

Segmentação de Imagens Celulares

Fotografia Computacional – UFRGS
Projeto Final

Bernardo Trevizan*

5 de Julho de 2018

Resumo

Segmentação de imagens celulares é um dos principais campos na área de análise computadorizada de imagens. Este trabalho apresenta um método híbrido de segmentação e classificação de células extraídas da mucosa bucal de pacientes. Observou-se que método possui melhor performance em imagens com células e núcleos bem definidos. Para os demais casos, o método apresentou falhas.

1 Introdução

Atualmente, as imagens das células orais são analisadas manualmente. Essa tarefa pode ser cansativa e muitas vezes podem levar a erros de contagem e classificação das células. Entretanto, algoritmos de segmentação e aprendizado de máquina podem oferecer um diagnóstico mais rápido, o qual pode ajudar na prevenção da doença e tratamento imediato do paciente.

Diversos métodos foram encontrados na literatura para segmentação de imagens celulares. Desde segmentação de células sanguíneas e ósseas, como apresentado em Malpica et al. [1997] até células cervicais [Bengtsson et al., 2004]. Entretanto, apesar de apresentarem bons resultados, estes trabalhos foram desenvolvidos para classes específicas de imagens celulares (sanguíneas, ósseas, cervicais). Bengtsson et al. [2004] apresenta um algoritmo de segmentação em três passos: (i) identificação dos núcleos através de máximas locais – $h\text{-maxima}$ –; (ii) remoção do plano de fundo com *threshold*; e (iii) aplicação do algoritmo de segmentação *Seeded-Watershed* para identificar células sobrepostas. Assim como Bengtsson et al. [2004], Malpica et al. [1997] utilizou $h\text{-maxima}$ para identificação dos núcleos celulares. No entanto, esse último computou um mapa de distâncias para aplicar o *Watershed*.

*btrevizan@inf.ufrgs.br

Pela natureza das imagens apresentadas neste trabalho, não foi possível replicar os trabalhos mencionados. Dessa forma, este projeto apresenta um método híbrido de segmentação e classificação de células extraídas da mucosa bucal de pacientes. Observamos que o método possui uma melhor performance em imagens com arestas e núcleos bem definidos do que nos demais casos. O método apresentado foca na identificação dos núcleos das células e, por isso, não segmenta bem células com núcleos mal definidos ou mesmo sem núcleos.

2 Metodologia

Observou-se, na Figura 1, as características de uma imagem celular como a propriedade escura e elíptica dos núcleos e as bordas bem definidas das células. A partir dessas observações, algoritmos foram propostos a fim de identificar cada célula individualmente e classificá-las por cor. O algoritmo possui vários passos (Figura 2), sendo a identificação dos núcleos o passo mais importante.

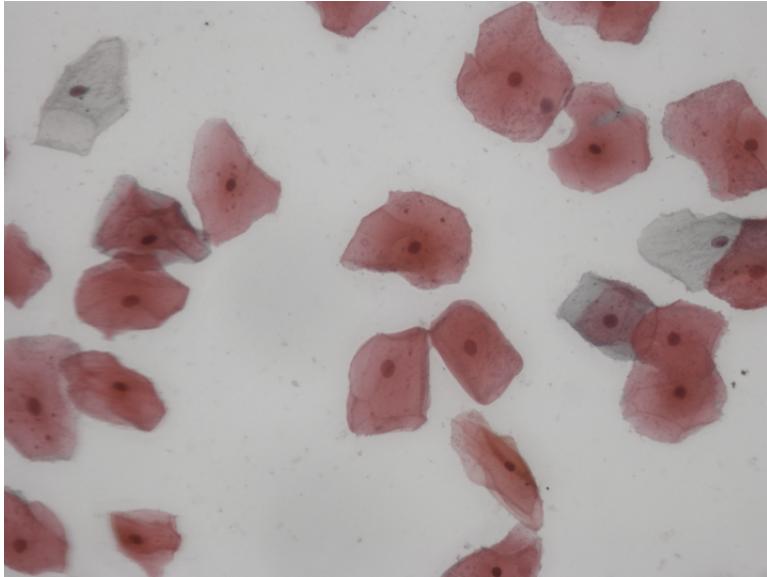


Figura 1: Imagem celular.

2.1 Pré-Processamento

Na fase de pré-processamento, a imagem é submetida ao filtro de Chambolle¹, apresentado em Duran et al. [2013]. O filtro de Chambolle não borra as bordas dos objetos da imagem. Dessa forma, ele é considerado um *edge-aware filter*.

¹A implementação do filtro utilizada está disponível na biblioteca Scikit-Image para linguagem Python.



Figura 2: Pipeline do algoritmo.

Dentre os filtros testados², o de Chambolle foi o que obteve o melhor resultado visualmente. Como apresentado na Figura 3, os ruídos da imagem, tanto nas células quanto no fundo são removidos de forma a manter a estrutura geral da figura. Por questões de desempenho, as imagens foram convertidas para escala de cinza.

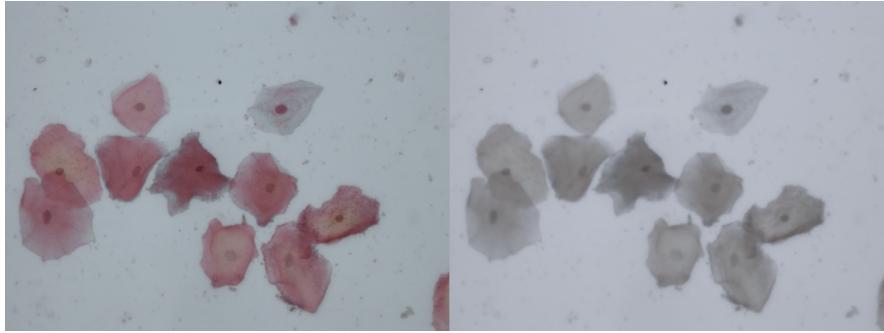


Figura 3: Imagem original (esquerda); imagem após a aplicação do filtro de Chambolle (direita).

2.2 Remoção do Plano de Fundo

Observa-se, nas imagens obtidas, um fundo majoritariamente cinza com alguns pontos mais escuros. Sendo assim, possível utilizar segmentação por ponto de corte (*threshold*). A Figura 4 demonstra a remoção do plano fundo com diferentes pontos de corte. O *threshold* de Yen e o mínimo não são capazes de detectar um ponto de corte que mantém as bordas das células. Em contrapartida, thresholds como o de Otsu, Isodata e Triangle são capazes de eliminar apenas o fundo da imagem. Dentres os três, o método Triangle, primeiramente proposto em Zack et al. [1977], calcula o melhor *threshold*, visto que mantém partes mais claras das células, contrário do Otsu e Isodata.

²http://scikit-image.org/docs/dev/auto_examples/filters/plot_denoise.html

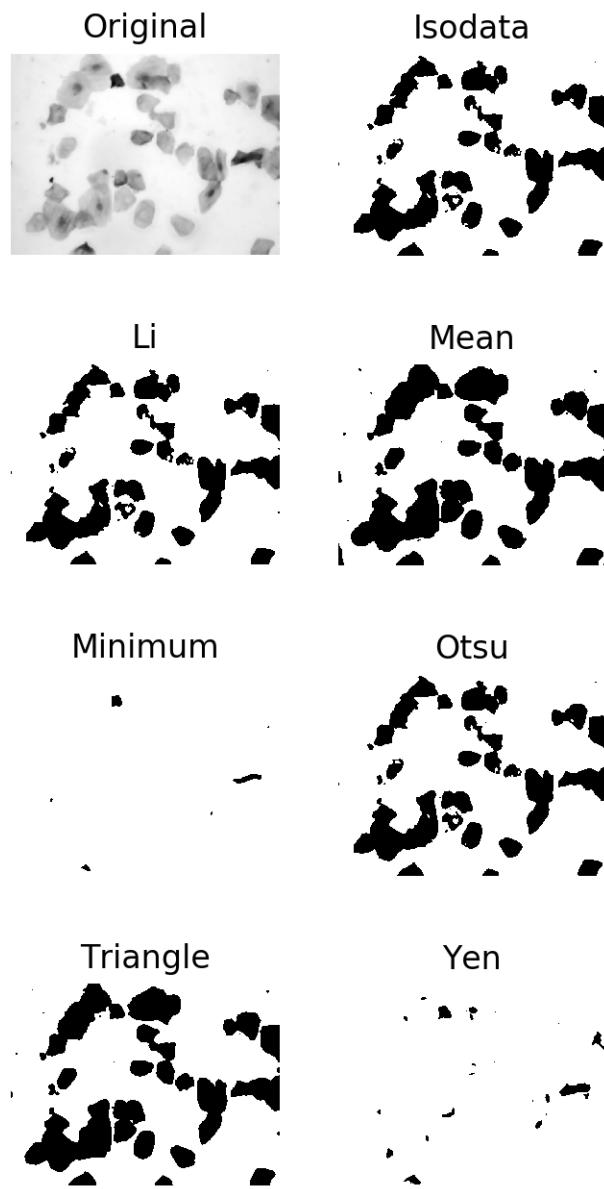


Figura 4: Diferentes técnicas de threshold aplicadas a mesma imagem (canto superior esquerdo).

O método Triangle projeta uma reta entre um pico numa das extremidades do histograma até o final da outra extremidade e, para cada intensidade no intervalo da reta, tenta maximizar a distância entre um ponto e a reta. O ponto

de corte final é somado por um deslocamento fixo (Figura 5). Esse método possui problemas quando a intensidade mais frequente não se encontra num dos extremos do histograma, o que pode gerar dois possíveis pontos de corte. Entretanto, esse não é o caso para as nossas imagens. Como o plano de fundo é predominantemente um tom de cinza, os histogramas das imagens celulares são semelhantes ao da Figura 6. Observa-se os picos nas extremidades, os quais favorecem o método Triangle. O resultado da remoção do plano de fundo utilizando o ponto de corte calculado pelo método Triangle pode ser visto na Figura 7, onde os pixels em preto representam o fundo da imagem.

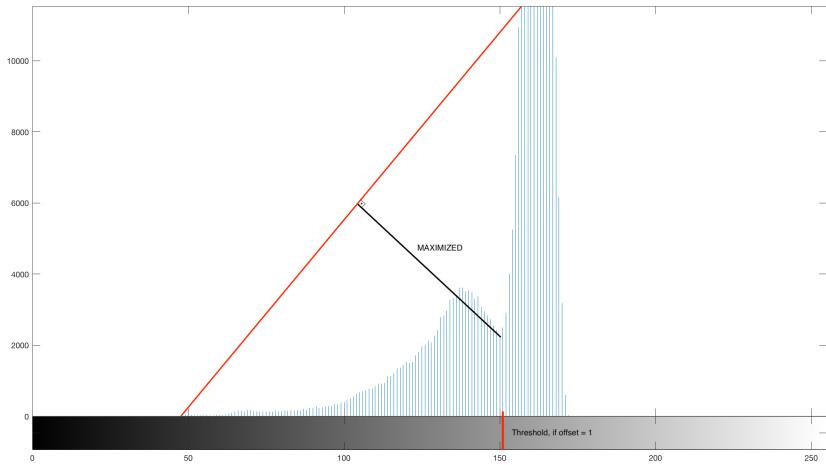


Figura 5: Ilustração do método Triangle.

2.3 Identificação dos Núcleos

Na remoção do plano de fundo, foi possível obter os segmentos da imagem (a partir de agora, vamos nos referir apenas por *segmentos celulares*), nos quais as células estão contidas (Figura 8). Entretanto, há células sobrepostas umas às outras e, assim, não é possível definir os limites de cada célula individualmente apenas com a remoção do fundo. Dessa forma, foi proposto um algoritmo que minimiza as regiões identificadas como células para regiões menores e com baixa intensidade média.

Para cada segmento celular, computa-se uma sequência de máscaras binárias B_i , para $i = \{1, 2, 3, \dots, 16\}$, através da Equação 1, onde $I(x, y)$ é a intensidade de um pixel.

$$B_i(x, y) = I(x, y) \leq T \quad (1)$$

T é um ponto de corte, e pode ser definido pela Equação 2, onde S_{MIN} e S_{MAX} são as intensidades mínima e máxima do segmento celular, respectivamente, e $\alpha = \{0.2, 0.25, 0.3, \dots, 1\}$.

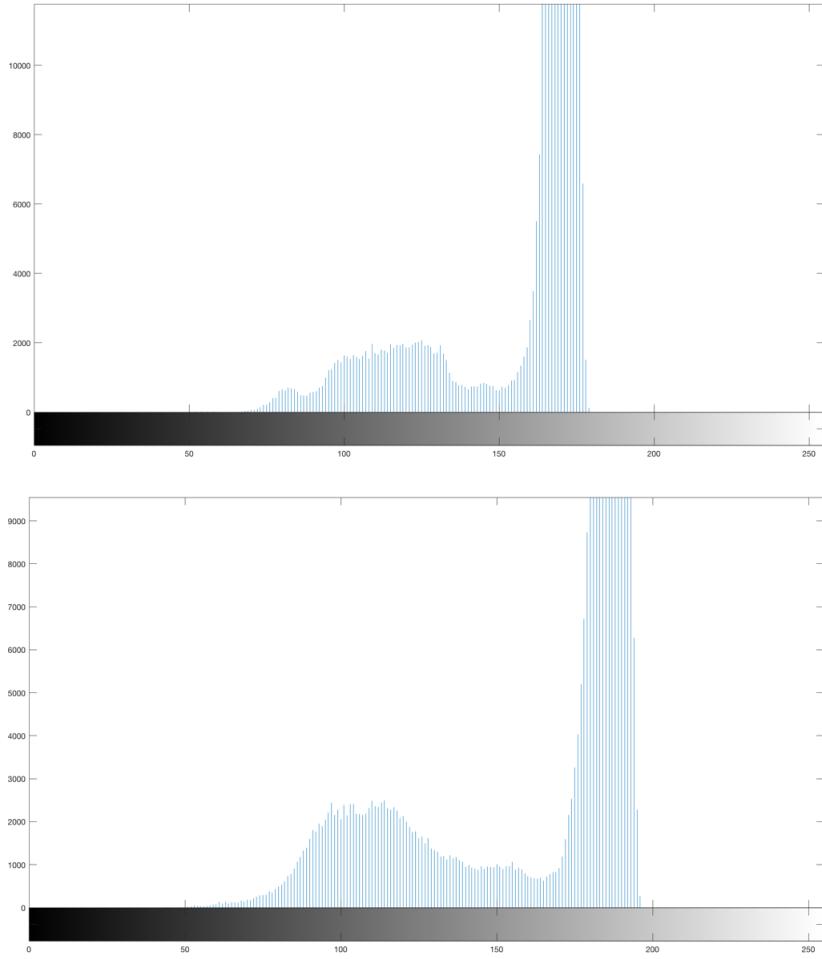


Figura 6: Histograma de duas imagens celulares.

$$T = \alpha_i S_{MIN} + (1 - \alpha_i) S_{MAX} \quad (2)$$

A Figura 9 demonstra uma das sequências geradas. Note que, ao aumentar o valor de α , as regiões mais escuras do segmento celular permanecem na máscara binária. Entretanto, nas máscaras da sequência, sujeiras da mucosa bucal são classificadas como núcleo. Para resolver esse problema, das 16 máscaras, seleciona-se aquele, a qual possui o maior número de regiões não conexas. A partir disso, cada região isolada torna-se um novo segmento celular e o processo se repete até encontrarmos uma sequência com no máximo um segmento.

Quando obtivermos uma sequência, como a apresentada na Figura 10, com apenas um segmento celular, a máscara, a qual contém a região com a melhor

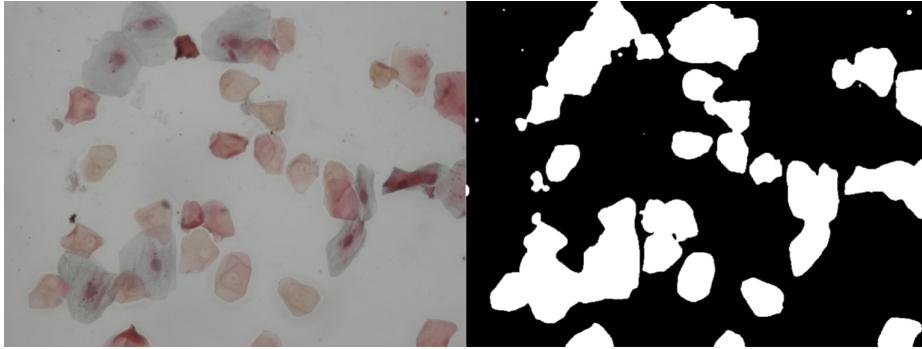


Figura 7: Imagem original (esquerda); máscara binária da segmentação (direita).

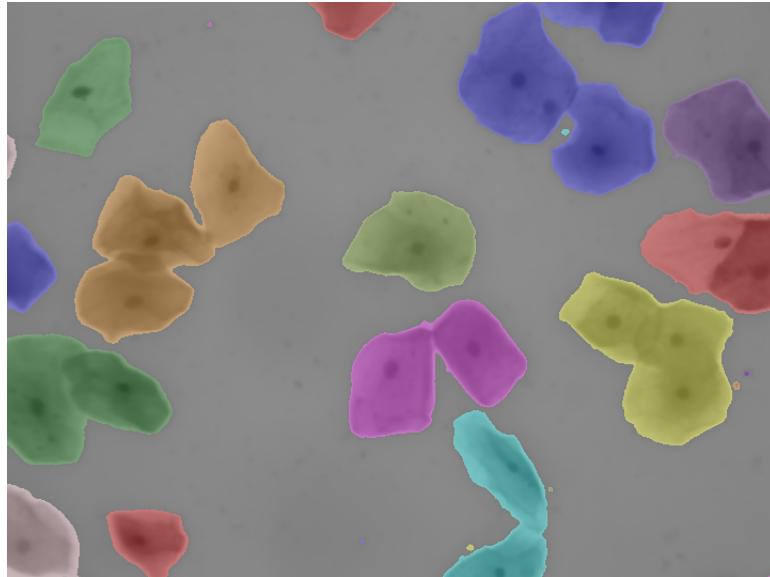


Figura 8: Cada segmentos colorido representa uma região a qual contém células.

ecentricidade é escolhida como o segmento final para aquela região. A ecentricidade é definida pela razão entre a diferença dos dois pontos focais (F_1 e F_2) de uma elipse e seu maior eixo (E_1), como demonstrado na Equação 3.

$$E = \frac{F_1 - F_2}{E_1} \quad (3)$$

Cada novo segmento celular não conexo recebeu uma identificação (*label*), criando, assim, uma nova imagem I' com as correspondentes identificações dos segmentos da imagem original. Regiões sem segmentos, como o fundo, recebe-

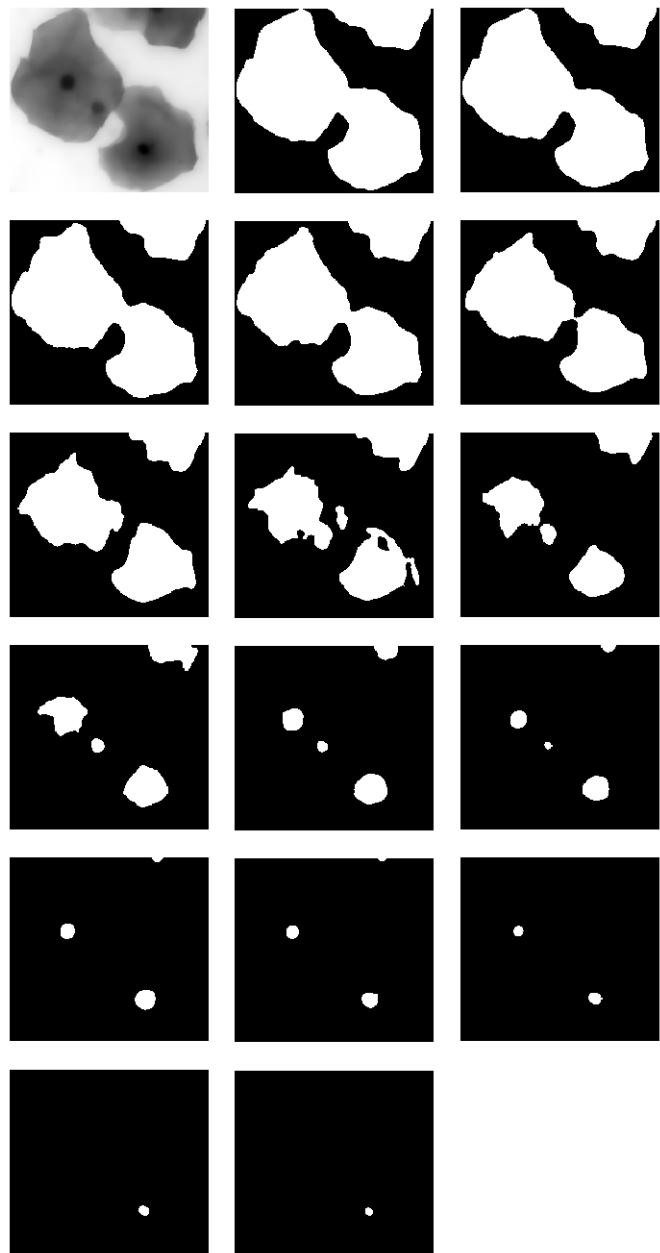


Figura 9: Segmento celular e 16 máscaras binárias.

ram identificação 0.

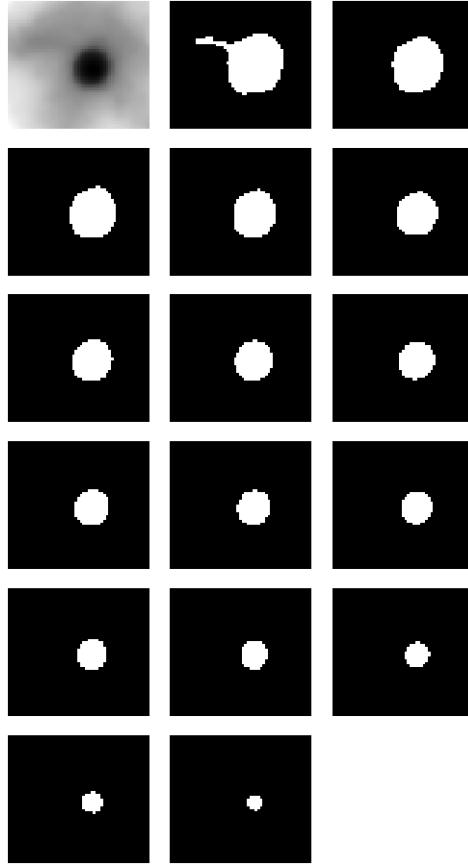


Figura 10: Sequência com, no máximo, um segmento celular.

2.4 Classificação das Células

Primeiramente, foi necessário expandir a segmentação dos núcleos para as bordas das células com o objetivo de obter as informações de toda a célula e não somente do núcleo. Para tanto, a partir da segmentação, para cada pixel P de I' , computou-se as distâncias $D_i(P, S_i)$, onde S é o centro de um segmento celular e $i = \{1, \dots, k\}$ para k segmentos celulares (Equação 4). Através da função D , foi possível encontrar a menor distância entre um pixel P e o centro de um segmento S_i e atribuir a identificação desse segmento para P .

$$D(P, S_i) = \sqrt{(P_x - S_i^x)^2 + (P_y - S_i^y)^2} \quad (4)$$

A fim de classificar as células por suas cores – vermelho, cinza e amarelo –, criou-se um dataset onde as instâncias são os segmentos celulares e as características são as intensidades médias de um segmento para cada canal de cor (R, G e B). Definiu-se uma classe para cada instância de acordo com o objetivo

da classificação: 0 para vermelho, 1 para cinza e 2 para amarelo. Dessa forma, foi possível treinar um modelo de classificação. Entre métodos com diferentes viéses indutivos, o algoritmo que apresentou maior acurácia foi o *K-Nearest Neighbors*³ e, por isso, foi o modelo escolhido.

3 Resultados

O resultado da identificação dos núcleos é mostrado na Figura 11. As regiões coloridas representam os novos segmentos celulares gerados. Observa-se que sujeiras ainda estão classificadas como núcleos, o que pode gerar classificações erradas. As células nas quais não há núcleos (geralmente, células amareladas), o algoritmo identifica bordas que caracterizam as regiões mais escuras.

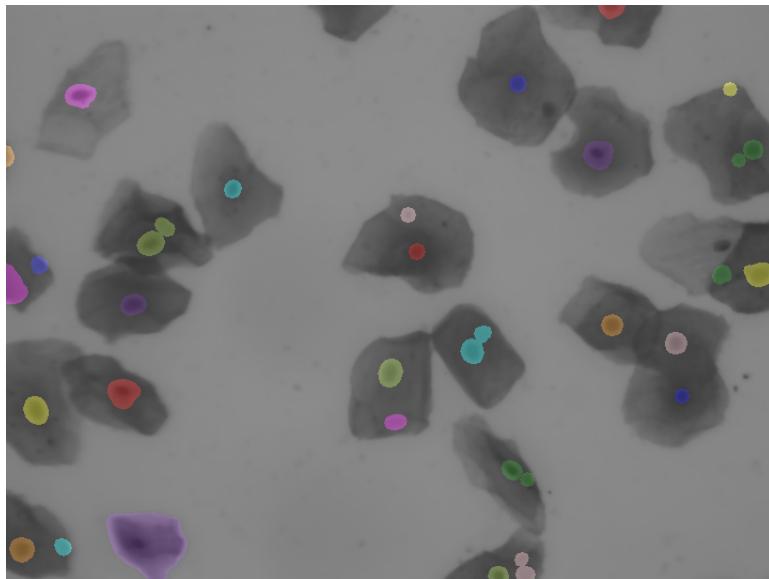


Figura 11: Imagem celular com os núcleos identificados pelas regiões coloridas.

A Figura 12 apresenta a segmentação apóis o algoritmo, o qual encontra a menor distância entre um pixel P e o centro de um segmento S e atribui a identificação do segmento mais próximo à P . Observa-se que a segmentação foi bem sucedida nas células sem muita sujeiro. Entretanto, em células com o núcleo mal definido ou inexistente, o método de identificação dos núcleos não foi capaz de segmentar corretamente as células, como demonstra a Figura 13.

A Tabela 1 mostra os resultados das classificações e contagens das células. Novamente, percebe-se que, na presença de células amareladas, a estimativa é significantemente diferente do valor real. Isso evidencia a falha do algoritmo em

³Implementado em Python e disponível na biblioteca do Scikit-Learning.

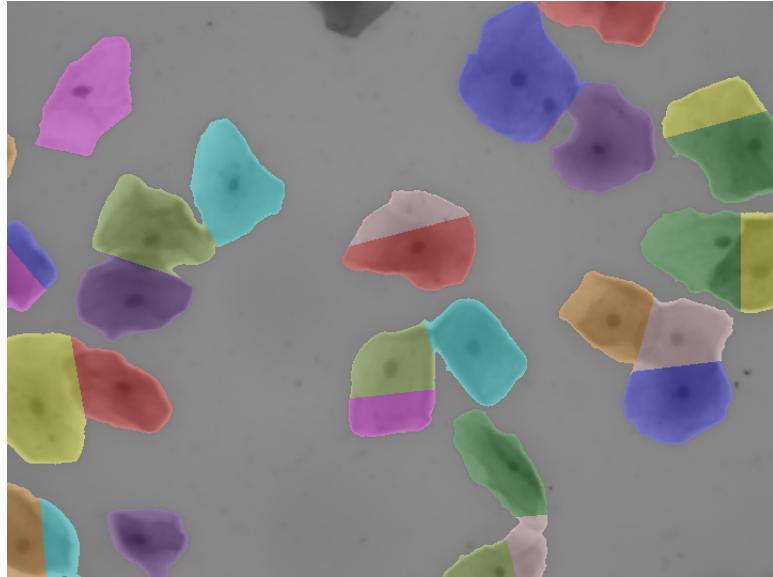


Figura 12: Imagem celular com as células identificadas pelas regiões coloridas.

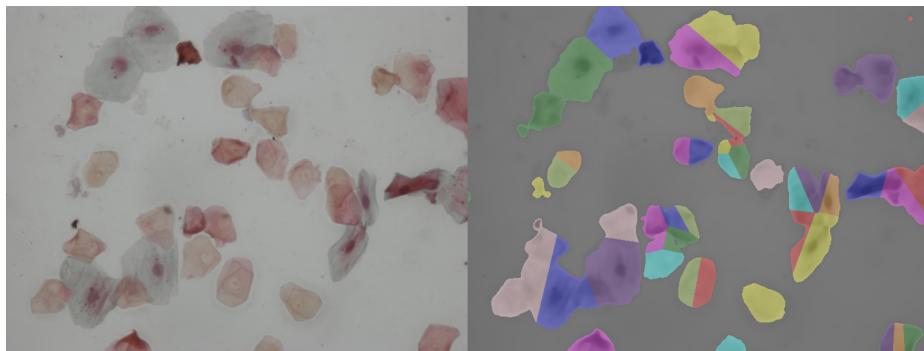


Figura 13: Image original (esquerda); (direita) regiões segmentadas identificadas pelas regiões coloridas.

identificar as células amareladas, o que também acaba impactando negativamente na contagem total de células.

4 Discussão

Esse método é semelhante do proposto por Malpica et al. [1997] e por Bengtsson et al. [2004]. Os autores desses trabalhos buscam por máximas locais na

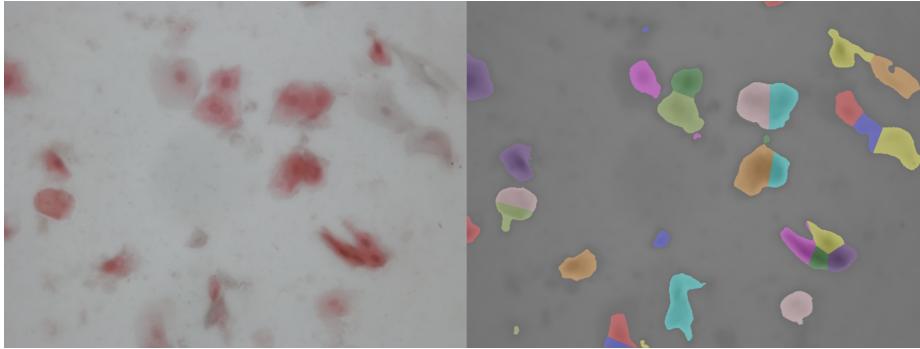


Figura 14: Imagem naturalmente borrada (esquerda); (direita) regiões segmentadas identificadas pelas regiões coloridas.

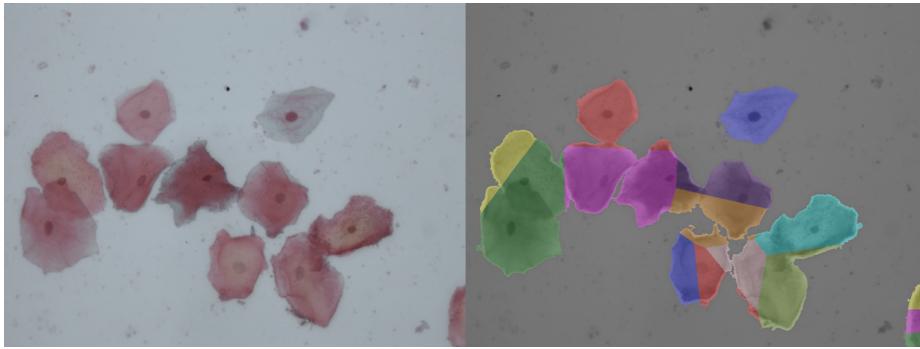


Figura 15: Imagem original (esquerda); (direita) regiões segmentadas identificadas pelas regiões coloridas.

imagem utilizando a definição de $h - maxima^4$. Uma $h - maxima$ é uma região local de pico, onde a intensidade dos pixels vizinhos é maior ou igual a h . Para $h - minima$ a região demarcada é um vale, onde a intensidade dos pixels vizinhos é menor ou igual a h . No cenário proposto pelas imagens do presente trabalho, não foi possível utilizar mínimas locais devido a sujeiras nas imagens que caracterizavam regiões de vale, identificadas como $h - minima$.

O algoritmo proposto consegue minimizar o efeito das sujeiras, mesmo em imagens borradass como a Figura 14. Porém, o método falha na segmentação de imagens com núcleos mal definidos ou inexistentes. Dessa forma, ele ainda não é confiável para obter dados de diagnóstico. Um método mais abrangente pode ser proposto a partir das observações e conclusões desse trabalho.

⁴No caso deles, os núcleos eram a parte mais clara de uma célula, por isso máxima e não mínima.

Figura	Contagem de Células (estimado / real)			
	Vermelhas	Cinzas	Amarelas	Total
12	29 / 22	2 / 2	0 / 0	31 / 24
13	14 / 7	11 / 8	26 / 16	51 / 31
14	14 / 6	10 / 7	1 / 0	25 / 23
15	13 / 6	2 / 1	0 / 4	15 / 1

Tabela 1: Resultados da classificação das células para cada imagem.

5 Conclusão

Ao longo do semestre, vimos diferentes métodos de segmentação. Para o problema de segmentação de células, a literatura possui uma ampla diversidade de métodos. Entretanto, esses métodos são desenvolvidos para imagens com características específicas do domínio do problema. Assim, foi necessário adaptar os métodos existentes para o domínio deste trabalho. No âmbito geral do problema, apenas alguns casos foram bem sucedidos. Assim, para este conjunto de imagens específicas, ainda não foi encontrada uma solução capaz de funcionar perfeitamente, deixando espaço para melhorias nos métodos apresentados ou criação de novos algoritmos, os quais solucionem o problema proposto.

Nem todos os resultados foram satisfatórios. Porém, é um passo dado para a solução do problema. Com os objetivos deste trabalho alcançados dentro do prazo previsto, o que fica é o conhecimento adquirido ao decorrer do desenvolvimento e será de grande valia para futuros trabalhos e pesquisas.

Referências

- E. Bengtsson, C. Wahlby, and J. Lindblad. Robust Cell Image Segmentation Methods. *Pattern Recognition and Image Analysis*, 14(2):157–167, 2004.
- Joan Duran, Bartomeu Coll, and Catalina Sbert. Chambolle’s Projection Algorithm for Total Variation Denoising. *Image Processing On Line*, (3):311–331, 2013.
- Norberto Malpica, Carlos Ortiz de Solorzano, Juan José Vaquero, André Santos, Isabel Vallcorba, José Miguel García-Sagredo, and Francisco del Pozo. Applying Watershed Algorithms to the Segmentation of Clustered Nuclei. *Cytometry*, 28:289–297, 1997.
- G. W. Zack, W. E. Rogers, and S. A. Latt. Automatic measurement of sister chromatid exchange frequency. *Journal of Histochemistry and Cytochemistry*, 25(7):741–753, 1977.