Atualização local automática de pesos de atributos para recuperação de nódulos pulmonares similares

David Jones Ferreira de Lucena*, Msc. José Raniery Ferreira Junior†,
Phd. Marcelo Costa Oliveira‡, Phd. Aydano Pamponet Machado§
Laboratório de Telemedicina e Informática Médica
Instituto de Computação - Universidade Federal de Alagoas - Maceió, Brasil
Email: *davidjones162@gmail.com, †jrfj@ic.ufal.br, ‡oliveiramc@ic.ufal.br, §aydano@ic.ufal.br

***Resumo*—Lung cancer is the third most common among the types of cancer existing in the world, staying back of prostate cancer in men and breast cancer in women. Computer-Aided (CAD) systems have been built in order to help experts identify and classify lung nodules. One type of CAD that has shown good results is the Content-Based Image Retrieval (CBIR). But one of the biggest challenges of CBIR is to define the appropriate measure for evaluating the similarity, other is to find a way to address the gap between the features used by experts to evaluate the images and attributes extracted from it segmentation. This work proposes a CBIR architecture to automatically calculate the weights of the attributes based on local learning to reflect the user interpretation in image retrieval process, reducing the semantic gap and improving performance in the recovery based on content.**

***Resumo*—O câncer de pulmão é o terceiro mais comum entre os tipos de câncer existentes no mundo, ficando atrás apenas do câncer de próstata nos homens e de mama nas mulheres. Muitos sistemas de auxílio computadorizado têm sido construídos com o propósito de ajudar os especialistas a identificar e classificar os nódulos pulmonares. Um tipo de sistemas que vem apresentando bons resultados é o *Content-Based Image Retrieval* (CBIR). Um dos grandes desafios dos sistemas CBIR é definir a medida apropriada para avaliar a similaridade, outro é encontrar uma forma de resolver o *gap* entre as características utilizadas pelos especialistas para avaliar as imagens e os atributos extraídos a partir da sua segmentação. Este trabalho propõe uma arquitetura CBIR capaz de calcular automaticamente os pesos dos atributos baseada em aprendizagem local para refletir a interpretação usuário no processo de recuperação de imagens, minimizando o *gap* semântico e melhorando a performance na recuperação baseada em conteúdo.**

***Keywords*-Content-based image retrieval; information retrieval; decision support; update weighing attributes; lung cancer.**

I. Introdução

O câncer de pulmão é o terceiro mais comum entre os tipos de câncer existentes no mundo, ficando atrás apenas do câncer de próstata nos homens e de mama nas mulheres [1]. Ela é uma doença com natureza agressiva e silenciosa. Isso faz com que os sintomas sejam percebidos quando a doença já está em estágio avançado. Aliado a isso, falhas na detecção das lesões pulmonares em estágios iniciais podem ocorrer por causa do pequeno tamanho dos nódulos e por estarem anexos à estruturas anatômicas complexas [2].

Nas últimas décadas, métodos de auxílio computadorizado têm sido desenvolvidos para a detecção e caracterização de nódulos pulmonares de imagens de tomografia computadorizada. Eles são baseados em técnicas de visão computacional, processamento de imagens e reconhecimento de padrões.

Com o propósito de fornecer uma segunda opinião ao radiologista no diagnóstico, surgiu o conceito de *Computer-Aided Diagnosis* (CAD). Este tipo de sistema não visa diagnosticar o paciente, mas auxiliar ao especialista fornecendo meios para que sejam obtidos melhores resultados na detecção de nódulos pulmonares [3].

Content-Based Image Retrieval (CBIR) é um *framework* CAD que provê suporte aos especialista recuperando imagens baseadas em conteúdo tanto da imagem quanto dos exames. É uma das áreas de pesquisa mais vivas no campo da visão computacional nas últimas décadas [4]. Ele tem sido utilizado para recuperar exames similares a um exame de referência com base em critérios de similaridade, permitindo que especialistas analisem outros casos já diagnósticados semelhantes a um novo ainda não diagnostiado [3].

Um dos grandes desafios dos sistemas CBIR é definir a medida apropriada para avaliar a similaridade que será utilizada para indexação do *ranking* das imagens recuperadas, outro é encontrar uma forma de resolver o *gap* entre as características utilizadas pelos especialistas para avaliar as imagens (informações semânticas extraídas pelo conhecimento adquirido do usuário) e os atributos extraídos pelas técnicas de análise de imagens [5].

Quando o usuário classifica um objeto, ele o faz com base em informações coletadas dos exames, imagens e quaisquer outros meios possíveis. Mas cada uma das informações utilizadas tem maior ou menor influência na tomada de decisão pelo especialista. Pensando desta forma, existem atributos com mais ou menos peso na classificação e isso introduz um fator semântico no problema da recuperação de imagens. Portanto, encontrar uma forma de mensurar o peso dos atributos permite que a recuperação seja melhorada através da atribuição de maior influência daqueles que melhor caracterizam os objetos em detrimento daqueles que não tem uma boa capacidade de representação [6].

O objetivo deste trabalho é apresentar um método capaz de calcular automaticamente os pesos dos atributos baseada em aprendizagem local com o com o objetivo de refletir a interpretação usuário no processo de recuperação de imagens.

II. IMPLEMENTAÇÃO

A. Base de imagens

A base de imagens utilizada neste trabalho é uma extensão da base de imagens do *Lung Image Database Consortium* (LIDC) [7] apresentada por Ferreira Junior & Oliveira [8]. Ela é uma base não-relacional pública, orientada a documentos, de nódulos pulmonares identificados e classificados por especialistas, caracterizados por Atributos de Textura 3D (ATs 3D). Cada um deles possui um atributo de probabilidade de malignidade com valores entre 1 e 5: *a*) 1 para probabilidade alta para ser benigno; *b*) 2 para probabilidade moderada para ser benigno; *c*) 3 para malignidade indeterminada do nódulo; *d*) 4 para probabilidade moderada para ser maligno; e *e*) 5 para probabilidade alta para ser maligno.

Foram descartados os nódulos com classificação 3, já que a probabilidade de malignidade é indeterminada, enquanto que as classificações 1, 2, 4 e 5 dão indícios ao radiologista de que o nódulo que está sendo analisado é benigno ou maligno.

Os ATs 3D de interesse neste trabalho estão definidos em [9] e são os seguintes: energia, entropia, contraste, momento de diferença inverso, matiz, proeminência, correlação, variância e homogeneidade. Eles são calculados a partir de uma extensão da Matriz de CoOcorrências (MCO 3D) para manipular volumes de imagens. Desta forma, os 9 ATs são calculados em 4 direções diferentes (0º, 45º, 90º e 135º) da MCO 3D.

B. Normalização

Cada atributo extraído possui o seu próprio intervalo de valores e que não são necessariamente coincidentes. Para utilizar métricas de similaridades baseadas em distância, faz-se necessária a normalização da base para que todos os dados se localizem em um intervalo de valores específico. O método de normalização aplicado neste trabalho foi o método Transformação Z, também conhecido como Normalização Estatística apresentado em Lucena *et al.* [9].

C. Métrica de similaridade

A medida de similaridade utilizada neste trabalho é a Distância Euclidiana Ponderada (DEP). Essa é uma variação da Distância Euclidiana (DE) que possui pesos associados às coordenadas dos vetores. Ela é utilizada para medir a similaridade entre dois nódulos através da comparação dos seus vetores de AT 3D. Quanto mais semelhantes as representações vetoriais dos nódulos forem, mais próximo de zero será o valor de sua distância, enquanto que quanto mais diferentes forem, maiores serão os valores da distância calculada.

A sua forma está descrita na Equação 1, onde $\vec{x} = [x_1, \ldots, x_{36}]$ como o vetor AT da imagem de referência, $\vec{y} = [y_1, \ldots, y_{36}]$ sendo o vetor AT das imagens que serão comparadas, $\vec{w} = [w_1, \ldots, w_{36}]$ como o vetor de pesos associados a cada um dos atributos.

$$d(\vec{x}, \vec{y}) = \sqrt{w_1(x_1 - y_1)^2 + \cdots + w_{36}(x_{36} - y_{36})^2} \quad (1)$$

Os pesos representam a influência dos atributos no processo de recuperação de nódulos semelhantes pela CBIR. Seu cálculo leva ao seguinte raciocínio: maiores valores serão associados aos atributos que têm valores homogênios, e menores valores serão associados aos atributos cujos valores possuam grande variabilidade. Identificar quais atributos carregam as informações mais relevantes para classificar as lesões pode ajudar sobremaneira no aumento da acurácia na avaliação médica, por propiciar resultados mais precisos dos algoritmos de recuperação de imagens [6].

D. Processo de atualização local de pesos automática

O processo é composto por duas fases que são executadas de forma sequencial e cíclica: uma Fase de Avaliação e uma Fase de Treinamento (Figura 1).

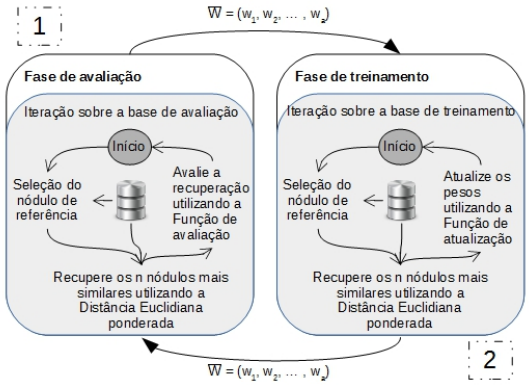

Figura 1. *Workflow* do processo de atualização de pesos

As fases possuem estruturas semelhantes que consistem basicamente em aplicar a técnica *Leave-One-Out* [10] para fazer iterações sobre as bases de dados correspondentes selecionando cada um dos nódulos armazenados e utilizando-os como nódulo de referência para a recuperação dos n nódulos mais semelhantes através de uma Métrica de Similaridade.

O ciclo se inicia com a Fase de Avaliação e em seguida a Fase de Treinamento (Figura 1). Este processo deve ser executado até que algum critério de parada seja alcançado.

E. Fase de Avaliação

Esta fase consiste em fazer uma iteração na base de dados de avaliação também com a técnica LOO utilizando cada um dos nódulos como nódulo de referência para recuperar os n mais similares. A cada recuperação feita é calculado o valor de avaliação v da recuperação através da Equação 2, que é uma função de decaimento exponencial, onde: $R_{n \times a}$ é a matriz ordenada com os nódulos recuperados na qual n corresponde ao número de nódulos similares que serão recuperados, e a é a quantidade de atributos. A ordem da tabela é determinada pela similaridade dos nódulos, os mais similares situam-se nas posições iniciais; s_i é o valor da recompensa associado à relevância do nódulo n da matriz R na posição i; $\{\gamma \in R \mid 0 < \gamma \leq 1\}$ é o fator de desconto, que ajusta a relevância das recompensas s dadas ao longo do *ranking* de recuperação.

$$f(R_{n\times a}) = \sum_{i=1}^{n} \gamma^i s_i \quad (2)$$

Esta Função de Avaliação foi adotada por ter a capacidade de representar a amortização das recompensas ao longo da ordem de recuperação, que é uma característica importante para a nossa proposta, pois, devido à grande quantidade de exames recuperados, os usuários tendem a avaliar os resultados melhores colocados e esses direcionarão os especialistas no diagnóstico [4], [11].

As recompensas aplicadas aos nódulos dependem da malignidade do nódulo de referência e da malignidade do nódulo recuperado. Os valores atribuídos são os seguintes: 4, se for altamente relevante; 2, se for moderadamente relevante; e 0 se for moderadamente ou altamente irrelevante.

F. Fase de Treinamento

A Fase de Treinamento tem por objetivo encontrar um conjunto W de pesos associados aos atributos que permitam uma recuperação na qual os nódulos sejam mais semelhantes.

Consideram-se nódulos semelhantes aqueles em que os valores de cada um dos atributos são muito próximos, ou até mesmo iguais.

Os pesos associados aos atributos refletem as diferentes contribuições dos descritores na caracterização do objeto. Não existe um mapeamento direto entre os critérios de classificação utilizados pelo usuário e a forma de representação dos objetos pela máquina. O que se busca aqui é uma adequação dos pesos de tal forma que seja possível alcançar melhores resultados na recuperação de nódulos semelhantes por meio da minimização da influência de atributos que possuem alto índice de dispersão em nódulos de mesma malignidade e aumento da influência dos atributos que possuem baixo índice de dispersão em nódulos de mesma malignidade.

Neste trabalho, é apresentada uma proposta de atualização de pesos baseada no desvio padrão para atualização dos pesos (Equação 4) e o ajuste se dará pela Equação 6 que ajusta o peso a cada iteração. Para nosso conhecimento, está é uma técnica inédita aplicada ao contexto de sistemas CBIR.

O desvio padrão (σ) é uma medida de dispersão estatística, ou seja, ele mede a dispersão dos dados de uma amostra com relação a sua média (Equação 3), onde: t é o tamanho da amostra e $\overline{x}$ é a média da amostra.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{t}(x_i - \overline{x})^2}{t-1}} \quad (3)$$

A proposta é baseada nas seguintes premissas: se todos os nódulos semelhantes têm valores similares para um determinado conjunto de atributos, isso significa que esses são bons indicadores para representar aqueles nódulos. Por outro lado, se os valores de um conjunto de atributos são muito diferentes, ou seja, muito dispersos, então eles não são bons indicadores. Logo, o inverso do desvio padrão (Equação 4) dos dados associados a um atributo pode ser considerada uma boa estimativa para o seu peso, porque quanto menor a variância, maior é o peso e vice-versa.

$$w(a) = \sigma^{-1} \quad (4)$$

Para ilustrar a atualização dos pesos dos atributos com base nos n nódulos recuperados observe a Figura 2, considerando: *a*) n o número de nódulos recuperados; *b*) f o número de atributos usados para representar cada nódulo; *c*) Π_f representando a projeção do atributo f na matriz de nódulos recuperados; *d*) $\sigma^{-1}(\Pi_f)$ representando a aplicação do inverso do desvio padrão sobre a amostra resultante da projeção Π_f; *e*) w_f como sendo o peso do atributo a_f.

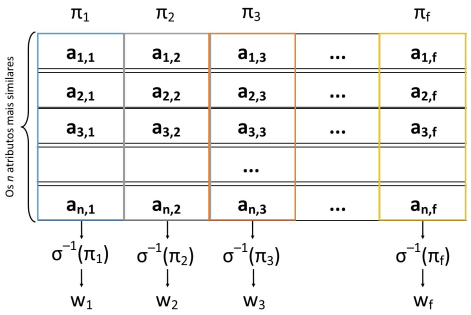

Figura 2. Método de atualização de pesos

Após a identificação de cada w_f associado aos atributos a_f é preciso aplicar a Equação 5, porque $w_f \in R_+^*$ e ele pode assumir valores muito grandes quando a amostra variar pouco, ou muito pequenos quando a amostra variar muito. Com isso, temos um novo peso $w'f$ com intervalo $(0, 1]$.

$$w'_f = \frac{w_f}{\sum_{i=1}^{f} w_i} \quad (5)$$

A cada iteração é encontrado um $WCorrente$, ele é usado para ajustar o conjunto de pesos W usando a Função de Ajuste (Equação 6), onde: *a*) W é o conjunto com os melhores pesos até o momento; *b*) $WCorrente$ é o conjunto de pesos da iteração corrente; *c*) α é o fator de ajuste; *d*) W^* é o novo conjunto de pesos ajustados.

$$W^* = W + \alpha(W - WCorrente) \quad (6)$$

III. Resultados e Discussão

Os gráficos 3 e 4 apresentam os resultados dos cálculos das precisões na recuperação dos nódulos de classificação 1 e 5, respectivamente, utilizando AT 3D aplicados à DEP com os pesos definidos por meio do processo de atualização de pesos proposto neste trabalho. Os nossos resultados estão sendo comparados com os do trabalho apresentado em nosso trabalho preliminar [9], onde foi utilizada a mesma arquitetura CBIR, entretanto, utilizando Distância Euclidiana (sem pesos).

A utilização do método aqui descrito proporcionou um aumento de 18,82% na precisão da recuperação de 10 nódulos

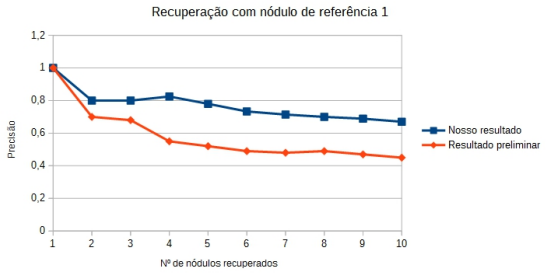

Figura 3. Resultados da recuperação de nódulos com malignidade 1

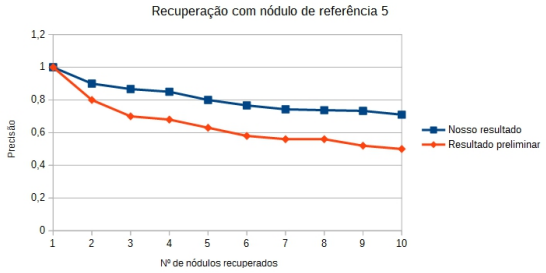

Figura 4. Resultados da recuperação de nódulos com malignidade 5

pulmonares com malignidade 1 e 15,77% na precisão da recuperação de 10 nódulos com maligindade 5 em comparação com os resultados em nosso trabalho preliminar [9]. Os pesos utilizados na DEP foram definidos após 10 iterações de treinamento sem acréscimos no valor de avaliação de recuperação com a taxa de aprendizagem (λ) igual a 0,3. O aumento médio na precisão foi de 17,3%.

Para poder comparar os resultados obtidos, a precisão foi calculada segundo a Equação 7, onde: *a*) P - vetor de tamanho i com as precisões da recuperação. Cada posição possue a precisão na ordem n; *b*) VP - é o número de Verdadeiros-Positivos obtidos até a ordem n; *c*) T - é o número total de nódulos recuperados até a ordem n; *d*) n - número da ordem na recuperação; e *e*) i - número de nódulos recuperados.

$$P_i = \sum_{n=1}^{i} \frac{VP_n}{T_n} \tag{7}$$

A. Limitação

Embora os resultados tenham mostrado um aumento médio de quase 20% na precisão da recuperação dos nódulos similares, a metodologia aqui apresentada não alcançou precisão acima de 90% que é um valor já alcançado por outros modelos de sistemas baseados em classificação como o apresentado por Han *et al.* [12]. Além disso, ainda não foi criada uma relação entre a atualização dos pesos e as probabilidades de malignidade e isso pode levar a melhores resultados por adicionar informações ao método de recuperação.

IV. Conclusão

Neste trabalho, para o nosso conhecimento, foi apresentada uma técnica inédita aplicada a uma arquitetura CBIR que permite identificar automaticamente os pesos dos atributos em sistemas que utilizam a DEP como métrica de similaridade. Os testes executados apresentaram o aumento médio da precisão na recuperação de nódulos com probabilidade de malignidade 1 e 5 de 17,3% em comparação com os resultados preliminares [9].

Referências

[1] R. Wender, E. T. H. Fontham, E. Barrera, G. A. Colditz, T. R. Church, D. S. Ettinger, R. Etzioni, C. R. Flowers, G. Scott Gazelle, D. K. Kelsey, S. J. LaMonte, J. S. Michaelson, K. C. Oeffinger, Y.-C. T. Shih, D. C. Sullivan, W. Travis, L. Walter, A. M. D. Wolf, O. W. Brawley, and R. A. Smith, "American cancer society lung cancer screening guidelines," *CA: A Cancer Journal for Clinicians*, vol. 63, no. 2, pp. 106–117, 2013. [Online]. Available: http://dx.doi.org/10.3322/caac.21172

[2] National Lung Screening Trial Research Team, D. R. Aberle, A. M. Adams, C. D. Berg, W. C. Black, J. D. Clapp, R. M. Fagerstrom, I. F. Gareen, C. Gatsonis, P. M. Marcus, and J. D. Sicks, "Reduced lung-cancer mortality with low-dose computed tomographic screening," *The New England journal of medicine*, vol. 365, no. 5, p. 395—409, August 2011. [Online]. Available: http://dx.doi.org/10.1056/NEJMoa1102873

[3] K. Doi, "Computer-aided diagnosis in medical imaging: Historical review, current status and future potential," *Computerized Medical Imaging and Graphics*, vol. 31, no. 4-5, pp. 198–211, 2007, computer-aided Diagnosis (CAD) and Image-guided Decision Support. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0895611107000262

[4] H. Müller, N. Michoux, D. Bandon, and A. Geissbuhler, "A review of content-based image retrieval systems in medical applications—clinical benefits and future directions," *International Journal of Medical Informatics*, vol. 73, no. 1, pp. 1 – 23, 2004. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1386505603002119

[5] C. Akgül, D. Rubin, S. Napel, C. Beaulieu, H. Greenspan, and B. Acar, "Content-based image retrieval in radiology: Current status and future directions," *Journal of Digital Imaging*, vol. 24, no. 2, pp. 208–222, 2011. [Online]. Available: http://dx.doi.org/10.1007/s10278-010-9290-9

[6] G. J. C. A. Faceli K., Lorena A. C., *Inteligência artificial - Uma abordagem de aprendizagem de máquina*. LTC, 2011.

[7] S. G. Armato III, G. McLennan, L. Bidaut, M. F. McNitt-Gray, C. R. Meyer, A. P. Reeves, B. Zhao, D. R. Aberle, C. I. Henschke, E. A. Hoffman *et al.*, "The lung image database consortium (lidc) and image database resource initiative (idri): a completed reference database of lung nodules on ct scans," *Medical physics*, vol. 38, no. 2, pp. 915–931, 2011.

[8] O. M. C. FERREIRA JUNIOR, J. R., "Banco de dados nosql público de nódulos pulmonares para auxílio à pesquisa e diagnóstico do câncer de pulmão," in *Publicado em anais do XXIV Congresso Brasileiro de Engenharia Biomédica (CBEB)*, 2014, pp. 177–180.

[9] F. J. J. R. O. M. C. LUCENA, D. J. F., "Avaliação da precisão de atributos de textura 3d normalizados aplicados à recuperação de nódulos pulmonares similares," in *Publicado em anais do XIV Congresso Brasileiro de Informática em Saúde (CBIS)*, 2014.

[10] T. Deselaers, D. Keysers, and H. Ney, "Features for image retrieval: A quantitative comparison," in *Pattern Recognition*. Springer, 2004, pp. 228–236.

[11] F. F. Faria, A. Veloso, H. M. Almeida, E. Valle, R. d. S. Torres, M. A. Gonçalves, and W. Meira, Jr., "Learning to rank for content-based image retrieval," in *Proceedings of the International Conference on Multimedia Information Retrieval*, ser. MIR '10. New York, NY, USA: ACM, 2010, pp. 285–294. [Online]. Available: http://doi.acm.org/10.1145/1743384.1743434

[12] F. Han, H. Wang, G. Zhang, H. Han, B. Song, L. Li, W. Moore, H. Lu, H. Zhao, and Z. Liang, "Texture feature analysis for computer-aided diagnosis on pulmonary nodules," *Journal of digital imaging*, pp. 1–17, 2014.