

Segmentação de Núcleos Celulares Baseada em Agrupamento e Características de Forma

Samuel P. B. D. Lélis¹, Lucas B. M. de Souza¹, Romuere R. V. Silva¹

¹ Departamento de Sistemas de informação - Universidade Federal do Piauí - UFPI

² Campus Senador Helvídio Nunes de Barros - CSHNB

³ Picos - PI - Brasil

slelis, lucasbezerraufpi527, romuere@ufpi.edu.br

Abstract. *Among the types of cancer, cervical cancer is the fourth most common worldwide among women. Diagnosis is made mainly through the pap smear, which offers a good detection rate in the primary stages of the disease, offering more chances of cure. Thus, automated results evaluation methods obtained a very valuable tool in combating this disease. This work evaluates the parameters for the combination of nuclei in pap smear images in the CRIC database. With the variation and study of the parameters we obtained an average data result of 0.7023 with the best combination of parameters.*

Resumo. *Dentre os tipos de câncer, o de colo de útero é a quarto mais comum em todo o mundo em mulheres. O diagnóstico é feito, principalmente por meio do exame de papanicolau, que oferece uma boa taxa de detecção em estágios primários da doença, oferecendo mais chances de cura. Desse modo, métodos de avaliação automatizados de exames constituem uma ferramenta muito valiosa no combate a essa doença. Este trabalho avalia os parâmetros para a segmentação de núcleos em imagens de papanicolau na base CRIC. Com a variação e estudo dos parâmetros obtivemos um resultado médio de dice em 0.7023 com a melhor combinação de parâmetros.*

1. Introdução

O câncer cervical é um tipo de tumor maligno que se aloja na parte inferior da reprodução do órgão feminino e é causado por lesões mal tratadas causadas pelo vírus do papiloma humano (HPV). Segundo a [WHO 2021], esta doença é o quarto tipo de câncer mais comum em mulheres no mundo. A ausência de sintomas nos estágios iniciais da doença dificulta a detecção em tempo hábil para a cura. O diagnóstico dessa doença pode ser feito através de exames de Papanicolau, que é o grande responsável pela detecção em estágios iniciais, que irão possibilitar o tratamento desde o início em caso positivo de câncer.

Conforme visto em [Silva et al. 2019], o processo de análise das imagens é feito por um especialista, que analisa visualmente as características das células no exame, buscando por alterações morfológicas nas células, como o tamanho do núcleo e citoplasma, o formato dos aglomerados de células e a distribuição de cromatina. A necessidade de um especialista para a detecção da doença prejudica a escalabilidade de programas públicos de saúde com foco na prevenção e detecção desse câncer, especialmente com o crescimento populacional acelerado em países de segundo e terceiro mundo.

Quando o câncer é detectado nos estágios iniciais as chances de recuperação e de sobrevivência dos pacientes são maiores. Nesse cenário, a computação vem cada vez mais trabalhando junto à saúde, sendo comum o termo *Computer-Aided-Detection* (CAD) em trabalhos científicos que tratam de computação aplicada à saúde. Esse tipo de sistema pode ser desenvolvido para detectar o câncer e agilizar o diagnóstico, com objetivo de obter laudos automáticos, com menor custo e maior eficiência.

Esse trabalho propõe o desenvolvimento de uma técnica para a segmentação de imagens de exames de papanicolau para o auxílio no diagnóstico destas. Para tanto, foram utilizadas técnicas de processamento digital de imagens e visão computacional.

2. Trabalhos Relacionados

O trabalho [Wang et al. 2019] propôs uma segmentação de núcleos celulares utilizando o algoritmo de agrupamento Mean-Shift, obtendo 0.9425 e 0.9345 de resultado para as métricas sensibilidade e especificidade respectivamente. Para classificação foi utilizado o classificador SVM [Cortes and Vapnik 1995], obtendo até mais 96% de acurácia fazendo o uso de métodos de extração de recursos como CAGA, valor P e MRmRC. Esses resultados foram obtidos com o uso de uma base proprietária.

No trabalho de [Silva et al. 2019], foram utilizadas as bases públicas de imagens CRIC e Herlev [Jantzen et al. 2005a], é proposto um descritor de características radiais. A classificação é feita via Random Forest e bootstrap. Como resultado foram obtidos os resultados para as bases CRIC: 0.02 de FNR e 0.89 de Kappa e Herlev: 0.14 de FNR e 0.79 de Kappa.

O trabalho de [Bandyopadhyay and Nasipuri 2020] propõe um processo de segmentação baseado no canal azul da imagem RGB, representado no color space HSV com a aplicação do algoritmo de agrupamento K-means e atingiu os melhores resultados de acurácia 0.9215 com a utilização do classificador Random Forest [Ho 1995] na subamostra de teste.

O trabalho de [William et al. 2019] propõe novos algoritmos de extração de características (*Simulated annealing and wrapper filter*) e classificação (fuzzy C-means [Talukdar et al. 2013]), com aplicação nas bases DTU [Jantzen et al. 2005b] e Herlev atingindo resultados de 0.9928, 0.9747 e 0.9888 para sensibilidade, especificidade e acurácia respectivamente.

O trabalho de [Zhao et al. 2021] propõe um método de melhoria seletiva de bordas com uso baseado no algoritmo de detecção de bordas canny [Canny 1986] e operações de morfologia matemática para aplicar a segmentação com acurácia média de 0.9590 em uma base de dados própria.

O trabalho de [Xiang et al. 2020] propõe um método diferente com o uso de redes neurais convolucionais que não conta com a etapa de segmentação, como é mais comum em trabalhos nesse segmento. O método conta com a rede YOLOv3 [Redmon and Farhadi 2018] como base para a detecção de células e aglomerados como objetos alvo e sem focar na extração manual de características, atingindo resultados de 0.975 e 0.678 de sensibilidade e especificidade, respectivamente na classificação em 10 tipos diferentes de células.

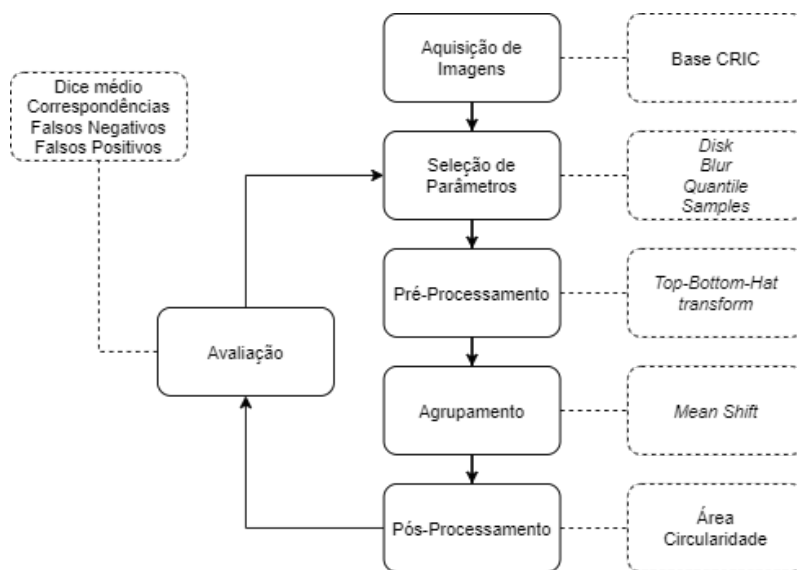


Figura 1. Diagrama representando a metodologia proposta, onde o processo principal de segmentação é repetido para busca de parâmetros.

3. Metodologia Proposta

O método proposto nesse trabalho está dividido em seis etapas: 1) na etapa de aquisição de imagens utilizamos a base de imagens de exames de papanicolau CRIC [Rezende et al. 2020]; 2) seleção manual de parâmetros dos algoritmos de 3) pré-processamento e 4) agrupamento; 5) o pós-processamento foi realizado com o objetivo de reduzir ruídos (Falsos Positivos) das etapas anteriores; 6) por fim na avaliação dos resultados foram selecionados os melhores conjunto de parâmetros para o método proposto. O fluxograma do método proposto é apresentado na Figura 1.

3.1. Base de Dados

A base de dados CRIC [Rezende et al. 2020] é composta por 169 exames de Papanicolau digitalizados com tamanho fixo obtidas em zoom 40X, totalizando mais de 2000 células cervicais individualizadas (VP), juntamente com algumas máscaras células delimitadas por especialistas. Neste trabalho utilizamos 12 das 169 imagens completas, pois estas possuem as máscaras de células individualizadas.

3.2. Pré-Processamento

Técnicas de pré-processamento de imagens consistem na aplicação de métodos, filtros ou transformações nas imagens de entrada de modo a melhorar o resultado das operações posteriores, nesta situação, as transformações aplicadas visam melhorar o resultado da etapa de segmentação. Nesta etapa, foram variados o parâmetro *disk*, que representa o raio do disco utilizado como elemento estruturante da operação de morfologia matemática *Top-Bottom Hat transform* [Wang et al. 2019] e o parâmetro *blur* que representa a intensidade da transformação gaussiana aplicada à imagem antes do processamento.

3.3. Agrupamento Mean-Shift

Trata-se de um algoritmo de agrupamento de vizinhanças que funciona selecionando centros na imagem e é calculada a média das posições de todos os pontos a um certo raio,

caso seja diferente da posição original do centro, este é movido para as coordenadas da média e o processo é repetido até que as coordenadas não se alterem ou estejam muito próximas [Cheng 1995]. Desse modo, os centros são movidos para as regiões de maior densidade de valores de pixel. Nessa etapa foram variados os parâmetros *Quantile* e *Samples* que são utilizados na estimativa da quantidade de agrupamentos para a imagem [Comaniciu and Meer 2002].

3.4. Pós-Processamento

Tendo em vista que o resultado do processamento anterior em vários casos retorna resultados com muitos falsos positivos, fez-se necessária uma etapa de pós-processamento para remoção destes. Para isso, fizemos uso de metodologia própria com inspiração em Wang et al. [Wang et al. 2019] e é composto pelas etapas:

- A média aritmética (Avg) da área (Ar) das regiões (componentes conectadas) é computada;
- São descartadas as regiões cujo perímetro (Pm) é igual a 0 ou a área está fora do intervalo aberto entre Avg/2 e Avg*4 [Wang et al. 2019];
- As regiões restantes que possuem a Circularidade (Equação 1) maior ou igual a 0.6 são computadas como núcleos únicos e as demais como aglomerados celulares [Wang et al. 2019].

$$Cr = 4 * \pi * \frac{Ar}{Pm^2}. \quad (1)$$

3.5. Métricas de Avaliação

Para a avaliação das segmentações, a métrica *dice* é calculada com base na média das correspondências, ou seja, regiões que possuem *dice* maior que 0,7 [Silva et al. 2019] entre o rótulo da máscara e o resultante da segmentação. A taxa FNR (False Negative Rate) representa percentualmente a quantidade de regiões segmentadas pelo especialista que não possuem correspondência com as regiões resultantes da segmentação automática. Por outro lado, a métrica FPR (False Positive Rate) representa o percentual de regiões resultantes da segmentação automática que não possuem correspondência com as regiões delimitadas pelo especialista.

4. Resultados

A variação dos parâmetros resultou em quase 700 experimentos. A Tabela 1 mostra os 20 melhores resultados ranqueados pelo *dice* médio das correspondências e parâmetro *quantile* fixo em 0,1, visto que todos os melhores resultados foram obtidos com esse valor no parâmetro.

A Tabela 2 representa o resultado individual de cada uma das 12 imagens utilizadas na melhor média presente na Tabela 1 ranqueados por número de correspondências. Como podemos observar, foram obtidos resultados diversos, desde um melhor resultado com 17 de 30 regiões de interesse encontradas a um pior resultado com 0 correspondências.

As Figuras 3 e ?? representam o resultado visual da segmentação com todas as regiões resultantes da metodologia proposta, sendo 3 o melhor resultado, ou seja, o Slide

Tabela 1. Melhores resultados no conjunto de dados classificados pela média dos dados dos slides, onde o *quantile* é fixado em 0,1 e variando os parâmetros *Disk*, *Blur* e *Samples*.

<i>Disk</i>	<i>Blur</i>	<i>Samples</i>	<i>Dice</i> Médio	FNR	FPR
3	5×5	100	0,7023	0,7574	0,8862
2	3×3	100	0,6634	0,6998	0,8589
3	3×3	1300	0,6404	0,7753	0,8174
3	3×3	1000	0,6352	0,7709	0,8092
2	5×5	100	0,5987	0,8469	0,8746
2	5×5	700	0,5983	0,8510	0,7894
2	3×3	400	0,5940	0,7406	0,7653
2	3×3	700	0,5935	0,7382	0,7632
3	3×3	100	0,5738	0,7283	0,8549
3	3×3	700	0,5737	0,7805	0,8084
3	3×3	400	0,5733	0,7782	0,8384
4	7×7	100	0,5697	0,7247	0,8944
4	5×5	100	0,5661	0,7844	0,9311
4	3×3	1300	0,5554	0,8322	0,8433
4	5×5	400	0,5497	0,8487	0,8969
2	5×5	1000	0,5435	0,8719	0,8865
3	5×5	1300	0,5146	0,8108	0,8954
3	5×5	400	0,5027	0,8306	0,9232
4	5×5	1000	0,5012	0,8283	0,8902
4	3×3	1000	0,4997	0,8193	0,8393

Tabela 2. Resultados de slide detalhados dos parâmetros em 3, 5 × 5, 100 e 0,1 valores de *Disk*, *Blur*, *Samples* e *Quantile*, respectivamente

Slide	<i>Dice</i> Médio	Correspondências	FNR	FPR
Slide 1	0,7771	17	0,4333	0,7017
Slide 2	0,7464	10	0,5652	0,8507
Slide 3	0,7960	7	0,6818	0,8679
Slide 4	0,7576	5	0,6429	0,8214
Slide 5	0,7275	5	0,6875	0,9576
Slide 6	0,8260	4	0,6000	0,8400
Slide 7	0,8421	2	0,9565	0,8461
Slide 8	0,8190	2	0,7778	0,8947
Slide 9	0,7180	2	0,8333	0,9798
Slide 10	0,6880	2	0,9429	0,8947
Slide 11	0,7296	1	0,9677	0,9800
Slide 12	0,0000	0	1,0000	1,0000

1, que obteve 17 correspondências de um total de 30 regiões, tendo estas uma média de 0,7771 Dice, 0,4333 FNR e 0,7017 FPR. Por outro lado, ?? representa o Slide 5 que teve um resultado médio em relação às correspondências, tendo 5 correspondências em um total de 16 regiões com média de 0,7275 de Dice, 0,6875 FNR e 0,9576 FPR.

A Figura 4 representa a variação do resultado principal (*average dice*) quando fixos os valores dos parâmetros *quantile* e *samples*. Os melhores resultados sempre foram com o valor do parâmetro *quantile* igual a 0,1 e foi observado, também, que o parâmetro *samples* teve impacto inexpressivo nos resultados, ou seja, quando fixados os 3 outros parâmetros, os *average dice* teve variação mínima.

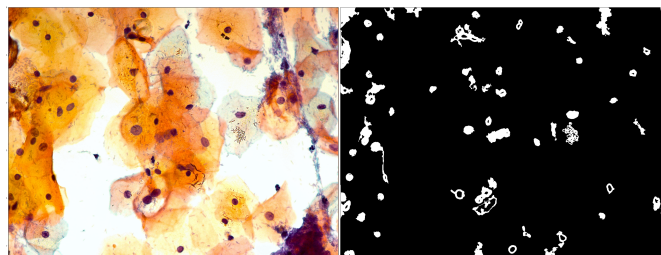


Figura 2. Representação dos resultados da segmentação no Slide 1 da Tabela 2.

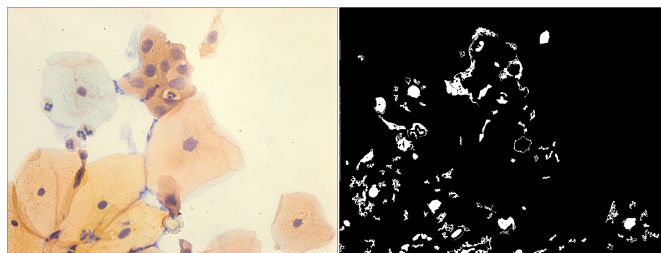


Figura 3. Representação dos resultados da segmentação no Slide 5 da Tabela 2.

Como visto nos resultados, o método possui limitações nos aspectos de FNR e FPR, apresentando altas taxas, mesmo entre as melhores combinações de parâmetros no quesito Dice Médio. Além disso, o método apresentou um alto custo computacional, levando em média 70 segundos para processar um slide em uma máquina com um processador AMD Threadripper 2950x.

5. Conclusões e Trabalhos Futuros

Este trabalho tem por objetivo a realização da busca por parâmetros do método, mostrando resultados promissores no quesito Dice médio das correspondências. No entanto, apresentou limitações no custo computacional e nos resultados em relação às taxas de erro FNR e FPR. Em relação aos parâmetros, como vemos no mapa de calor, principalmente, os melhores parâmetros do pré-processamento são 3 para o *disk* e 5×5 para o *Blur* e, com base nos demais resultados, pode-se observar um impacto mínimo dos demais parâmetros.

O alto custo computacional apresentado no método tem como principal causador o clusterizador Mean-Shift, que possui um alto custo computacional, visto que o *estimate*

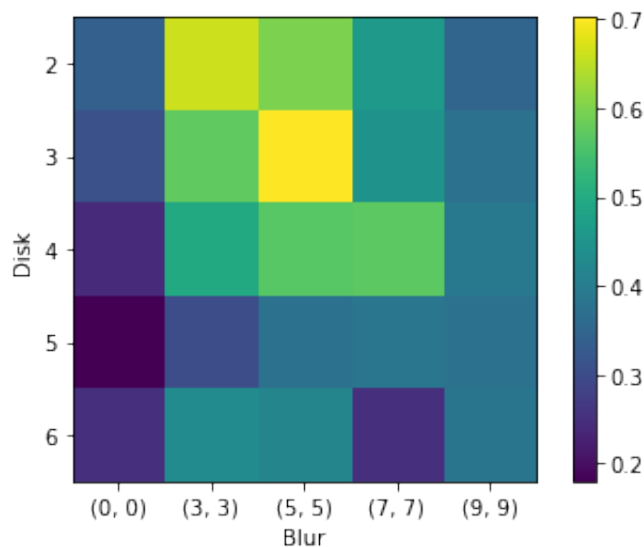


Figura 4. Mapa de calor que representa o valor médio dos dados na variação dos valores do disco e desfoque enquanto fixa os valores do *quantile* e do *samples* em 0,1 e 100, respectivamente.

bandwidth apresenta custo computacional $O(n^2)$, em que n representa o parâmetro *Samples* além do custo relativo ao agrupador quando comparado com outros algoritmos como o *k-means* que apresenta um menor custo computacional.

Como trabalhos futuros, as limitações deste podem ser contornadas com a utilização de redes neurais na etapa de pós processamento, visando atingir percentuais de acerto mais altos e consistentes, reduzindo, desse modo, a taxa de falsos negativos, além de possibilitar o uso de capacidades de GPU, aumentando a eficiência e reduzindo o custo computacional.

Além disso, a taxa de falsos negativos pode ser melhorada com o uso de técnicas mais diversas nas etapas de pré-processamento e clusterização, podendo obter mais regiões de interesse na segmentação e, com associação de redes neurais no pós processamento, melhorar o desempenho geral do método.

Referências

- Bandyopadhyay, H. and Nasipuri, M. (2020). Segmentation of pap smear images for cervical cancer detection. In *2020 IEEE Calcutta Conference (CALCON)*, pages 30–33. IEEE.
- Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698.
- Cheng, Y. (1995). Mean shift, mode seeking, and clustering. *IEEE transactions on pattern analysis and machine intelligence*, 17(8):790–799.
- Comaniciu, D. and Meer, P. (2002). Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 24(5):603–619.

- Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.
- Ho, T. K. (1995). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE.
- Jantzen, J., Norup, J., Dounias, G., and Bjerregaard, B. (2005a). Pap-smear benchmark data for pattern classification. *Nature inspired Smart Information Systems (NiSIS 2005)*, pages 1–9.
- Jantzen, J., Norup, J., Dounias, G., and Bjerregaard, B. (2005b). Pap-smear benchmark data for pattern classification. *Nature inspired Smart Information Systems (NiSIS 2005)*, pages 1–9.
- Redmon, J. and Farhadi, A. (2018). Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*.
- Rezende, M. T., Silva, R., Tobias, A. H. G., Oliveira, P., Sombra de Medeiros, F. N., Ushizima, D., Carneiro, C. M., and Bianchi, A. G. C. (2020). Cric cervix classification.
- Silva, R. R., Araujo, F. H., Ushizima, D. M., Bianchi, A. G., Carneiro, C. M., and Medeiros, F. N. (2019). Radial feature descriptors for cell classification and recommendation. *Journal of Visual Communication and Image Representation*, 62:105–116.
- Talukdar, J., Nath, C. K., and Talukdar, P. (2013). Fuzzy clustering based image segmentation of pap smear images of cervical cancer cell using fcm algorithm. *markers*, 3(1):460–2.
- Wang, P., Wang, L., Li, Y., Song, Q., Lv, S., and Hu, X. (2019). Automatic cell nuclei segmentation and classification of cervical pap smear images. *Biomedical Signal Processing and Control*, 48:93–103.
- WHO (2021). Cervical cancer. https://www.who.int/health-topics/cervical-cancer/#tab=tab_1. Accessed on 2021-06-11.
- William, W., Ware, A., Basaza-Ejiri, A. H., and Obungoloch, J. (2019). Cervical cancer classification from pap-smears using an enhanced fuzzy c-means algorithm. *Informatics in Medicine Unlocked*, 14:23–33.
- Xiang, Y., Sun, W., Pan, C., Yan, M., Yin, Z., and Liang, Y. (2020). A novel automation-assisted cervical cancer reading method based on convolutional neural network. *Biocybernetics and Biomedical Engineering*, 40(2):611–623.
- Zhao, M., Wang, H., Han, Y., Wang, X., Dai, H.-N., Sun, X., Zhang, J., and Pedersen, M. (2021). Seens: Nuclei segmentation in pap smear images with selective edge enhancement. *Future Generation Computer Systems*, 114:185–194.