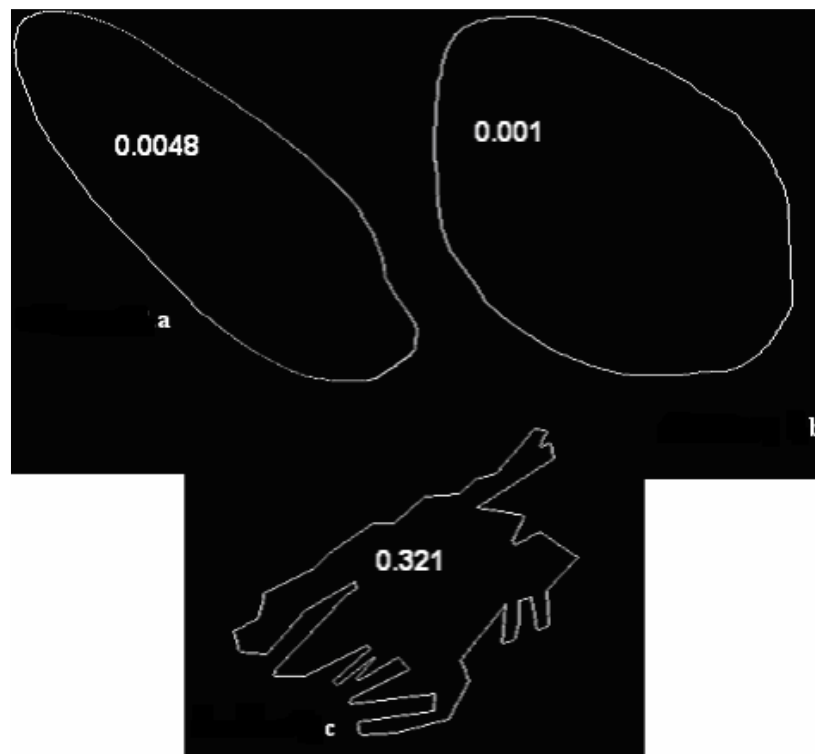


Figura 8: Valor do índice complexidade (CI) de alguns contornos.

4. O EXTRATOR *ELLIPTIC VARIANCE* (Ev)

O extrator *Elliptic Variance* (Ev) tem como objetivo medir o quanto um dado contorno de lesão se aproxima de uma forma elíptica e é uma aplicação do método proposto por Peura e Iivarinen [14].

Algumas definições fazem-se necessárias:

- $pi = \begin{pmatrix} xi \\ yi \end{pmatrix}$ é o i -ésimo ponto do contorno.
- $P = \{ pi \}, 1 \leq i \leq N$, é a representação de um contorno com perímetro igual a N .
- $M = \frac{1}{N} \sum_{i=1}^N pi$, é o centróide do contorno.
- A^T é a matriz transposta da matriz A .
- A^{-1} é a matriz inversa da matriz A .

$$C = \frac{1}{N} \sum_{i=1}^N ((pi - M)(pi - M)^T)$$

, é a matriz de covariância.

Dadas estas definições, são propostas as Equações 7, 8 e 9. O descritor Ev é variância elíptica, ou seja, quanto menor o valor de Ev mais parecida com uma elipse o contorno é. Por conseguinte, quanto maior o valor de Ev menos parecido com uma elipse o contorno é.

$$Bi = \sqrt{(pi - M)^T C^{-1} (pi - M)} \quad (7)$$

$$1 \leq i \leq N$$

$$M_{rc} = \frac{1}{N} \sum_{i=1}^N Bi \quad (8)$$

$$E_v = \frac{1}{NM_{rc}} \sum_{i=1}^N (Bi - M_{rc})^2 \quad (9)$$

Apesar da aparente complexidade este é um algoritmo muito simples, com ordem de complexidade linear.

O resultado (Az) do extrator Ev foi de 0.8532. A curva ROC deste extrator pode ser vista na Figura 9.

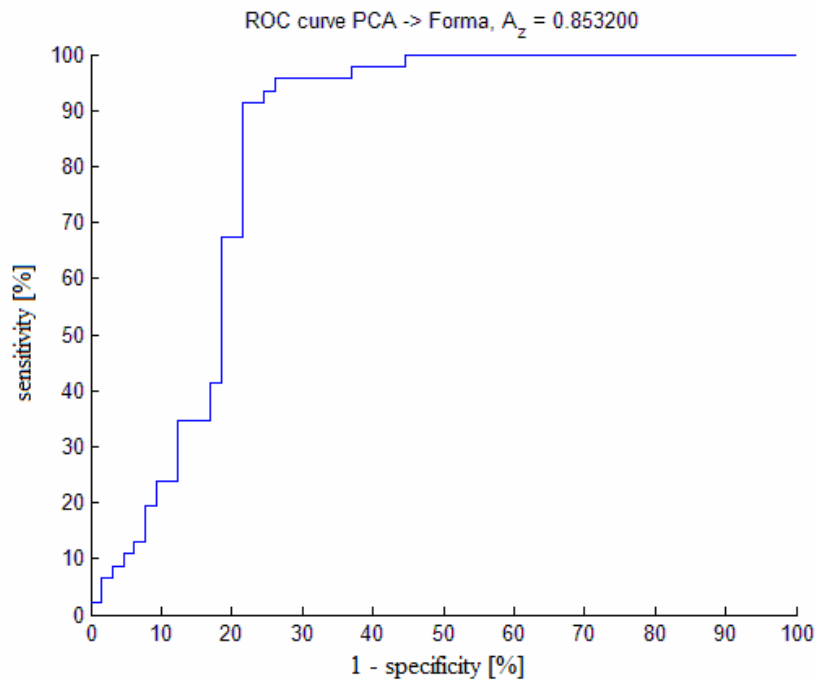


Figura 9: Curva ROC para o extrator Ev.

O objetivo principal deste extrator é medir o quão parecido com uma elipse um dado contorno é. Essa característica mostrou-se útil na distinção entre lesões malignas e benignas, visto que as primeiras, como possuem contornos muito mais irregulares, possuem uma forma muito menos elíptica do que as segundas. Resumindo, em geral o valor de Ev é muito menor em lesões benignas do que malignas. A Figura 10 mostra alguns exemplos.

5. DISCUSSÕES DE RESULTADOS

A Tabela 2 mostra os resultados obtidos pelos extratores de características

apresentados neste artigo, bem como de outros extratores.

Os resultados obtidos pelos dois extratores apresentados neste artigo (CI e Ev) podem ser considerados bons, especialmente o extrator Complexity Index (CI), que como pode ser verificado na Tabela 2 possui o terceiro melhor desempenho

6. CONCLUSÃO

O extrator de características *Complexity Index* (CI) apresentado neste artigo pode ser aplicado para medir o índice de complexidade de qualquer objeto poligonal e não apenas de


BENIGNAS	MALIGNAS
 <p data-bbox="405 846 587 875">$E_v = 0.002747$</p>	 <p data-bbox="884 786 1070 815">$E_v = 0.240215$</p>
 <p data-bbox="252 1272 438 1301">$E_v = 0.002196$</p>	 <p data-bbox="740 1361 927 1391">$E_v = 0.166942$</p>
 <p data-bbox="236 1733 422 1762">$E_v = 0.003175$</p>	 <p data-bbox="794 1697 981 1727">$E_v = 0.148761$</p>

Figura 10: Valor E_v de algumas lesões.

Tabela 2: Comparação dos extratores CI e Ev com outros extratores.

Extrator	Resultado (Az)
Fourier Factor [9]	0.77
Elliptic Variance (Ev)	0.85
Compactness [9]	0.87
Fcc [9]	0.88
FD [9]	0.89
Complexity Index (CI)	0.91
FD _{ta} [10]	0.92
CX _{ta} [10]	0.93

contornos de lesões de mama. Ele será integrado ao sistema AMDI, onde já estão incorporados vários outros extratores de características. Os três parâmetros descritos (*Feq*, *Ampl* e *Dev*) podem ser combinados de maneiras diferentes da mostrada na Equação 6, e os pesos dados aos parâmetros também devem ser ajustados de acordo com a aplicação e a base de dados utilizada. Da mesma maneira, o extrator *Elliptic Variance* também é um extrator de propósito geral podendo ser aplicado a qualquer tipo de contorno, não apenas lesões de mama. Apesar de serem de uso geral, os dois extratores se mostraram úteis no caso específico de tumores de mama

7. AGRADECIMENTOS

Agradecemos à FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais) e a PROPP-UFU (Pró-reitoria de pesquisa e pós-graduação da Universidade Federal de Uberlândia), pelo apoio financeiro.

8. BIBLIOGRAFIA

- [1] Gupta, S.; Chyn P.F. and Markey, M.K. Breast cancer CADx based on BI_RADS™ descriptors from two mammographic views. *Med. Phys.* 33(6):1810-1817. June 2006.
- [2] Boyle P. and Ferlay J. Cancer incidence and mortality in Europe, 2004. *Annals of Oncology* 16, 481. 2005.
- [3] Câncer Facts and Figures 2005. American Câncer Society. 2005.
- [4] INCA – Instituto Nacional do Câncer. Estimativa 2005 – Incidência de Câncer no Brasil. 2005. <http://www.inca.gov.br>. Visitado em 10/02/2009.
- [5] Rangayyan, R.M., Ayres, F.J.; Leo Desautels, J.E. A review of computer-aided diagnosis of breast cancer: Toward the detection of subtle signs. *Journal of The Franklin Institute*. 2007. In press.

Available online at www.sciencedirect.com.

[6] Giger, M. Computer Aided Diagnosis of Breast Lesions in Medical Images. *Computing in Medicine*, pp: 39-45, September/October 2000.

[7] Huo, Z., Giger, M.L. Vyborny, C.J., Wolverton D.E and Metz C.E. Computadorized Classification of benign and malignant masses on digitized mammograms: A study of robustness. *Academic Radiology*, 7: 1077-1084, 2000.

[8] Huo, Z., Giger, M.L. and Vyborny, C.J., Metz, C.E. Breast cancer: effectiveness of computer-aided diagnosis – Observer study with independent database of mammograms. *Radiology*, pp: 224-256, 2002.

[9] Rangayyan RM, Nguyen TM: Fractal analysis of contours of breast masses in mammograms. *J Digital Imaging*, 2007, in press 10.1007/s10278-006-0860-9

[10] Guliato, D.; Rangayyan, Rangaraj M ; Carvalho, J. D. ; Santiago S. A., 2008, “Poligonal approximation of contours based on the turning angle function”. *Journal of Electronic Imaging* (Springfield), v. 17, p. 023016-1-023016-14.

[11] Metz, C. E., Herman, B. A., Shen J. H., 1998, “Maximum-likelihood estimation of receiver operating characteristic (ROC) curves from continuously distributed data”. *Stat Med*; 17:1033:1053.

[12] Brinkhoff, T., Kriegel, H. P., Schneider, R., e Braun, A., 1995, “Measuring the Complexity of Polygona Objects”. In *Proceedings of ACM International Workshop on Advances in Geographic Information Systems*, Baltimore, MD, USA.

[13] Graham, R.L., 1972, “An Efficient Algorithm for Determining the Convex Hull of a Finite Planar Set”. *Information Processing Letters* 1, 132-133.

[14] Peura M, Iivarinen J (1997) Efficiency of simple shape descriptors.

In: Arcelli C, Cordella LP, Sanniti di Baja
G (eds.) Aspects of visual form
processing. World Scientific, Singapore,
pp 443–451