UNIVERSIDADE FEDERAL DO CEARÁ

CENTRO DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA DE TELEINFORMÁTICA

PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA DE TELEINFORMÁTICA

ROMUERE RODRIGUES VELOSO E SILVA

DETECÇÃO AUTOMÁTICA DE CÉLULAS CERVICAIS COM BASE EM ATRIBUTOS RADIAIS

FORTALEZA

2018

ROMUERE RODRIGUES VELOSO E SILVA

DETECÇÃO AUTOMÁTICA DE CÉLULAS CERVICAIS COM BASE EM ATRIBUTOS RADIAIS

Tese apresentada ao Curso de Doutorado em Engenharia de Teleinformática do Programa de Pós-Graduação em Engenharia de Teleinformática do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do título de doutor em Engenharia de Teleinformática. Área de Concentração: Sinais e Sistemas

Orientadora: Prof[a]. Dr[a]. Fátima Nelsizeuma Sombra de Medeiros

Coorientadora: Dr[a]. Daniela Mayumi Ushizima

FORTALEZA

2018

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

S583d Silva, Romuere Rodrigues Veloso e.
Detecção Automática de Células Cervicais com Base em Atributos Radiais / Romuere Rodrigues Veloso e Silva. – 2018.
84 f. : il. color.

Tese (doutorado) – Universidade Federal do Ceará, Centro de Tecnologia, Programa de Pós-Graduação em Engenharia de Teleinformática, Fortaleza, 2018.
Orientação: Profa. Dra. Fátima Nelsizeuma Sombra de Medeiros.
Coorientação: Profa. Dra. Daniela Mayumi Ushizima.

1. Descritor de Atributos Radiais. 2. Papanicolau. 3. Classificação. 4. CBIR. I. Título.

CDD 621.38

ROMUERE RODRIGUES VELOSO E SILVA

DETECÇÃO AUTOMÁTICA DE CÉLULAS CERVICAIS COM BASE EM ATRIBUTOS RADIAIS

Tese apresentada ao Curso de Doutorado em Engenharia de Teleinformática do Programa de Pós-Graduação em Engenharia de Teleinformática do Centro de Tecnologia da Universidade Federal do Ceará, como requisito parcial à obtenção do título de doutor em Engenharia de Teleinformática. Área de Concentração: Sinais e Sistemas

Aprovada em: 03 de Julho de 2018

BANCA EXAMINADORA

Profª. Drª. Fátima Nelsizeuma Sombra de Medeiros (Orientadora)
Universidade Federal do Ceará (UFC)

Drª. Daniela Mayumi Ushizima (Coorientadora)
University of California - Berkeley

Profª. Drª. Andrea Gomes Campos Bianchi
Universidade Federal de Ouro Preto (UFOP)

Prof. Dr. Danielo Gonçalves Gomes
Universidade Federal do Ceará (UFC)

Prof. Dr. Rodrigo de Melo Souza Veras
Universidade Federal do Piauí (UFPI)

Anselmo Cardoso de Paiva
Universidade Federal do Maranhão (UFMA)

Aos meus pais Conceição de Maria de Sousa Silva e Renato José Rodrigues da Silva.

AGRADECIMENTOS

Agradeço primeiramente a Deus, pelo dom da vida. À minha esposa, Lilian Silva, por todo apoio e compreensão. Agradeço aos meus pais, Renato José Rodrigues da Siva e Conceição de Maria de Sousa Silva, por me apoiarem em todos os momentos da minha vida. À minha irmã Rayssa Maria por estar sempre comigo.

À minha orientadora, Fátima Medeiros, pelos ensinamentos e conselhos nesse período. À minha coorientadora Daniela Ushizima por me acolher durando o doutorado sanduíche.

Aos meus amigos do LABVIS em especial Alan Braga, Alixandre Ávila, Allan Carneiro, Alcilene Dalília, Daniel Silva, Deborah Maria, Flávio Henrique, Geovani Martins, José Gerardo, Marcelo Marques, Ricardo Nobre e Régia Araújo.

Aos colaboradores do NUPEB e do LAPAC da Universidade Federal de Ouro Preto (UFOP) por contribuírem no desenvolvimento desta pesquisa e por formarem o CRIC, em especial a professora Dra. Andrea Bianchi, Paulo Calaes, Dra. Alessandra Tobias, professora Dra. Cláudia Carneiro e Mariana Trevisan. Gostaria de agradecer ao Programa Âmbar que trata de desafios e ações em saúde da mulher da UFOP.

Ao *Berkeley Institute for Data Science* e ao *Lawrence National Berkeley Laboratory* em especial a Kevin Koy, Fernando Peréz e James Sethian por toda ajuda no período de doutorado sanduíche.

Ao PPGETI pela oportunidade de participar deste doutorado e à Universidade Federal do Piauí (UFPI) por ter me liberado para capacitação.

Aos demais amigos Antonio Oseas, Christiano Rodrigues, Ítalo Linhares, Thalisson Oliveira.

Ao CNPq pelo apoio financeiro que através do projeto Pesquisador Visitante Estrangeiro (PVE) do Programa Ciência sem Fronteiras (CsF) possibilitou a realização do doutorado sanduíche no exterior, tal fato foi determinante para o desenvolvimento desta tese e crescimento profissional e pessoal.

“Education is the most powerful weapon which you can use to change the world.”

(Nelson Mandela)

RESUMO

A quantificação microscópica de propriedades de células cervicais tem sido utilizada por décadas na detecção de lesões pré-cancerígenas do colo do útero. A abordagem tradicional é baseada na busca visual de células cervicais em lâminas do exame Papanicolau, onde o objetivo é encontrar padrões que sejam correlacionados com células anormais. O maior desafio na análise de células cervicais em programas de saúde pública é a inspeção manual, uma tarefa que não é escalável com o crescimento da população. Esta tese apresenta ferramentas para classificação de células e recuperação de imagens, incluindo o descritor de atributos radiais (*Radial Feature Descriptor* - RFD) para categorizar padrões normais e anormais de células cervicais. O RFD define retas igualmente espaçadas ao redor do núcleo que são responsáveis por capturar variações de intensidade na membrana citoplasmática. Ele combina os atributos de intensidade com atributos de textura provenientes do cálculo da distribuição de cromatina no núcleo sem a necessidade de segmentação do citoplasma. A avaliação dos resultados utilizou duas bases de imagens: Herlev, uma base pública; e CRIC, uma base de imagens apresentada nesta tese. As células individuais da base CRIC foram obtidas pelo uso de um algoritmo de segmentação de núcleos proposto. Realizamos experimentos de classificação de imagens e de recuperação de imagens baseada em conteúdo (*Content-Based Image Retrieval* - CBIR). Nós classificamos as células com o algoritmo *Random Forest* e utilizamos a metodologia *bootstrap* para criar os conjuntos de treinamento e teste. Os experimentos CBIR foram realizados utilizando a distância cosseno como métrica de similaridade. As principais contribuições desta tese são: a) um novo método para segmentação de núcleos de células cervicais b) RFD, um extrator de atributos de células cervicais; c) pyCBIR, uma nova ferramenta para recuperação de imagens.Nossos experimentos de classificação foram mensurados em termos do índice Kappa (κ) e da taxa de falso negativos (*False Negative Rate* - FNR), calculamos o *Mean Average Precision* (MAP) para avaliar os experimentos CBIR. Os resultados obtidos mostraram que o descritor proposto permitiu a discriminação de células cervicais para a classificação em normais e anormais, alcançando κ = 0,89 e FNR = 0,02 para a base de imagens Herlev e κ = 0,78 e FNR = 0,14 para a base CRIC. Em relação aos experimentos CBIR obtivemos MAP = 0,84 e MAP = 0,82 para as bases de imagens Herlev e CRIC, respectivamente. Em ambos experimentos alcançamos resultados melhores quando comparados com métodos do extração de atributos do estado-da-arte.

Palavras-chave: Descritor de Atributos Radiais. Papanicolau. Classificação. CBIR.

ABSTRACT

Microscopic quantification of cervical cell properties has been used for the early detection of precancerous lesions from Pap smears for decades. The traditional approach relies on the visual screening of a Pap smear glass slide to search for patterns correlated to abnormal cells. The major challenge in cervical cell screening for public health programs is the reliance upon manual inspection by different pathologists, a task that does not scale to the population growth. This work introduces cell classification and image retrieval computational tools, including a new radial feature description (RFD) to distinguish normal and abnormal patterns. The key idea lies in defining evenly interspaced segments around the cell nucleus, and proportional to the convexity of the nuclear boundary. The main advantage of the proposed RFD is the sensitivity to the intensity variation around the nuclear membrane, without cytoplasm outlining, and combining chromatin distribution through texture features. For performance evaluation, we applied two databases: the Herlev and a higher resolution one called BHS, both with thousands of samples. We create the BHS database by applying a new nucleus segmentation method that we proposed here. Then, we classify cells with Random Forest and bootstrap, and perform content-based image retrieval with the cosine similarity, comparing our methodology to other cell recognition techniques. The main contributions are: a) a new method for cervical cell nuclei segmentation; b) RFD as a fast and an accurate cervical cell descriptor, without prior cytoplasm segmentation; b) the BHS database with high-resolution images and ground-truth; c) pyCBIR, a new tool for image retrieval; d) classification and CBIR results using 14 different algorithms including the proposed RFD and two convolutional neural networks. Our results show that the proposed RFD allows accurate discrimination between normal and abnormal cervical cells, achieving the highest accuracy measures in terms of Kappa for both Herlev and BHS cell image sets.

Keywords: Radial feature description. Pap smear. CBIR.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ASM	*Angular Second Moment*
CBIR	*Content-Based Image Retrieval*
CNN	*Convolutional Neural Network*
CRIC	*Cell Recognition for the Inspection of the Cervix*
EOH	*Edge Orientation Histogram*
FN	Falso Negativo
FNR	*False Negative Rate*
FP	Falso Positivo
GLCM	*Gray-Level Co-occurrence Matrix*
GLN	*Gray Level Nonuniformity*
GLRLM	*Gray-Level Run Length Matrix*
HGRE	*High Gray Level Run Emphasis*
HOG	*Histogram of Oriented Gradients*
IGR	*Information Gain Ratio*
KNN	*K-Nearest Neighbor*
LBP	*Local Binary Pattern*
LGRE	*Low Gray Level Run Emphasis*
LRE	*Long Run Emphasis*
LRHGE	*Long Run High Gray Level Emphasis*
LRLGE	*Long Run Low Gray Level Emphasis*
LSH	*Locality Sensitive Hashing Forest*
MAP	*Mean Average Precision*
PCA	*Principal Component Analysis*
RF	*Random Forest*
RFD	*Radial Feature Descriptor*
RH	*Radial Histogram*
RLN	*Run Length Nonuniformity*
ROC	*Receiver Operating Characteristic*
RP	*Run Percentage*
SIFT	*Scale Invariant Feature Transform*
SRE	*Short Run Emphasis*

SRHGE *Short Run High Gray Level Emphasis*

SRLGE *Short Run Low Gray Level Emphasis*

SURF *Speeded-Up Robust Features*

SUS Sistema Único de Saúde

SVM *Support Vector Machine*

SUMÁRIO

1 INTRODUÇÃO

Esta tese apresenta algoritmos para caracterização de células cervicais do colo do útero utilizando imagens do exame Papanicolau em meio convencional. Nesse capítulo, apresentamos o problema, os objetivos da tese, bem como as principais contribuições alcançadas e a produção científica.

1.1 Contextualização

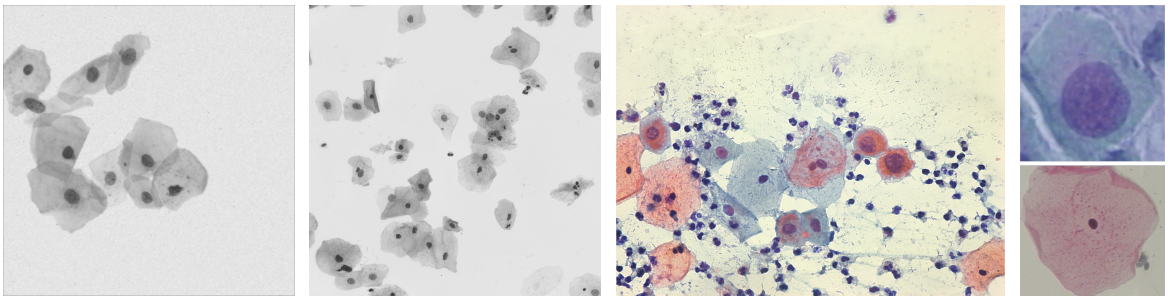
Muitos tipos de câncer, incluindo câncer cervical do colo do útero, possuem cura caso sejam detectados nos estágios iniciais. O câncer cervical afeta mulheres em todas as partes do mundo, sendo um dos mais comuns em países em desenvolvimento como Brasil (Instituto Nacional de Câncer, 2018). Mundialmente, o método mais comum para identificação de lesões pré-cancerígenas em estágios iniciais é o esfregaço convencional (exame Papanicolau). No exame, é coletada uma amostra do tecido do colo do útero, essa amostra é preparada em uma lâmina para posterior análise, que é feita por um citologista utilizando um microscópio com aumento variando de $20\times$ a $40\times$. Com isso, é possível obter imagens de campos dessa lâmina. Durante a análise conduzida pelo citologista, é feita uma busca visual por células anormais, focando em características da célula que são associadas com alterações morfológicas, tais como: tamanho de núcleo e citoplasma, distribuição de cromatina no núcleo, formas de aglomerados de células, razão entre as áreas do núcleo e citoplasma, dentre outras.

Sistemas para detecção automática de células cervicais em meio convencional do exame Papanicolau buscam: 1) separar células em classes (normais e anormais), ou várias classes, nesse caso, cada classe representa um grau da lesão; ou 2) recuperar imagens baseadas em conteúdo -*Content-Based Image Retrieval* (CBIR).

Na literatura, existem vários métodos para processamento e análise automática de células cervicais (IRSHAD *et al.*, 2014; LU *et al.*, 2016), a maioria deles tem como foco a segmentação de células, extração de características, classificação e CBIR. Apesar dos vários trabalhos associados, pesquisas em detecção de câncer cervical continuam com carência em bases de imagens públicas de qualidade. Na literatura, pode-se observar que as bases de imagens mais utilizadas são: 1) ISBI (LU *et al.*, 2015), nessa base são disponibilizadas imagens simuladas (Figura 1 (a)) e reais em meio líquido[1] (Figura 1 (b)). Essa base apresenta somente imagens de

[1] Exame mais avançado, porém apresenta custos muito elevados se comparado com o meio convencional. O meio

células saudáveis, com isso, seu uso é viabilizado somente para o desenvolvimento de algoritmos de segmentação de núcleo e citoplasma. 2) Herlev (JANTZEN *et al.*, 2005), essa base é composta por imagens reais de células recortadas (Figura 1 (d)), que podem ser células saudáveis e anormais. Essa base é largamente utilizada em experimentos de classificação. Entretanto a realidade do citologista são imagens de exames em meio convencional do exame Papanicolau (Figura 1 (c)).



(a) (b) (c) (d)

Figura 1 – Tipos de imagens de células cervicais. (a) Imagem sintética. (b) Meio líquido. (c) Meio convencional. (d) Exemplos da base Herlev.

A Figura 2 mostra um fluxograma geral para detecção automática de imagens de células cervicais. O passo inicial é a aquisição de imagens, que é feita pelo exame Papanicolau, com isso é criada a base de imagens. O passo seguinte é a aplicação de algoritmos de segmentação com o objetivo de encontrar regiões onde serão extraído atributos de forma e/ou textura. Os atributos são utilizados como entrada para algoritmos de classificação ou CBIR com o objetivo de diferenciar imagens de células normais e anormais.

Algoritmos de segmentação de células cervicais têm por objetivo separar núcleo, citoplasma e fundo da imagem. Entretanto, resultados com nível elevado de acurácia para segmentação de citoplasma estão restritos a: 1) imagens sintéticas (Figura 1 (a)) (USHIZIMA *et al.*, 2014; LU *et al.*, 2016) e isso se deve ao elevado nível de sobreposição de células que é obtido em imagens de esfregaço convencional. A sobreposição dificulta a definição das bordas do citoplasma até para especialistas; e 2) imagens reais com células sem sobreposição (Figura 1 (b)) (LI *et al.*, 2012), tais imagens são as obtidas através da citologia em meio líquido.

A segmentação de núcleos em imagens reais vem apresentando resultados promissores em vários trabalhos da literatura (LI *et al.*, 2012; IRSHAD *et al.*, 2014). Tais avanços podem beneficiar a análise automática de células cervicais, visto que, a partir da segmentação somente

líquido ainda não é uma realidade em países em desenvolvimento como Brasil e Índia.

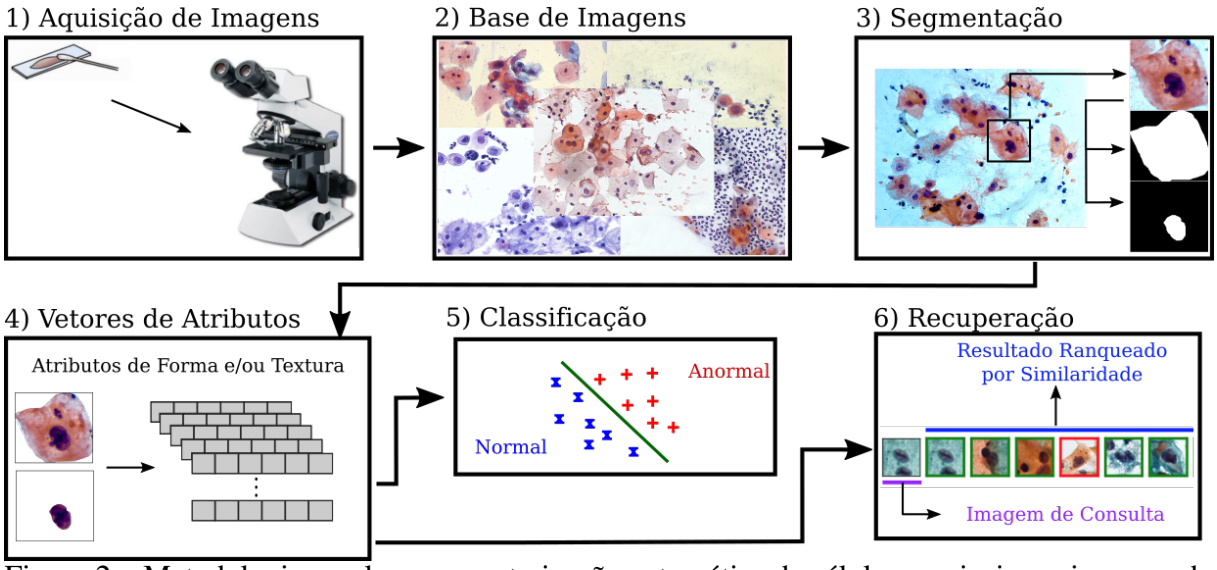


Figura 2 – Metodologia geral para caracterização automática de células cervicais em imagens do exame Papanicolau.

do núcleo é possível obter valiosas informações relativas à presença ou não de anormalidades na célula (PLISSITI; NIKOU, 2012). Essas informações podem ser calculadas por meio de algoritmos de descrição de imagens.

Descritores de forma fornecem importantes informações, tais como: quantificação do tamanho da célula, relação de área entre núcleo e citoplasma, circularidade e alongamento da célula. Atributos de forma podem ser utilizados em tarefas de classificação de células normais e anormais (MARINAKIS *et al.*, 2009). Em contrapartida, são necessários algoritmos de segmentação com altas taxas de acurácia para as regiões do núcleo e citoplasma. Além disso, atributos de textura vêm sendo utilizados para representar padrões de distribuição de cromatina no núcleo das células. Tais padrões, quando apresentam irregularidades, são frequentemente associados a células cancerígenas (EINSTEIN *et al.*, 1998; BEJNORDI *et al.*, 2013).

Após a extração dos atributos necessários para representar a célula, a detecção automática das células cervicais pode ser feita, usualmente, de duas formas: 1) classificação e 2) recuperação de imagens baseada em conteúdo.

Sistemas de classificação supervisionada de imagens tem por objetivo dividir as amostras em classes, utilizando os atributos calculados em amostras previamente identificadas para treinamento, ou seja, obter valores ótimos de parâmetros que dividam o espaço de atributos em classes. Com isso, é possível obter um modelo que generalize as informações para novas amostras (teste).

Em relação às células cervicais, é possível observar que sistemas de classificação são utilizados, na maioria dos casos, para separar as imagens nas classes normal e anormal (MARI-

NAKIS *et al.*, 2009; PLISSITI; NIKOU, 2012; MARIARPUTHAM; STEPHEN, 2015). Apesar de ser possível utilizar a classificação multi-classes (níveis normalidade/anormalidade), existem diversas desvantagens com esse tipo de sistema: a) o aumento no número de classes geralmente acarreta maior número de dados e tempo de processamento; b) deve-se incluir atributos específicos que diferenciam cada nível de normalidade e anormalidade, parte dos quais ainda são controversos entre diferentes patologistas; c) um sistema viável para aplicação em um cenário real precisa auxiliar o citologistas e não substituí-lo durante o pré-escrutínio; ou seja, o sistema deve sugerir ao citologista/patologista a localização de regiões onde provavelmente estão células anormais e o mesmo será responsável por identificar o nível de anormalidade presente no exame. Diante do exposto, podemos inferir que um sistema de classificação binária (normal e anormal) de células cervicais é a alternativa mais viável para o auxílio à detecção de lesões pré-cancerígenas do colo do útero.

Além da classificação, é possível realizar a detecção automática de células cervicais com o uso de sistemas CBIR, onde a informação visual de uma imagem é utilizada para calcular o grau de similaridade entre amostras. Assim, sistemas CBIR são úteis para catalogar células por ranqueamento e sugerir uma classificação para novos exemplos. Ao apresentar uma imagem de consulta de uma célula saudável ao sistema, por exemplo, é desejável que o mesmo apresente como resultado as imagens saudáveis mais similares à imagem de consulta que estão em uma base de imagens catalogada. Com isso, é possível sugerir uma classificação para a imagem apresentada na consulta. Apesar de possuírem aplicações em diversas áreas da ciências, são raros os sistemas CBIR disponíveis publicamente e de código aberto para aplicações isoladas ou de propósito geral.

Analisando algoritmos da literatura para segmentação de citoplasma em imagens obtidas pelo exame em meio convencional, constatamos a escassez de sistemas que produzam resultados em tempo hábil. As principais dificuldades encontradas nesse problema foram: 1) poucas bases de imagens públicas para o desenvolvimento e teste de novas metodologias em meio convencional; 2) a maioria dos métodos de descrição de células dependem da segmentação do núcleo e citoplasma com nível de acerto compatível com a verdade-terrestre; e 3) sistemas CBIR utilizando métodos de redes neurais profundas continuam restritos a aplicações proprietárias e/ou de código fechado em sua grande maioria.

1.2 Objetivos

Este trabalho tem como objetivo principal a proposição de uma metodologia para categorização de células cervicais do colo do útero. Tal metodologia engloba as etapas de segmentação, descrição, classificação e/ou CBIR.

Dentre os objetivos específicos a serem atingidos nesta tese estão:

- Organizar uma nova base de imagens de células cervicais obtida em meio convencional;
- Implementar uma abordagem para segmentação de núcleo em células cervicais;
- Desenvolver um extrator de atributos de células cervicais, considerando a relação entre custo computacional e acurácia;
- Propor um descritor de células cervicais com base em atributos radiais e que dependa apenas da segmentação do núcleo;
- Desenvolver uma ferramenta CBIR versátil que seja capaz de suportar aplicações em diferentes áreas do conhecimento;
- Aplicar os métodos propostos em outras bases de imagens reais a fim de avaliar sua capacidade de generalização;
- Comparar a metodologia proposta com outras na literatura.

1.3 Contribuições

A análise de células cervicais do colo do útero engloba as etapas de: aquisição de imagens, segmentação de estruturas de interesse, descrição de regiões para obtenção de atributos, que serão utilizados na caracterização das imagens. Tal caracterização pode ser feita por meio de algoritmos de classificação ou de sistemas de recuperação de imagens baseada em conteúdo. Dessa forma, as contribuições durante o doutorado são listadas a seguir:

1. Disponibilização de uma base de imagens reais que foi obtida via colaboração com o projeto Âmbar[2], que foi criado com o intuito de identificar os desafios relacionados à saúde da mulher e elaborar ações para a prevenção e a proteção da mulher e tem acesso a dados do Sistema Único de Saúde (SUS). A base *Cell Recognition for the Inspection of the Cervix* (CRIC) é composta por 164 imagens contendo 2470 células cervicais. Além disso, a base contém a verdade-terrestre com diagnóstico das células;
2. Desenvolvimento de um método de segmentação de núcleos utilizando o algoritmo k-

[2] http://www.ambar.net.br

médias hierárquico para as imagens da base CRIC;

3. Proposta do Descritor de Atributos Radiais (*Radial Feature Descriptor* (RFD)), para capturar informações ao redor e dentro do núcleo de modo independente da segmentação do citoplasma;
4. Desenvolvimento e disponibilização de uma ferramenta (`pyCBIR`) para experimentos CBIR. Essa ferramenta foi desenvolvida para facilitar a recuperação de imagens científicas em diversas aplicações;
5. Validação do descritor RFD, o qual foi usada na classificação e no CBIR utilizando as bases de imagens Herlev (JANTZEN *et al.*, 2005) e CRIC.

1.4 Produção Científica

A produção científica durante o doutorado resultou na publicação e submissão de artigos científicos a periódicos e congressos, capítulos de livros, além de um registro software. A seguir estão listadas as publicações diretamente relacionadas à pesquisa descrita neste documento:

1. **SILVA, R. R. V.**; ARAÚJO, F. H. D.; REZENDE, M. T.; CALAES, P. H.; MEDEIROS, F. N. S.; VERAS, R. M. S.; USHIZIMA, D. M.. *Searching for Cell Signatures in Multidimensional Feature Spaces*. International Journal of Biomedical Engineering and Technology, p. 1–20, Março, 2018.
2. **SILVA, R. R. V.**; ARAÚJO, F. H. D.; MEDEIROS, F. N. S.; BIANCHI, A. G. C.; CARNEIRO, C. M.; USHIZIMA, D. M.. *Radial Feature Description for Cell Classification*. Submetido ao Journal of Visual Communication and Image Representation, 2018.
3. SILVA, R. R. V.; **ARAUJO, F. H. D.**; MEDEIROS, F. N. S.; USHIZIMA, D. M.. *pyCBIR: a content-based image retrieval in python*. Propriedade intelectual: OI2018-01676, Fevereiro, 2018.
4. ARAÚJO, F. H. D.; **SILVA, R. R. V.**, MEDEIROS, F. N. S.; PARKINSON, D. D.; HEXEMER, A.; CARNEIRO, C. M.; USHIZIMA, D.. *Reverse Image Search for Scientific Data within and beyond the Visible Spectrum*. Expert Systems with Applications, v. 109, p. 35–48, Maio, 2018.

Também geramos publicações relacionadas à segmentação, descrição e classificação em outras bases de imagens, além da aplicação de outras técnicas de processamento de imagens às células cervicais. São listadas a seguir essas publicações:

1. **SILVA, R. R. V.**; LOPES, J. G. F.; ARAÚJO, F. H. D.; Medeiros, F. N. S.; USHIZIMA, D..

Visão Computacional em Python Utilizando as Bibliotecas Scikit-image e *Scikit-learn*. In: III Escola Regional de Informática do Piauí. (Org.). Livro Anais - Artigos e Minicursos. 1ed.: Sociedade Brasileira de Computação, v. 1, p. 407-428, 2017.

2. OLIVEIRA, P.; MOREIRA, G.; USHIZIMA, D.; CARNEIRO, C.; MEDEIROS, F. N. S.; ARAÚJO, F. H. D.; **SILVA, R. R. V.**; BIANCHI, A.. *A Multi-objective Approach for Calibration and Detection of Cervical Cells Nuclei*. In: IEEE Congress on Evolutionary Computation, 2017, Donostia.
3. USHIZIMA, D.; YANG, C.; VENKATAKRISHNAN, S.; ARAÚJO, F.; **SILVA, R.**; TANG, H.; MASCARENHAS, J. V.; HEXEMER, A.; PARKINSON, D.; SETHIAN, J.. *Convolutional Neural Networks at the Interface of Physical and Digital Data*. In: IEEE Applied Imagery Pattern Recognition Workshop (AIPR), Washington. p. 1, 2016.
4. CARNEIRO, A. C.; LOPES, J. G. F.; ARAÚJO, F. H. D.; **SILVA, R. R. V.**; PASSARINHO, C. J. P.; ROCHA NETO, J. F. S.; MEDEIROS, F. N. S.. *Análise de Fotografias de Pílulas por Redes Neurais Convolucionais*. VIII Simpósio de Instrumentação e Imagens Médicas e VII Simpósio de Processamento de Sinais, São Bernardo do Campo, 2017.
5. ARAÚJO, F. H. D.; **SILVA, R. R. V.**; MEDEIROS, F. N. S.; ROCHA NETO, J.F.; OLIVEIRA, P. H. C.; BIANCHI, A. C. G.; USHIZIMA, D. M.. *Active Contours for Overlapping Cervical Cell Segmentation*. International Journal of Biomedical Engineering and Technology, p. 1–21, Janeiro, 2018.
6. **ARAUJO, F. H. D.**; SILVA, R. R. V.; MEDEIROS, F. N. S.; RESENDE, M. T.; CARNEIRO, C. M.; USHIZIMA, D. M.. *Deep learning for cell image segmentation and ranking*. Submetido a Computerized Medical Imaging and Graphics, 2018.

1.5 Organização da Tese

A estrutura desta tese está organizada da seguinte forma:

- **Capítulo 2**: apresenta a fundamentação teórica, com os principais métodos para descrição de imagens, tais como: descritores de forma, textura, e Redes Neurais Convolucionais (*Convolutional Neural Network* (CNN)). Além disso, mostra o algoritmo de classificação utilizado e o fluxo geral de um sistema CBIR;
- **Capítulo 3**: mostra o processo de aquisição de imagens e detalha o algoritmo de segmentação de núcleos proposto para criação de uma nova base de imagens;
- **Capítulo 4**: introduz o RFD, um algoritmo para extração de atributos em imagens de

células cervicais baseado somente na segmentação do núcleo e apresenta a metodologia de classificação e sua avaliação qualitativa e quantitativa aplicada às bases de imagens Herlev e CRIC;

- **Capítulo 5**: detalha a ferramenta desenvolvida para experimentos CBIR: `pyCBIR` e exibe e discute os resultados obtidos para recuperação de imagens utilizando as bases de imagens Herlev e CRIC;
- **Capítulo 6**: apresenta as principais conclusões alcançadas bem como suas limitações e possibilidades de melhorias.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo apresentamos as principais ferramentas e técnicas implementadas nesta tese. A primeira seção trata da descrição de imagens, a seção seguinte faz uma breve introdução do algoritmo de classificação utilizado, o capítulo encerra apresentando a estrutura de um sistema CBIR.

2.1 Descrição de Imagens

As propriedades de objetos e regiões em uma imagem podem ser mensurados utilizando informações de forma, cor ou textura. Usualmente, essas informações são representadas por um conjunto de escalares, denominado descritor de uma imagem, vetor de atributos ou assinatura da imagem. Neste sentido, cada objeto ou região é representado por um ponto em um espaço R^n, para n atributos.

Dentre os vários conjuntos de descritores aplicados a células cervicais, merecem destaque aqueles baseados em textura e forma. Descritores de textura são obtidos por representações estatísticas da distribuição e magnitude dos pixels. Os descritores de forma são computados a partir de bordas calculadas, na maioria das vezes, por algoritmos de segmentação. Além disso, existem os atributos obtidos pelo treinamento de Redes Neurais Convolucionais (CNN), que vêm se mostrando eficazes na descrição dos mais diversos tipos de imagens.

Nas subseções a seguir apresentamos os descritores de textura, os descritores de forma e a estrutura básica das CNNs mais comuns disponíveis na literatura.

2.1.1 Descritores de Textura

A descrição de um objeto ou região em uma imagem pode ser obtida através da análise de textura, e está relacionada a informações de suavidade, rugosidade e regularidade (GONZALEZ; WOODS, 2000). A seguir detalhamos os algoritmos mais utilizados para descrição de textura em imagens de células.

2.1.1.1 Matriz de Coocorrência de Níveis de Cinza

A matriz de coocorrência de níveis de cinza, do inglês *Gray-Level Co-occurrence Matrix* (GLCM) é uma técnica que é utilizada na análise de textura em imagens. Na GLCM

são analisadas as coocorrências existentes entre pares de pixels através de algum padrão. A matriz GLCM é sempre quadrada e armazena as informações das intensidades relativas dos pixels. Por este motivo, as imagens utilizadas são sempre em tons de cinza (HARALICK *et al.*, 1973). A matriz GLCM tem sido utilizada na descrição de células cervicais (CHEN *et al.*, 2014), citologia (WANG *et al.*, 2016), mamografia (KANADAM; CHEREDDY, 2016), além de outras áreas.

As probabilidades de coocorrências em uma matriz P são calculadas entre dois níveis de cinza i e j, utilizando uma orientação θ e uma distância conhecida como espaçamento entre pares de pixels. Essa orientação pode assumir os valores 0^o, 45^o, 90^o ou 135^o. Para cada relacionamento espacial possível (distância e orientação) existe uma matriz de coocorrência. Desse modo, as informações sobre a textura de uma imagem estarão contidas nessa matriz (BARALDI; PARMIGGIANI, 1995). Na Figura 3 é mostrado um exemplo de cálculo da GLCM.

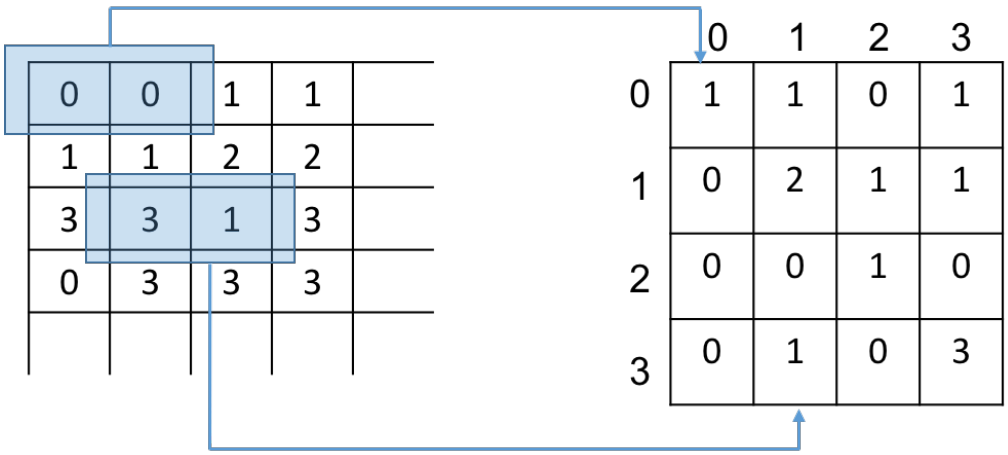


Figura 3 – Exemplo de cálculo da matriz GLCM com espaçamento entre pares de pixels igual a 1 e $\theta = 0^o$. (a) Intensidade de pixels de uma imagem representada por 3 bits, (b) GLCM.

Haralick et al. (HARALICK *et al.*, 1973) definiram 14 características significativas para a GLCM há mais de quarenta anos, e outras medidas derivadas dessas foram acrescentadas (SABINO *et al.*, 2004) ao longo das últimas décadas. Contudo, a utilização de algumas dessas características podem obter melhor desempenho do que a utilização de todas. Assim, nesse trabalho são feitos cálculos dos seguintes atributos de textura: contraste (Equação 2.1), dissimilaridade (Equação 2.2), homogeneidade (Equação 2.3), energia (Equação 2.20), correlação

(Equação 2.5) e segundo momento angular (*Angular Second Moment* (ASM)) (Equação 2.6).

$$Contraste = \sum_{i,j=0}^{L-1} P_{ij}(i-j)^2, \tag{2.1}$$

onde L corresponde à profundidade do pixel (número de níveis de cinza) e P é a GLCM.

$$Dissimilaridade = \sum_{i,j=0}^{L-1} P_{ij}|i-j|, \tag{2.2}$$

onde |.| é o valor absoluto.

$$Homogeneidade = \sum_{i,j=0}^{L-1} \frac{P_{ij}}{1+(i-j)^2}. \tag{2.3}$$

$$Energia = \sqrt{\sum_{i,j=0}^{L-1} (P_{ij})^2}. \tag{2.4}$$

$$Correlação = \sum_{i,j=0}^{L-1} \frac{(i-\mu_i)(j-\mu_j)P_{ij}}{\sqrt{(\sigma_i)^2(\sigma_j)^2}}, \tag{2.5}$$

em que μ é a média e σ é o desvio padrão.

$$ASM = \sum_{i,j=0}^{L-1} (P_{ij})^2. \tag{2.6}$$

.

2.1.1.2 Gray-Level Run Length Matrix

Similar à GLCM, a *Gray-Level Run Length Matrix* (GLRLM) é um histograma bidimensional **P** de elementos onde cada elemento $p(i, j)$ contém o número total de ocorrências de um conjunto de pixels com o mesmo valor em determinada direção $\theta = \{0°, 45°, 90°, 135°\}$ (GALLOWAY, 1975) (TANG, 1998). Esse descritor é comumente utilizado em análise de imagens de solo (AJDADI *et al.*, 2016) e na quantificação de biomarcadores (DIJK *et al.*, 2016). A Figura 4 mostra como calcular a matriz P desse descritor.

A partir de P, calculamos os seguintes atributos: *Short Run Emphasis* (SRE)(Equação 2.7), *Long Run Emphasis* (LRE) (Equação 2.8), *Gray Level Nonuniformity* (GLN) (Equação 2.9),

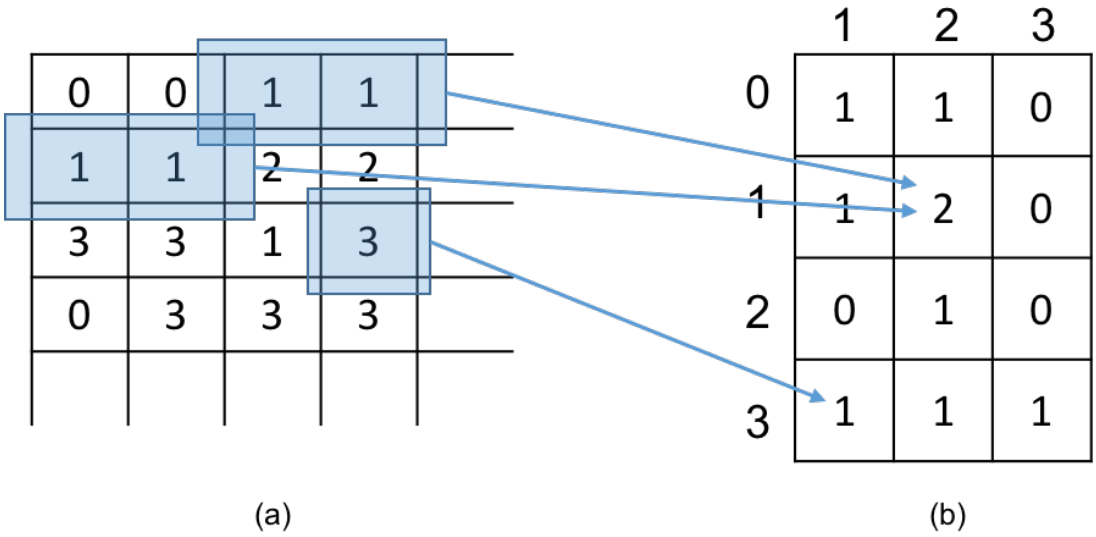


Figura 4 – Cálculo da GLRLM. (a) Imagens de níveis de cinza com valores de intensidade variando entre 0 e 3 (imagem de 2 bits). (b) Matriz GLRLM.

Run Length Nonuniformity (RLN) (Equação 2.10), *Run Percentage* (RP) (Equação 2.11), *Low Gray Level Run Emphasis* (LGRE) (Equação 2.12), *High Gray Level Run Emphasis* (HGRE) (Equação 2.13), *Short Run Low Gray Level Emphasis* (SRLGE) (Equação 2.14), *Short Run High Gray Level Emphasis* (SRHGE)(Equação 2.15), *Long Run Low Gray Level Emphasis* (LRLGE) (Equação 2.16), *Long Run High Gray Level Emphasis* (LRHGE) (Equação 2.17).

$$SRE = \frac{1}{n_r}\sum_{i=1}^{L}\sum_{j=1}^{R}\frac{P(i,j|\theta)}{j^2}, \tag{2.7}$$

onde n_r é o número de ocorrências, L é o número de níveis de cinza, R é o tamanho máximo de um conjunto de ocorrências.

$$LRE = \frac{1}{n_r}\sum_{i=1}^{L}\sum_{j=1}^{R} j^2 P(i,j|\theta). \tag{2.8}$$

$$GLN = \frac{1}{n_r}\sum_{i=1}^{L}\left(\sum_{j=1}^{R} P(i,j|\theta)\right)^2. \tag{2.9}$$

$$RLN = \frac{1}{n_r}\sum_{j=1}^{R}\left(\sum_{i=1}^{L} P(i,j|\theta)\right)^2. \tag{2.10}$$

$$RP = \frac{n_r}{n_p}, \tag{2.11}$$

onde n_p é o número de pixels na imagem.

$$LGRE = \frac{1}{n_r}\sum_{i=1}^{L}\sum_{j=1}^{R}\frac{P(i,j|\theta)}{i^2}. \tag{2.12}$$

$$HGRE = \frac{1}{n_r}\sum_{i=1}^{L}\sum_{j=1}^{R}i^2P(i,j|\theta). \tag{2.13}$$

$$SRLGE = \frac{1}{n_r}\sum_{i=1}^{L}\sum_{j=1}^{R}\frac{P(i,j|\theta)}{i^2j^2}. \tag{2.14}$$

$$SRHGE = \frac{1}{n_r}\sum_{i=1}^{L}\sum_{j=1}^{R}\frac{i^2P(i,j|\theta)}{j^2}. \tag{2.15}$$

$$LRLGE = \frac{1}{n_r}\sum_{i=1}^{L}\sum_{j=1}^{R}\frac{j^2P(i,j|\theta)}{i^2}. \tag{2.16}$$

$$LRHGE = \frac{1}{n_r}\sum_{i=1}^{L}\sum_{j=1}^{R}i^2j^2p(i,j|\theta). \tag{2.17}$$

A GLRLM pode extrair até 44 atributos, visto que são extraídos 11 atributos para cada direção possível $(0^\circ, 45^\circ, 90^\circ, 135^\circ)$.

2.1.1.3 Atributos de Histograma

Atributos de textura de 1^a ordem ou atributos de histograma consistem no cálculo das seguintes medidas: média (μ) (Equação 2.18), entropia (μ_e) (Equação 2.19), energia (μ_n) (Equação 2.20), variância (σ^2) (Equação 2.21), assimetria (μ_s) (Equação 2.22) e curtose (μ_k) (Equação 2.23). Recentemente, atributos de histograma foram utilizados na detecção automática de tumores no cérebro (NABIZADEH; KUBAT, 2015) e detecção de sinais de vibração em sistemas mecânicos (JEGADEESHWARAN; SUGUMARAN, 2015).

$$\mu = \sum_{i=1}^{nBins}\frac{H[i]}{nBins}, \tag{2.18}$$

onde *nBins* é a quantidade de níveis de cinza utilizados para representar a imagem e H é o histograma da imagem.

$$\mu_e = -\sum_{i=1}^{nBins} \Big(H[i] \times \log_2(H[i]) \Big). \tag{2.19}$$

onde $\times$ representa o produto.

$$\mu_n = \sum_{i=1}^{nBins} \frac{H[i]^2}{nBins}. \tag{2.20}$$

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{nBins} (H[i] - \mu)^2. \tag{2.21}$$

$$\mu_s = \sqrt[3]{\frac{1}{N} \Big(\sum_{i=1}^{nBins} (H[i] - \mu)^3 \Big)} \tag{2.22}$$

$$\mu_k = N \frac{\sum_{i=1}^{nBins} (H[i] - \mu)^4}{\Big(\sum_{i=1}^{nBins} (H[i] - \mu)^2 \Big)^2}. \tag{2.23}$$

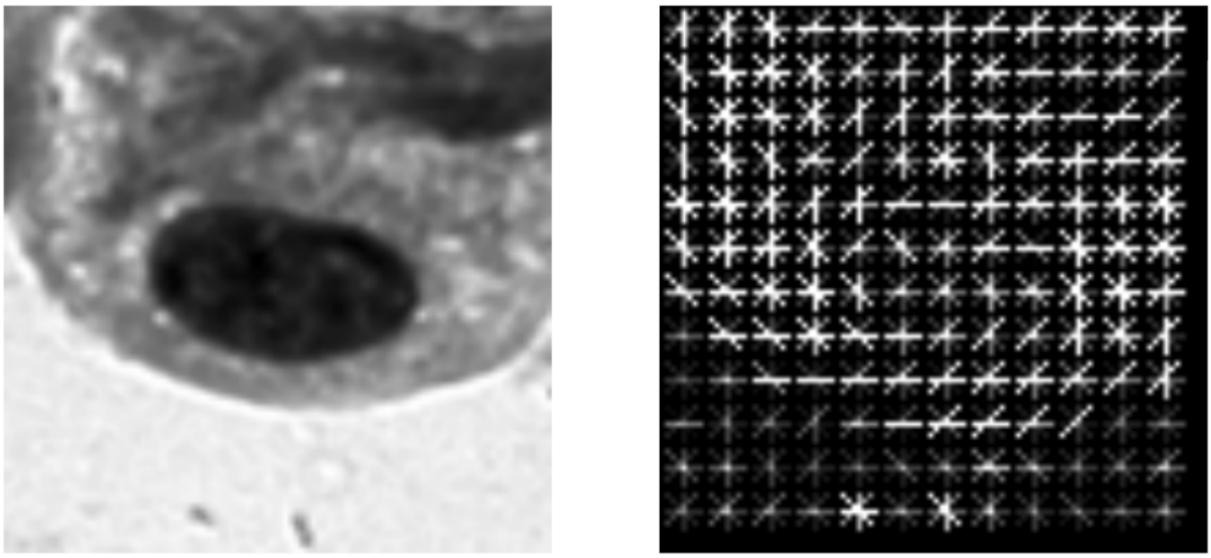
2.1.1.4 Histograma de Gradientes Orientados

O algoritmo de Histograma de Gradientes Orientados (*Histogram of Oriented Gradients* (HOG)) (DALAL; TRIGGS, 2005) calcula o histograma da orientação dos gradientes na imagem. Esse algoritmo se baseia na ideia de que a forma e a aparência de um objeto podem ser descritas pela intensidade dos gradientes ou da direção das bordas. Ele foi primeiramente utilizado para detecção de pedestres (DALAL; TRIGGS, 2005) e é amplamente utilizado no reconhecimento de faces (DÉNIZ *et al.*, 2011) e na detecção de objetos (VASHAEE *et al.*, 2016).

O primeiro passo do método consiste em converter a imagem para escala de cinzas seguido da computação dos gradientes utilizando o operador de Sobel (GONZALEZ; WOODS, 2000). Dada uma imagem I, os gradientes Gx e Gy são calculados, assim como a magnitude e a orientação de cada vetor.

Em seguida, é feita uma divisão da imagem em regiões espaciais chamadas blocos. Cada bloco é sub-dividido em regiões menores chamadas sub-blocos. O tamanho dos blocos e dos sub-blocos são parâmetros essenciais do algoritmo. Para cada célula é calculado um histograma local 1-D das orientações sobre os pixels da célula.

A Figura 5 mostra um exemplo do cálculo das orientações dos histogramas em uma imagem. O descritor nada mais é do que uma lista dos histogramas calculados a partir de todos os blocos e sub-blocos. Após calculado, o descritor é normalizado.



(a) (b)

Figura 5 – Exemplo do descritor Histograma de Gradientes Orientados (HOG) aplicado a uma imagem de célula cervical. (a) Imagens em níveis de cinza de uma célula cervical. (b) Gradientes resultantes do descritor HOG.

2.1.1.5 Padrão Binário Local

Utilizamos o descritor Padrão Binário Local (*Local Binary Pattern* (LBP)) da forma como foi proposta em (OJALA; PIETIKAINEN, 1997). O descritor LBP forma rótulos para os pixels da imagem utilizando uma limiarização com base no pixel central de janelas 3×3. Como resultado é obtido um número binário, que corresponde ao rótulo do pixel. Como utilizamos janelas de tamanho 3×3, isso resultou em 2^8 diferentes possibilidades de rótulos, o histograma desses rótulos é utilizado como descritor de textura. Estudos recentes (YLIOINAS *et al.*, 2016; TIWARI; TYAGI, 2016) mostram o uso do LBP na categorização de materiais, reconhecimento de face e categorização de sequências de estruturas dinâmicas. A Figura 6 ilustra o processo.

2.1.2 Descritores de Forma

Descritores de forma podem ser representados pela região (características internas) ou pelo contorno. A representação por região é adequada quando é necessário calcular informações

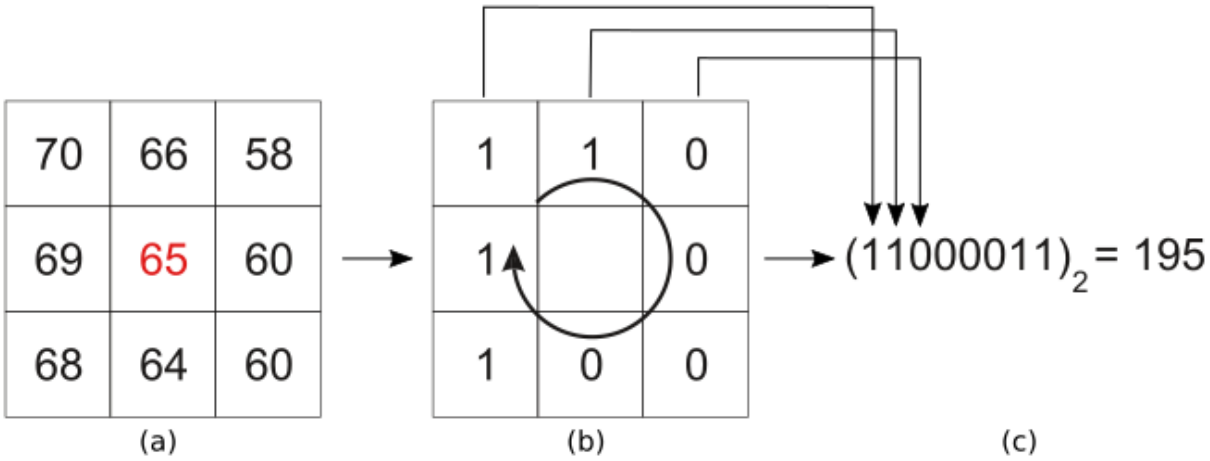


Figura 6 – Exemplo de cálculo do descritor LBP. (a) Exemplo de janela. (b) Limiarização baseada no pixel central. (c) Padrão calculado a partir do resultado da limiarização.

internas como área e diâmetro. A representação pelo contorno é adequada quando o interesse são características externas, tais como cantos e pontos de alta curvatura da forma (COSTA; CESAR JR., 2000).

Os descritores de forma mais utilizados na análise de células cervicais aplicados às regiões do núcleo e do citoplasma são: área, razão entre as áreas do núcleo e do citoplasma, menor e maior diâmetro, alongamento, circularidade, perímetro, centróide, compacidade, maior e menor eixo principal, razão entre eixos, homogeneidade e centro de gravidade (MARINAKIS *et al.*, 2009; CHEN *et al.*, 2014; SARWAR *et al.*, 2015).

Células saudáveis usualmente possuem: 1) núcleo pequeno em relação ao citoplasma e em formato circular; 2) núcleo posicionado próximo ao centro da célula e com bordas bem definidas; e 3) cromatina regularmente distribuída. Devido a esses fatos, os descritores de formas são bastante utilizados para descrição de células, geralmente de forma híbrida agrupados com descritores de textura.

2.1.3 Redes Neurais Convolucionais para Extração de Atributos

As Redes Neurais Convolucionais (CNN) se tornaram um novo paradigma em visão computacional e são fáceis de treinar, quando existe uma grande quantidade de amostras rotuladas representando as diferentes classes de interesse. Algumas das vantagens da utilização de CNNs são: (a) capacidade de extrair características relevantes através de aprendizado de transformações (*kernels*); (b) depender de menor número de parâmetros de ajustes do que redes totalmente conectadas com o mesmo número de camadas ocultas. Como cada unidade de uma camada não é conectada com todas as unidades da camada seguinte, há menos pesos para serem

atualizados, facilitando assim o treinamento; (c) capacidade de obter informações relevantes sem a necessidade da segmentação de estruturas do objeto.

Dentre as diversas opções de CNNs, exploramos nesta tese duas diferentes arquiteturas que utilizam transferência de aprendizagem:

1. LeNet (LECUN *et al.*, 1998), consiste de um arranjo de neurônios composto por camadas complemente conectadas e convolucionais permitindo aplicações em tempo real. Devido à sua simplicidade, o treinamento muitas vezes tem um bom desempenho com uma pequena quantidade de dados em comparação às CNNs mais profundas. Entretanto, a LeNet é mais imprecisa quando o foco são problemas de reconhecimento mais complexos, como aqueles com muitas classes e imagens similares entre as classes. Um exemplo de problema complexo é a base de imagens ImageNet (DENG *et al.*, 2009) que possui cerca de um milhão de imagens dividas em mil classes;
2. Inception-ResNet-v2 (SZEGEDY *et al.*, 2016a) é uma arquitetura profunda formada por múltiplas subredes, obtendo padrões de classificação mais elaborados. Esse modelo necessita de aproximadamente duas vezes mais memória e poder de processamento que a sua versão anterior (Inception v3 (SZEGEDY *et al.*, 2016b)). Por outro lado, ela aparenta ser mais robusta e eficiente em relação à experimentos de classificação que os demais modelos da literatura.

A Figura 7 exibe uma arquitetura de CNN. A imagem de entrada passa por sequências de convoluções intercaladas por camadas de redução de dimensionalidade, geralmente a *Maxpool* é utilizada. Após isso, uma camada completamente conectada tem por objetivo calcular os atributos de saída da rede.

2.2 Algoritmo de Classificação *Random Forest*

O objetivo de um classificador é dividir o espaço de atributos em regiões de decisão. Dessa forma, os vetores de atributos que estiverem contidos na mesma região de decisão compartilham a mesma classe.

A entrada do classificador é um vetor de atributos que representa uma célula cervical, e a saída é a classe a qual a célula pertence (normal ou anormal, ou o grau da lesão). Aqui utilizamos o classificador *Random Forest* (RF) (BREIMAN, 2001).

O algoritmo RF é uma combinação de predições de diversas árvores de decisão em que cada árvore depende dos valores de um vetor independente, amostrados aleatoriamente e

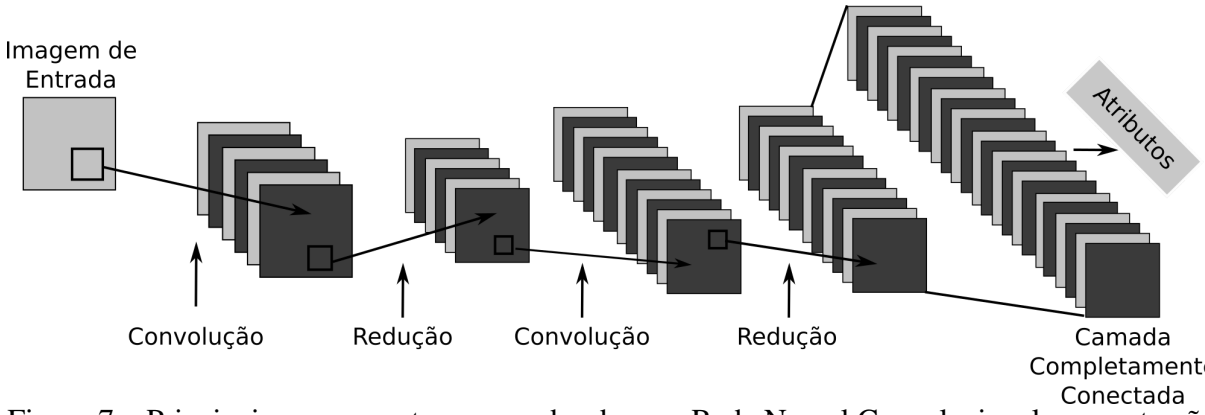


Figura 7 – Principais componentes e camadas de uma Rede Neural Convolucional para extração de atributos em imagens: cada etapa ilustra a transformação de uma imagem. A partir de uma imagem de entrada, são aplicadas convoluções intercaladas por transformações de redução de dimensionalidade (camadas *Maxpool*), por fim a camada totalmente conectada tem como saída o vetor de atributos calculado.

com a mesma distribuição para todas as árvores do conjunto ou floresta, a denominação para uma série de árvores de decisão.

A árvore de decisão é um tipo de classificador supervisionado baseado em modelos estatísticos. A capacidade de classificação de uma árvore de decisão vem da divisão recursiva do espaço de atributos em sub-espaços, ao final cada sub-espaço é associado a uma classe.

Após a geração de um grande número de árvores, as classes com maior número de votos são eleitas. Além disso, a usabilidade do mesmo é relativamente simples pois o algoritmo RF é menos sensível ao ajuste de parâmetros (BREIMAN, 2001) quando comparado com outros classificadores da literatura.

A Figura 8 mostra os passos do classificador RF. A partir de um vetor de atributos (D), são gerados outros vetores de atributos ($D_1, \cdots, D_n$, onde n é a quantidade de árvores na floresta), que são embaralhados em relação ao vetor original. É gerado um vetor para cada árvore do RF. Em seguida, os vetores de atributos são passados como parâmetro para as árvores de decisão. Cada árvore irá gerar um resultado para a classificação e, os resultados são combinados obtendo uma saída unificada. Esse classificador foi escolhido por apresentar desempenho superior em várias aplicações, inclusive imagens médicas (NAGARAJ *et al.*, 2018).

2.3 Sistemas de Recuperação de Imagens Baseada em Conteúdo

Sistemas de Recuperação de Imagens Baseado em Conteúdo (*Content-Based Image Retrieval* - CBIR) realizam buscas utilizando propriedades pictoriais de imagens em bases de imagens. Tais sistemas visam recuperar imagens que sejam do interesse do usuário mediante um

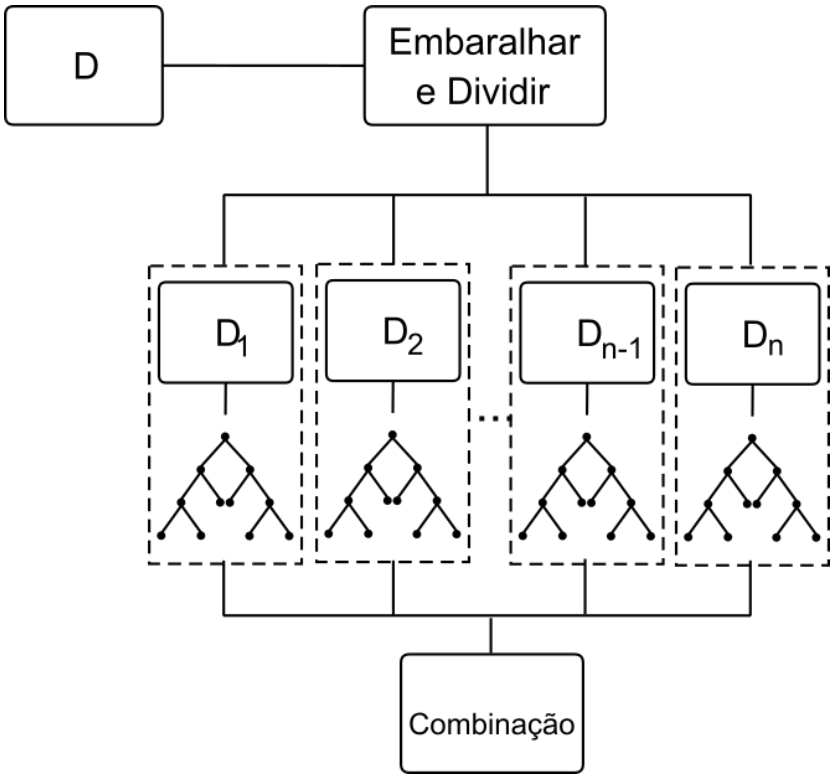


Figura 8 – Passos do classificador *Random Forest*, onde D é o vetor de atributos original, D_1, $D_2 \ldots$, D_{n-1}, D_n são os vetores de atributos gerados a partir de D.

padrão de consulta especificado por meio de uma figura.

A busca de imagens em grandes bases de dados é um dos serviços mais importantes disponibilizados pelos sistemas de gerenciamento de informação na atualidade. O método clássico para se prover tal serviço emprega a rotulação textual por palavras-chave. Atualmente, essa abordagem se tornou inviável devido ao grande volume de informação multimídia disponível, e sem rotulo. Ademais, o processo de descrição textual é impreciso e sujeito a erros, uma vez que diferentes indivíduos tendem a interpretar e descrever uma mesma imagem utilizando diferentes palavras-chave na descrição. Os sistemas CBIR foram propostos em resposta a essas dificuldades. Nesses sistemas o processo de busca utiliza o conteúdo visual das imagens ao invés da rotulação textual.

A Figura 9 apresenta a arquitetura de um sistema *CBIR* clássico. A partir de uma base de imagens, são criados conjuntos de treinamento e teste. A base de treinamento serve para obter o melhor conjunto de parâmetros. Para isso, várias metodologias podem ser empregadas, como, por exemplo, algoritmos evolucionários (SOUZA *et al.*, 2016). Por outro lado, a base de teste é utilizada para que as buscas por similaridade sejam feitas. Os passos de pré-processamento e extração de atributos são aplicados nas bases de teste e treino para criação da base de assinaturas. A recuperação de imagens é feita utilizando uma imagem de consulta, nessa imagem são

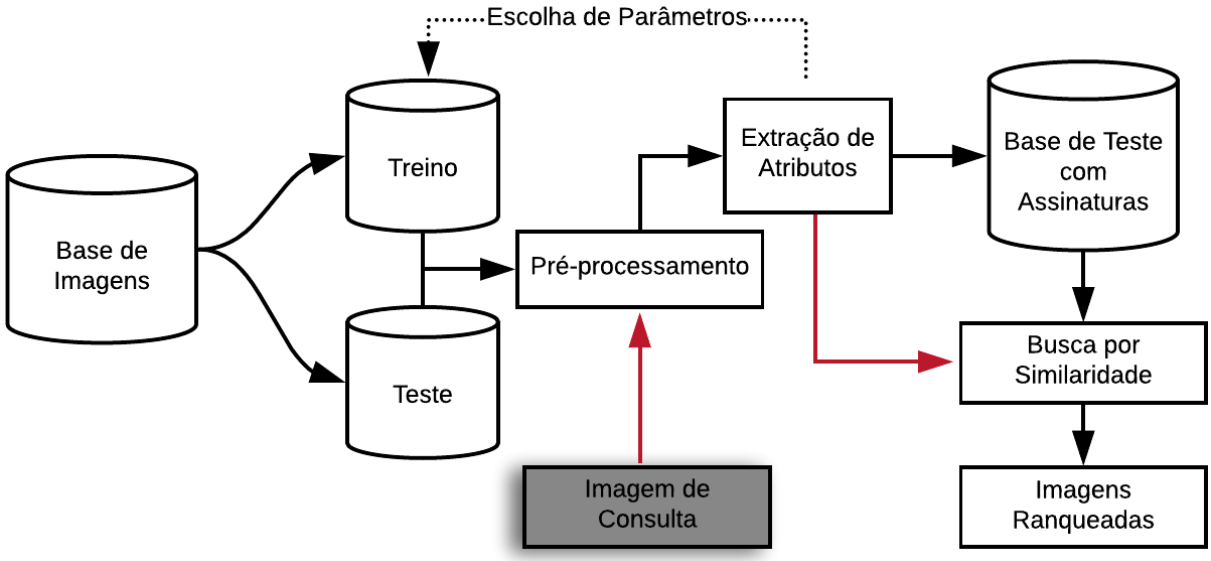


Figura 9 – Arquitetura de um sistema *CBIR*. A partir de uma base de imagens, que é dividida em conjuntos de treino e teste, é feito o pré-processamento e a extração de atributos gerando a base de assinaturas. Com isso, é possível buscar imagens com conteúdo similar ao de uma imagem de referência e ranquear o resultado por nível de similaridade. As setas vermelhas mostram o fluxo da imagem de consulta.

aplicados os mesmos processos de pré-processamento e extração de atributos. A partir do vetor de características obtido, é feita uma busca na base de assinatura de teste que irá retornar as assinaturas mais similares àquela calculada para a imagem de consulta. O cálculo de similaridade é feito utilizando funções de distância - como a euclidiana, cosseno ou Manhattan, por exemplo. Por fim, o sistema terá como saída as imagens da base de dados de teste ranqueadas por similaridade dos atributos que os representam.

A usabilidade de um sistema CBIR é determinada pela rapidez em que as informações são buscadas e retornadas. Entretanto uma busca linear por uma base de dados de milhões de imagens pode levar o usuário a um tempo de espera inaceitável. Para resolver esses problemas, muitas vezes são necessários algoritmos de busca mais avançados, como os baseados em árvores de busca, além de formas de reduzir a quantidade de atributos que representa uma imagem, tais como: Análise de Componentes Principais (*Principal Component Analysis* (PCA)) (PEARSON, 1901) e razão do ganho de informação (*Information Gain Ratio* (IGR)) (QUINLAN, 1993).

2.4 Considerações Finais

A análise de descritores é uma etapa fundamental na análise de células cervicais do colo do útero. Alguns passos dessa análise irão depender da escolha do descritor a ser utilizado. Descritores de textura ou baseados em CNNs podem ser obtidos sem o uso de pré-processamentos

ou segmentação, ou seja, a imagem completa é utilizada. Observamos nos trabalhos da literatura que a maioria dos descritores de textura são utilizados em uma região de interesse, geralmente o núcleo ou o citoplasma da célula. Por outro lado, os descritores de forma dependem da etapa de segmentação para obter as bordas da região de interesse.

Em relação à classificação das imagens, foram testados outros algoritmos, além do RF, com o objetivo de validar sua eficiência, a saber: K-Vizinhos Mais Próximos (*K-Nearest Neighbor* (KNN)) e Máquina de Vetor de Suporte (*Support Vector Machine* (SVM)). Contudo, o RF se mostrou eficiente tanto em relação ao custo computacional quanto na capacidade de categorizar as imagens.

Sistemas CBIR são usualmente propostos como metodologias, e não como ferramentas. Dessa forma, vimos a oportunidade de propor uma ferramenta com esse propósito com uma interface gráfica amigável, escalável, e de propósito geral que é detalhada no Capítulo 5

No próximo capítulo, mostramos como é realizada a aquisição de imagens pelo exame Papanicolau. Além disso, apresentamos o método proposto para segmentação de núcleos que terá como resultado uma nova base de imagens de células cervicais.

3 AQUISIÇÃO E SEGMENTAÇÃO DE IMAGENS

Este capítulo apresenta o processo de aquisição de imagens de células do colo do útero e introduz uma base de imagens obtidas pelo SUS: *Cell Recognition for the Inspection of the Cervix* (CRIC). Ele ainda descreve o método de segmentação de imagens desenvolvido para a esta base. Tal método aplica o algoritmo k-médias na segmentação de núcleos, os quais serão utilizados posteriormente como entrada para o algoritmo de extração e descrição de células.

3.1 Aquisição de Imagens

A aquisição de imagens de células cervicais é feita por meio das lâminas obtidas pelo exame Papanicolau. A Figura 10 mostra as principais etapas para aquisição de imagens desse exame. No exame Papanicolau é coletado material do colo do útero, que é preparado em uma lâmina para posterior análise. As imagens obtidas (Figura 10 (d)) são campos/regiões dessa lâmina. Em uma única lâmina são obtidos, aproximadamente, 15.000 campos. Ainda estima-se que uma lâmina de citologia convencional possa conter de 15.000 até 150.000 células cervicais (ARAUJO, 2012). Logo, é necessário o desenvolvimento de algoritmos de baixo custo computacional para que essa análise seja feita em um curto espaço de tempo, tornando a solução viável para ser aplicada em um ambiente real. Ao mesmo tempo, os métodos desenvolvidos devem ser robustos o suficiente para discriminar células evitando a ocorrência de falsos negativos que são células anormais não detectadas.

A base de imagens CRIC foi obtida por exames provenientes do SUS. Ela é composta por 164 imagens digitalizadas de exames de Papanicolau, possuindo verdade-terrestre para segmentação de células (núcleo e citoplasma) e para classificação (células normais e anormais). A base está dividida da seguinte forma:

- Subconjunto de treino: a base CRIC possui 12 imagens totalizando 270 células com verdade-terrestre para: 1) segmentação de núcleos e citoplasmas; e 2) classificação das células. A Figura 11 mostra um exemplo da verdade-terrestre disponível na base.
- Subconjunto de teste: composto por 152 imagens, que possuem verdade-terrestre para classificação das células. A Figura 12 exibe uma imagem da base bem como sua marcação de células normais e anormais.

As imagens da base CRIC possuem 1392×1040 pixels, adquiridas com um microscópio Carl Zeiss e uma câmera Zeiss AxioCam MRc com magnificação de $40\times$. Essa base

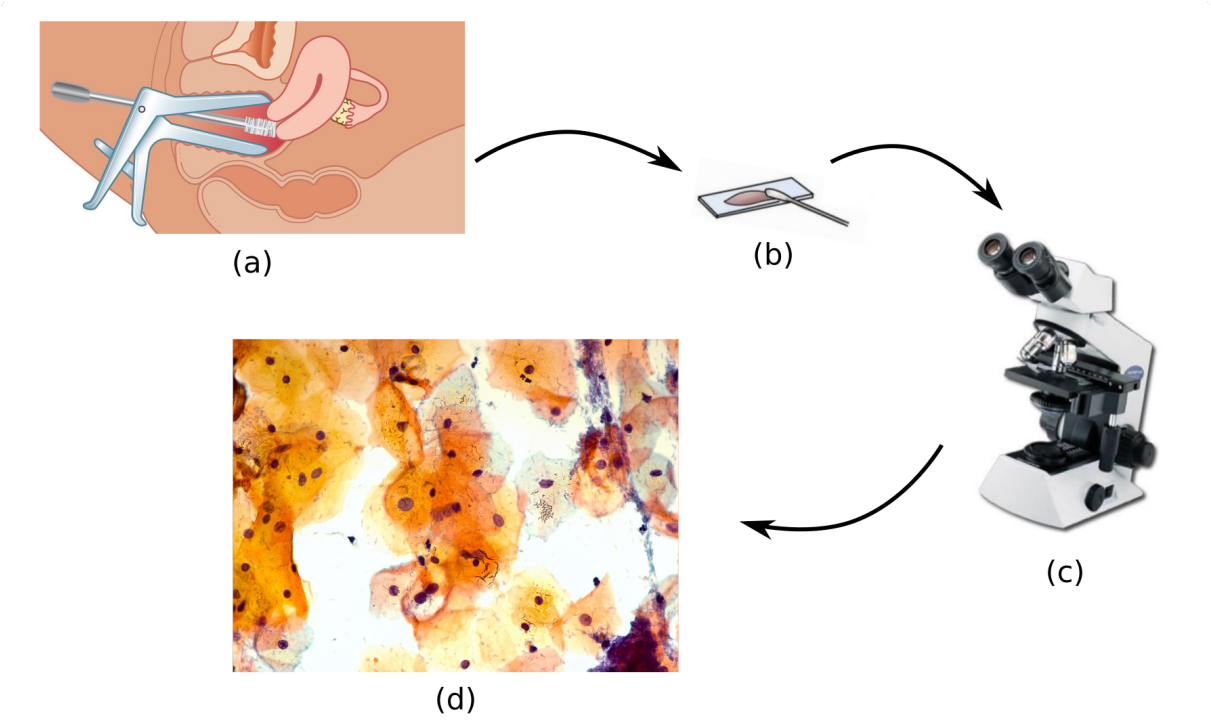
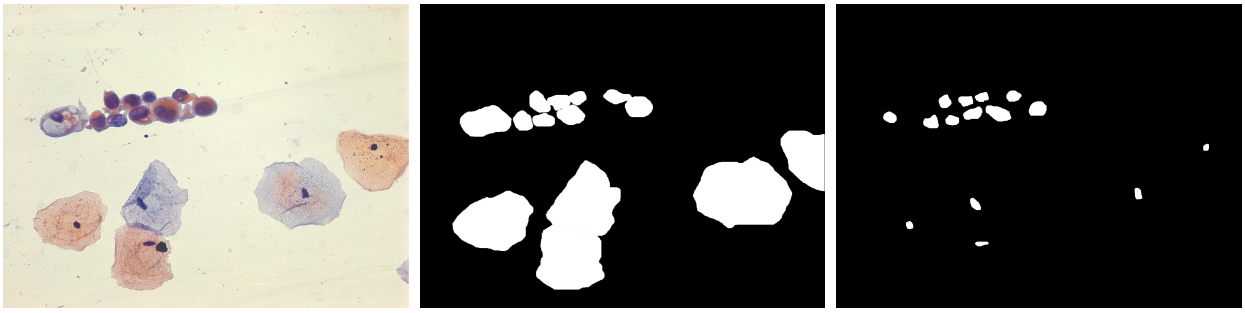


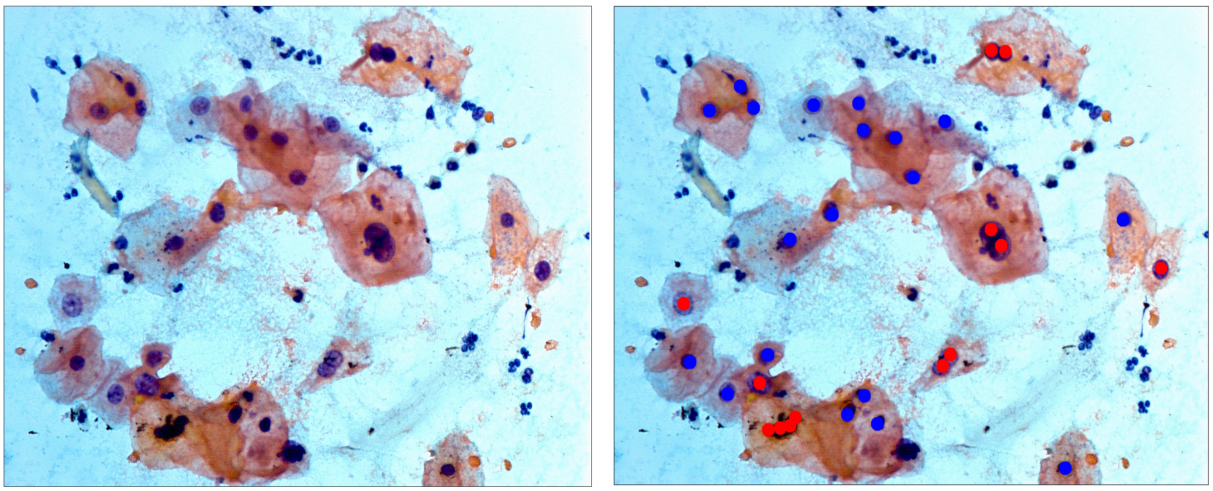
Figura 10 – Fluxograma da aquisição de imagens pelo exame Papanicolau. (a) No exame é feita uma coleta do tecido do colo do útero. (b) O material coletado é preparado em uma lâmina. (c) A lâmina é analisada pelo citologista em um microscópio com aumento variando entre 20 e 40 vezes. (d) As imagens são obtidas através de câmeras acopladas ao microscópio.



(a) (b) (c)

Figura 11 – Exemplo de uma imagem do subconjunto de treino da base CRIC. (a) Imagem original. Verdade-terrestre para segmentação do (b) citoplasma e do (c) núcleo .

apresenta várias características desejáveis: as amostras são de uma ampla diversidade racial, que é um traço marcante da população brasileira. As imagens apresentam células cervicais obtidas a partir de exames convencionais de Papanicolau, que incluem células com sobreposição, artefatos tais como neutrófilos e hemácias, e outros achados que são inerentes a essas amostras.



(a) (b)

Figura 12 – Exemplo de uma imagem do subconjunto de teste da base CRIC. (a) Imagem original com padrões de células normais e anormais. (b) Ilustração das informações disponíveis na base, pontos azuis e vermelhos fazem referência ao centro do núcleo de células normais e anormais, respectivamente.

3.2 Segmentação de Imagens

Na literatura observamos que na análise automática de exames de Papanicolau as células são processadas separadamente, ao invés de serem analisadas as imagens por completo (PLISSITI *et al.*, 2011; GENÇTAV *et al.*, 2012). Isso acontece devido à quantidade de células presente em cada imagem, ou seja, uma mesma imagem pode conter células saudáveis e anormais. Dessa forma, faz-se necessário processar uma célula por vez. Para isso, são desenvolvidos algoritmos para segmentação de células.

A segmentação de células cervicais tem sido tópico de vários trabalhos nos últimos anos (USHIZIMA *et al.*, 2014; LU *et al.*, 2016), tendo como objetivo principal separar núcleo e citoplasma. Algoritmos para segmentação do núcleo tem apresentado soluções viáveis tanto em tempo de execução dos algoritmos quanto nos resultados obtidos, o mesmo não ocorre em relação ao citoplasma, mostrando-se um tema mais desafiador. Apesar dos avanços obtidos, bons resultados foram obtidos em imagens sintéticas (USHIZIMA *et al.*, 2014; LU *et al.*, 2016) e em imagens reais sem sobreposição de células (LI *et al.*, 2012), e isso não condiz com a realidade dos exames citológicos (Figura 12). Além disso, algoritmos como o proposto por LU *et al.* apresentam um alto custo computacional, tornando-se uma solução inviável de ser implementada em um ambiente real.

Muitos métodos para segmentação de núcleos em imagens de células cervicais foram propostos (IRSHAD *et al.*, 2014). Eles podem ser divididos em três grupos: segmentação de

único núcleo, segmentação múltipla de núcleos, e segmentação de núcleos que se tocam (ZHANG *et al.*, 2014). Os métodos para segmentação de único núcleo utilizam informações de contorno ou cor para aplicar modelos de contorno ativo (BAMFORD; LOVELL, 1998; LI *et al.*, 2012).

Para segmentação múltipla de núcleos é possível utilizar técnicas de limiarização (HARANDI *et al.*, 2010), a transformada de Hough (BERGMEIR *et al.*, 2012), morfologia (*watershed*) (GENÇTAV *et al.*, 2012; PLISSITI *et al.*, 2011; PLISSITI *et al.*, 2011), e *level sets* (LU *et al.*, 2013). Finalmente, as técnicas utilizadas para segmentar núcleos que se tocam incluem erosão e agrupamento não-supervisionado (JUNG; KIM, 2010; JUNG *et al.*, 2010; GUVEN; CENGIZLER, 2014).

Li et al. (LI *et al.*, 2012) introduziu um método para segmentação de núcleos que utiliza o espaço de cores CIELAB (HUNTER, 1948). Após isso, é extraído o canal L e aplicado um filtro da média para redução de ruídos, produzindo uma entrada melhor ao algoritmo de clusterização k-médias. Esse algoritmo é responsável por extrair o contorno inicial dos núcleos. Esse contorno é aprimorado pelo uso do algoritmo de contorno ativo *Radiating Gradient Vector Flow* (LI *et al.*, 2012). Os resultados obtidos mostraram uma taxa de 91,97% utilizando a métrica de similaridade Zijdenbos (ZIJDENBOS *et al.*, 1994) para a base de imagens Herlev. Um ponto negativo desse trabalho é o uso de algoritmo de contorno ativo, o que torna o custo computacional do método bastante elevado.

Propomos um método de segmentação de núcleos aplicado à base de imagens CRIC, sendo os principais motivos: 1) redução da complexidade computacional da tarefa considerando que o custo computacional da segmentação de citoplasma em imagens de células cervicais em meio convencional é alto; e 2) possibilidade de utilizar somente a borda do núcleo para extrair informações que discriminem imagens de células normais de anormais.

3.3 Método Proposto

O algoritmo de segmentação proposto nesta tese tem por objetivo encontrar somente núcleos. O motivo da proposta desse segmentador de núcleos consiste em utilizá-lo como entrada do descritor proposto no Capítulo 4, que descreve as imagens de células sem a necessidade da identificação da borda do citoplasma. Na literatura somente Plissiti et al. (PLISSITI; NIKOU, 2012) reportaram o uso somente do núcleo para extrair atributos de células.

Nosso método de segmentação automática de núcleos foi baseado no uso do algoritmo de agrupamento k-médias (XU; MANDAL, 2015). É válido ressaltar que esse algoritmo

possui um baixo custo computacional, com a complexidade em $O(p)$, em que p é a quantidade de objetos, no caso de imagens, p é a quantidade de pixels. Utilizamos atributos de cor como entrada durante o agrupamento e atributos de forma para identificação das regiões/núcleos. O método foi dividido em três etapas: agrupamento de regiões, pré-classificação de regiões, e identificação de células individuais. A estimação de todos os parâmetros do método de segmentação de núcleos foi feita com o subconjunto de treino da base CRIC que é composto com 12 imagens totalizando 270 células.

Na etapa de agrupamento de regiões, é feito um pré-processamento aplicando o filtro da média com uma janela 5×5, que tem por objetivo remover ruídos inerentes à captura das imagens. Após isso, é realizado um agrupamento de pixels utilizando o algoritmo k-médias para separação das regiões baseado em atributos de cor, que são os níveis de cinza de cada pixel. Utilizamos o valor de $k = 2$, visto que o objetivo é segmentar os núcleos nas imagens de células cervicais, ou seja, ao final do processamento a imagem é dividida em duas regiões: núcleos e fundo. Os resultados experimentais incluíram testes com três diferentes modelos de cores, a saber: Lab, RGB, e HSV. Foram feitos testes com os diferentes canais de cores de cada modelo, por fim o canal verde (G) do sistema de cores RGB se mostrou o mais apropriado e de melhor contraste. O próximo passo é a aplicação de um operador morfológico de abertura com um disco de tamanho 10 como elemento estruturante. Observamos que nenhum núcleo seria perdido com esse elemento estruturante. A operação morfológica foi utilizada para remoção de ruídos resultantes da aplicação do algoritmo k-médias.

A segunda etapa é responsável por extrair os atributos de forma, área e compacidade, que são utilizados para a pré-classificação de regiões. Utilizamos atributos de forma por serem computacionalmente simples e apresentarem resultados condizentes com o objetivo do algoritmo. Foram testados outros atributos de forma como perímetro, circularidade e diâmetro. Esses atributos apresentaram taxas de acerto semelhantes às obtidas pela área e compacidade. Entretanto, optamos por estes por apresentarem menor tempo de processamento. Os atributos de forma são calculados em todas as regiões resultantes do agrupamento com o uso do k-médias e são responsáveis por separar essas regiões em:

1. Agrupamento de células;
2. Candidatos a núcleo; e
3. Artefatos, ex. glóbulos brancos.

A partir da estimação de parâmetros utilizando o subconjunto de treino da base CRIC,

obtivemos alguns limiares para definir esses regiões. Dessa forma, caso a região encontrada possua menos que 600 pixels, ou sua compacidade seja menor que $0,3$ essa região é considerada um artefato e é descartada. Caso a região possua mais que 4000 pixels a região é considerada um agrupamento de células. Nesse caso o k-médias é aplicado novamente somente nessa região com o objetivo de separar esse agrupamento. Por fim, caso a região não obedeça as regras anteriores ela é considerada um candidato a núcleo.

A última etapa consiste em detectar células individuais e criar a base de imagens de células livres, ou seja, uma célula por imagem. As células são recortadas considerando o centro do núcleo encontrado e utilizando como padrão o tamanho 100×100 para os referidos recortes.

A Figura 13 ilustra os principais passos para o processo de segmentação proposto. Após a metodologia de segmentação, obtivemos, a partir do subconjunto de teste da base CRIC, um total de 2470 imagens de células com tamanho 100×100, sendo 1004 anormais e 1466 normais com suas respectivas máscaras de segmentação do núcleo. Essas imagens foram utilizadas para validação dos métodos de descrição, classificação e recuperação de imagens propostos.

O subconjunto de treino da base CRIC possui 3782 marcações de células. Dessa forma, observamos que o método de segmentação proposto encontrou 65,31% dos núcleos que compõe a base.

3.4 Considerações Finais

O método de segmentação foi desenvolvido para a identificação de núcleos. É válido ressaltar que um dos maiores desafios no processamento e análise de células cervicais é a acurácia dos algoritmos de segmentação de citoplasma, devido, principalmente, aos níveis de sobreposição das células. Em muitos campos de imagens é impossível identificar todas as bordas dos citoplasmas na imagem, o que dificulta a criação de uma base de imagens com verdade-terrestre para imagens reais em meio convencional, sendo necessário soluções alternativas para esse problema.

O próximo capítulo apresenta uma nova proposta para descrição de imagens de células que independe da segmentação do citoplasma. Tal proposta pode ser utilizada em imagens com grande quantidade de células com sobreposição além de apresentar uma baixa complexidade computacional.

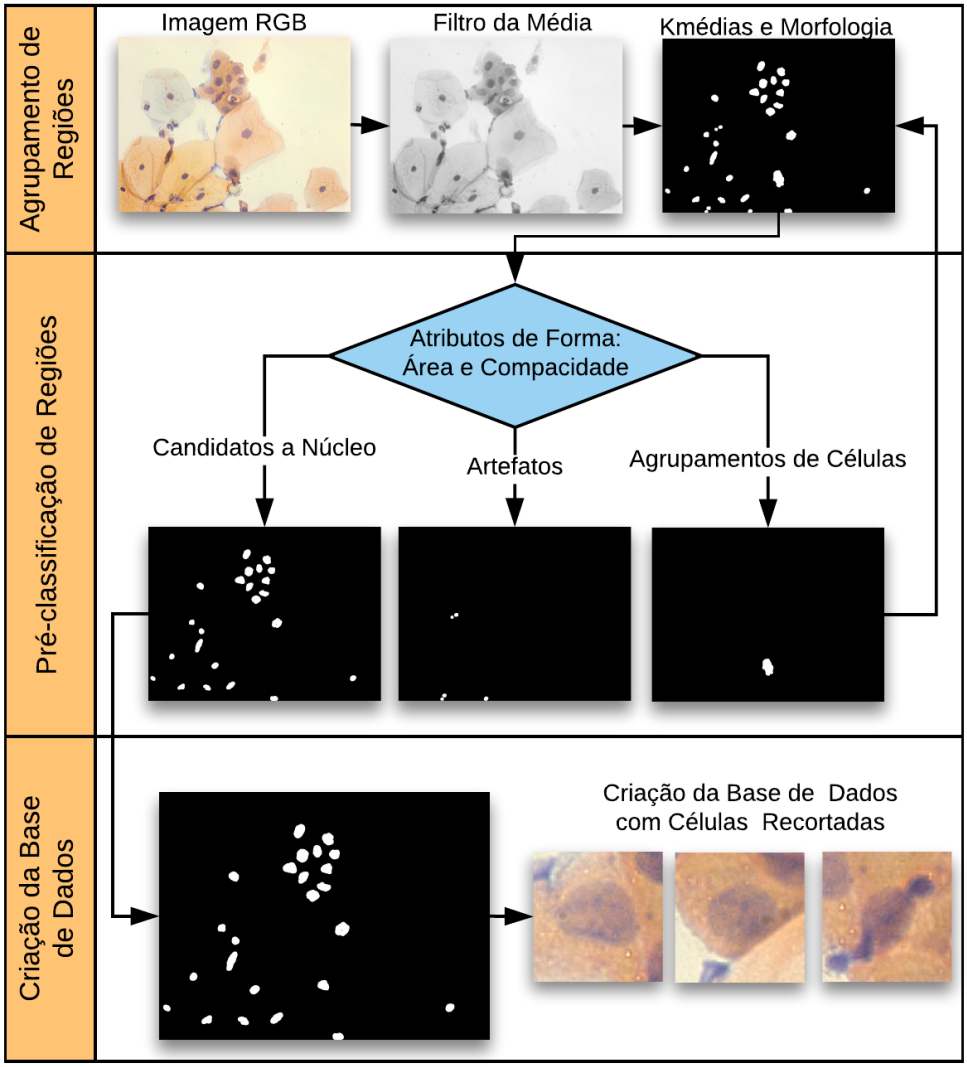


Figura 13 – Fluxograma do método proposto para segmentação de núcleos. O método é dividido em três etapas: agrupamento de regiões, pré-classificação de regiões de células; e criação da base de imagens.

4 DESCRIÇÃO DE CÉLULAS

Neste capítulo apresentamos os principais trabalhos relacionados à descrição de imagens de células cervicais bem como um novo descritor baseado na textura interna e ao redor do núcleo: *Radial Feature Descriptor*.

4.1 Trabalhos Relacionados

A maioria dos métodos para classificação de células presentes na literatura utilizam descritores híbridos que combinam atributos de forma e textura. Apresentaremos os principais métodos propostos na literatura e os atributos utilizados na análise comparativa desta tese.

Os métodos de extração de atributos em imagens de células cervicais podem ser divididos em dois grupos de acordo com os tipos de atributos utilizados, são eles: 1) informações somente do núcleo, como o proposto em (PLISSITI; NIKOU, 2012); e 2) informações de toda a célula (núcleo e citoplasma) (BEJNORDI *et al.*, 2013; MARINAKIS *et al.*, 2009). Alguns autores têm apresentado métodos promissores de diagnóstico assistido por computador para identificar células ou núcleos com base em atributos de forma (PLISSITI; NIKOU, 2012; MARINAKIS *et al.*, 2009) e textura (SA; BACKES, 2014), além de métodos híbridos (USHIZIMA *et al.*, 2014; BEJNORDI *et al.*, 2013; KALE; AKSOY, 2010; MARIARPUTHAM; STEPHEN, 2015).

Marinakis et al. (MARINAKIS *et al.*, 2009) extraíram um conjunto de 20 atributos para o núcleo e citoplasma para classificar imagens da base Herlev (JANTZEN *et al.*, 2005). Esse conjunto de atributos é composto por características de forma como área, diâmetro, e alongamento, além de atributos de intensidade como brilho, máxima e mínima[1]. Marinakis et al. utilizaram um algoritmo genético para buscar o melhor conjunto de atributos e obtiveram Falso Negativo (FN) e Falso Positivo (FP) de 2,66% e 10,74% para a classificação de células em normais e anormais utilizando a metodologia *10-fold cross-validation* e o classificador KNN.

Plissiti e Nikou (PLISSITI; NIKOU, 2012) extraíram um conjunto com 9 atributos de forma e intensidade da região do núcleo, dentre eles: área, diâmetro, brilho, intensidades máxima e mínima. Esses autores obtiveram um *H-mean* (média harmônica entre a sensitividade e a especificidade) de 0,74 na classificação de células em duas classes (normais e anormais) utilizando o k-médias Fuzzy (DUNN, 1973) como classificador e a base de imagens Herlev.

Bejnordi et al. (BEJNORDI *et al.*, 2013) apresentaram um conjunto de atributos de

[1] Esse atributos são calculados utilizando o valor máximo e mínimo de intensidade em uma janela 3×3 em uma área específica.

textura com o objetivo de quantificar padrões de cromatina no núcleo em células cervicais do colo do útero. Os resultados foram obtidos em uma base de imagens privada e mostraram que os atributos estruturais de textura foram os mais relevantes nos experimentos de classificação. A combinação de atributos estruturais e convencionais de textura resultou numa classificação com acurácia de 95,4% considerando as classes de células normais e anormais.

A combinação de atributos de forma e textura foram reportados por Mariarputham e Stephen (MARIARPUTHAM; STEPHEN, 2015), que utilizaram sete conjuntos de atributos que incluíram: a relação do tamanho entre núcleo e citoplasma, o intervalo dinâmico[2], os quatro primeiros momentos relativos a intensidade (média, variância, assimilaridade, e curtose), posição do núcleo em relação ao citoplasma, GLCM (HARALICK *et al.*, 1973), LBP (OJALA *et al.*, 1996), atributos de Tamura (TAMURA *et al.*, 1978), e Histograma de Orientação das Bordas (*Edge Orientation Histogram* (EOH)) (FREEMAN *et al.*, 1994). Mariarputham e Stephen utilizaram o classificador SVM (CORTES; VAPNIK, 1995) na base de imagens Herlev e obtiveram uma precisão de 97,38% para a classe normal escamosa, 93,89% para a escamosa intermediária, 86,90% para a colunar, 87,33% para a displasia leve, 58,52% para a displasia severa, 84,72% para a carcinoma, e 83,62% para a displasia moderada. Apesar de obterem altas taxas de acerto, a combinação de descritores proposta por Mariarputham e Stephen possui um custo computacional elevado devido à quantidade de atributos extraídos.

Os trabalhos apresentados mostram que a extração de atributos de textura é uma importante tarefa na classificação de células cervicais. Da mesma forma, atributos de forma extraídos do núcleo e do citoplasma representam informações relevantes na análise de células.

4.2 Descritor de Atributos Radiais

Nesta tese propomos um descritor baseado na borda do núcleo que calcula intensidades ao redor do núcleo utilizando informações de textura das células.

De acordo com os resultados observados nos trabalhos relacionados, os atributos extraídos de toda a célula são mais relevantes do que aqueles extraídos somente do núcleo (PLISSITI; NIKOU, 2012; MARINAKIS *et al.*, 2009). Nós testamos essa premissa e também propomos o descritor RFD que combina informação do núcleo e citoplasma, utilizando somente a segmentação do núcleo. Esse descritor é composto pelo histograma radial (*Radial Histogram* (RH)) e o GLRLM (GALLOWAY, 1975) com o objetivo de obter informação de variação de

[2] Diferença entre os pixels de maior e menor intensidade.

intensidade na área citoplasmática. A Figura 14 mostra as etapas para o cálculo do RFD. A partir da base de imagens é realizada a segmentação do núcleo. As entradas do RFD são a imagem da célula em níveis de cinza e a máscara de segmentação do núcleo. Utilizamos nesse trabalho a componente em níveis de cinza resultante de uma combinação dos canais de cores do modelo RGB e na saída são obtidos os atributos que são utilizados para categorizar as células em normais e anormais.

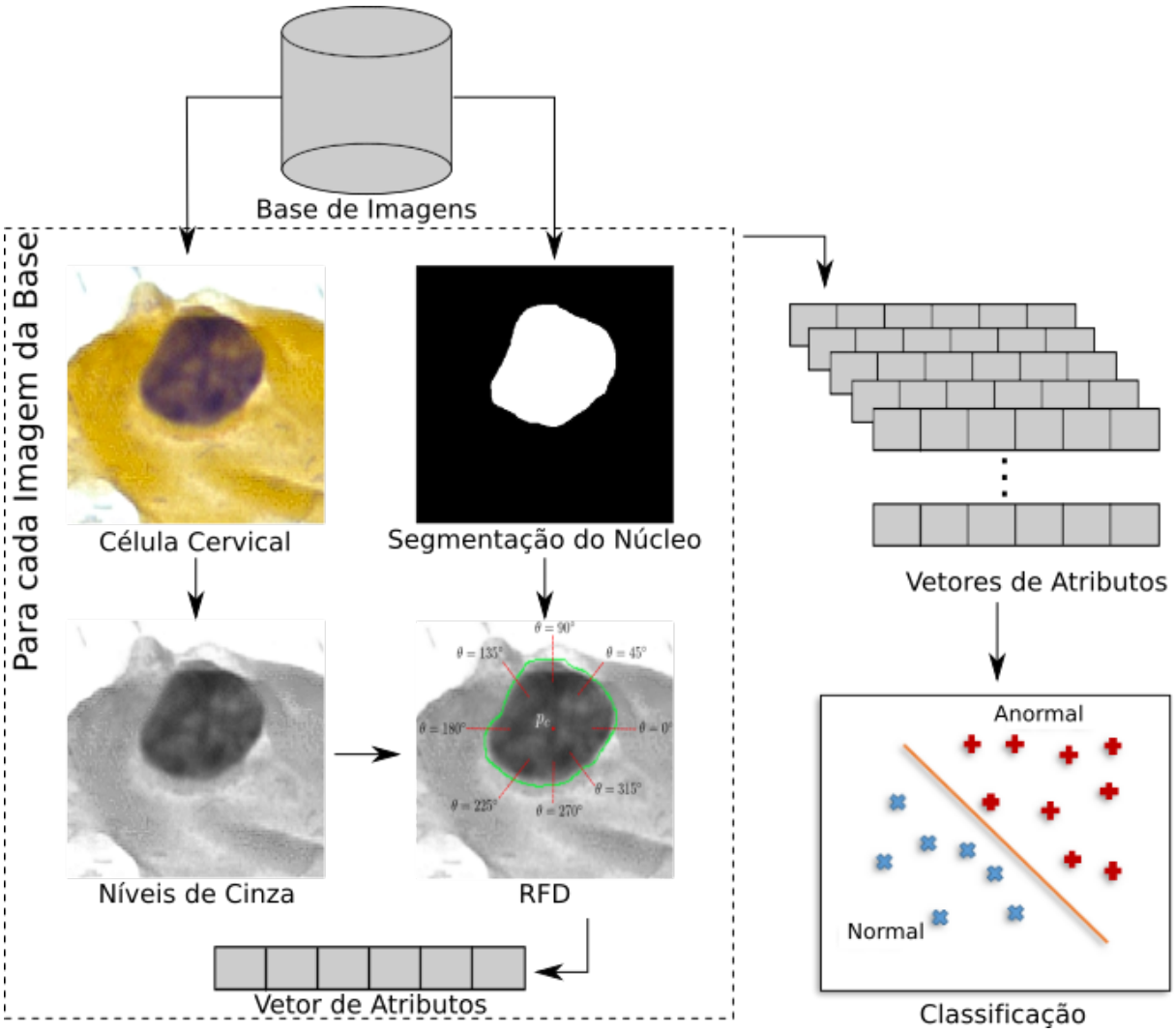


Figura 14 – Etapas para classificação de células: transformação de cor em níveis de cinza, detecção de bordas do núcleo, histograma radial, processamento para a base de dados, método de classificação.

4.3 Histograma Radial

A análise da textura do núcleo é uma importante tarefa para a extração de atributos em células cervicais. A Figura 15 apresenta o histograma na região do núcleo de uma célula

normal e outra anormal. O histograma da célula normal possui uma menor distribuição de intensidades em relação à célula anormal, isso se deve ao fato de não existir nenhum tipo de deformação na célula. Células anormais, geralmente, possuem o núcleo aumentado em relação ao citoplasma, com isso é possível observar uma maior irregularidade na textura nesse tipo de célula. Além disso, os pixels próximos à borda do núcleo de células anormais possuem uma transição mais suave em relação a relação a células normais.

A variação de intensidade e a irregularidade da textura no núcleo de células cervicais é conhecido como distribuição de cromatina (WATANABE *et al.*, 2004). Tal distribuição é largamente utilizada por citologistas para classificar células cervicais. O uso de descritores de textura no interior do núcleo tem por objetivo mensurar essa distribuição.

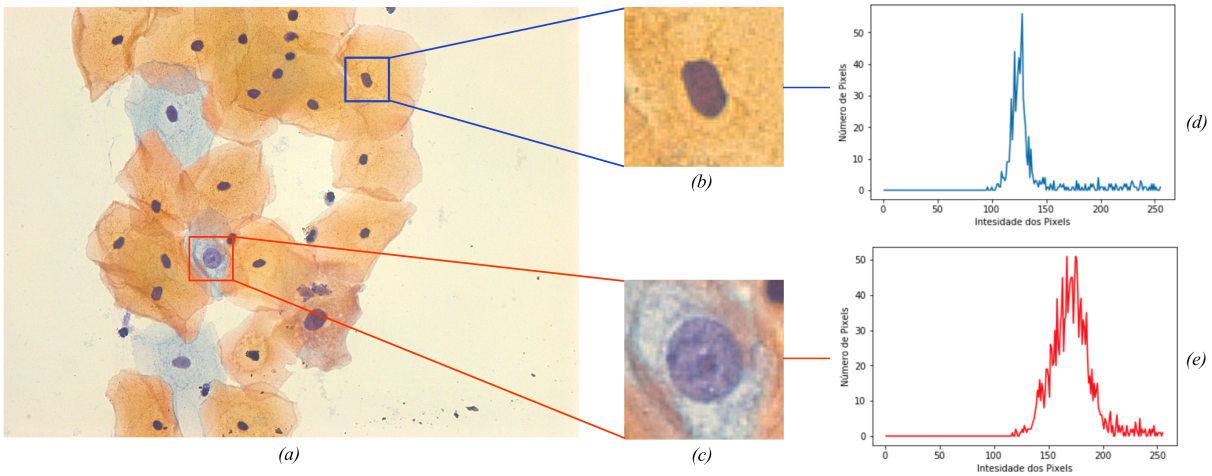


Figura 15 – Distribuição de intensidade em imagens de núcleos de células cervicais. (a) Imagem original. Núcleo de uma célula saudável (b) e anormal (c). (d) e (e) Histogramas da região do núcleo de (b) e (c), respectivamente.

O RH define retas com origem dentro do núcleo e se estendem até o citoplasma. Elas são distribuídas circularmente ao redor da borda do núcleo em vários ângulos para calcular um histograma de intensidades ao longo dessas retas. As retas cruzam a célula com início no núcleo e se estendem até o citoplasma. O cálculo do descritor utiliza a máscara de segmentação do núcleo das imagens de células cervicais disponível nas bases de imagens, o cálculo de cada reta utiliza um ponto da borda do núcleo. Dessa forma, o número de retas (n_r) é proporcional à quantidade de pontos que a borda do núcleo possui (p_e). Essa proporção é definida pelo parâmetro $0 < n \leq 1$, que é definido empiricamente, como mostra a Equação 4.1. Caso $n = 1$, todos os pixels da borda do núcleo serão utilizados, ou seja, $n_r = p_e$. Entretanto, caso $n \approx 0$,

poucas retas serão obtidas pelo cálculo do RH.

$$n_r = n \times p_e, \tag{4.1}$$

onde $\times$ é o produto escalar.

As retas do RH são igualmente espaçadas ao redor do núcleo, ou seja, a distância entre retas consecutivas é definida por um ângulo que é calculado dividindo o espaço circular (2π) pela quantidade de retas, como mostra a Equação 4.2.

$$\hat{a}ngulo = \frac{2\pi}{n_r}. \tag{4.2}$$

A partir da máscara de segmentação do núcleo, extraímos os pontos da borda $\{p_e\}$ e então calculamos o centro de massa (p_c) da região do núcleo. Selecionamos n_r pontos a partir do conjunto $\{p_e\}$, i.e., $\{p^1, p^2, \cdots, p^{n_r}\}$, seguindo o seguinte critério:

$$p^j = p \in \{p_e\} | \theta_j = \arctan(\Delta y/\Delta x) \simeq j \times \frac{2\pi}{n_r}, \tag{4.3}$$

onde, $\Delta x = x_p - x_{p_c}$, $\Delta y = y_p - y_{p_c}$ e $j = 1 \cdots n_r$.

Os valores x_p, y_p são as coordenadas de um ponto qualquer no conjunto p_e e x_{p_c}, y_{p_c} são as coordenadas de p_c. Nosso algoritmo faz uma busca por n_r pontos de borda separados pelo mesmo ângulo relacionados com o centro de massa do contorno do núcleo. Na sequência, nós definimos um ponto externo (p^j_+) e um ponto interno (p^j_-) para cada p^j calculado anteriormente. Esses pontos são utilizados para calcular a variação de intensidade em células cervicais e são definidos como pontos a uma distância D_j passando por p^j indo de p^j_- até p^j_+. O valor de D_j é definido em razão do parâmetro de distância chamado de d, que é definido empiricamente.

A distância é dada por:

$$D_j = \left| Dt(p_c, p^j) \right| \times d, \tag{4.4}$$

em que $|\cdot|$ denota o valor absoluto, $Dt(\alpha, \beta)$ corresponde ao cálculo da distância euclidiana entre dois pontos α e β e d é o parâmetro que controla o tamanho de D_j.

Para cada imagem da base de dados, criamos um histograma utilizando a intensidade de todos os pixels que estão no intervalo $[p^j_+, p^j_-]$, $j = 1 \cdots n_r$. Por fim, o histograma é normalizado. O descritor utiliza um imagens em níveis de cinza calculado pelo balanceamento das componentes do modelo de cor RGB (Equação 4.5).

$$cinza = 0.299 * R + 0.587 * G + 0.114 * B, \tag{4.5}$$

onde R, G e B são as componentes do modelo de cor RGB.

A Figura 16 mostra a variação de intensidade do núcleo para o citoplasma em células normais e anormais, indicando a diferença entre esses padrões. A curva representa o RH normalizado calculado com $n = 0,7$ e $d = 0,5$. Como podemos observar nas Figuras 16 (a) e (c), as retas do RH tem início dentro do núcleo cruzando o mesmo e chega ao citoplasma. O histograma da célula saudável é visivelmente dividido em duas regiões, que representam os pixels pertencentes ao núcleo (mais escuro), e ao citoplasma (mais claro). Em relação à célula anormal observamos que não há contraste e a fronteira não é bem definida.

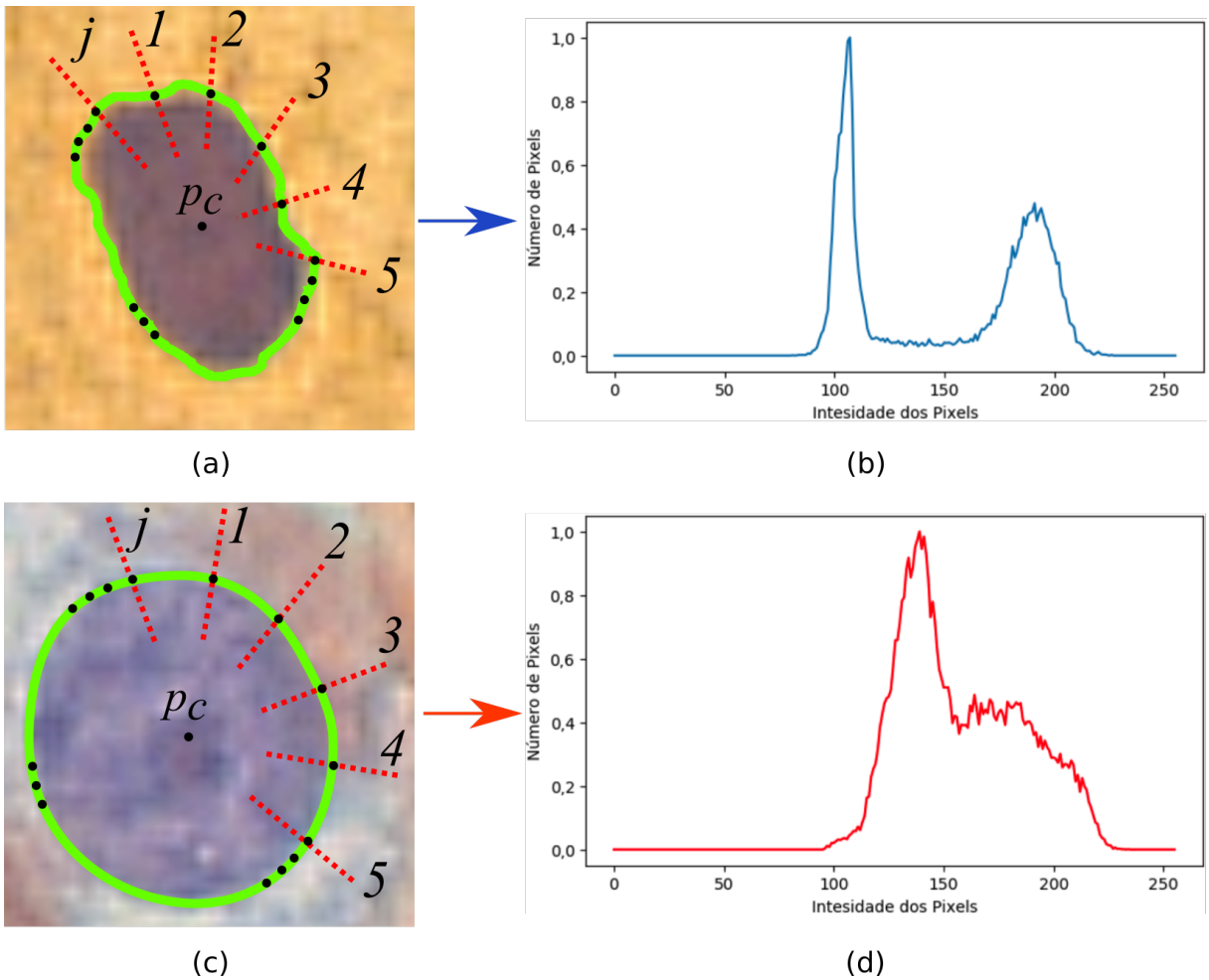


Figura 16 – Cálculo do histograma radial. (a) Imagem do núcleo de uma célula saudável. (b) Histograma Radial de (a). (c) Imagem do núcleo de uma célula anormal. (d) Histograma Radial de (c).

4.3.1 Gray-Level Run Length Matrix

O cálculo da GLRLM ocorre dentro da região do núcleo da célula. Nossa hipótese é que a GLRLM é o melhor descritor para calcular a distribuição de cromatina no núcleo da célula.

Como mostramos anteriormente, a distribuição de cromatina é uma importante informação para separar as células em normais e anormais.

Calculamos 11 atributos para cada direção $(0°, 45°, 90°, 135°)$ da GLRLM, são eles: SRE, LRE,
glsGLN, RLN, RP, LGRE, HGRE, SRLGE, SRHGE, LRLGE, LRHGE. A GLRLM totalizou 44 atributos.

A etapa final para o cálculo do RFD é a concatenação do RH e dos atributos calculados a partir da GLRLM, formando um único vetor de atributos. O RH possui 256 atributos pois é gerado a partir de imagens de 8-bits. Dessa forma, o vetor de atributos consiste em 300 atributos.

4.4 Experimentos de Classificação de Células Cervicais

Os testes foram realizados utilizando duas bases de imagens. Além da base CRIC, apresentada no Capítulo 3, utilizamos a base de imagens Herlev (JANTZEN *et al.*, 2005) que é composta por 917 imagens de células e classificadas em sete níveis de lesão. A base Herlev possui verdade-terrestre de segmentação do núcleo e citoplasma para todas as células. Em virtude disso, ela tornou-se largamente utilizada para testes de métodos automáticos para caracterização de células (PLISSITI; NIKOU, 2012; MARINAKIS *et al.*, 2009; SA; BACKES, 2014; MARIARPUTHAM; STEPHEN, 2015; CHANKONG *et al.*, 2014; SARWAR *et al.*, 2015; GENÇTAV *et al.*, 2012). A Figura 17 apresenta exemplos de imagens da base Herlev: as bordas brancas representam a verdade-terrestre da segmentação feita por um especialista. As imagens da Herlev estão assim distribuídas: 675 imagens com células anormais (*squamous cell carcinoma in situ intermediate, mild squamous non-keratinizing dysplasia, moderate squamous non-keratinizing dysplasia, e severe squamous non-keratinizing dysplasia*) e 242 imagens de células normais (*columnar epithelial*, *intermediate squamous epithelial*, *superficial squamous epithelial*).

Foram realizados dois tipos de testes nos experimentos de classificação: 1) classificação binária onde a base de dados é dividida em células normais e anormais e 2) classificação multiclasse (graus de normalidade e anormalidade) onde dividimos a base de dados em sete classes, esse teste foi realizado somente na Herlev, visto que ela possui subdivisões dentro das classes normal e anormal.

Dessa forma, utilizamos o algoritmo RF (BREIMAN, 2001) com a metodologia

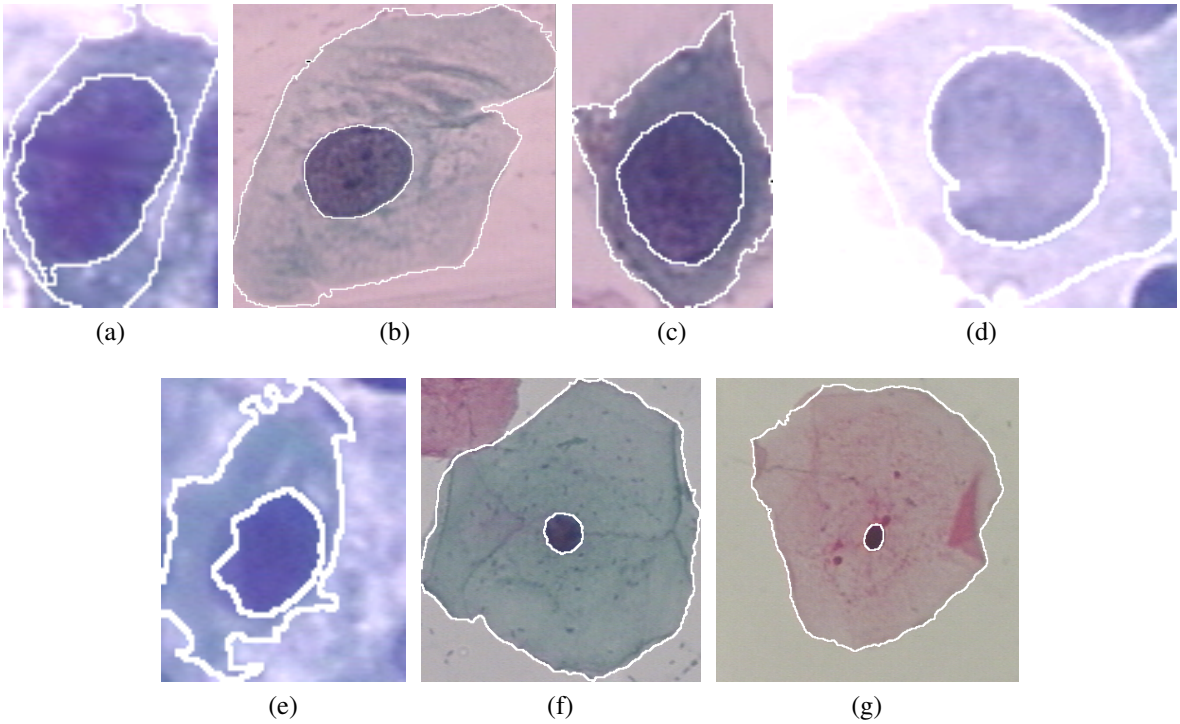


Figura 17 – Exemplos de imagens de células cervicais da base Herlev. (a) *Intermediate squamous cell carcinoma in situ*. (b) *Mild squamous non-keratinizing dysplasia*. (c) *Moderate squamous non-keratinizing dysplasia*. (d) *Severe squamous non-keratinizing dysplasia*. (e) *Columnar epithelial*. (f) *Intermediate squamous epithelial*. (g) *Superficial squamous epithelial*. As bordas brancas correspondem às máscaras de segmentação do núcleo e citoplasma disponíveis.

de classificação *bootstrap* 0,632 para classificação de imagens de células. Foram realizados testes preliminares com outros classificadores como SVM (CORTES; VAPNIK, 1995), KNN, e *Multilayer Perceptron* (HAYKIN, 1998). Entretanto, os mesmos não apresentaram resultados superiores aos do classificador RF.

4.4.1 Métricas de Avaliação dos Resultados

A classificação de imagens nas bases Herlev e CRIC em normais e anormais, considerou que a base Herlev contém sete diferentes níveis de normalidade e anormalidade e que essas bases apresentam diferentes quantidades de imagens por classe, ou seja, ambas são desbalanceadas. Devido a esse desbalanceamento, escolhemos o coeficiente Kappa (κ) (LANDIS; KOCH, 1977) para avaliar o desempenho dos experimentos de classificação. O κ é calculado pela Equação 4.6 e são estabelecidos os seguintes níveis de acurácia na classificação: Ruim,

Razoável, Bom, Muito Bom e Excelente, como mostra a Tabela 2.

$$\kappa = \frac{\beta_1 - \beta_2}{1 - \beta_2}, \tag{4.6}$$

onde,

$$\beta_1 = \frac{VP + VN}{VP + VN + FP + FN}, \tag{4.7}$$

$$\beta_2 = \frac{[(VP + FN)(VP + FP)] + [(VN + FN)(VN + FP)]}{(VP + VN + FP + FN)^2}, \tag{4.8}$$

e VP, VN, FP e FN são os valores para Verdadeiro Positivo, Verdadeiro Negativo, Falso Positivo e Falso Negativo.

Tabela 2 – Nível de classificação da acurácia de acordo com o índice Kappa.

Índice Kappa (κ)	Qualidade
$\kappa < 0{,}2$	Ruim
$0{,}2 \leq \kappa < 0{,}4$	Razoável
$0{,}4 \leq \kappa < 0{,}6$	Bom
$0{,}6 \leq \kappa < 0{,}8$	Muito Bom
$\kappa \geq 0{,}8$	Excelente

Diferente da curva *Receiver Operating Characteristic* (ROC) (BRADLEY, 1997), o coeficiente Kappa não depende do balanceamento dos dados. Alguns trabalhos em classificação de células cervicais também adotaram o índice Kappa como métrica de avaliação (CHANKONG *et al.*, 2014; GENÇTAV *et al.*, 2012; WANG *et al.*, 2010). Além disso, utilizamos a taxa de falsos negativos (*False Negative Rate* (FNR)) como métrica de avaliação uma vez que esta é uma importante métrica na avaliação de sistemas aplicados à área médica. O FNR (Equação 4.9) calcula a razão entre o número de células anormais classificadas como células normais, ou seja, ela reflete os casos em que o paciente é diagnosticado como saudável, mas apresenta patologia.

$$FNR = \frac{FN}{FN + VN}. \tag{4.9}$$

Nos experimentos multiclasse da base Herlev utilizamos a metodologia todos-contra-um. Dessa forma, o cálculo das métricas de avaliação é feito da mesma forma nesses experimentos.

4.4.2 Metodologia de Classificação com Bootstrap 0,632

Utilizamos o método *bootstrap* (EFRON, 1983) nesta tese para criar os conjuntos de treino e teste. Esse método simula várias bases de dados a partir dos dados originais sem utilizar informação prévia dos dados. Dada uma base de dados $\mathbf{x} = (x_1, x_2, ..., x_N)$ com N amostras (vetores de atributos), geramos M distribuições aleatórias a partir de $\mathbf{x}$. Cada amostra, definida por $\mathbf{x}^* = x_1^*, x_2^*, ..., x_N^*$, é composta por N amostras do conjunto original com reposição de dados.

Considerando que $\varepsilon(\mathbf{x}_{treino}^{*^m}, \mathbf{x}_{teste}^{*^m})$ é a taxa de acerto para o treinamento com $\mathbf{x}_{treino}^{*^m}$ e teste com $\mathbf{x}_{teste}^{*^m}$. Utilizando a técnica *bootstrap*, geramos M conjuntos de treino $(\mathbf{x}_{treino}^{*^1}, \mathbf{x}_{treino}^{*^2}, ..., \mathbf{x}_{treino}^{*^M})$ onde cada $\mathbf{x}_{treino}^{*^m} = x_1^{*^m}, x_2^{*^m}, ..., x_N^{*^m}$ é obtido pela escolha de N vetores de atributos, com reposição de dados, a partir da base de dados original $\mathbf{x}$. Definimos ainda o conjunto de teste por $\mathbf{x}_{test}^{*^m}$ e esse conjunto possui vetores de atributos que não aparecem em $\mathbf{x}_{treino}^{*^m}$. O método *bootstrap* $0,632$ foi introduzido em (EFRON, 1983) e pode ser descrito por:

$$\varepsilon_{0,632} = \varepsilon(\mathbf{x}, \mathbf{x}) - \hat{w}_{0,632}^m, \tag{4.10}$$

em que $\varepsilon(\mathbf{x}, \mathbf{x})$ é o índice para o treinamento/teste com a base de dados original $\mathbf{x}$ e:

$$\hat{w}_{0,632}^m = 0,632[\varepsilon(\mathbf{x}, \mathbf{x}) - \varepsilon(\mathbf{x}_{treino}^{*^m}, \mathbf{x}_{teste}^{*^m})]. \tag{4.11}$$

A estimação para o método 0,632 é encontrado pela Equação 4.11 com M amostras. O valor de $\varepsilon_{0,632}$ é dado por:

$$\varepsilon_{0,632} = 0.368\varepsilon(\mathbf{x}, \mathbf{x}) + \frac{0,632}{M} \sum_{m=1}^{M} \varepsilon(\mathbf{x}_{treino}^{*^m}, \mathbf{x}_{teste}^{*^m}) \tag{4.12}$$

4.4.3 Estimação de Parâmetros

A estimação de parâmetros para um classificador visa obter o melhor desempenho em uma tarefa de classificação. A partir do cálculo dos vetores de atributos de uma base de imagens, obtivemos uma base de vetores de atributos que foi dividida em dois subconjuntos: o primeiro corresponde aos vetores associados ao subconjunto de treino da base CRIC e é utilizado para estimação de parâmetros; o segundo corresponde aos vetores associados ao subconjunto de teste da base CRIC e é utilizado nos experimentos de classificação. Em relação à base Herlev, utilizamos 30% dos dados para estimação de parâmetros e 70% para avaliar os resultados.

As seguintes variáveis do classificador RF são partes da estimação de parâmetros: o número de árvores utilizadas, a função que mede a qualidade das divisões dos nós da árvore,

a profundidade máxima da árvore, o número máximo de atributos quando a melhor divisão é analisada, a quantidade mínima de amostras para dividir um nó interno, e a quantidade mínima de amostras necessárias para estar em um nó folha. A Tabela 3 mostra o intervalo de valores que nós utilizamos para a parametrização do RF. Além disso, essa tabela mostra os valores obtidos para a otimização dos parâmetros utilizando os atributos do descritor RFD.

A estimação de parâmetros para os descritores foi feita empiricamente, utilizando um intervalo de valores e calculando as métricas de acerto. A Figura 18 mostra alguns dos valores testados para os descritores GLCM, GLRLM e HOG.

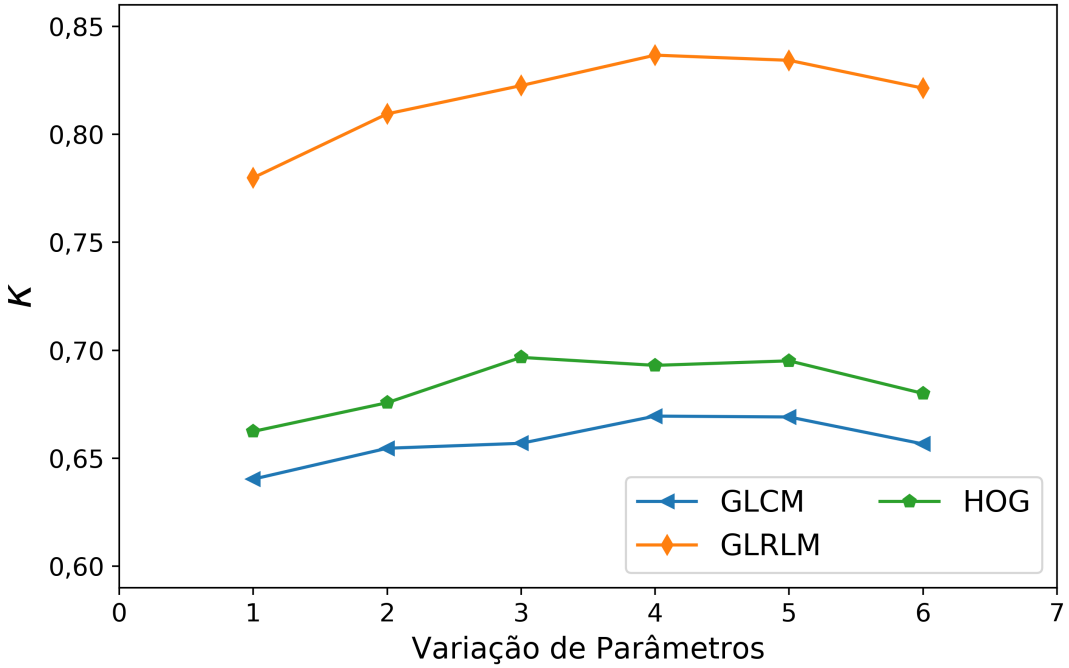


Figura 18 – Estimação de parâmetros para extração de atributos utilizando a métrica de avaliação Kappa. Esse gráfico mostra 5 diferentes parâmetros para cada descritor. Em relação à GLCM são exibidos resultados para os seguintes valores de distância: {1,3,5,7,9,11}, respectivamente. A curva da GLRLM mostra os resultados obtidos para os seguintes intervalos de níveis de cinza: {8,16,32,64,128,256}. Por fim, no descritor HOG, fixamos o número de blocos em 20 e variamos o número de sub-blocos com os seguintes valores: {9,10,11,12,13,14}.

4.4.4 Resultados Quantitativos para Classificação de Imagens de Células Cervicais

Em nossos testes utilizamos os parâmetros n e d do descritor RFD, variando no intervalo $[0,1;1,0]$ com passos de $0,2$. Além disso, comparamos nossos resultados com métodos da literatura. Reportamos somente os resultados para o melhor conjunto de parâmetros atribuídos

Tabela 3 – Parâmetros do classificador *Random Forest*, onde s é a quantidade de atributos em cada vetor (seu valor depende do descritor utilizado), e $range(\alpha, \beta, \delta)$ é uma função que retorna valores entre α e β sendo δ o intervalo entre esse valores.

Parâmetros	Intervalo de Valores	Valores encontrados para o RFD
Número de árvores	$range(10, 1000, 50)$	810
Profundidade máxima	$range(1, 100, 1)$	43
Número máximo de atributos	$range(1, s, 1)$	10
Mínimo de amostras (nós internos)	$range(1, s, 1)$	7
Mínimo de amostras (nós folha)	$range(1, s, 1)$	2

ao RH e ao RFD. Utilizamos $n = 0,7$ e $d = 0,5$. Dessa forma, observamos que valores de n e d próximos a 0 não exploram a total capacidade do RH em relação às taxas de classificação. Por exemplo, utilizando $n = 0,1$ e $d = 0,1$ obtivemos um valor de $\kappa = 0,77$ para a base Herlev. De fato, com $n = 0,1$ somente 10% dos pixels da borda do núcleo são utilizadas para definir as retas que compõe o descritor. Da mesma forma, com $d = 0,1$ somente 10% do tamanho do raio do núcleo será utilizado para definir o tamanho das retas. Concluímos que esses valores de n e d não tornam possível uma boa descrição da célula, visto que informações importantes da célula não serão capturadas pelo descritor. Essa mesma análise pode ser feita em relação a valores desses parâmetros próximos a 1. Utilizando $n = 0,9$ e $d = 0,9$, por exemplo, obtivemos um valor de $\kappa = 0,81$ para a base Herlev. Valores próximos a 1 de n retornam informações redundantes devido à baixa resolução de algumas imagens das bases Herlev e CRIC, em algumas delas o núcleo possue menos de 360 pixels na sua borda. O mesmo se aplica ao parâmetro d, que pode adicionar informações irrelevantes ao descritor, tais como características de regiões fora do citoplasma.

A Tabela 4 mostra o valor de FNR e o κ obtidos com as base de dados Herlev e CRIC. O método *bootstrap* 0,632 utilizou M=500 para criar os conjuntos de treinamento e teste. O melhor resultado obtido pelo método proposto para a base de imagens Herlev foi $\kappa = 0,89$. Apesar do FNR alcançado pelo RFD não ter sido o melhor (FNR$= 0,02 \pm 0,01$), esse valor foi próximo ao melhor e com um desvio padrão próximo de zero, o que torna o resultado mais confiável.

Em relação à base de dados CRIC, o melhor resultado foi obtido utilizando somente o RH, e o RFD obteve o segundo melhor resultado. Isso pode ser explicado pela baixa taxa de acurácia que o descritor GLRLM obteve. Vale ressaltar que as máscaras de segmentação da base CRIC foram geradas pelo nosso algoritmo de segmentação. Dessa forma, a borda do núcleo não possui o mesmo nível de acurácia como as da base Herlev. Com isso, o resultado da classificação

alcançou uma taxa de acurácia menor para a base CRIC.

Tabela 4 – FNR e κ para cada método utilizando verdade-terrestre da base Herlev e do método proposto utilizando a base CRIC.

	Herlev		CRIC	
	FNR	κ	FNR	κ
GLRLM	**0,02±0,01**	0,86±0,04	0,19±0,02	0,73±0,02
RH	**0,02±0,01**	0,86±0,04	**0,14±0,03**	**0,78±0,03**
RFD	**0,02±0,01**	**0,89±0,04**	0,15±0,03	0,77±0,03

*Em negrito estão os melhores resultados.

Com o objetivo de avaliar nossos resultados de classificação, comparamos o método proposto com 7 métodos de extração de atributos da literatura e outros 6 de propósito geral, são eles: GLCM, atributos de histograma, HOG, LBP e duas CNNs. Para obter uma comparação justa, utilizamos a mesma metodologia de classificação para todos os métodos. Isso significa que testamos somente os atributos propostos por cada método. Os métodos foram implementados de acordo com as informações disponíveis nos respectivos trabalhos. Em relação aos métodos de propósito geral, utilizamos a região do núcleo para obter os atributos. Fizemos isso para comparar o desempenho do RFD, que utiliza somente o núcleo para extrair os atributos, com os demais.

Foram testadas duas CNNs por serem as técnicas mais utilizadas atualmente para descrição/classificação de imagens. A Inception-Resnet-v2 é uma rede pré-treinada com a base ImageNet (RUSSAKOVSKY *et al.*, 2015) e possui desempenho superior à maioria dos descritores do estado da arte em várias aplicações de imagens médicas (LITJENS *et al.*, 2017; ABIDIN *et al.*, 2018). A LeNet foi treinada com as bases Herlev e CRIC. Como CNNs necessitam de mais amostras que os demais algoritmos para serem treinadas, dividimos as bases de treino e teste em 50%. Em seguida, fizemos uma operação de *augmentation*, que tem por objetivo aumentar a base de treino através de transformações como translação, rotação e adição de ruído. Essa técnica é bastante utilizada para o treinamento de CNNs com bases que possuem menos de 5000 imagens, como é o caso da Herlev e CRIC. A rede LeNet foi treinada por 600 épocas, a condição de parada foi o decaimento do erro, ou seja, o treinamento parou quando o erro se alterou.

A Tabela 5 apresenta a análise comparativa utilizando as imagens das bases Herlev e BHS. O método proposto obteve um valor de $\kappa = 0,89$, superando os algoritmos do estado

da arte aplicados a células cervicais e os de propósito geral como as CNNs. Em segundo lugar estão os atributos propostos por Sarwar *et al*. (SARWAR *et al.*, 2015) que utiliza como atributos: área, diâmetro e perímetro do núcleo e citoplasma, dentre outros. Em relação ao FNR, a melhor taxa de sucesso foi obtida pelo método proposto por Mariarputham *et al*. (MARIARPUTHAM; STEPHEN, 2015) que utiliza atributos de textura, tais como GLCM, LBP e Tamura - e de forma, como a relação entre as áreas do núcleo e citoplasma. Por outro lado, Mariarputham *et al*. obtiveram uma menor taxa de sucesso em relação ao índice Kappa.

Nosso principal argumento em relação aos resultados apresentados é que os métodos do estado da arte, com exceção de Plissiti e Nikou (PLISSITI; NIKOU, 2012), utilizam resultados de segmentação do núcleo e citoplasma. De fato, a literatura reporta que não existe algoritmo para segmentação de citoplasma em imagens de exame de Papanicolau no modo convencional, como realizamos para a CRIC.

É válido mencionar que a base de imagens CRIC não possui verdade-terrestre em relação à segmentação do núcleo e do citoplasma, entretanto possui dados que indicam quais células possuem alguma patologia associada. Por isso, propomos um método para segmentação de núcleos. Com isso, comparamos nossos resultados para a base de dados CRIC com o resultado obtido pelo método introduzido em (PLISSITI; NIKOU, 2012) que utiliza a mesma entrada que o algoritmo proposto (segmentação do núcleo) e com os demais de propósito geral. Plissiti e Nikou (PLISSITI; NIKOU, 2012) obtiveram um $\kappa = 0,75 \pm 0,03$ e um FNR$= 0,17 \pm 0,03$, ou seja, desempenho inferior em relação ao índice Kappa ($\kappa = 0,77 \pm 0,03$) e a Taxa de Falso Negativos (FNR$= 0,14 \pm 0,03$) do método proposto. O baixo valor de FNR pode ter sido causado pela qualidade da segmentação das células. Espera-se que quanto melhor a segmentação do núcleo, melhores são os resultados obtidos pelo método proposto.

Além dos resultados para a classificação binária, apresentamos na Tabela 6 os resultados para as sete classes da base Herlev. Adotamos a metodologia de classificação um-contra-todos e podemos observar que esses métodos alcançaram resultados de κ abaixo de 0,60 em comparação à classificação binária. Isso se deve ao fato da base Herlev possuir um quantidade desbalanceada de imagens em cada classe e de existir uma grande diversidade de estruturas nucleares na classe das células anormais.

Comparamos nosso resultado com os obtidos pelo algoritmo proposto por Plissiti *et al*. (PLISSITI; NIKOU, 2012), nossa metodologia alcançou o melhor valor de κ em cinco das sete classes. Em relação ao FNR, os atributos de Plissiti *et al*. (PLISSITI; NIKOU, 2012) obtiveram

Tabela 5 – Análise comparativa para os experimentos de classificação: FNR e κ utilizando as bases de imagens Herlev e CRIC.

Descritores dos Métodos	Herlev		CRIC	
	FNR	κ	FNR	κ
CHANKONG *et al.*	0,02±0,02	0,84±0,04	-	-
CHEN *et al.*	0,02±0,02	0,86±0,04	-	-
GENÇTAV *et al.*	0,03±0,02	0,83±0,04	-	-
MARIARPUTHAM; STEPHEN	**0,01±0,01**	0,82±0,05	-	-
MARINAKIS *et al.*	0,03±0,02	0,82±0,04	-	-
PLISSITI; NIKOU	0,05±0,01	0,76±0,03	0,18±0,02	0,74±0,02
SARWAR *et al.*	0,02±0,02	0,88±0,04	-	-
GLCM	0,07±0,03	0,56±0,06	0,22±0,05	0,50±0,03
Histograma	0,02±0,01	0,83±0,04	0,32±0,06	0,50±0,03
HOG	0,26±0,04	0,77±0,04	0,50±0,05	0,43±0,04
LBP	0,21±0,06	0,75±0,05	0,27±0,04	0,65±0,03
Inception	**0,01±0,01**	0,77±0,05	**0,04±0,02**	0,74±0,05
LeNet	**0,01±0,01**	0,78±0,04	0,10±0,02	0,72±0,03
RFD	0,02±0,01	**0,89±0,04**	0,14±0,03	**0,77±0,03**

− : Esse símbolo nos campos da tabela indica os métodos que dependem da segmentação do citoplasma. Em negrito estão os melhores resultados.

melhor resultado que o RFD. De modo geral, os algoritmos que utilizaram a segmentação do citoplasma como entrada para o cálculo dos atributos apresentaram melhores resultados, como é o caso dos atributos propostos em (CHEN *et al.*, 2014), (MARINAKIS *et al.*, 2009) e (SARWAR *et al.*, 2015).

O algoritmo proposto foi implementado na linguagem de programação python e o tempo total de processamento para extrair os atributos utilizando um código não otimizado levou 0,06 segundos (s) em média para a base de imagens Herlev em um PC com um processador Intel Core i7 com 3,1 GHz e 16 GB RAM. Comparamos ainda o tempo de processamento do nosso algoritmo com o proposto por Sarwar *et al*. (SARWAR *et al.*, 2015) visto que o mesmo alcançou o segundo melhor resultado em relação aos valores de Kappa e FNR. O tempo médio de computação desse algoritmo foi de 0,04s, entretanto ele requer um tempo médio de 5,68s para realizar a segmentação do citoplasma aplicando o algoritmo proposto em por Li *et al*. (LI *et al.*, 2012). Apesar do RFD requerer mais tempo de processamento para extrair os atributos, ele não necessita da segmentação do citoplasma em contrapartida ao método de Sarwar *et al*. Com isso o RFD superou o algoritmo de Sarwar *et al*. em relação ao tempo de processamento.

Tabela 6 – Análise comparativa do FNR e κ utilizando a verdade-terreste para as sete classes da base de imagens Herlev.

	CI	LD	MD	NC	NI	NS	SD
				CHANKONG *et al.*			
κ	0,58±0,07	0,70±0,05	0,49±0,06	0,65±0,08	0,89±0,05	0,90±0,05	0,47±0,06
FNR	0,36±0,08	0,24±0,07	0,45±0,07	0,33±0,09	0,11±0,08	0,10±0,07	0,43±0,06
				CHEN *et al.*			
κ	**0,66±0,05**	0,72±0,04	**0,53±0,05**	**0,76±0,08**	0,88±0,05	**0,92±0,04**	0,55±0,06
FNR	**0,32±0,07**	**0,23±0,06**	0,46±0,05	**0,26±0,09**	0,10±0,08	0,08±0,07	**0,40±0,06**
				GENÇTAV *et al.*			
κ	0,64±0,06	0,72±0,05	0,50±0,07	0,73±0,07	0,88±0,07	0,91±0,05	0,50±0,06
FNR	0,33±0,08	0,24±0,05	0,47±0,07	0,27±0,09	0,12±0,08	0,08±0,07	0,44±0,06
				MARIARPUTHAM; STEPHEN			
κ	0,51±0,07	0,65±0,05	0,41±0,04	0,44±0,07	0,69±0,10	0,76±0,10	0,46±0,05
FNR	0,53±0,05	0,39±0,05	**0,61±0,03**	0,58±0,05	0,37±0,11	0,31±0,11	0,55±0,05
				MARINAKIS *et al.*			
κ	0,64±0,05	0,74±0,05	0,52±0,06	0,71±0,06	0,89±0,04	0,91±0,05	**0,51±0,05**
FNR	0,34±0,07	**0,23±0,06**	0,47±0,06	0,30±0,07	0,11±0,07	**0,08±0,06**	0,45±0,06
				PLISSITI; NIKOU			
κ	0,46±0,06	0,55±0,06	0,49±0,05	0,61±0,07	0,66±0,08	0,86±0,04	0,43±0,06
FNR	0,47±0,06	0,37±0,07	0,44±0,06	0,37±0,08	0,32±0,11	0,13±0,07	0,47±0,06
				SARWAR *et al.*			
κ	0,65±0,07	**0,75±0,04**	0,52±0,07	0,72±0,06	**0,90±0,05**	0,91±0,06	0,50±0,06
FNR	0,32±0,08	**0,23±0,06**	0,48±0,07	0,30±0,08	**0,10±0,07**	0,10±0,07	0,45±0,06
				RFD			
κ	0,51±0,05	0,66±0,05	0,44±0,05	0,55±0,08	0,74±0,09	0,87±0,06	0,50±0,05
FNR	0,53±0,04	0,34±0,05	0,58±0,03	0,50±0,06	0,33±0,10	0,16±0,08	0,52±0,04

CI: *carcinoma in situ*, LD: *light dysplastic*, MD: *moderate dysplastic*. NC: *normal columnar*, NI: *normal intermediate*, NS: *normal superficial*, SD: *severe dysplastic*. Em negrito estão os melhores resultados.

4.4.5 Resultados Qualitativos para Classificação de Imagens de Células Cervicais

A base de imagens proposta (CRIC), ao contrário da Herlev, não possui verdade-terrestre para segmentação. Dessa forma, não foi possível avaliar quantitativamente a eficiência do algoritmo de segmentação proposto. Entretanto, o resultado do algoritmo foi avaliado por especialistas da área de citologia (citopatologistas). Regiões erroneamente segmentadas como núcleos, mas que não eram núcleo, foram inseridas na classe de imagens normais. Com isso, no treinamento do algoritmo de classificação o algoritmo pode aprender padrões de classificação erradas. A Figura 19 mostra alguns núcleos obtidos pela segmentação de uma imagem de células cervicais.

Para mostrar a acurácia do classificador mostramos todos os resultados de uma imagem completa na Figura 20. As bordas amarelas correspondem a todos os resultados de segmentação obtidos, as setas verdes indicam regiões saudáveis e corretamente classificadas, as

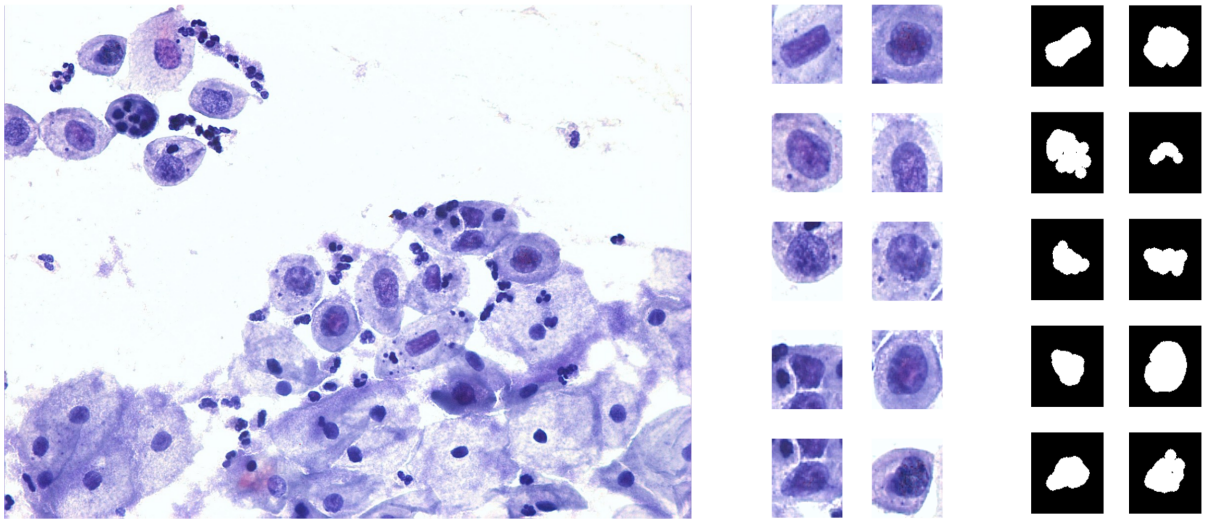


Figura 19 – Resultado utilizando o algoritmo de segmentação proposto na base de imagens CRIC: (a) imagem original, (b)imagens recortadas obtidas a partir de (a), e (c) as máscaras de segmentação correspondentes.

setas azuis mostram regiões anormais corretamente classificadas, e as setas vermelhas mostram regiões normais incorretamente classificadas. É válido mencionar que nenhuma região anormal foi incorretamente classificada, isso é refletido nas taxas de FNR obtidas pela metodologia proposta, FNR = 0,02 e FNR = 0,14 para as bases Herlev e CRIC, respectivamente.

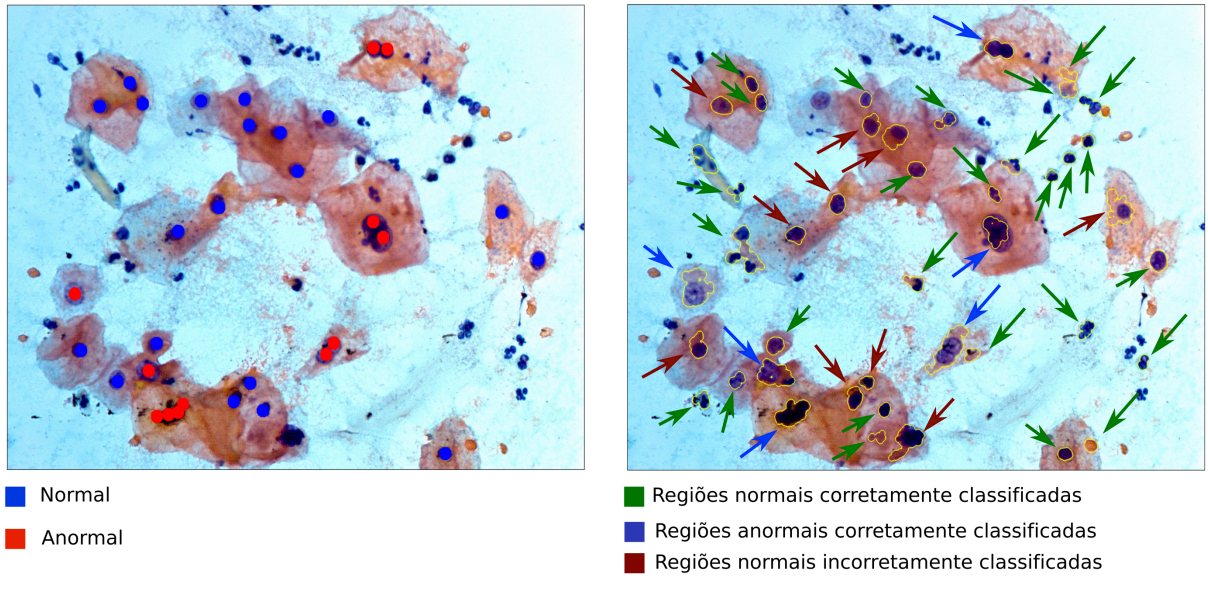


Figura 20 – Resultado para os algoritmos de segmentação e classificação aplicados a uma imagem da base de dados CRIC: (a) imagem original com a classe de cada núcleo, e (b) classificação de cada região segmentada. As bordas amarelas correspondem ao resultado de segmentação.

4.5 Considerações Finais

A descrição de células é uma tarefa importante na automatização de processos citológicos. Para se obter um sistema viável em um ambiente real, é necessário que o mesmo seja: 1) robusto à sobreposição de células; e 2) baixo custo computacional, visto que uma grande quantidade de células são analisadas em apenas uma lâmina do exame.

O descritor proposto possui algumas vantagens em relação aos demais da literatura: 1) não necessita da segmentação do citoplasma para o cálculo do histograma radial; 2) baixo custo computacional em relação a outros métodos; 3) viável em imagens com altos níveis de sobreposição de células. Por outro lado podemos destacar os seguintes pontos negativos: 1) seu desempenho é dependente da segmentação dos núcleos; e 2) requer estimação de dois parâmetros.

O próximo capítulo apresenta o `pyCBIR`, uma ferramenta para recuperação de imagens baseado em conteúdo. O `pyCBIR` possui uma interface gráfica e algoritmos clássicos e do estado da arte para extração de atributos em células.

5 PYCBIR: UMA FERRAMENTA DE RECUPERAÇÃO DE IMAGENS EM PYTHON

Nesta tese apresentamos uma ferramenta de Recuperação de Imagens Baseada em Conteúdo (*Content-Based Image Retrieval* - CBIR) chamada `pyCBIR`. Tal ferramenta foi desenvolvida na linguagem python e possui propósito geral, podendo ser utilizada em diferentes domínios da ciência com os mais diversos tipos de imagens, desde as citológicas às imagens de raio-x. Aqui mostramos as principais funcionalidades do `pyCBIR`, bem como seu uso aplicado ao problema do reconhecimento de células cervicais.

5.1 Contextualização

O termo CBIR foi introduzido em 1992 por Kato (KATO, 1992; HIRATA; KATO, 1992), e tem sido associado com sistemas que disponibilizam recuperação de imagens através de características visuais. Um dos maiores desafios para reconhecimento de imagens consiste em desempenhar tarefas que são de fácil realização por seres humanos, mas difíceis de descrever formalmente (GOODFELLOW *et al.*, 2016). Essa situação acontece frequentemente entre cientistas de vários domínios do conhecimento (DONATELLI *et al.*, 2015), que são treinados visualmente para identificar padrões de seus dados experimentais, apesar de muitas vezes serem incapazes de descrever matematicamente as primitivas que constroem esses padrões.

Vários esforços foram feitos com o objetivo de otimizar tarefas de busca de imagens em sistemas explorando algoritmos de visão computacional e aprendizado de máquina para representar imagens (YU *et al.*, 2017; TZELEPI; TEFAS, 2018; KHATAMI *et al.*, 2018). Trabalhos com recuperação de formas são encontrados em aplicação de catalogação de folhas (SOUZA *et al.*, 2016) e análise de fotografias de pílulas (CARNEIRO *et al.*, 2017).

A partir de uma imagem de consulta, essa abordagem permite encontrar outras amostras por similaridade. Algumas ferramentas gratuitas para CBIR são utilizadas em serviços de *e-commerce* (YU *et al.*, 2016; SHAMOI *et al.*, 2015), entretanto seus códigos-fonte permanecem fechado, e poucas ferramentas foram desenvolvidos para uso em imagens científicas. Nesse contexto, propomos o `pyCBIR` a fim de disponibilizar uma ferramenta para uso prático na comunidade científica.

Algoritmos baseados em dados que aprendem por experiência acumulada, como os implementados no `pyCBIR`, podem prover soluções para o ranqueamento de conjuntos de imagens levando em consideração: (a) as dificuldades de se obter conhecimento específico necessário para

modelagem, (b) limitação de aprendizado das especialidades de todos os domínios da ciência.

Nessa seção discutimos como o `pyCBIR` utiliza vetores de atributos para prover aprendizado automático de assinaturas que representem cada imagem. Uma etapa fundamental para a organização de bases de imagens é a construção de modelos para diferentes problemas científicos em conjunto com algoritmos para melhorar o desempenho da busca.

A Figura 21 mostra a interface gráfica do `pyCBIR` e um resultado de ranqueamento para a base de imagens CRIC utilizando a CNN LeNet para o cálculo das assinaturas. A primeira coluna no resultado corresponde às imagens de consulta, e as outras colunas são os resultados ranqueados para cada consulta. Casos em que a base de imagens esteja identificada, as bordas verdes correspondem a imagens corretamente recuperadas, e as vermelhas a imagens incorretamente retornadas.

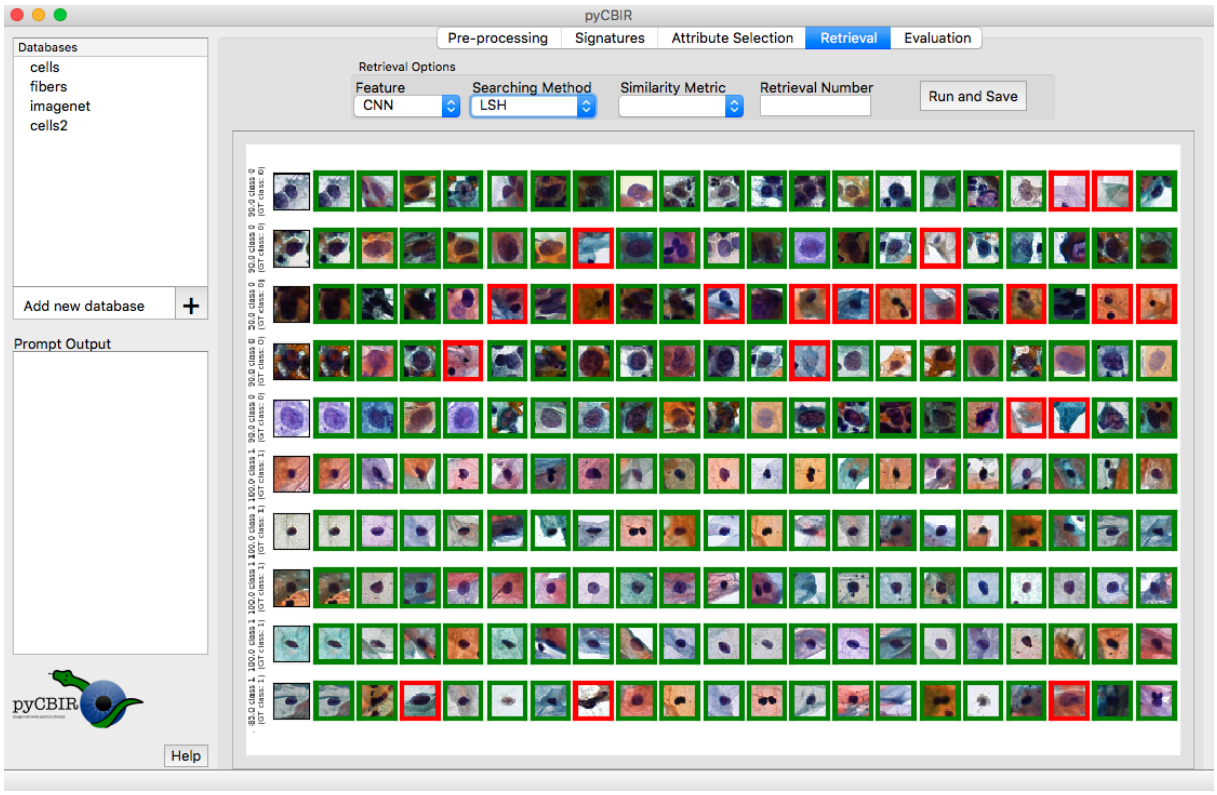


Figura 21 – Interface do `pyCBIR`: módulos implementados (cima), base de dados cadastradas (esquerda), e resultados de ranqueamento (centro) com imagens de consulta (primeira coluna) e resultados (demais colunas); bordas verdes indicam acerto, e vermelhas indicam erro.

O `pyCBIR` está organizado em cinco módulos: `Pre-processing`, `Signatures`, `Attribute Selection`, `Retrieval` e `Evaluation`. A seguir, descrevemos as principais funcionalidades e algoritmos que compõem cada um desses módulos.

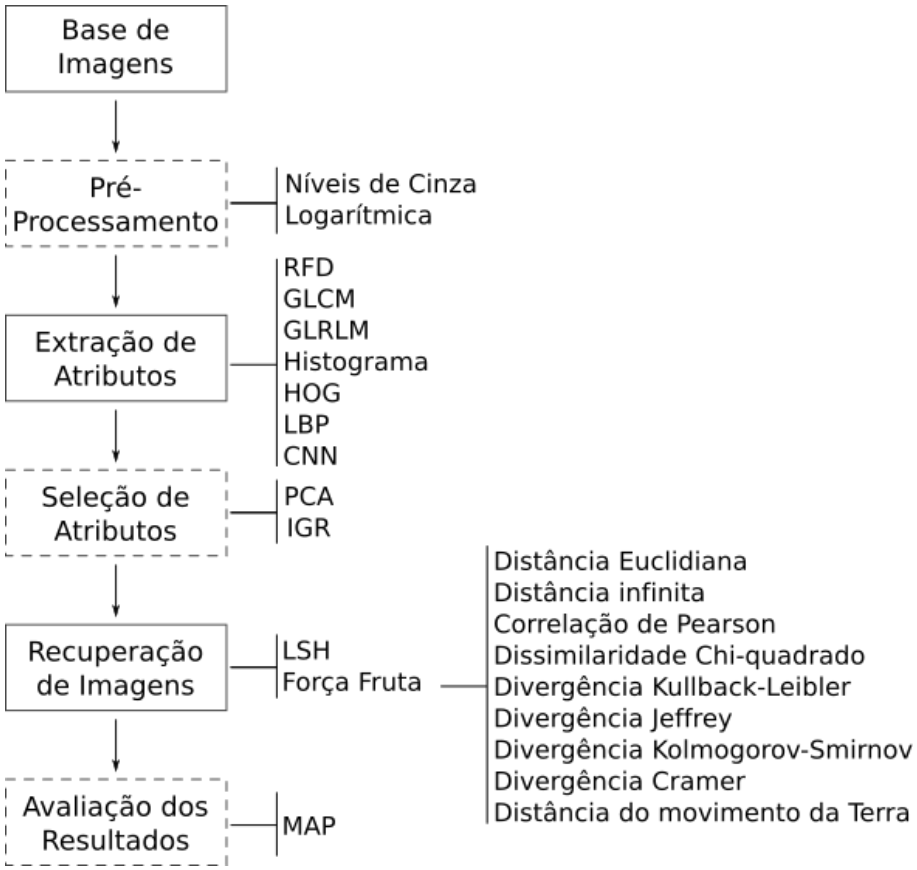


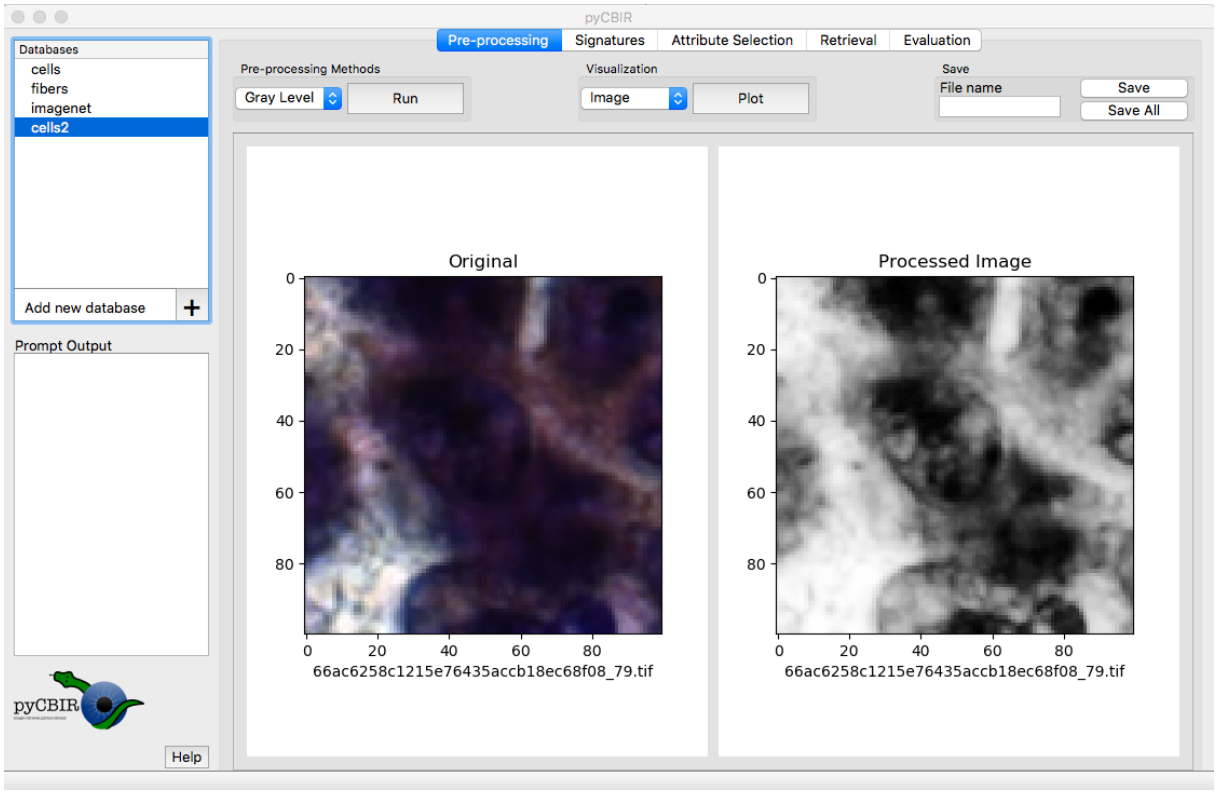
Figura 22 – Módulos do `pyCBIR` e seus métodos. As caixas pontilhadas representam etapas não obrigatórias para o processo de recuperação.

5.1.1 Pré-processamentos e Cálculo das Assinaturas

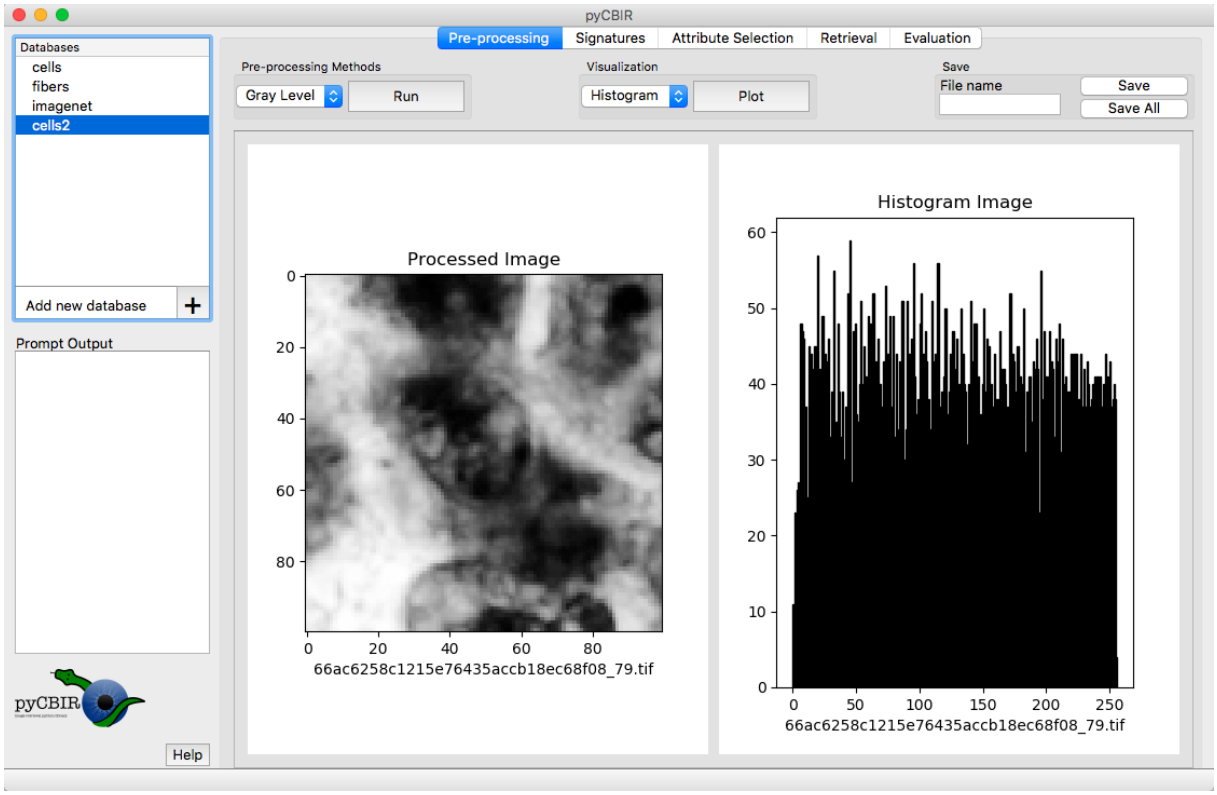
O pré-processamento pode ser feito utilizando o módulo *Pre-processing* (Figura 23). Esse módulo possui por padrão dois métodos de pré-processamento: transformação em níveis de cinza (*Gray Level*) e transformação logarítmica (*Log Transformation*). Esses métodos foram escolhidos por serem as principais técnicas utilizadas nos experimentos realizados. Além disso, é possível adicionar outros métodos que apresentem melhor resultado para determinado campo de pesquisa.

Existem ainda as opções para visualização do resultado como imagem ou histograma além do armazenamento dos resultados obtidos. O `pyCBIR` utiliza processamento paralelo para otimizar a tarefa de pré-processamento e cálculo de assinaturas.

No módulo `Signatures` estão implementos alguns descritores da literatura. Estão disponíveis no `pyCBIR`, além do RFD, os seguintes descritores: GLCM (HARALICK *et al.*, 1973), HOG (DALAL; TRIGGS, 2005), Atributos de Histograma, LBP (OJALA *et al.*, 1996) e duas CNNs (LeNet (LECUN *et al.*, 1998) e Inception-Resnet-v2 (SZEGEDY *et al.*, 2016a)). Esses métodos foram incluídos por serem largamente utilizado para reconhecimento de padrões



(a) Pré-processamento utilizando transformação em níveis de cinza.



(b) Visualização do histograma da imagem após o pré-processamento.

Figura 23 – O módulo *Pre-processing* do pyCBIR com suas respectivas formas de visualização dos dados.

em vários domínios da ciência. Quando o cálculo das assinaturas é realizado por CNNs é possível utilizar processamento paralelo com placas gráficas (*Graphic Process Unit* - GPU), reduzindo o tempo necessário para o processamento e ranqueamento de imagens de uma base. Nesse módulo é possível ainda adicionar novos descritores que melhor se adequem ao problema tratado e armazenar os vetores de atributos (assinaturas) em disco.

5.1.2 Seleção de Atributos e Recuperação de Imagens

Em alguns casos se torna necessário uma seleção ou redução de atributos. Os principais motivos são: 1) em determinados tipos de imagens alguns atributos possuem pouca ou nenhuma relevância, com isso a exclusão desse atributos não acarretará na perda de desempenho do algoritmo; 2) o tamanho do vetor de atributos é diretamente proporcional ao tempo necessário para o processo de recuperação das imagens, para se obter uma recuperação em aplicações de tempo real, por exemplo, é necessária uma representação mais compacta das assinaturas.

O módulo `Attribute Selection` possui dois métodos para seleção/redução de atributos: PCA (PEARSON, 1901) e IGR (QUINLAN, 1993). É possível ainda visualizar e armazenar as assinaturas obtidas nessa etapa. O módulo `Attribute Selection` possui interface similar ao *Pre-processing*.

No módulo `Retrieval` (Figura 21) é possível executar a busca por similaridade utilizando assinaturas previamente calculadas. O `pyCBIR` possui, por padrão, dois métodos de busca por similaridade: o *Locality Sensitive Hashing Forest* (LSH) (BAWA *et al.*, 2005) e força bruta. Caso a opção força bruta seja escolhida é possível utilizar: distância Euclidiana, distância infinita, correlação de Pearson, dissimilaridade Chi-quadrado, divergência Kullback-Leibler, divergência Jeffrey, divergência Kolmogorov-Smirnov, divergência Cramer, e distância do movimento da Terra (JONES *et al.*, 2001). Tais métodos para o cálculo de similaridade foram escolhidos por serem os mais presentes em trabalhos da literatura. O campo *retrieval number* informa ao sistema quantas imagens deverão ser retornadas na busca para cada imagens de consulta. Como resultado, o `pyCBIR` retorna uma imagem onde a primeira coluna corresponde às imagens de consulta, as demais representam o resultado ranqueado. Quando a base de imagens possuir verdade-terrestre para as classes da base de imagens, é possível sugerir classes para a imagem de consulta.

Reportamos nossos resultados utilizando o método de busca por força bruta e a distância cosseno como métrica de similaridade. Verificamos que para as bases de imagens

de imagens de células disponíveis a escolha do método de busca e da distância não trouxeram mudanças significantes nos resultados.

Com o objetivo de sugerir categorias (classes) a partir de características prévias, denotamos Y como as classes de uma base de imagens X que é representa por assinaturas. Essa base é composta por m imagens x_i, onde $1 \leq i \leq n$. As buscas ocorrem em um espaço h-dimensional, onde h é a dimensionalidade do extrator de atributos.

O módulo `Retrieval` do `pyCBIR` reconhece amostras similares através de uma função de similaridade S, dessa forma $S(x_i, x_q)$ retorna itens relevantes assim como as respectivas medidas de similaridade (distância). Em outras palavras, o mecanismo retorna as k (campo *Retrieval Number* do módulo) imagens mais similares, seu respectivo y_j para uma imagem de referência x_q, e $S(x_i, x_q)$, que para a distância cosseno, por exemplo, é definido da seguinte forma:

$$S(x_i, x_q) = \frac{x_i \cdot x_q}{||x_i|| \times ||x_q||}, \tag{5.1}$$

onde $\cdot$ é o produto interno e $\times$ é o produto.

5.1.3 Métricas de Avaliação dos Resultados

O `pyCBIR` possui como métrica de avaliação de desempenho o *Mean Average Precision* (MAP) (WANG *et al.*, 2015), que avalia a qualidade da recuperação de imagens. Para calcular o MAP, a precisão média $AP(Q)$ é calculada para cada imagem de consulta Q:

$$AP(Q) = \frac{\sum_{n=1}^{M}(P(n) \times f(n))}{N}, \tag{5.2}$$

e,

$$P(n) = \frac{VP}{VP + FN}, \tag{5.3}$$

em que o símbolo $\times$ representa o produto escalar, $P(n)$ é a precisão até a posição n do ranqueamento, $f(n)$ é igual a 1 se a imagem n do ranqueamento pertence à mesma classe da imagem de referência, e 0 caso contrário. M é o número de imagens no ranqueamento e N é o número de imagens da mesma classe obtidas no ranqueamento. O MAP é obtido pela média da AP em todas as imagens do ranqueamento. Quanto mais próximo de 1 o valor do MAP melhor o desempenho.

É possível calcular o MAP para a base completa bem como para cada classe, dessa forma é possível observar o desempenho do sistema em diferentes classes.

5.2 Experimentos de Recuperação de Imagens Baseada em Conteúdo

Esta seção apresenta os resultados para os experimentos CBIR utilizando a ferramenta `pyCBIR`. Os experimentos foram realizados com o objetivo de explorar a capacidade de generalização do RFD. Para isso, utilizamos as bases de imagens Herlev e CRIC. Os vetores de atributos foram os mesmos dos experimentos de classificação. Os resultados foram reportados utilizando força bruta como método de busca e a distância cosseno como métrica de similaridade.

A métrica utilizada na avaliação dos resultados foi o MAP (WANG *et al.*, 2015), e o cálculo do MAP considerou: 1) toda a base de imagens (MAP), 2) somente imagens de células normais (MAP_n), e 3) somente imagens de células anormais (MAP_{an}).

5.2.1 Resultados para Recuperação de Imagens de Células Cervicais

Nos testes CBIR, utilizamos os mesmos parâmetros em relação ao RFD nos experimentos de classificação: $n = 0,7$ e $d = 0,5$. A Tabela 7 apresenta os resultados obtidos utilizando as bases de imagens Herlev e CRIC com o descritor proposto. Os melhores resultados foram obtidos pelo RFD. Para a base de imagens Herlev, tal descritor alcançou um MAP = 0,84. É válido observar que o MAP_n = 0,81 foi inferior ao MAP_{an} = 0,88 para a base Herlev exibindo uma maior capacidade de recuperação de imagens anormais. Entretanto para a base CRIC, apesar do MAP = 0,82 ser próximo ao da Herlev o descritor apresentou melhores resultados para o MAP_n = 0,86 em relação ao MAP_{an} = 0,77.

Analisando os resultados da base Herlev, observamos que o principal motivo da diferença de resultados entre as classes normais e anormais é a similaridade entre dois grupos: *columnar epithelial* (normal) e *carcinoma* (anormal). A Figura 24 mostra a média dos histogramas radias para esses dois grupos de imagens. É possível observar a similaridade entre as duas curvas, com isso ao utilizar uma amostra do grupo *carcinoma* como imagem de consulta, parte dos resultados ranqueados retornados são do grupo *columnar epithelial*.

Observamos um menor desempenho da base de imagens CRIC em relação à base Herlev, o mesmo ocorre nos experimentos de classificação. Acreditamos que tal comportamento ocorra em virtude do algoritmo de segmentação utilizado, visto que na base Herlev é utilizada a verdade-terrestre. Erros na segmentação do núcleo podem levar à perda de informações importantes, ou seja, a GLRLM não será capaz de extrair os atributos da distribuição de cromatina, da mesma forma as variações de intensidade ao redor do núcleo não serão capturadas pelo RH.

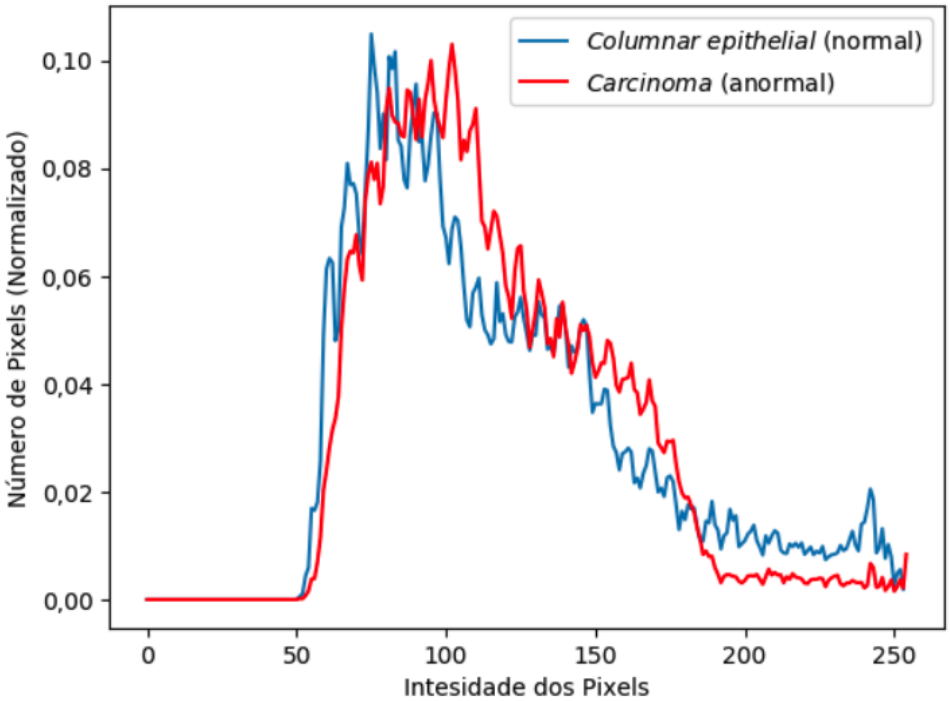


Figura 24 – Média dos histogramas radiais para duas classes da base de imagens Herlev: *columnar epithelial* (normal) e *carcinoma* (anormal).

De forma análoga aos experimentos de classificação, observamos que a junção dos descritores GLRLM e RH produziram melhores resultados que seu uso separado. Isso pode ser observado no resultado para a base CRIC onde os descritores GLRLM e RH obtiveram MAP = 0,71 e MAP = 0,70, respectivamente, enquanto que a concatenação alcançou um MAP = 0,82. O mesmo comportamento é observado para a base Herlev, mostrando a relevância das informações extraídas por ambos descritores.

A Figura 25 mostra o resultado CBIR utilizando o RFD e as amostras da base CRIC, onde as bordas verdes e vermelhas representam as imagens corretamente e incorretamente retornadas, respectivamente. Foram utilizadas 16 imagens de consulta, as 8 primeiras correspondem a células anormais e as demais às saudáveis. Esse resultado destaca que existem mais retornos incorretos para imagens de células anormais, refletindo em um menor valor de MAP obtido para essa classe na base CRIC. É possível observar que erros de segmentação afetaram a recuperação de imagens anormais, por exemplo, na segunda linha de resultados da Figura 25 ocorreram dois ranqueamentos incorretos (bordas vermelhas), observa-se também que as imagens incorretas são recortes sem núcleo, o mesmo pode ser observado para outros exemplos da classe de imagens anormais.

Comparamos os resultados obtidos pelo RFD com outros descritores da literatura sendo que os experimentos CBIR foram conduzidos utilizando os mesmos vetores de atributos

Tabela 7 – Resultados da medida MAP para os experimentos CBIR utilizando o descritor proposto com as bases Herlev e CRIC.

	MAP	MAP_n	MAP_{an}
	Herlev		
GLRLM	0,76±0,16	0,68±0,20	0,84±0,12
RH	0,79±0,16	0,76±0,18	0,81±0,13
RFD	**0,84±0,14**	**0,81±0,18**	**0,88±0,10**
	CRIC		
GLRLM	0,71±0,18	0,79±0,13	0,63±0,23
RH	0,70±0,16	0,80±0,12	0,61±0,20
RFD	**0,82±0,15**	**0,86±0,12**	**0,77±0,17**

*Em negrito estão os melhores resultados.

aplicados nos experimentos de classificação. Esses resultados são apresentados nas Tabelas 8 e 9 para as bases de imagens Herlev e CRIC, respectivamente. Alguns métodos não foram calculados para a base CRIC devido a necessidade da segmentação do citoplasma.

Analisando os resultados obtidos na base de imagens Herlev, podemos observar que o descritor RFD superou os demais vetores de atributos para a recuperação de imagens utilizando toda a base (MAP) e somente as imagens de células normais (MAP_n). Entretanto, em relação às imagens anormais, o método proposto por Plissiti e Nikou (PLISSITI; NIKOU, 2012) alcançou o melhor resultado para a base Herlev com MAP_{an} = 0,93. Apesar disso, o método em questão não conseguiu manter o mesmo desempenho em relação às células normais obtendo MAP = 0,69, que corresponde a uma acurácia abaixo de métodos como Marinakis *et al*. (MARINAKIS *et al.*, 2009), Sarwar *et al*. (SARWAR *et al.*, 2015) e o RFD.

Em relação à base de imagens CRIC, os resultados foram semelhantes aos obtidos na base Herlev, ou seja, o descritor RFD obteve os melhores resultados em relação ao MAP e ao MAP_n. Diferente da base Herlev, o vetor de atributos proposto por Plissiti e Nikou (PLISSITI; NIKOU, 2012) não alcançou taxa de acerto próxima às melhores. Entretanto, a CNN LeNet (LECUN *et al.*, 1998) obteve o melhor resultado para a classe de células anormais com MAP_{an} = 0,87. Além disso, o resultado para a base completa foi bem próximo do obtido pelo RFD com MAP = 0,82.

Ainda, os vetores de atributos propostos por Plissiti e Nikou (PLISSITI; NIKOU, 2012) e Lecun (LECUN *et al.*, 1998) não obtiveram o mesmo desempenho em ambas as bases. Isso mostra a robustez do RFD que, mesmo não obtendo as melhores métricas em todos os casos, apresentou melhores resultados quando as bases completas foram analisadas, ou seja, para o cálculo do MAP.

Figura 25 – Resultado gráfico para um experimento CBIR utilizando o descritor RFD. A primeira coluna são as imagens de consulta e as demais são os resultados ranqueados. Bordas verdes representam imagens corretamente retornadas e vermelhas representam as incorretamente retornadas.

5.3 Considerações Finais

O `pyCBIR` é uma ferramenta de propósito geral, podendo ser empregado nas mais diversas áreas da ciência. Por ser uma ferramenta de código aberto, é possível que usuários com conhecimentos de programação adicionem novas funcionalidades, tais como: métodos de pré-processamento (por exemplo, filtrar a imagem), extratores de atributos, seleção e redução de atributos, métodos de busca, visualização e formas de avaliação da recuperação de imagens.

Tabela 8 – Análise comparativa para os experimentos CBIR utilizando a base Herlev.

Métodos	Herlev		
	MAP	MAP_n	MAP_{an}
CHANKONG *et al.*	0,71±0,21	0,68±0,36	0,85±0,07
CHEN *et al.*	0,79±0,18	0,67±0,28	0,91±0,07
GENÇTAV *et al.*	0,77±0,21	0,69±0,36	0,85±0,07
MARIARPUTHAM; STEPHEN	0,72±0,18	0,62±0,30	0,82±0,06
MARINAKIS *et al.*	0,81±0,19	0,74±0,31	0,88±0,07
PLISSITI; NIKOU	0,81±0,17	0,69±0,25	**0,93±0,10**
SARWAR *et al.*	0,81±0,19	0,74±0,31	0,88±0,07
HARALICK *et al.*	0,57±0,08	0,34±0,13	0,79±0,03
NABIZADEH; KUBAT	0,63±0,14	0,45±0,18	0,81±0,10
DALAL; TRIGGS	0,60±0,06	0,40±0,09	0,79±0,02
OJALA *et al.*	0,55±0,06	0,23±0,08	0,86±0,03
SZEGEDY *et al.*	0,70±0,16	0,56±0,27	0,85±0,06
LECUN *et al.*	0,69±0,15	0,53±0,26	0,86±0,05
RFD	**0,84±0,14**	**0,81±0,18**	0,88±0,10

Tabela 9 – Análise comparativa para os experimentos CBIR utilizando a base CRIC.

Métodos	CRIC		
	MAP	MAP_n	MAP_{ab}
CHANKONG *et al.*	-	-	-
CHEN *et al.*	-	-	-
GENÇTAV *et al.*	-	-	-
MARIARPUTHAM; STEPHEN	-	-	-
MARINAKIS *et al.*	-	-	-
PLISSITI; NIKOU	0,70±0,20	0,73±0,19	0,66±0,20
SARWAR *et al.*	-	-	-
HARALICK *et al.*	0,58±0,14	0,64±0,18	0,51±0,10
NABIZADEH; KUBAT	0,57±0,13	0,64±0,11	0,50±0,15
DALAL; TRIGGS	0,61±0,08	0,79±0,06	0,44±0,10
OJALA *et al.*	0,56±0,08	0,52±0,07	0,60±0,09
SZEGEDY *et al.*	0,69±0,14	0,62±0,17	0,77±0,10
LECUN *et al.*	0,81±0,21	0,75±0,25	**0,87±0,18**
RFD	**0,82±0,15**	**0,86±0,12**	0,77±0,17

− : Esse símbolo nos campos da tabela indica os métodos que dependem da segmentação do citoplasma.
Em negrito estão os melhores resultados.

Além disso, o `pyCBIR` possui módulos que processam as imagens em paralelo, melhorando o desempenho da ferramenta.

Os resultados obtidos para recuperação de imagens se mostraram bastante promissores quando comparados aos métodos presentes no estado da arte como as redes neurais convolucionais. O RFD obteve as melhores taxas de acerto para a base completa MAP = 0,84 e MAP = 0,82 para as bases Herlev e CRIC, respectivamente. Entretanto, observou-se que o descritor RFD foi superado nos resultados para recuperação de imagens anormais.

6 CONCLUSÕES E TRABALHOS FUTUROS

Essa tese discutiu metodologias para descrição e categorização de células, com foco em classificação de células cervicais. Foram propostos um segmentador de núcleos baseado em clusterização e um descritor de atributos radiais (RFD). Em seguida, foram realizados experimentos de recuperação de imagens baseada em conteúdo através de uma nova ferramenta computacional, chamada `pyCBIR`, que permite recuperação de imagens; todos esse algoritmos e programas de computadores foram desenvolvidos e publicados como parte dos trabalhos dessa tese.

O método proposto para segmentação de núcleos foi aplicado na base de imagens CRIC para obtenção das máscaras de segmentação dos núcleos. Apesar do alto grau de sobreposição de células presente em exames de meio convencional do Papanicolau, mostramos que a segmentação somente do núcleo é viável. Ainda, o uso da borda do núcleo é capaz de prover informações que possibilitam a diferenciação entre células normais e anormais. Além do método de segmentação, apresentamos uma nova base de imagens composta por: campos do exame Papanicolau, imagens de células recortadas e o diagnóstico de cada célula. Acredita-se que essa base é de grande utilidade pública como referência para criação de trabalhos futuros, em alternativa às bases sintéticas, de meio líquido, ou que possuem somente recortes.

Salienta-se que existem algumas limitações em relação à base e ao método de segmentação proposto: 1) a avaliação quantitativa não é viável visto que o subconjunto de teste da base CRIC não possui verdade-terrestre para segmentação; 2) os parâmetros utilizados no método foram obtidos em um conjunto de 12 imagens com 270 células (subconjunto de treino da base CRIC), logo, esses valores possivelmente necessitam de ajustes em outras bases de imagens; 3) visualmente (Figura 20), observamos que o método necessita de aprimoramentos, com isso, provavelmente, melhores índices de classificação e recuperação serão alcançados, visto que os atributos serão extraídos de uma região mais próximo ao que realmente é o núcleo da célula.

O RFD foi desenvolvido para caracterizar células registradas em imagens digitais de microscopia, embora se tenha focado em células cervicais. Esse descritor tem a vantagem de não necessitar da segmentação do citoplasma, e utilizar segmentação do núcleo para obter informações radiais de textura ao redor do citoplasma. Com isso, o tempo de processamento, que é um componente crucial na análise de imagens de células cervicais, tem uma redução em relação aos demais da literatura que necessitam da segmentação do citoplasma. As principais conclusões do método proposto são: 1) a variação na distribuição de cromatina dentro do núcleo

indica anormalidade na célula; 2) células normais apresentam menor variação de intensidade dentro do núcleo; 3) em células anormais a transição entre núcleo e citoplasma é mais suave; 4) é possível classificar e recuperar células cervicais mesmo quando existe sobreposição de células.

As principais limitações do RFD são: 1) a necessidade do ajuste dos parâmetros n e d, que controlam quantos pontos da borda serão utilizados e o tamanho de cada reta, respectivamente; 2) a dependência de um algoritmo de segmentação de núcleos, onde a qualidade da segmentação possivelmente irá afetar as taxas de acerto.

Após a segmentação e cálculo do RFD, fizemos experimentos de classificação com as bases Herlev e CRIC. Os experimentos foram realizados com o classificador *Random Forest* e com a técnica de treinamento/teste *bootstrap* 0.632. O vetor de atributos RFD trouxe melhorias para a classificação em diferentes bases de imagens, em termos de índice Kappa e FNR quando comparados com outros 13 métodos da literatura, incluindo duas Redes Neurais Convolucionais.

Além dos experimentos de classificação, esta tese apresentou a ferramenta `pyCBIR`, a qual foi desenvolvida com o objetivo de realizar experimentos de recuperação de imagens baseada em conteúdo. Tal ferramenta é capaz de auxiliar cientistas que trabalham com os mais diversos tipos de imagens. A principal limitação da ferramenta está na dependência de um especialista do problema, ou seja, para cada problema apresentado à ferramenta é necessária a identificação da melhor forma de extrair os atributos. Apesar da versatilidade da versão corrente do `pyCBIR`, alguns módulos com novos métodos de visão computacional para segmentação de imagens, combinação de atributos e estimação de parâmetros ainda estão por ser desenvolvidos.

O `pyCBIR` utilizou os mesmos vetores de atributos e métodos dos experimentos de classificação. O cálculo de similaridade entre vetores de atributos foi feito pela distância cosseno. Os resultados foram obtidos em termos de MAP, o RFD alcançou os melhores resultados para o conjunto de células normais, e resultados próximos aos melhores métodos considerando o MAP para ambas as classes.

6.1 Trabalhos Futuros

Como trabalhos futuros, investigaremos métricas a serem calculadas com o histograma radial, tais como: entropia, assimetria, energia, e curtose, essas métricas podem permitir uma representação mais compacta dos atributos extraídos das imagens. Considerando que o tempo de processamento para classificação de uma célula é diretamente proporcional ao tamanho do vetor de atributos, uma representação mais compacta pode reduzir esse tempo. Também

serão feitos testes com a utilização do descritor RFD em multi-escala, visto a possibilidade de extrair informações importantes de textura em diferentes níveis de decomposição *wavelet*, por exemplo. Ainda, analisaremos o desempenho do RFD com diferentes algoritmos de segmentação de núcleos disponíveis na literatura.

Em relação ao `pyCBIR`, pretendemos adicionar novos módulos que tornarão a ferramenta mais genérica abrangendo outros problemas. Dentre os módulos a serem incluídos estão: 1) `Bag of Features` - com a adição desse módulo será possível trabalhar com descritores como *Scale Invariant Feature Transform* (SIFT) e *Speeded-Up Robust Features* (SURF), que dependem de métodos para concatenar os vários vetores de atributos gerados para cada imagem, o mesmo se aplica no caso de descritores multi-escala; 2) `Segmentação` - nesse módulo teremos algoritmos de segmentação como k-médias, CNNs e contorno ativo, que resultarão em regiões de interesse que serão aplicadas como base para o cálculo das assinaturas.

REFERÊNCIAS

ABIDIN, A. Z.; DENG, B.; DSOUZA, A. M.; NAGARAJAN, M. B.; COAN, P.; WISMüLLER, A. Deep transfer learning for characterizing chondrocyte patterns in phase contrast x-ray computed tomography images of the human patellar cartilage. **Computers in Biology and Medicine**, v. 95, p. 24 – 33, 2018.

AJDADI, F. R.; GILANDEH, Y. A.; MOLLAZADE, K.; HASANZADEH, R. P. Application of machine vision for classification of soil aggregate size. **Soil and Tillage Research**, v. 162, p. 8 – 17, 2016.

ARAUJO, S. **Citologia Cervicovaginal - Passo A Passo**. 3. ed. Brasil: DI LIVROS, 2012. ISBN 9788580530261.

BAMFORD, P.; LOVELL, B. Unsupervised cell nucleus segmentation with active contours. **Signal Processing**, v. 71, n. 2, p. 203 – 213, 1998. ISSN 0165-1684.

BARALDI, A.; PARMIGGIANI, F. An investigation of the textural characteristics associated with gray level cooccurrence matrix statistical parameters. **IEEE Transactions on Geoscience and Remote Sensing**, v. 33, n. 2, p. 293–304, Mar 1995.

BAWA, M.; CONDIE, T.; GANESAN, P. LSH forest: Self-tuning indexes for similarity search. In: **Fourteenth International World Wide Web Conference**. Chiba: [s.n.], 2005.

BEJNORDI, B. E.; MOSHAVEGH, R.; SUJATHAN, K.; MALM, P.; BENGTSSON, E.; MEHNERT, A. Novel chromatin texture features for the classification of pap smears. In: **Proc. SPIE**. [S.l.: s.n.], 2013. v. 8676, p. 8.

BERGMEIR, C.; SILVENTE, M. G.; BENíTEZ, J. M. Segmentation of cervical cell nuclei in high-resolution microscopic images: A new algorithm and a web-based software framework. **Computer Methods and Programs in Biomedicine**, v. 107, n. 3, p. 497 – 512, 2012. ISSN 0169-2607.

BRADLEY, A. P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. **Pattern Recognition**, Elsevier Science Inc., New York, NY, USA, v. 30, n. 7, p. 1145–1159, Jul 1997.

BREIMAN, L. Random forests. **Machine Learning**, Kluwer Academic Publishers, Hingham, MA, USA, v. 45, n. 1, p. 5–32, out. 2001. ISSN 0885-6125.

CARNEIRO, A. C.; LOPES, J. G. F.; ARAÚJO, F. H. D.; SILVA, R. R. V.; PASSARINHO, C. J. P.; NETO, J. F. S. R.; MEDEIROS, F. N. S. Análise de fotografias de pílulas por redes neurais convolucionais. In: **VIII Simpósio de Instrumentação e Imagens Médicas e VII Simpósio de Processamento de Sinais**. [S.l.: s.n.], 2017. p. 14.

CHANKONG, T.; THEERA-UMPON, N.; AUEPHANWIRIYAKUL, S. Automatic cervical cell segmentation and classification in pap smears. **Computer Methods and Programs in Biomedicine**, Elsevier North-Holland, Inc., New York, NY, USA, v. 113, n. 2, p. 539–556, fev. 2014. ISSN 0169-2607.

CHEN, Y.-F.; HUANG, P.-C.; LIN, K.-C.; LIN, H.-H.; WANG, L.-E.; CHENG, C.-C.; CHEN, T.-P.; CHAN, Y.-K.; CHIANG, J. Semi-automatic segmentation and classification of pap smear

cells. **IEEE Journal of Biomedical and Health Informatics**, v. 18, n. 1, p. 94–108, Jan 2014. ISSN 2168-2194.

CORTES, C.; VAPNIK, V. Support-vector networks. In: **Machine Learning**. [S.l.: s.n.], 1995. p. 273–297.

COSTA, L. d. F. D.; CESAR JR., R. M. **Shape Analysis and Classification: Theory and Practice**. 1st. ed. Boca Raton, FL, USA: CRC Press, Inc., 2000. ISBN 0849334934.

DALAL, N.; TRIGGS, B. Histograms of oriented gradients for human detection. In: **Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2005. p. 886–893.

DENG, J.; DONG, W.; SOCHER, R.; LI, L. jia; LI, K.; FEI-FEI, L. Imagenet: A large-scale hierarchical image database. In: **In CVPR**. [S.l.: s.n.], 2009.

DIJK, L. V. van; BROUWER, C. L.; SCHAAF, A. van der; BURGERHOF, J. G.; BEUKINGA, R. J.; LANGENDIJK, J. A.; SIJTSEMA, N. M.; STEENBAKKERS, R. J. CT image biomarkers to improve patient-specific prediction of radiation-induced xerostomia and sticky saliva. **Radiotherapy and Oncology**, p. 185–191, 2016. ISSN 0167-8140.

DONATELLI, J.; HARANCZYK, M.; HEXEMER, A.; KRISHNAN, H.; LI, X.; LIN, L.; MAIA, F.; MARCHESINI, S.; PARKINSON, D.; PERCIANO, T.; SHAPIRO, D.; USHIZIMA, D.; YANG, C.; SETHIAN, J. Camera: The center for advanced mathematics for energy research applications. **Synchrotron Radiation News**, v. 28, n. 2, p. 4–9, 2015.

DUNN, J. C. A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. **Journal of Cybernetics**, v. 3, n. 3, p. 32–57, 1973.

DÉNIZ, O.; BUENO, G.; SALIDO, J.; TORRE, F. Face recognition using histograms of oriented gradients. **Pattern Recognition Letters**, v. 32, n. 12, p. 1598 – 1603, 2011. ISSN 0167-8655.

EFRON, B. Estimating the error rate of a prediction rule: improvement on cross-validation. **Journal of the American Statistical Association**, American Statistical Association, v. 78, n. 382, p. 316–331, 1983. ISSN 01621459.

EINSTEIN, A. J.; WU, H.; GIL, J. Self-affinity and lacunarity of chromatin texture in benign and malignant breast epithelial cell nuclei. **Physical Review Letters**, v. 80, n. 2, p. 397 – 400, 1998.

FREEMAN, W. T.; FREEMAN, W. T.; ROTH, M.; ROTH, M. Orientation histograms for hand gesture recognition. In: **In International Workshop on Automatic Face and Gesture Recognition**. [S.l.: s.n.], 1994. p. 296–301.

GALLOWAY, M. M. Texture analysis using gray level run lengths. **Computer Graphics and Image Processing**, v. 4, n. 2, p. 172 – 179, 1975. ISSN 0146-664X.

GENÇTAV, A.; AKSOY, S.; ONDER, S. Unsupervised segmentation and classification of cervical cell images. **Pattern Recognition**, v. 45, n. 12, p. 4151 – 4168, 2012. ISSN 0031-3203.

GONZALEZ, R. C.; WOODS, R. E. **Processamento de imagens digitais**. 2. ed. São Paulo: Addison Wesley, 2000.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. **Deep Learning**. MIT Press, 2016. Disponível em: <http://www.deeplearningbook.org>.

GUVEN, M.; CENGIZLER, C. Data cluster analysis-based classification of overlapping nuclei in pap mear samples. **Biomedical Engineering OnLine**, v. 13, 12 2014.

HARALICK, R.; SHANMUGAM, K.; DINSTEIN, I. Textural features for image classification. **IEEE Transactions on Systems, Man and Cybernetics**, SMC-3, n. 6, p. 610–621, Nov 1973. ISSN 0018-9472.

HARANDI, N.; SADRI, S.; MOGHADDAM, N.; AMIRFATTAHI, R. An automated method for segmentation of epithelial cervical cells in images of thinprep. **Journal of Medical Systems**, Springer US, v. 34, n. 6, p. 1043–1058, 2010. ISSN 0148-5598.

HAYKIN, S. **Neural Networks: A Comprehensive Foundation**. 2nd. ed. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1998. ISBN 0132733501.

HIRATA, K.; KATO, T. Query by visual example - content based image retrieval. In: **Proceedings of the 3rd International Conference on Extending Database Technology: Advances in Database Technology**. London, UK: Springer-Verlag, 1992. (EDBT '92), p. 56–71. ISBN 3-540-55270-7.

HUNTER, R. S. Photoelectric color-difference meter. **Journal of the Optical Society of America**, OSA, v. 38, n. 7, p. 651–651, Jul 1948.

Instituto Nacional de Câncer. **Tipos de Câncer - Colo do Útero**. 2018. <http://www2.inca.gov.br/wps/wcm/connect/tiposdecancer/site/home/colo_utero>. Online; acessado em 07 Julho 2018.

IRSHAD, H.; VEILLARD, A.; ROUX, L.; RACOCEANU, D. Methods for nuclei detection, segmentation, and classification in digital histopathology: A review-current status and future potential. **IEEE Reviews in Biomedical Engineering**, v. 7, p. 97–114, 2014. ISSN 1937-3333.

JANTZEN, J.; NORUP, J.; DOUNIAS, G.; BJERREGAARD, B. Pap-smear benchmark data for pattern classification technical University of Denmark. **Nature inspired Smart Information Systems**, p. 1–9, 2005.

JEGADEESHWARAN, R.; SUGUMARAN, V. Fault diagnosis of automobile hydraulic brake system using statistical features and support vector machines. **Mechanical Systems and Signal Processing**, v. 52–53, p. 436 – 446, 2015. ISSN 0888-3270.

JONES, E.; OLIPHANT, T.; PETERSON, P. *et al.* **SciPy: Open source scientific tools for Python**. 2001. <http://www.scipy.org/>. Acessado em 05 de abril de 2018.

JUNG, C.; KIM, C. Segmenting clustered nuclei using h-minima transform-based marker extraction and contour parameterization. **IEEE Transactions on Biomedical Engineering**, v. 57, n. 10, p. 2600–2604, Oct 2010. ISSN 0018-9294.

JUNG, C.; KIM, C.; CHAE, S. W.; OH, S. Unsupervised segmentation of overlapped nuclei using bayesian classification. **IEEE Transactions on Biomedical Engineering**, v. 57, n. 12, p. 2825–2832, Dec 2010. ISSN 0018-9294.

KALE, A.; AKSOY, S. Segmentation of cervical cell images. In: **20th International Conference on Pattern Recognition (ICPR)**. [S.l.: s.n.], 2010. p. 2399–2402. ISSN 1051-4651.

KANADAM, K. P.; CHEREDDY, S. R. Mammogram classification using sparse-roi: A novel representation to arbitrary shaped masses. **Expert Systems with Applications**, v. 57, p. 204 – 213, 2016. ISSN 0957-4174.

KATO, T. Database architecture for content-based image retrieval. In: **Proc. of SPIE Image Storage and Retrieval Systems**. San Jose, CA, USA: [s.n.], 1992. v. 1662, p. 112–123.

KHATAMI, A.; BABAIE, M.; TIZHOOSH, H.; KHOSRAVI, A.; NGUYEN, T.; NAHAVANDI, S. A sequential search-space shrinking using cnn transfer learning and a radon projection pool for medical image retrieval. **Expert Systems with Applications**, v. 100, p. 224 – 233, 2018. ISSN 0957-4174.

LANDIS, J.; KOCH, G. The measurement of observer agreement for categorical data. **Biometrics**, v. 33, n. 1, p. 159–174, 1977.

LECUN, Y.; BOTTOU, L.; BENGIO, Y.; HAFFNER, P. Gradient-based learning applied to document recognition. In: **Proceedings of the IEEE**. [S.l.: s.n.], 1998. p. 2278–2324.

LI, K.; LU, Z.; LIU, W.; YIN, J. Cytoplasm and nucleus segmentation in cervical smear images using radiating gvf snake. **Pattern Recognition**, v. 45, n. 4, p. 1255 – 1264, 2012. ISSN 0031-3203.

LITJENS, G.; KOOI, T.; BEJNORDI, B. E.; SETIO, A. A. A.; CIOMPI, F.; GHAFOORIAN, M.; LAAK, J. A. van der; GINNEKEN, B. van; SáNCHEZ, C. I. A survey on deep learning in medical image analysis. **Medical Image Analysis**, v. 42, p. 60 – 88, 2017.

LU, Z.; CARNEIRO, G.; BRADLEY, A. Automated nucleus and cytoplasm segmentation of overlapping cervical cells. In: **Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013**. [S.l.]: Springer Berlin Heidelberg, 2013, (Lecture Notes in Computer Science, v. 8149). p. 452–460. ISBN 978-3-642-40810-6.

LU, Z.; CARNEIRO, G.; BRADLEY, A. An improved joint optimization of multiple level set functions for the segmentation of overlapping cervical cells. **IEEE Transactions on Image Processing**, v. 24, n. 4, p. 1261–1272, April 2015. ISSN 1057-7149.

LU, Z.; CARNEIRO, G.; BRADLEY, A.; USHIZIMA, D.; NOSRATI, M. S.; BIANCHI, A.; CARNEIRO, C.; HAMARNEH, G. Evaluation of three algorithms for the segmentation of overlapping cervical cells. **IEEE Journal of Biomedical and Health Informatics**, p. 1–11, 2016. ISSN 2168-2194.

MARIARPUTHAM, M. E. J.; STEPHEN, A. Nominated texture based cervical cancer classification. **Comp. Math. Methods in Medicine**, v. 2015, p. 1 – 10, 2015.

MARINAKIS, Y.; MARINAKI, M.; DOUNIAS, G.; JANTZEN, J.; BJERREGAARD, B. Intelligent and nature inspired optimization methods in medicine: the pap smear cell classification problem. **Expert Systems**, Blackwell Publishing Ltd, v. 26, n. 5, p. 433–457, 2009. ISSN 1468-0394.

NABIZADEH, N.; KUBAT, M. Brain tumors detection and segmentation in {MR} images: Gabor wavelet vs. statistical features. **Computers & Electrical Engineering**, v. 45, p. 286 – 301, 2015. ISSN 0045-7906.

NAGARAJ, Y.; ASHA, C.; A., H. S. T.; NARASIMHADHAN, A. V. Carotid wall segmentation in longitudinal ultrasound images using structured random forest. **Computers & Electrical Engineering**, 2018. ISSN 0045-7906.

OJALA, T.; PIETIKAINEN, M. Unsupervised texture segmentation using feature distributions. In: **International Conference on Image Analysis and Processing**. London: Springer-Verlag, 1997. p. 311–318. ISBN 3-540-63507-6.

OJALA, T.; PIETIKAINEN, M.; HARWOOD, D. A comparative study of texture measures with classification based on featured distributions. **Pattern Recognition**, v. 29, n. 1, p. 51 – 59, 1996. ISSN 0031-3203.

PEARSON, K. On lines and planes of closest fit to systems of points in space. **Philosophical Magazine**, v. 2, n. 6, p. 559–572, 1901.

PLISSITI, M.; NIKOU, C.; CHARCHANTI, A. Automated detection of cell nuclei in pap smear images using morphological reconstruction and clustering. **IEEE Transactions on Information Technology in Biomedicine**, v. 15, n. 2, p. 233–241, March 2011. ISSN 1089-7771.

PLISSITI, M. E.; NIKOU, C. Cervical cell classification based exclusively on nucleus features. In: CAMPILHO, A.; KAMEL, M. (Ed.). **Image Analysis and Recognition**. [S.l.]: Springer, 2012, (Lecture Notes in Computer Science, v. 7325). p. 483–490. ISBN 978-3-642-31297-7.

PLISSITI, M. E.; NIKOU, C.; CHARCHANTI, A. Combining shape, texture and intensity features for cell nuclei extraction in pap smear images. **Pattern Recognition Letters**, v. 32, n. 6, p. 838 – 853, 2011. ISSN 0167-8655.

QUINLAN, J. R. **C4.5: Programs for Machine Learning**. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1993. ISBN 1-55860-238-0.

RUSSAKOVSKY, O.; DENG, J.; SU, H.; KRAUSE, J.; SATHEESH, S.; MA, S.; HUANG, Z.; KARPATHY, A.; KHOSLA, A.; BERNSTEIN, M.; BERG, A. C.; FEI-FEI, L. Imagenet large scale visual recognition challenge. **International Journal of Computer Vision**, v. 115, n. 3, p. 211–252, 2015.

SA, J.; BACKES, A. A color texture analysis method based on a gravitational approach for classification of the pap-smear database. In: **2014 IEEE International Conference on Image Processing (ICIP)**. [S.l.: s.n.], 2014. p. 2280–2284.

SABINO, D. M. U.; COSTA, L. da F.; RIZZATTI, E. G.; ZAGO, M. A. A texture approach to leukocyte recognition. **Real-Time Imaging**, v. 10, n. 4, p. 205 – 216, 2004. ISSN 1077-2014. Imaging in Bioinformatics: Part {III}.

SARWAR, A.; SHARMA, V.; GUPTA, R. Hybrid ensemble learning technique for screening of cervical cancer using papanicolaou smear image analysis. **Personalized Medicine Universe**, v. 4, p. 54 – 62, 2015. ISSN 2186-4950.

SHAMOI, P.; INOUE, A.; KAWANAKA, H. Deep color semantics for e-commerce content-based image retrieval. In: **Conf. Fuzzy Logic in Artificial Intelligence**. [S.l.: s.n.], 2015. p. 14–20.

SOUZA, M. M. de; MEDEIROS, F. N.; RAMALHO, G. L.; JR., I. C. de P.; OLIVEIRA, I. N. Evolutionary optimization of a multiscale descriptor for leaf shape analysis. **Expert Systems with Applications**, v. 63, p. 375 – 385, 2016.

SZEGEDY, C.; IOFFE, S.; VANHOUCKE, V. Inception-v4, inception-resnet and the impact of residual connections on learning. **Computing Research Repository**, abs/1602.07261, 2016.

SZEGEDY, C.; VANHOUCKE, V.; IOFFE, S.; SHLENS, J.; WOJNA, Z. Rethinking the inception architecture for computer vision. In: **IEEE Conference on Computer Vision and Pattern Recognition**. Las Vegas: [s.n.], 2016. p. 2818–2826.

TAMURA, H.; MORI, S.; YAMAWAKI, T. Textural features corresponding to visual perception. **IEEE Transactions on Systems, Man, and Cybernetics**, v. 8, n. 6, p. 460–473, Junho 1978. ISSN 0018-9472.

TANG, X. Texture information in run-length matrices. **IEEE Transactions on Image Processing**, IEEE Press, Piscataway, NJ, USA, v. 7, n. 11, p. 1602–1609, nov. 1998. ISSN 1057-7149.

TIWARI, D.; TYAGI, V. A novel scheme based on local binary pattern for dynamic texture recognition. **Computer Vision and Image Understanding**, v. 150, p. 58 – 65, 2016. ISSN 1077-3142.

TZELEPI, M.; TEFAS, A. Deep convolutional learning for content based image retrieval. **Neurocomputing**, v. 275, p. 2467 – 2478, 2018. ISSN 0925-2312.

USHIZIMA, D.; BIANCHI, A.; CARNEIRO, C. Segmentation of subcellular compartiments combining superpixel representation with voronoi diagrams. In: **Overlapping Cervical Cytology Image Segmentation Challenge - IEEE ISBI**. [S.l.: s.n.], 2014. p. 1–2.

VASHAEE, A.; JAFARI, R.; ZIOU, D.; RASHIDI, M. M. Rotation invariant HOG for object localization in web images. **Signal Processing**, v. 125, p. 304 – 314, 2016. ISSN 0165-1684.

WANG, B.; BROWN, D.; GAO, Y.; SALLE, J. L. March: Multiscale-arch-height description for mobile retrieval of leaf images. **Information Sciences**, v. 302, p. 132 – 148, 2015. ISSN 0020-0255.

WANG, H.; FENG, Y.; SA, Y.; LU, J. Q.; DING, J.; ZHANG, J.; HU, X.-H. Pattern recognition and classification of two cancer cell lines by diffraction imaging at multiple pixel distances. **Pattern Recognition**, 2016. ISSN 0031-3203.

WANG, X.; ZHENG, B.; ZHANG, R. R.; LI, S.; CHEN, X.; MULVIHILL, J. J.; LU, X.; PANG, H.; LIU, H. Automated analysis of fluorescent in situ hybridization (fish) labeled genetic biomarkers in assisting cervical cancer diagnosis. **Technology in Cancer Research and Treatment**, v. 9, n. 3, p. 231–242, 2010.

WATANABE, S.; IWASAKA, T.; YOKOYAMA, M.; UCHIYAMA, M.; KAKU, T.; MATSUYAMA, T. Analysis of nuclear chromatin distribution in cervical glandular abnormalities. **Acta cytologica**, v. 48, n. 4, p. 505—513, 2004. ISSN 0001-5547.

XU, H.; MANDAL, M. Epidermis segmentation in skin histopathological images based on thickness measurement and k-means algorithm. **EURASIP Journal on Image and Video Processing**, v. 2015, n. 1, p. 18, 2015. ISSN 1687-5281.

YLIOINAS, J.; POH, N.; HOLAPPA, J.; PIETIKäINEN, M. Data-driven techniques for smoothing histograms of local binary patterns. **Pattern Recognition**, v. 60, p. 734 – 747, 2016. ISSN 0031-3203.

YU, Q.; LIU, F.; SONG, Y.-Z.; XIANG, T.; HOSPEDALES, T.; LOY, C. C. Sketch me that shoe. In: **Computer Vision and Pattern Recognition**. [S.l.: s.n.], 2016. p. 799–807.

YU, W.; YANG, K.; YAO, H.; SUN, X.; XU, P. Exploiting the complementary strengths of multi-layer CNN features for image retrieval. **Neurocomputing**, v. 237, p. 235 – 241, 2017. ISSN 0925-2312.

ZHANG, L.; KONG, H.; CHIN, C. T.; LIU, S.; CHEN, Z.; WANG, T.; CHEN, S. Segmentation of cytoplasm and nuclei of abnormal cells in cervical cytology using global and local graph cuts. **Computerized Medical Imaging and Graphics**, v. 38, n. 5, p. 369 – 380, 2014. ISSN 0895-6111.

ZIJDENBOS, A. P.; DAWANT, B. M.; MARGOLIN, R. A.; PALMER, A. C. Morphometric analysis of white matter lesions in mr images: method and validation. **IEEE Transactions on Medical Imaging**, v. 13, n. 4, p. 716–724, Dec 1994. ISSN 0278-0062.