

CAMILA VELOSO TRÓPIA

Orientador: Andrea Gomes Campos Bianchi

Co-orientador: Cláudia Martins Carneiro

**ANÁLISE DE IMAGENS DE MICROSCOPIA DO TECIDO
CARDÍACO**

Ouro Preto

Dezembro de 2019

UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

ANÁLISE DE IMAGENS DE MICROSCOPIA DO TECIDO CARDÍACO

Monografia apresentada ao Curso de Bacharelado em Ciência da Computação da Universidade Federal de Ouro Preto como requisito parcial para a obtenção do grau de Bacharel em Ciência da Computação.

CAMILA VELOSO TRÓPIA

Ouro Preto
Dezembro de 2019

T753a Trópia, Camila Veloso.
Análise de imagens de microscopia do tecido cardíaco
[manuscrito] / Camila Veloso Trópia. -
2020.

40f.: il.: color; grafs.

Orientadora: Profª. Drª. Andrea Gomes Campos Bianchi.
Coorientadora: Profª. Drª. Cláudia Martins Carneiro.

Monografia (Graduação). Universidade Federal de Ouro Preto. Instituto de
Ciências Exatas e Biológicas. Departamento de Computação.

1. Aprendizado do computador. 2. Processamento de imagens. I. Bianchi,
Andrea Gomes Campos. II. Carneiro, Cláudia Martins. III. Universidade
Federal de Ouro Preto. IV. Título.

CDU: 004.85

Catálogo: ficha.sisbin@ufop.edu.br

RELATÓRIO DE AVALIAÇÃO FINAL DO TCC – MONOGRAFIA II

ALUNO(A): <i>Camila Veloso Trópica</i>
TÍTULO DO TCC: <i>Análise de Imagens de Microscopia do Teido Lendíneo</i>

AVALIAÇÃO DA BANCA EXAMINADORA

BANCA EXAMINADORA	NOTAS ATRIBUÍDAS
Professor Orientador/ Co-orientador: <i>Andrea G. Campos Bianchi/ Cláudia Martins Carneiro</i>	<i>9,0/8,9</i>
1º Membro: <i>Taléria de Carvalho Santos</i>	<i>8,6</i>
2º Membro: <i>Liliana Naxa Lima</i>	<i>9,4</i>

MÉDIA FINAL: *9,0* (*nove*)

Ouro Preto, *18* de *dezembro* de 20 *17*

ASSINATURAS DA BANCA EXAMINADORA

Andrea G. Campos Bianchi / Cláudia Martins Carneiro
Professor Orientador / Co-orientador

Taléria de Carvalho Santos
1º Membro

Liliana Naxa Lima
2º Membro

Resumo

Este trabalho consiste em desenvolver um algoritmo para segmentação do núcleo de imagens do tecido cardíaco. Para os pesquisadores na microscopia, obter e analisar as imagens é algo rotineiro, mas atualmente são tantas imagens que o maior desafio dessa área não se concentra apenas em obter e processar imagens, mas em reconstruí-las apropriadamente, extrair medidas precisas e também correlacionar os resultados obtidos com a proposição de modelos computacionais para compreender os diversos fenômenos biológicos investigados. O primeiro passo, e talvez o mais complexo para conseguir fazer essa análise computacionalmente, é segmentar as imagens, com o objetivo de encontrar regiões de interesse e aplicar nelas as análises necessárias.

Para atingir tal objetivo, inicialmente foi investigado como essas imagens de tecido de coração são obtidas e quais corantes são usados na lâmina para facilitar a identificação dos núcleos celulares. Sabe-se que esses pigmentos permitem a pigmentação de estruturas biológicas de forma diferenciada, logo o uso desses corantes permitem uma segmentação baseada nos critérios biológicos. Em seguida, agrupamos *pixels* da imagem para obtermos regiões de interesse semelhantes e identificar e rotular os núcleos, gerando uma segmentação automática dos núcleos na imagem. Essa segmentação é importante para o patologista porque o número de núcleos em uma lâmina pode indicar inflamações ou anomalias que estão ocorrendo no tecido analisado.

A base de dados usada foi disponibilizada pelo Núcleo de Pesquisas em Ciências Biológicas - NUPEB / UFOP, e é conjunto de 804 imagens de cortes histológicos do tecido do coração. As imagens foram divididas em dois grupos, um para o átrio direito e outro para o ventrículo esquerdo. Além das imagens, também foi disponibilizada uma tabela com o número de células para cada imagem.

Os agrupamentos dos *pixels* foram feitos de forma não supervisionada e usando o algoritmo K-means, diferentes valores de k foram investigados. Os resultados de número de núcleos foram comparados com a contagem feita por um patologista, e esses resultados se aproximaram bastante do resultado obtido pelo profissional, o erro relativo médio foi de 37,74% para $k = 4$ e de 19,84% para $k=5$, o que significa que o método automático pode dar resultados semelhantes aos de um patologista e assim, auxiliá-lo na análise.

Abstract

This work consists of developing an algorithm for nucleus segmentation of cardiac tissue images. For microscopic researchers, obtaining and analyzing images is routine. However, nowadays there are so many images that the biggest challenge in this area is not just about obtaining and processing images, but about properly reconstructing them, extracting accurate measurements and also correlating them. Results obtained by proposing computational models to understand the various biological phenomena investigated. The first, and perhaps the most complex, step to perform this analysis computationally, is to segment the images in order to find regions of interest and apply the necessary analyzes to them.

To achieve this goal, it was initially investigated how these heart tissue images are obtained and which dyes are used on the slide to facilitate the identification of cell nuclei. These pigments are known to allow the pigmentation of biological structures in different ways, so the use of these dyes allows segmentation based on biological criteria. We then group pixels on the image to get similar regions of interest and identify and label the cores, generating automatic segmentation of the cores in the image. This segmentation is important for the pathologist because the number of nuclei in a slide may indicate inflammation or anomalies that are occurring in the tissue analyzed.

The database used was made available by the Center for Research in Biological Sciences - NUPEB / UFOP, and is a set of 804 images of histological sections of the heart tissue. The images were divided into two groups, one for the right atrium and one for the left ventricle. In addition to the images, a table with the number of cells for each image was also provided.

Pixel groupings were done unsupervised and using the K-means algorithm, different k values were investigated. The nucleus number results were compared with the count made by a pathologist, and these results were very close to the results obtained by the professional, the average relative error was 37.74 % for $k = 4$ and 19.84 % to $k = 5$, which means that the automatic method can give results similar to those of a pathologist and thus assist him in the analysis.

Dedico este trabalho a minha família e amigos, pelas alegrias, tristezas e dores compartilhadas. Com vocês, as pausas entre um parágrafo e outro de produção melhora tudo o que tenho produzido na vida.

Agradecimentos

Gostaria de agradecer à minha mãe, Mônica, e irmã, Nathália por me dar carinho e apoio incondicional, não medindo esforços para me ajudar chegar até aqui.

Aos amigos da Voluta Soluções Digitais, por serem uma segunda família e pela motivação em continuar nesta empresa jr. que me orgulho de fazer parte, à Natália, a Camila Beatriz e o Fagner Bernardo, por todas as horas de estudo que se tornaram numa amizade sincera e duradoura, por nunca desistirem de mim, por aguentarem o meu mau-humor, birutices e me motivar a continuar quando eu quis desistir.

Aos amigos da Dti Digital, um time que me apoiou e ensinou muito esse semestre.

À Prof. Andrea Gomes Campos Bianchi e Prof. Cláudia Martins Carneiro, pela orientação, motivação, criatividade, grande desprendimento em ajudar, incentivo, ajuda com o fornecimento de material para a realização deste trabalho e amizade sincera.

À todos os outros amigos, professores e colegas de curso que não citei aqui o nome, por fazerem parte desta jornada.

Sumário

1	Introdução	1
1.1	Justificativa	2
1.2	Objetivos	2
2	Revisão de Literatura	3
3	Fundamentação Teórica	8
3.1	Deconvolução de cores	8
3.2	K-means	11
3.3	Avaliação	12
3.3.1	Distância Euclidiana	12
3.3.2	Operação morfológica: Abertura	12
3.3.3	Componentes conectados	12
4	Metodologia	14
4.1	Base de dados	14
4.2	Metodologia	15
4.2.1	Representação das cores dos pigmentos em RGB	15
4.2.2	Deconvolução de cores	16
4.2.3	Clusterização por cores da nova imagem	16
4.2.4	Avaliação de qualidade	17
5	Resultados	18
6	Conclusão	25
6.1	Lista de Siglas	27
	Referências Bibliográficas	28

Lista de Figuras

3.1	A imagem acima é a imagem original, e abaixo, temos a imagem depois de passar pela abertura.	13
4.1	Imagens do coração disponibilizadas pelo NUPEB, átrio direito (esquerda) e ventrículo esquerdo (direita).	14
4.2	Exemplos de tabelas contendo informações das imagens microscópicas disponibilizadas pelo NUPEB.	15
4.3	Metodologia proposta para a segmentação das imagens do tecido cardíaco: 1) Representação das cores dos pigmentos em RGB; 2) Deconvolução de cores da imagem original; 3) grupoização por cores da nova imagem; 4) Avaliação de qualidade	15
4.4	Mapa de cores de a) Hematoxilina, b) DAB e c) Eosina, respectivamente	16
5.1	a) Mapa de cores da hematoxilina e b) Intensidades	18
5.2	a) Mapa de cores da eosina e b) Intensidades	18
5.3	a) Mapa de cores do DAB e b) Intensidades	18
5.4	a) Imagem original, b) imagem apenas com pigmento da hematoxilina, c) imagem apenas com pigmento da eosina e d) imagem apenas com pigmento do DAB	19
5.5	Resultado k-means com $k = 3$	19
5.6	Resultado k-means com $k = 4$	20
5.7	Resultado k-means com $k = 5$	20
5.8	Resultado k-means com $k = 6$	20
5.9	Resultado do agrupamento com apenas o grupo do núcleo, para $k=4$	21
5.10	Resultado da operação de abertura na imagem segmentada	21
5.11	Rotulação de total de números de células para a Figura 5.10	22
5.12	Tabela de resultados do diretório 67-17 do átrio direito	22
5.13	Tabela com médias dos erros relativos percentuais de cada diretório em AD e VE	23
5.14	Gráfico comparando medidas dos resultados encontrados por um patologista, $k=4$ e $k=5$ para as imagens do átrio direito	23
5.15	Gráfico comparando medidas dos resultados encontrados por um patologista, $k=4$ e $k=5$ para as imagens do ventrículo esquerdo	24

Capítulo 1

Introdução

A análise de dados cumpre um papel importante em aplicações nas mais diversas áreas científicas e industriais dada a enorme quantidade de informações que são capturadas e armazenadas diariamente pelos inúmeros sensores disponíveis. Além disso, as informações não se restringem aos dados clássicos, mas envolvem diferentes tipos de informação (não estruturadas), como por exemplo, numérica, textual, espacial, multimídia e relacional. Logo, o desenvolvimento de técnicas automáticas capazes de analisar, integrar e gerar modelos de tomada de decisões vem se tornando mais importantes a cada dia.

Neste trabalho, as informações são restritas a imagens microscópicas de células do tecido cardíaco. Para extrair os dados das imagens, usa-se processamento digital de imagens, cujas técnicas e aparelhos permitem que a visão de uma imagem seja capaz de ser feita cobrindo quase todo o espectro eletromagnético, podendo obter melhores resultados que a visão humana, que são limitados à banda visual do espectro. O processamento de imagens digitais "pode trabalhar com imagens geradas por fontes que os humanos não estão acostumados a associar com imagens. Essas fontes incluem ultrassom, microscopia eletrônica e imagens geradas por computador. Desta forma, o processamento digital de imagens inclui um amplo e variado campo de aplicações." [Gonzalez e Woods \(2010\)](#).

Particularmente, nas imagens de microscopia, os pesquisadores encontram-se limitados às análises e aos softwares que acompanham seus respectivos microscópios. Se por um lado estes equipamentos proporcionaram inovações científicas nas áreas de pesquisa, por outro, resultam na utilização indiscriminada de suas imagens e na extração de informações de uma pequena amostragem das imagens coletadas. Isto é, o maior desafio dessa área atualmente não se concentra apenas em obter e processar imagens, mas em reconstruí-la apropriadamente, extrair medidas precisas e também correlacionar os resultados obtidos com a proposição de modelos computacionais para compreender os diversos fenômenos biológicos investigados. Desse modo, a construção de metodologias robustas e eficazes de análise de imagens e aprendizagem de máquina são ferramentas importantes no processo de transformação dos dados em informação.

1.1 Justificativa

O reconhecimento de padrões é uma área bastante explorada na Ciência da computação devido a sua grande quantidade de aplicações. Visão computacional e *Machine Learning* vem sendo utilizados nas áreas de medicina (diagnóstico auxiliado por computador), reconhecimento de fala, reconhecimento facial e de escrita.

Além disso, reconhecer está ligado ao ato de perceber algo. Ao dar ao computador a capacidade de perceber objetos ao seu redor pode-se criar também uma máquina mais autônoma com melhor interatividade com os usuários. No caso de exames de células, as análises são feitas por especialistas de forma manual, portanto extremamente custosas. Nos exames, além do processo de triagem manual ser bastante rigoroso e exigir um protocolo específico de revisão de lâminas, ele pode ser afetado pela fadiga do operador após períodos prolongados de avaliação de lâminas, pela habituação visual, e pela variabilidade entre os citologistas. Um exemplo de como a análise pode ser fadigante é que aproximadamente 80% de um milhão de biópsias de próstatas feitas nos Estados Unidos, todos os anos, é benigno (Metin N. Gurcan e Yener (2009)). Ou seja, são analisadas um milhão de biópsias e apenas 20% do tempo dessas análises é dedicado às situações realmente importantes, o caso de células malignas. Além disso, os exames envolvem subjetividade e não permitem a detecção de mudanças sutis em grandes populações. Desse modo, análises quantitativas sistemáticas podem aperfeiçoar os resultados relativos à prevenção, diagnóstico e tratamento.

Este trabalho visa segmentar núcleos de células de imagens de cortes histológicos do tecido do coração. Encontrar anormalidades no núcleo das células é parte do processo de diagnóstico do paciente e fazer isto com ajuda da tecnologia, neste caso com a segmentação automática para entender características baseadas em tamanho e forma do núcleo, pode, como foi estabelecido anteriormente, aperfeiçoar o diagnóstico.

1.2 Objetivos

Este trabalho tem como objetivo principal realizar a segmentação do núcleo das células das imagens de microscopia do tecido cardíaco.

São objetivos específicos:

- Separar canais de cores da imagem de acordo com os pigmentos utilizados nas lâminas;
- Segmentação da imagem em regiões de interesse (ROI): Serão investigadas as diferentes metodologias de segmentação, rotulação e extração de medidas morfométricas;
- Implementação de extração de atributos e medidas morfológicas das imagens de microscopia;
- Avaliação de resultados gerados.

Capítulo 2

Revisão de Literatura

Devido ao grande crescimento de descobertas tecnológicas nas últimas décadas, muitas áreas têm mudado graças às novas tecnologias e a praticidade que elas proporcionam. Dessas tecnologias, pode-se destacar o aumento de poder computacional que fez surgir poderosas ferramentas de análise de imagem que podem ser usadas para a área de diagnósticos na patologia.

A citologia é o estudo de células em termos de estrutura, funcionalidade e química. Segundo [A. J. Mendez e Vidal \(1998\)](#), a imagem citológica é a mais comumente usada tanto para a triagem de doenças quanto para a biópsia. Na citologia, existe hoje um ramo chamado de "citologia digital", isto se dá pelo fato de que imagens digitais podem ser usadas para uma variedade de aplicações em citologia, incluindo interpretações rápidas, diagnóstico primário e segundas opiniões, educação e testes de proficiência, [Wilbur \(2011\)](#). Logo, muitas ferramentas foram e estão sendo desenvolvidas por pesquisadores para tornar mais fácil o estudo e a análise da citologia para vários tipos de células, tecidos e afins.

Para analisar as imagens da citologia digital, podem ser usadas variadas técnicas, e para entender os diferentes métodos utilizados, é preciso entender qual área estará sendo analisada e que tipo de imagem se terá acesso. Tudo depende da área que se está sendo analisada, assim algumas imagens podem conter mais textura, um volume de células maiores, vasos sanguíneos, vasos de gordura, etc. E algo recorrente é o fato das imagens obtidas pelo microscópio óptico terem interferências de luz, sujeira da lâmina ou até interferências humanas como tremor.

A segmentação das estruturas encontradas em imagens da citologia digital tornou-se um pré-requisito para a análise dessas imagens. Assim, a presença, a extensão, o tamanho e o formato dessas estruturas segmentadas podem indicar alguma anomalia. Existem muitos trabalhos na literatura que visam fazer a segmentação de estruturas. [Veillard \(2013\)](#) propôs um método para extração do núcleo de células de imagens de histopatologia do câncer de mama baseado na criação de uma modalidade que consiste em um mapa de escala de cinza, onde o valor de cada pixel mostra a probabilidade deste pertencer ao núcleo da célula. Este mapa de probabilidade é calculado a partir de informações da textura e escala com a adição

de intensidades de cor de pixel. Contudo, este método mostrou-se eficaz apenas em imagens em que os limites dos núcleos são muito salientes e outras características irrelevantes são suavizadas, o que não acontece na maioria das imagens de microscopia.

Em outro estudo, [Lei He \(2011\)](#) faz uma revisão de outros trabalhos que analisam imagens para detecção e classificação de carcinomas em próstata, mama e colo do útero. O autor aponta que as características usadas para a classificação são o tamanho e a distribuição do formato do núcleo e das glândulas, que provavelmente devem ser segmentadas antes de classificar. E para classificar, o autor cita métodos de classificação usados por outros autores, como o *Markov Random Field* (MRF), AdaBoost, *Support Vector Machine* (SVM) e *Bayesian Classifier*. Para o câncer de mama, o autor cita muitas ferramentas usadas por outros autores para classificação da grade *Bloom-Richardson* (BR) que determina a severidade do câncer. Entre essas classificações tem-se o Latent Dirichlet Allocation (LDA), *forward and backward search*, linear, quadrática, redes neurais, árvores de decisão, *Principal component analysis* (PCA), *Bayesian classifier* e SVM. O que realmente interessa é como é feita a segmentação do núcleo feita antes de aplicar qualquer método de classificação. Para isso, o autor cita o trabalho de [A. Brook \(2006\)](#), que usa um crescimento de regiões baseado em *watershed*. Por fim, para o colo do útero, há uma grade para diagnosticar o câncer, a grade *cervical intraepithelial neoplasia* (CIN). Além do estado normal das células, a grade classifica três níveis de CIN, baseado na severidade do câncer. O objetivo do autor é classificar o câncer de acordo com o CIN, mas antes de classificar, ele precisa segmentar os núcleos. Para isso, ele usa a técnica de *thresholding*, que usa um valor de intensidade para separar objetos do fundo da imagem. O *thresholding* é baseado na intensidade da imagem para encontrar bordas. Depois, para classificar, são consideradas dezoito características para classificação e a ferramenta supervisionada de classificação LDA (incorpora informações de rotulação de dados para encontrar as projeções que irão maximizar a razão entre a variação entre classes e a variação dentro das classes) é aplicada para diferenciar o caso normal com cada nível de CIN. O melhor caso obtido foi da classificação obtida entre o normal e o CIN 3 (o caso mais agravante), com a precisão de 98.7%.

Assim, em células cervicais, [Keenan \(2000\)](#) classificou automaticamente lesões cervicais. A segmentação do núcleo feita em imagens de tecido foi baseada em um processo chamado de *thresholding* iterativo: os pixels pertencentes ao núcleo são selecionados de acordo com valores do *threshold* em cada iteração do algoritmo. O processo restante gera uma rede sobre os núcleos segmentados usando métricas relacionadas à triângulos resultantes da aplicação da Triangulação Delaunay, que estabelece conexões entre triplas de pontos. Este método da triangulação também usado por [Miranda \(2012\)](#) em imagens binárias, encontra regiões de interesse, cria camadas e depois segmenta lesões baseado em atributos de interesse, caracterizados pelo grau médio da entropia e pela soma das áreas dos triângulos de uma camada em particular, dividido pela soma das áreas de todos os triângulos da rede.

Um trabalho muito interessante foi o de [Barros \(2017\)](#), que desenvolveu uma ferramenta computacional designada à ajudar patologistas a ensinar e pesquisar sobre doenças dos rins. Chamada de PathoSpotter-K, a ferramenta foi projetada para detectar lesões nefrológicas em imagens digitais de rins usando processamento de imagens e métodos de reconhecimento de padrão para detectar lesões glomerulares proliferativas. O motivo deste trabalho se destacar, é por apesar de se tratar de regiões diferentes, afinal, [Barros \(2017\)](#) estava segmentando rins e este trabalho visa a segmentação de imagens do tecido cardíaco, estas imagens tem uma semelhança que é a pigmentação usada para extrair a imagem digital.

A metodologia de [Barros \(2017\)](#) foi baseada em fazer o processamento das imagens, segmentação, extração de características, classificação e validação. No processamento de imagens foram utilizadas duas operações: para intensificar as áreas de interesse das imagens e para reduzir ruídos. Na primeira, foi necessário mudar o espaço de cor das imagens utilizando um método de deconvolução de cor para representar imagens glomerulares em canais, separando os corantes que foram usados para adquiri-las. Assim, foi utilizado o espaço de cor HED, que armazena informação da hematoxilina, eosina e DBA em três camadas diferentes.

Para remover o ruído, foi usada a camada H (matriz H) que representa a hematoxilina, que tem ruídos relacionados à informação de textura, o que pode complicar na segmentação das regiões. Assim, para diminuir a interferência causada por estes ruídos, foram tratados os fragmentos do núcleo (que poderiam acabar sendo classificados erroneamente como um núcleo inteiro) aplicando nestes um filtro de suavização baseado no filtro da mediana, com tamanho de janela de 3x3 pixels, o que elimina o ruído enquanto preserva as bordas dos núcleos.

Para [Barros \(2017\)](#) fazer a segmentação, foi usada a operação morfológica de fechamento junto com o método de Otsu para segmentar o núcleo das células. Inicialmente, foi utilizado o fechamento usando o operador *top-hat* para intensificar o núcleo das células nas imagens. Logo depois, é aplicado o método de Otsu para gerar imagens binárias e então é feito novamente o fechamento para melhorar a representação do núcleo e reduzir o ruído introduzido pela binarização.

Feita a segmentação, o próximo passo foi fazer a extração de características. Foram testados três métricas candidatas como discriminantes: 1) o número de regiões com núcleo, 2) aglomeração de núcleos e 3) densidade dos pixels. Para fazer a contagem de 1), foram usados métodos de detecção de bordas e crescimento de regiões. Cada região foi caracterizada por um grupo de pixels com o mesmo brilho. Em 2), foi utilizado o método de *Laplacian of Gaussian* (LoG) que funciona como um método de detecção de bolha. Este método foi utilizado para identificar o número de núcleos das células vizinhas que estavam perto o suficiente mas não necessariamente ligadas umas as outras para formar um cluster. E em 3) foi dividido o número de pixels brancos pelo número de pixels pretos.

Enfim, para classificar as imagens segmentadas, [Barros \(2017\)](#) dividiu seu conjunto em 90% para teste e 10% para validação e usou um classificador binário baseado no algoritmo

k-Nearest Neighbor (k-NN). Para parametrizar o kNN, foram ajustadas as métricas de distância (testando distância Euclidiana, de Manhattan e Minkowski) e o parâmetro k (numa escala de 3 a 80, onde o valor máximo era correspondente ao número total de amostras em cada pasta, e o mínimo corresponde ao número de classes mais 1). O modelo de classificação e parametrização foi avaliado analisando a taxa de erro. O menor erro encontrado foi correspondente ao uso da distância de Manhattan e $k = 11$. Como resultado, a ferramenta proposta obteve para o conjunto de teste 92,3% de precisão, 88% de revocação e 88,3% de acurácia, e para o conjunto de validação 88% de precisão, 88% de revocação e 85% de acurácia, uma média comparável aos trabalhos relacionados à área.

F. A. Peres (2010) desenvolveu um modelo automático de *thresholding* multi-level que leva em consideração os vales do histograma, que são os lugares onde se concentram os *thresholds*, logo são as divisões da imagem. Para encontrar os vales, os autores parametrizaram um número que é o grupo específico de *pixels* na escala de 265 níveis de cinza. O histograma é dividido por esse número de grupos e para identificar os vales, compara-se o primeiro valor do grupo, que foi determinado na segmentação do histograma, com o último do mesmo grupo. "Se o primeiro valor é menor, significa que os valores do histograma estão aumentando e o sinal se torna positivo. Caso contrário, os valores do histograma estão em caído e, portanto, o sinal é negativo".

Assim, quando F. A. Peres (2010) passa de uma transição negativa para positiva entre esses grupos, quer dizer que um vale pode ter sido encontrado. Para garantir que a transição é mesmo um vale, os autores definem como parâmetro também uma porcentagem mínima, e ela é comparada com o cálculo da inclinação alcançada pela diferença da média dos três últimos valores do grupo com os três primeiros. Se a média for maior, a verificação do histograma pára pois o grupo com um vale foi encontrado. Com esse grupo em análise, é calculada a entropia máxima e onde o nível de cinza indicar a maior entropia, é aí que está o *threshold*. Depois de encontrá-lo, o algoritmo volta a analisar os outros grupos. O resultado do algoritmo comparado ao método de Otsu (1978) teve uma diferença de 26,95% de *thresholds* encontrados.

Em 2012, dois anos depois, Pritee Gupta (2012) usou a mesma estratégia de segmentação de F. A. Peres (2010). Foram usadas imagens do tecido cardíaco, feito o cálculo do histograma, a quantização parametrizável dos grupos, a detecção de possíveis vales pelo cálculo da inclinação dos valores do grupo em comparação com uma porcentagem também parametrizável e cálculo da entropia. Os trabalhos são bastante semelhantes. No entanto, Pritee Gupta (2012) chega à conclusão que este método só é eficaz quando a imagem e o histograma são bem definidos, ou seja, quando a imagem tem ruídos, distorção ou histogramas não normalizados, o método não consegue encontrar um limiar efetivo.

De todos os trabalhos citados, Barros (2017) foi o que mais se aproximou com o problema que queremos tratar aqui. Outros autores já segmentaram o tecido cardíaco, contudo seus métodos foram bem próximos, usando *thresholding* para fazer a segmentação.

Nota-se que cada vez mais, a análise de imagens torna-se uma importante ferramenta na patologia pela sua praticidade e dependência destas imagens. Muitos estudos mostram diferentes abordagens para diversas imagens de áreas variadas, cânceres, lesões, carcinomas, etc. O que estas pesquisas partilham é o interesse em extrair as características que são significativas para os diagnósticos das diferentes imagens que analisam e é relevante exaltar que pode-se extrair medidas morfológicas comuns em cada pesquisa, tais como área, tamanho, limite, formato, cor, intensidade de cores, histogramas e entropia. Sumarizando todas informações, averigua-se que usar métodos de limiarização ou classificação pode ser o caminho a ser seguido para segmentar essas imagens. Contudo, como esses métodos apresentaram resultados consideravelmente bons apenas para imagens bem definidas, pode ser que seja interessante introduzir a deconvolução de cores usada por Barros (2017) antes de aplicar qualquer método de segmentação, para obter apenas a camada da imagem que contém o núcleo, a camada da hematoxilina, e então testar alguns métodos de segmentação em cima da imagem obtida apenas por essa camada, e é isso que trataremos neste trabalho.

Capítulo 3

Fundamentação Teórica

O resultado de uma segmentação é, geralmente, um conjunto de regiões ou objetos de interesse definidos por algum critério para a análise de imagens. Pode-se considerar a segmentação a tarefa mais básica para a análise de imagens e a partir do resultado dela pode-se extrair características, fazer análises e reconhecimentos do escopo em questão. Por isso, pode-se afirmar que a precisão da fase de segmentação determina o sucesso ou falha dos procedimento de análise de imagem. E apesar de ser uma tarefa básica na análise de imagens, a segmentação também é uma das tarefas mais difíceis no processamento de imagens, então para ter uma boa precisão nessa fase, [Gonzalez e Woods \(2010\)](#) diz que deve-se fazer uso de todo e qualquer conhecimento prévio sobre o problema. Por isso, este capítulo se tratará de definir o método de deconvolução de cores, usado como principal conhecimento prévio das imagens analisadas, além dos métodos de suavização e operações morfológicas para pré e pós processamento das imagens e os métodos de segmentação.

3.1 Deconvolução de cores

As imagens analisadas são pigmentadas por três colorações diferentes: a hematoxilina, que é absorvida (ou como vamos tratar aqui, "mancha") pelo núcleo, a eosina, que mancha o tecido e o DAB, que mancha estruturas proteicas.

O método criado por [Ruifrok e Johnston \(2001\)](#) é dividido por duas etapas: a representação de cores dos pigmentos que produziram manchas na imagem em uma matriz RGB e a deconvolução de cores. A representação de cores é baseada em imagens de uma câmera RGB onde o nível de cinza em cada canal é linear com brilho T , com T sendo I/I_0 , onde I_0 é a luz incidente e I a luz transmitida pelo material. Cada mancha é gerada por um fator de absorção específico C para a luz em cada um dos canais RGB. As intensidades detectadas de luz transmitidas através de uma amostra e a quantidade (A) do pigmento com fator de absorção C é descrita pela lei de LambertBeer que se refere à relação entre a transmissão de luz e a espessura da camada do meio absorvente:

$$I_C = I_{0,C} \exp(-Ac_C)$$

Com $I_{0,C}$ sendo a intensidade da luz que entra na amostra, I_C a intensidade da luz transmitida e subscripto C indicando o canal de detecção. Isso significa que a transmissão da luz e os valores de cinza de cada canal dependem da concentração da mancha de cada pigmento de maneira não linear.

No modelo RGB, as intensidades I_R , I_G , e I_B são obtidas pela câmera para cada pixel, porque a intensidade relativa em cada um dos canais depende da concentração de mancha de forma não linear, os valores de intensidade da imagem não podem ser usados diretamente para separação e medição de cada uma das manchas. No entanto, a densidade óptica (OD) para cada canal pode ser definido como

$$OD_c = -\log(I_C/I_{0,C}) = A * c_c$$

Como pode ser visto, o OD para cada canal é linear com a concentração de absorção e, portanto, pode ser usado para separar a contribuição de várias manchas em uma amostra.

Cada mancha pura será caracterizada por uma densidade óptica específica para a luz em cada um dos os três canais RGB, que podem ser representados por um vetor OD de tamanho 3 por 1, descrevendo a mancha no espaço de cores RGB. Por exemplo, medições de uma amostra coradas apenas com hematoxilina resultaram em valores de OD de 0,18, 0,20 e 0,08 para os canais R, G e B, respectivamente.

O comprimento do vetor será proporcional à quantidade de mancha, enquanto o valor relativo aos valores do vetor descrevem o OD real para os canais de detecção. No caso de três canais, o sistema de cores pode ser descrito como uma matriz da forma:

$$\begin{bmatrix} p11 & p12 & p13 \\ p21 & p22 & p23 \\ p31 & p32 & p33 \end{bmatrix}$$

No qual cada linha representando uma mancha específica (primeira linha - Hematoxilina, segunda linha - Eosina e terceira linha - DAB) e todas as colunas que representam o OD detectadas por o canal vermelho (primeira coluna), verde (segunda coluna) e azul (terceira coluna) para cada mancha.

Valores específicos de mancha para o OD em cada um dos três canais podem ser facilmente determinados medindo absorção relativa para vermelho, verde e azul que pigmentaram a imagem. Um exemplo do OD matriz para a combinação de hematoxilina, eosina e DAB é:

$$\begin{bmatrix} 0.18 & 0.20 & 0.08 \\ 0.01 & 0.13 & 0.01 \\ 0.10 & 0.21 & 0.29 \end{bmatrix}$$

Assim, chegamos à deconvolução das cores. Para realizar a separação das manchas, temos que fazer uma transformação ortonormal das informações do RGB para obter informações independentes sobre a contribuição de cada mancha. A transformação tem que ser ortogonal para obter informações independentes sobre cada uma das manchas; e ainda, a transformação tem que obter o equilíbrio correto do fator de absorção para cada mancha separada.

Para a normalização, divide-se cada vetor OD pelo seu tamanho total.

$$\hat{p}_{11} = p_{11} / \sqrt{p_{11}^2 + p_{12}^2 + p_{13}^2}$$

$$\hat{p}_{21} = p_{21} / \sqrt{p_{21}^2 + p_{22}^2 + p_{23}^2}$$

$$\hat{p}_{31} = p_{31} / \sqrt{p_{31}^2 + p_{32}^2 + p_{33}^2}$$

E assim por diante, resultando em uma matriz OD normalizada, M :

$$M = \begin{bmatrix} \hat{p}_{11} & \hat{p}_{12} & \hat{p}_{13} \\ \hat{p}_{21} & \hat{p}_{22} & \hat{p}_{23} \\ \hat{p}_{31} & \hat{p}_{32} & \hat{p}_{33} \end{bmatrix}$$

A matriz OD normalizada, M , para a combinação acima de hematoxilina, eosina e DAB é:

$$M = \begin{bmatrix} 0.65 & 0.70 & 0.29 \\ 0.07 & 0.99 & 0.11 \\ 0.27 & 0.57 & 0.78 \end{bmatrix}$$

Sendo C um o vetor 3×1 para quantidades das três manchas em um pixel específico, a intensidade detectada nesse pixel é $y = CM$.

Pelo exposto acima, fica claro que $C = M^{-1}[y]$. Isso significa que a multiplicação da imagem OD com o inverso da matriz OD , que definimos como a matriz de deconvolução de cores, D , resulta na representação ortogonal das manchas formadas na imagem: $C = D[y]$.

D , correspondente a M para a matriz hematoxilina, eosina e DAB, é:

$$D = \begin{bmatrix} 1.88 & -0.07 & -0.60 \\ -1.02 & 1.13 & -0.48 \\ -0.55 & -0.13 & 1.57 \end{bmatrix}$$

Nesta matriz, os elementos diagonais são maiores que a unidade, enquanto os elementos fora da diagonal são negativos. A matriz acima implica que os valores corrigidos do nível de OD para cada mancha são formados por:

- Hematoxilina: OD verde e OD azul, subtraídos pelo OD vermelho;

- Eosina: OD vermelho e OD azul, subtraídos pelo OD verde;
- DAB: OD vermelho e OD verde, subtraídos pelo OD azul.

Caso as manchas utilizadas sejam vermelhas, verdes e azuis puras, a matriz acima seria a matriz unitária.

Com essa informação, conseguimos converter uma imagem RGB em HED (hematoxilina-eosina-DAB), separar os canais dela e usar apenas o canal do pigmento que mancha aquilo que temos interesse.

3.2 K-means

O K-means é um algoritmo de agrupamento de aprendizado de máquina não supervisionado, isto é, ele faz inferências a partir de conjuntos de dados usando apenas vetores de entrada sem se referir a resultados conhecidos ou rotulados. O objetivo do K-means é descobrir padrões subjacentes, e para atingir esse objetivo, ele procura um número fixo k de grupos em um conjunto de dados. Um grupo é uma coleção de pontos de dados agregados devido à certas semelhanças e cada grupo tem um centróide. Cada ponto de dados é alocado para cada um dos grupos, reduzindo a soma dos quadrados no grupo.

Pode-se resumir o algoritmo em 4 passos:

1. Escolhe-se aleatoriamente k centróides nos pontos de amostra como centros iniciais do grupo.
2. Atribui-se cada amostra ao centróide mais próximo $\mu^j, j \in \{1, \dots, k\}$.
3. Move-se os centróides para o centro das amostras que foram atribuídas a ele.
4. Repete-se as etapas 2 e 3 até que as atribuições do *grupo* não sejam alteradas ou até que seja atingida uma tolerância definida pelo usuário ou o número máximo de iterações.

Para medir a semelhança entre os objetos, define-se a similaridade como o oposto da distância. Uma distância comumente usada para agrupar amostras com recursos contínuos é a distância euclidiana ao quadrado entre dois pontos x e y no espaço m -dimensional:

$$d(x, y)^2 = \sum_{j=1}^m (x_j - y_j)^2 = \|x - y\|_2^2$$

Com base nessa métrica de distância euclidiana, pode-se descrever o algoritmo k-means como um problema de otimização simples, uma abordagem iterativa para minimizar a soma dos erros quadrados (SSE) dentro do *grupo*, que às vezes também é chamada de inércia do *grupo*:

$$SSE = \sum_{i=1}^n \sum_{j=1}^k w^{(i,j)} \|x^i - \mu^j\|_2^2$$

Onde μ^j é a posição do centróide do grupo j , j se refere à j -ésima dimensão dos pontos de amostra x e y . Usa-se os sobrescritos i e j para referir ao índice de amostra e ao índice de grupo, respectivamente. e

$$w^{(i,j)} = \begin{cases} 1, & \text{se o objeto } x^i \text{ está no grupo } j \\ 0, & \text{caso contrário} \end{cases}$$

3.3 Avaliação

A avaliação será feita baseada na comparação estatística do resultado do algoritmo com o resultado obtido por um patólogo. Para a avaliação, serão necessários alguns algoritmos simples:

3.3.1 Distância Euclidiana

Será necessário calcular a distância Euclidiana (d) das intensidades em comparação com a intensidade do preto ($[0, 0, 0]$):

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}$$

Onde, x_1 , y_1 e z_1 são intensidades de uma cor e x_2 , y_2 e z_2 são intensidades do preto ($[0, 0, 0]$).

3.3.2 Operação morfológica: Abertura

Esta é uma operação que geralmente suaviza o contorno de uma imagem, quebra istmos e elimina protuberâncias finas. A Abertura de uma imagem f por um elemento estruturante s , definida por $f \circ s$ é simplesmente uma erosão seguida por uma dilatação (Cámara-Chávez (2019)):

$$f \circ s = (f \ominus s) \oplus s$$

Um exemplo do comportamento da abertura pode ser visto na Figura 3.1:

3.3.3 Componentes conectados

Este algoritmo é baseado em aplicar busca em largura (BFS) ou a busca em profundidade (DFS) a partir de cada vértice não visitado e assim, obter todos os componentes fortemente conectados. Abaixo estão as etapas baseadas no DFS:

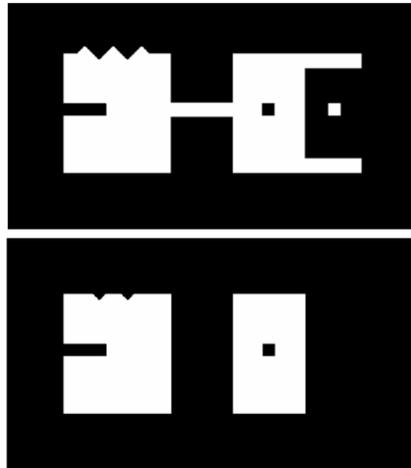


Figura 3.1: A imagem acima é a imagem original, e abaixo, temos a imagem depois de passar pela abertura.

1. Inicialize todos os vértices como não visitados.
2. Faça o seguinte para cada vértice 'v'.
 - (a) Se 'v' não for visitado antes, chame a função DFSUtil(v)
 - (b) Imprimir novo caractere

Algoritmo 1: DFSUtil

Entrada: v

```

1 início
2   Marque 'v' como visitado;
3   Imprimir 'v';
4   repita
5     if 'u' não for visitado then
6       Chame recursivamente DFSUtil(u);
7       próximo se torna o atual;
8   até para todos os 'u' adjacentes de 'v';
9 fim
```

Capítulo 4

Metodologia

O desenvolvimento deste trabalho se baseia na segmentação de imagens do tecido cardíaco. Por isso, esta sessão está dividida em duas partes: uma sobre a base de dados usada e outra sobre o método para segmentar as imagens dessa base.

4.1 Base de dados

Neste trabalho, serão consideradas imagens microscópicas do tecido muscular cardíaco. As imagens foram disponibilizadas pelo Núcleo de Pesquisas em Ciências Biológicas (NUPEB) da Universidade Federal de Ouro Preto (UFOP).

As imagens estão divididas de acordo com o átrio direito e ventrículo esquerdo. São imagens difíceis de analisar por serem heterogêneas, com textura rugosa e ruídos, mostradas na Figura [4.1](#).

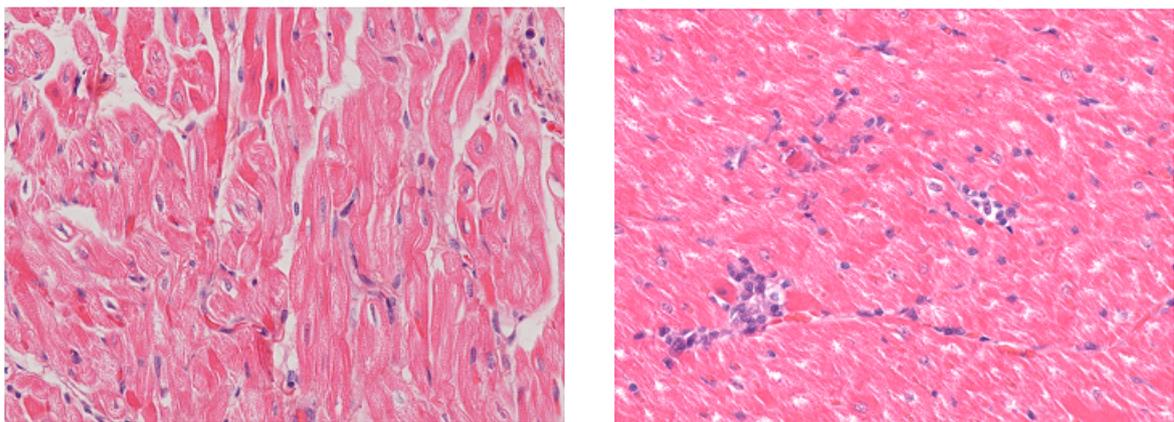


Figura 4.1: Imagens do coração disponibilizadas pelo NUPEB, átrio direito (esquerda) e ventrículo esquerdo (direita)

Foram disponibilizadas 401 imagens microscópicas do átrio direito, 403 imagens microscó-

picas do ventrículo esquerdo e um documento contendo informações de cada uma das imagens, obtidas no Laboratório Multiusuários de Microscopia do NUPEB, tais como o número de células, a área em branco e a área de tecido, como mostra a Figura 4.2.

61-17					62-17				
Campo	Número de Células	Área branco	Área Tecido	Número de Células Normalizado	Campo	Número de Células	Área branco	Área Tecido	Número de Células Normalizado
1	139	1095,25	75725,05	141	1	170	8351,01	68469,29	191
2	173	5472,75	71347,55	186	2	131	13457,45	63362,85	159
3	114	3389,6	73430,7	119	3	88	16569,84	60250,46	112
4	110	2869,67	73950,63	114	4	57	10387,31	66432,99	66
5	101	4143	72677,3	107	5	74	4577,87	72242,43	79
6	165	7734,56	69085,74	183	6	127	2678,48	74141,82	132
7	84	10392,67	66427,63	97	7	175	6814,85	70005,45	192
8	103	7210,23	69610,07	114	8	129	5562	71258,3	139
9	103	4637,35	72182,95	110	9	114	5433	71387,3	123
10	76	4582,54	72237,76	81	10	103	5143,9	71676,4	110
11	75	3517,75	73302,55	79	11	147	11660,59	65159,71	173
12	78	1644,63	75175,67	80	12	137	15302,82	61517,48	171
13	58	1318,17	75502,13	59	13	60	11203,81	65616,49	70
14	137	3499,81	73320,49	144	14	100	10323,89	66496,41	116
15	56	1415,07	75405,23	57	15	239	5517,64	71302,66	257
16	70	1094,51	75725,79	71	16	198	7574,25	69246,05	220
17	109	1103,26	75717,04	111	17	129	4367,15	72453,15	137
18	75	2177,39	74642,91	77	18	160	3754,68	73065,62	168
19	74	2535,47	74284,83	77	19	152	3483,78	73336,52	159
20	85	2554,84	74265,46	88	20	87	4209,44	72610,86	92
<i>Média</i>	<i>99</i>			<i>105</i>	<i>Média</i>	<i>129</i>			<i>143</i>

Figura 4.2: Exemplos de tabelas contendo informações das imagens microscópicas disponibilizadas pelo NUPEB

Note na Figura 4.1 os elementos das imagens: os pequenos pontos roxos são os núcleos das células, a forma rosa que ocupa grande parte da imagem é tecido e em branco há o fundo.

4.2 Metodologia

Baseado nas informações que desejamos extrair da imagem, foi proposta a metodologia descrita na Figura 4.3. Ela foi dividida em quatro etapas etapas: representação das cores dos pigmento em RGB, deconvolução de cores da imagem original, grupoização por cores da nova imagem e avaliação de qualidade. Cada uma das etapas será apresentada nas subseções seguintes:



Figura 4.3: Metodologia proposta para a segmentação das imagens do tecido cardíaco: 1) Representação das cores dos pigmentos em RGB; 2) Deconvolução de cores da imagem original; 3) grupoização por cores da nova imagem; 4) Avaliação de qualidade

4.2.1 Representação das cores dos pigmentos em RGB

As imagens foram pigmentadas com três compostos diferentes, a hematoxilina, que vai pigmentar o núcleo; a eosina, que pigmenta o tecido; e a diaminobenzidina (DAB), que pig-

menta estruturas proteicas. O primeiro passo é criar um mapa de cores artificiais para cada pigmento da imagem original.

O mapa de cor é uma sequência de cores que transitam uniformemente e variam intensidades. Logo, para a hematoxilina o mapa seria do branco ao azul marinho, a eosina do violeta escuro ao branco e o DAB do branco ao marrom areia. Pode-se ver um exemplo dos três mapas de cores na Figura 4.4

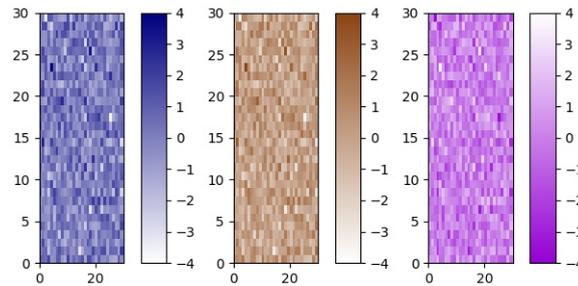


Figura 4.4: Mapa de cores de a) Hematoxilina, b) DAB e c) Eosina, respectivamente

4.2.2 Deconvolução de cores

Com cada mapa de cor definido, aplica-se o método de deconvolução de cores para converter a imagem original RGB em HED. Assim, temos três novos canais, H para a hematoxilina, E para eosina e D para DAB, e cada canal representa de forma mais específica as estruturas celulares.

Para visualizar cada camada separadamente, cria-se novas imagens com o canal e mapa de cores correspondente de cada pigmento.

Por fim, a nossa imagem de interesse é composta pela camada H, que é o pigmento que mancha os núcleos das células.

4.2.3 Clusterização por cores da nova imagem

Muitos trabalhos da revisão bibliográfica usam limiarização nas imagens. Contudo aplicar limiarização nessas imagens não nos dá um resultado muito preciso, devida a concentração de pigmentos de eosina e hematoxilina que ficam nas imagens, e causam manchas escuras no tecido. Essas manchas podem ser facilmente confundidas no limiar, pois a área ocupada por essas manchas é muito grande. Por isso, foi pensado um algoritmo que agrupasse as cores da imagem para fazer o agrupamento. Assim, foi utilizado o algoritmo não supervisionado K-means, disponibilizado por Pedregosa et al. (2011). Esse algoritmo tem como parâmetros de entrada o número de grupos, o método usado para escolher os centróides iniciais, o número de vezes que o algoritmo será executado com diferentes centróides, o número máximo de iterações do algoritmo para uma única execução, a tolerância relativa em relação à inércia para declarar

convergência, um booleano para computar distâncias e um estado randômico que determina a geração aleatória de números para inicialização do centroide. Como parâmetros de saída temos: as coordenadas dos centroides, as rotulações para cada ponto, a soma das distâncias quadradas das amostras ao centro de grupo mais próximo e o número de iterações executadas.

4.2.4 Avaliação de qualidade

Como não temos disponível o *ground truth* das regiões analisadas, vamos propor dois diferentes valores de k e comparar as medidas extraídas de números de células pelo método e o número observado pelo patologista, Figura [4.2](#)

Essa comparação será feita para os resultados do K-means de $k = 4$ e $k = 5$. Para isso, é preciso identificar qual dos *grupos* resultante do K-means é o *grupo* referente à intensidade do núcleo. O K-means retorna os rótulos de cada *grupo*, cada rótulo é uma intensidade diferente e é sabido que o núcleo vai ter a intensidade mais próxima à $[0, 0, 0]$. Por isso, calculamos a distância Euclidiana de cada rótulo à intensidade do preto: $[0, 0, 0]$. O rótulo que estiver à menor distância do preto, é o do núcleo.

Com o rótulo do núcleo descoberto, podemos recriar a imagem original mostrando apenas os núcleos. Como as imagens ainda terão algum ruído, também é aplicada a operação morfológica de abertura, que irá suavizar o contorno dos núcleos, quebra istmos e elimina ruídos.

Assim, a próxima fase é contar o número de núcleos na nova imagem, o que é feito através de um algoritmo de componentes conectados. Este algoritmo aplica a busca em profundidade (DFS) a partir de cada vértice (pixel) não visitado e assim, obtém todos os componentes fortemente conectados. Então, para cada componente conectado, teremos uma rotulação de um núcleo. Por fim, teremos os rótulos de todos os núcleos, basta contá-las para saber o total de núcleos na imagem para comparar com o número de núcleos obtido pelo patologista.

Capítulo 5

Resultados

Ao processar cada imagem, separamos os canais de cores para cada pigmento: hematoxilina (Figura 5.1), eosina (Figura 5.2) e DAB (Figura 5.3), e aplicamos a deconvolução de cores na imagem original RGB para obter cada canal HED separadamente, e com cada canal e seu respectivo mapa de cor, criamos novas imagens, como pode ser visto na Figura 5.4.

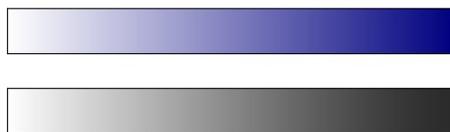


Figura 5.1: a) Mapa de cores da hematoxilina e b) Intensidades

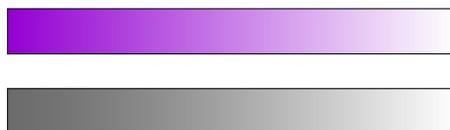


Figura 5.2: a) Mapa de cores da eosina e b) Intensidades

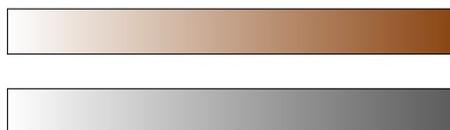


Figura 5.3: a) Mapa de cores do DAB e b) Intensidades

Após obter as imagens, escolhemos aquela que é do nosso interesse, que é a da hematoxilina, Figura 5.4 b). Em seguida, nessa imagem, aplicamos o algoritmo K-means com $k = 3, 4, 5,$

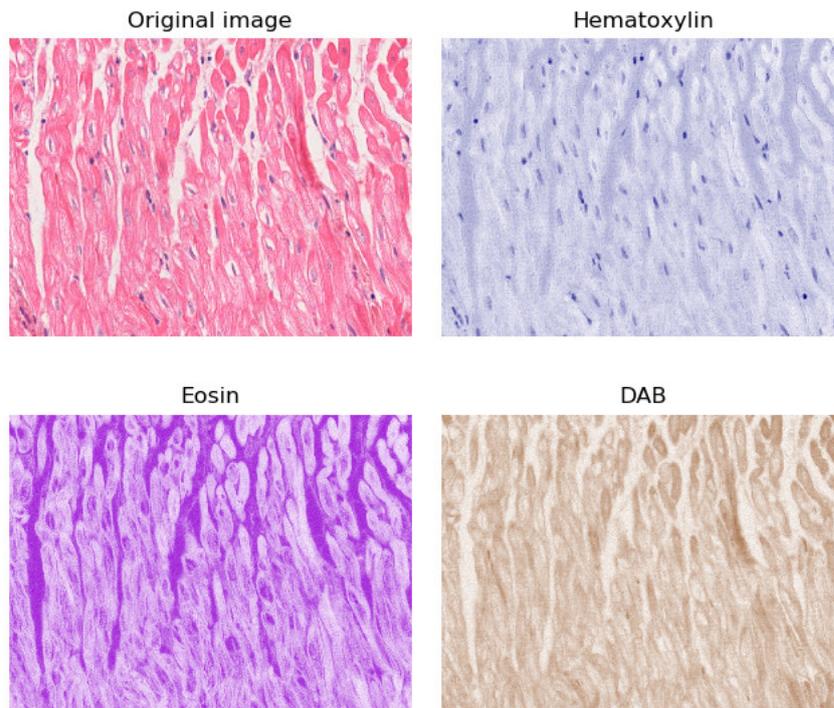


Figura 5.4: a) Imagem original, b) imagem apenas com pigmento da hematoxilina, c) imagem apenas com pigmento da eosina e d) imagem apenas com pigmento do DAB

e 6, afim de agrupar a imagem pelas suas cores (que pertencem ao mapa de cores da Figura 5.2).

Pode-se ver o resultado para uma mesma imagem com $K = 3$ na Figura 5.5, $K = 4$ na Figura 5.6, $K = 5$ na Figura 5.7 e $K = 6$ na Figura 5.8.

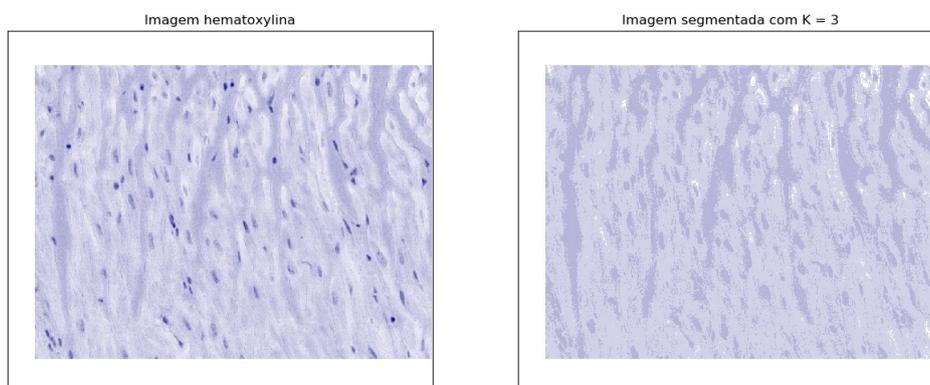
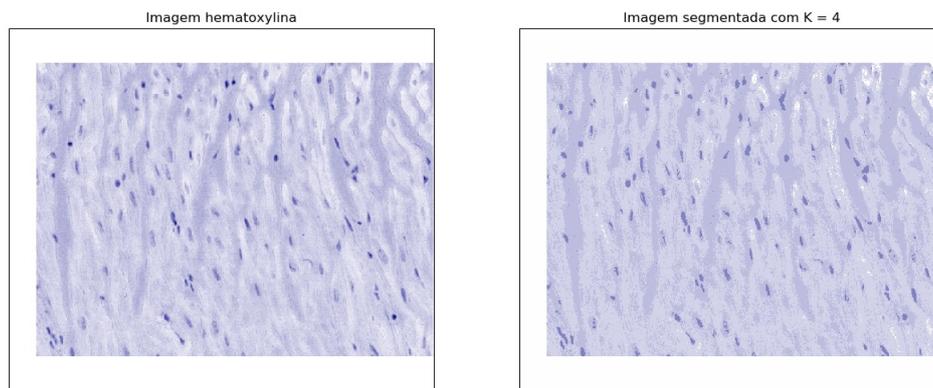
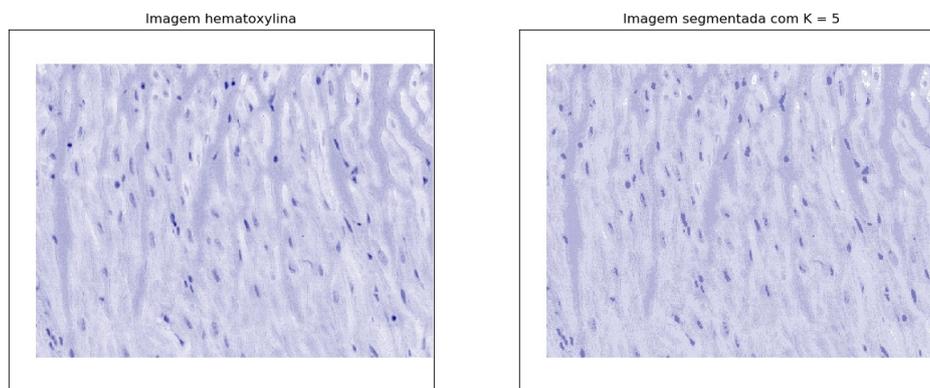
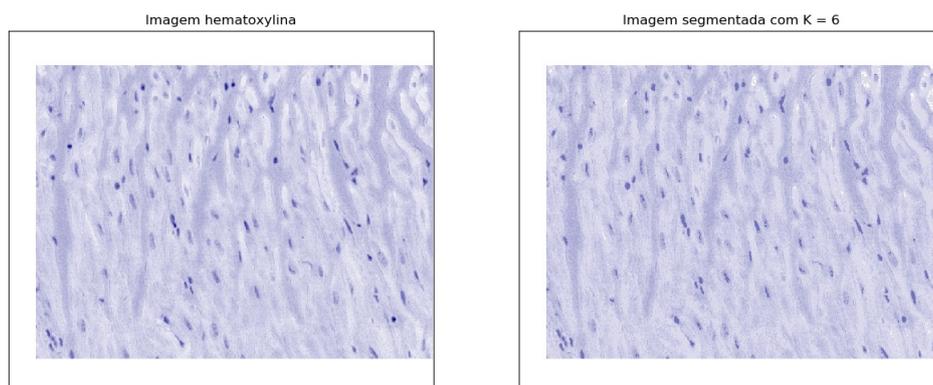


Figura 5.5: Resultado k-means com $k = 3$

Figura 5.6: Resultado k-means com $k = 4$ Figura 5.7: Resultado k-means com $k = 5$ Figura 5.8: Resultado k-means com $k = 6$

Aqui, já temos uma segmentação dada pelos *grupos*, que está separando as nossas regiões

de interesse. Para continuar a análise, vamos ignorar os resultados de $k = 3$, por não ter conseguido *agrupar* o núcleo, e também o $k = 6$, que por ter mais *grupos*, também tem mais ruído (além de ter ficado muito próximo à imagem da hematoxilina original).

Então, para avaliar os resultados, criamos uma nova imagem apenas com o *grupo* do núcleo, que foi descoberto pelo valor do rótulo de intensidade dele, como pode ser visto na Figura [5.9](#).

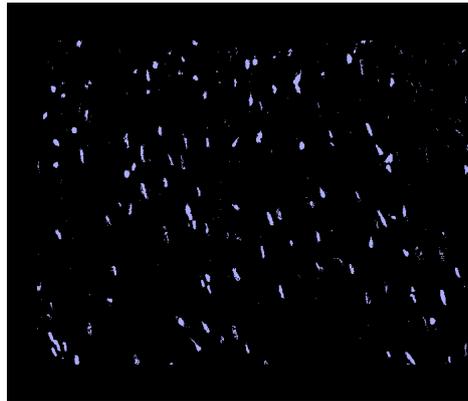


Figura 5.9: Resultado do agrupamento com apenas o grupo do núcleo, para $k=4$

Como pode ser visto, a imagem ainda tem alguns ruídos. Por isso, aplicamos uma abertura para definir melhor os núcleos, usando um *slider* de tamanho 3×3 *pixels*. A Figura [5.10](#) mostra o resultado da abertura.



Figura 5.10: Resultado da operação de abertura na imagem segmentada

Nesse resultado da abertura, aplicamos o algoritmo de componentes conectados, fazendo com que cada núcleo receba um rótulo e com isso, obtemos o número total de núcleos da imagem, Figura [5.11](#).

Finalmente, podemos comparar os resultados finais para $k = 4$ e $k = 5$ com o resultado obtido pelo patologista. No total, foram comparadas 360 imagens do total de 401 do átrio direito (AD), e 360 imagens do total de 403 imagens do ventrículo esquerdo (VE). Esse valor

É evidente que 20 imagens é uma porção muito pequena do total de imagens analisadas, por isso, a tabela mostrada na Figura 5.13, mostra as médias de erro relativo para cada um dos diretórios:

Diretório	AD		VE	
	K=4	K=5	K=4	K=5
61-17	53,87	29,25	28,90	16,54
62-17	21,08	12,35	31,08	22,45
63-17	28,40	16,72	20,63	19,42
64-17	38,82	23,09	4,53	4,93
65-17	48,91	42,54	23,11	17,33
66-17	28,03	17,41	54,05	19,08
67-17	23,28	7,68	22,45	14,06
68-17	40,08	24,97	38,37	25,51
69-17	30,10	21,83	43,13	31,66
70-17	54,87	32,70	233,41	13,47
71-17	22,81	8,78	31,30	23,82
72-17	44,81	15,69	13,17	13,45
73-17	24,46	20,03	34,44	20,03
74-17	32,05	10,80	53,37	36,74
75-17	30,60	22,91	18,79	17,98
76-17	25,67	18,20	18,31	14,93
77-17	35,17	20,11	39,89	26,00
78-17	34,62	14,39	31,93	17,35

Figura 5.13: Tabela com médias dos erros relativos percentuais de cada diretório em AD e VE

Com os dados dos erros relativos, fica evidente que $k = 5$ produziu melhores resultados que $k = 4$, e outra forma de visualizar isto é através das Figuras 5.14 e 5.15, que mostram os gráficos *boxplots* do número de núcleos encontrados pelo patologista, $k = 4$ e $k = 5$, para os conjuntos AD e VE, respectivamente.

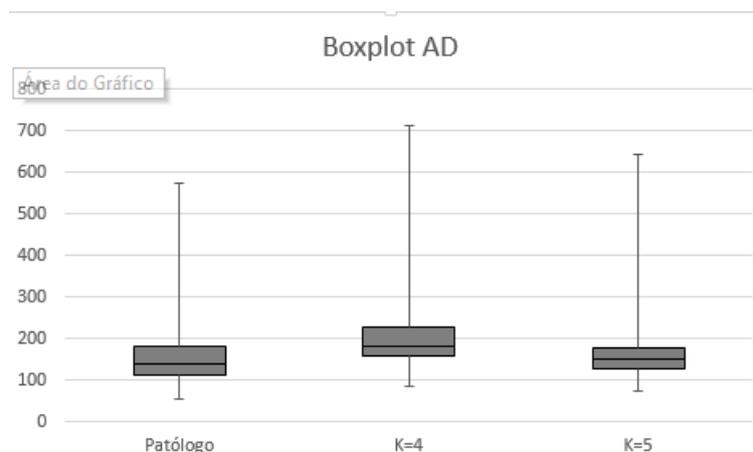


Figura 5.14: Gráfico comparando medidas dos resultados encontrados por um patologista, $k=4$ e $k=5$ para as imagens do átrio direito

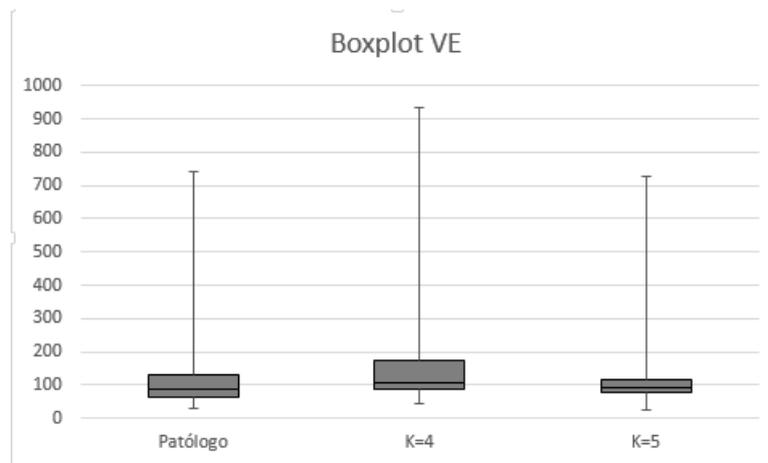


Figura 5.15: Gráfico comparando medidas dos resultados encontrados por um patologista, $k=4$ e $k=5$ para as imagens do ventrículo esquerdo

Através destes gráficos, pode-se observar que os valores médios do número de células detectadas de $k = 4$ e $k = 5$ se aproximaram da média do patologista, e ainda, esses valores médios dos k s estão no erro (variância) do patologista, isto é, a medida produzida por $k = 4$ e $k = 5$, podem ser ambas, embora $k = 5$ ainda seja melhor que $k = 4$, válidas para a identificação dos núcleos, embora $k = 5$ seja melhor que $k = 4$.

Capítulo 6

Conclusão

A citologia digital é uma área em crescimento com a capacidade de transformar a maneira que se leciona, trabalha e vê a citologia. Novas aplicações são desenvolvidas a todo momento, disponibilizando cada vez mais recursos de análise de imagens voltado à microscopia e melhorando a vida dos profissionais que as utilizam. Alguns dos maiores problemas é a escassez de *ground truth* de imagens para o desenvolvimento de melhores tecnologias do tipo, justamente pela dificuldade na aquisição de base de dados. Além disso, avaliação feita por patologistas é qualitativa e subjetiva, o que dificulta o caso da falta de *ground truth* também. Contudo, é possível concluir que algoritmos para a análise de imagens de microscopia podem contribuir para o crescimento da área, tanto para patologistas que poderão utilizá-los como consulta para formar diagnósticos quanto para acadêmicos, estudantes e professores, por fazer com que as disciplinas lecionadas sejam mais dinâmicas com o acesso à tecnologia.

O objetivo deste trabalho é segmentar os núcleos de células das imagens de microscopia do tecido cardíaco. Para isso, extraímos o pigmento que nos interessava de cada imagem, usando a deconvolução de cores, agrupamos a imagem para obter regiões de interesse e rotulamos cada núcleo, assim cumprindo o objetivo proposto.

Os resultados não teriam sido os mesmos se não tivesse sido usada a deconvolução de cores. Separar os pigmentos foi a estratégia mais importante neste trabalho, pois eliminou muitos pré-processamentos que teriam que ser feitos nas imagens e nos entregou uma imagem com definição dos núcleos melhor do que a original.

Com os resultados obtidos ⁰ para o número de núcleos detectados no tecido do coração, conseguimos evidenciar a hipótese de que o método automático pode dar resultados semelhantes aos de um patologista e assim, auxiliá-lo na análise. Ainda existem melhorias que podem ser investigadas, como ruídos que permaneceram nas imagens e confundiram a contagem de núcleos, núcleos que ficaram juntos e por isso dão um resultado menor do que deveria e núcleos que não foram contados porque foram eliminados como ruídos. Contudo, os resultados

⁰Todos os resultados de imagens e tabelas podem ser encontrados em: <http://bit.ly/resultadosSegmentacaoCoracao>.

mostrados pelos gráficos nas Figuras 5.14 e 5.15 indicam que os valores médios da contagem automática encontra-se dentro do erro, variância, da medida feita pelo patologista.

Apêndice

6.1 Lista de Siglas

- ROI: regiões de interesse (regions of interest);
- MRF: algoritmo Markov random field;
- SVM: algoritmo Support Vector Machine;
- BR: grade Bloom-Richardson para graduação de carcinoma de mama;
- LDA: algoritmo Latent Dirichlet Allocation;
- PCA: análise de componentes principais (principal component analysis);
- LoG: método Laplacian of Gaussian;
- k-NN: algoritmo k-Nearest Neighbor;
- OD: densidade óptica;
- BFS: busca em largura;
- DFS: busca em profundidade;
- UFOP: Universidade Federal de Ouro Preto;
- NUPEB: Núcleo de Pesquisas em Ciências Biológicas;

Referências Bibliográficas

- A. Brook, R. El-Yaniv, E. I. R. K. R. M. D. P. (2006). Breast cancer diagnosis from biopsy images using generic features and svms. *Technical report, Technion - Israel Institute of Technology*.
- A. J. Mendez, P. G. Tahoces, M. J. L. M. S. e Vidal, J. J. (1998). Computer-aided diagnosis: Automatic detection of malignant masses in digitized mammograms. *Med Phys, vol 25*.
- Barros, G. O. (2017). Pathspotter-k: A computational tool for the automatic identification of glomerular lesions in histological images of kidneys. *Scientific Reports, 7:46769, DOI: 10.1038/srep46769*.
- Cámara-Chávez, G. (2019). Morfologia matemática. *Material de aula consultado em 30 de Novembro de 2019, disponível em <http://www.decom.ufop.br/guillermo/BCC326/slides/Processamento%20de%20Imagens%20-%20Morfologia%20Matematica.pdf>*.
- F. A. Peres, F. R. Oliveira, L. A. N. M. F. G. (2010). Automatic segmentation of digital images applied in cardiac medical images. *PAN AMERICAN HEALTH CARE EXCHANGES – PAHCE. CONFERENCE, IEEE CATALOG NUMBER CFP1018G-ART*.
- Gonzalez, R. C. e Woods, R. E. (2010). *Processamento Digital de Imagens*. Pearson.
- Keenan, S. (2000). An automated machine vision system for the histological grading of cervical intraepithelial neoplasia (cin). *The Journal of pathology vol. 192 no. 3 pp. 351-362 2000*.
- Lei He, L.Rodney Long, S. A. G. R. T. (2011). Histology image analysis for carcinoma detection and grading. *Elsevier Ireland Ltd | DOI/; 10.1016*.
- Metin N. Gurcan, Laura E. Boucheron, A. C. A. M. N. M. R. e Yener, B. (2009). Histopathological image analysis: A review. *IEEE REVIEWS IN BIOMEDICAL ENGINEERING, VOL 2*.
- Miranda, G. H. B. (2012). Structural analysis of histological images to aid diagnosis of cervical cancer. *Graphics, Patterns and Images (SIBGRAPI), 2012 25th SIBGRAPI Conference on | DOI: 10.1109/SIBGRAPI.2012.51*.

- Otsu, A. N. (1978). A threshold selection method from gray-level histogram. *IEEE Transactions on Systems, Man, and Cybernetics*, pp. 62-66.
- Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M. e Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.
- Pritee Gupta, Vandana Malik, M. G. (2012). Implementation of multilevel threshold method for digital images used in medical image processing. *International Journal of Advanced Research in Computer Science and Software Engineering*. ISSN: 2277 128X.
- Ruifrok, A. e Johnston, D. A. (2001). Quantification of histochemical staining by color deconvolution. *Analytical and quantitative cytology and histology / the International Academy of Cytology [and] American Society of Cytology*.
- Veillard, A. (2013). Cell nuclei extraction from breast cancer histopathological images using colour, texture, scale and shape information. *Diagnostic Pathology* | DOI: 10.1186/1746-1596-8-S1-S5.
- Wilbur, D. C. (2011). Digital cytology: Current state of the art and prospects for the future. *Acta Cytologica* 2011;55:227–238.